Toward a Spanish SKA Regional Centre fully engaged with open science

Julián Garrido[a,*] **Laura Darriba**[a] **Susana Sánchez-Expósito**[a] **Manuel Parra-Royón**[a] **Javier Moldón**[a] **María Á. Mendoza**[a] **Sebastián Luna-Valero**[a,b] **Antxón Alberdi**[a] **Isabel Márquez**[a] **and Lourdes Verdes-Montenegro**[a]

[a]Instituto de Astrofísica de Andalucía (CSIC), Granada, Spain
[b]EGI Foundation, Amsterdam, The Netherlands

Abstract. The Square Kilometre Array Observatory (SKAO) will build the most sensitive radio telescopes on Earth. To address fundamental questions in astrophysics, fundamental physics, and astrobiology, it will require processing and handling complex and extremely massive data close to the exascale, hence constituting a technological challenge for the next decade. Approximately 600 Peta-bytes (PB) of calibrated data will be delivered to the network of SKA Regional Centers (SRCs) worldwide. As a world-leading scientific instrument, SKAO aims to pursue the best practices in scientific methodology. Remarkably, it includes the reproducibility of its data as a metric of success. We present the Spanish prototype of an SRC (SPSRC), which supports preparatory scientific activities for the future SKA projects. These include science with SKA precursors and pathfinders while promoting Open Science practices as a way to enable scientific reproducibility. We describe the key developments and components of the SPSRC that align with these objectives. In particular, we describe the performed work on hardware and cloud computing infrastructure, science archive, software and services, user support and training, and collaboration with other SRCs. The resulting SPSRC platform is flexible enough to host heterogeneous projects while being scalable toward the demanding SKA requirements. [DOI: 10.1117/1.JATIS.8.1.011004]

Keywords: SKA Regional Centre; open science; cloud computing; data handling; archive; science platform.

Paper 21093SS received Aug. 17, 2021; accepted for publication Oct. 14, 2021; published online Nov. 12, 2021.

1 Introduction

The Square Kilometre Array Observatory (SKAO) is an international effort to build the largest and most sensitive radio telescopes. It is designed as a "physics machine" for the 21st century that will address fundamental scientific questions in astrophysics, fundamental physics, and astrobiology (see further details in the SKA science book[1]). The SKAO will be a single observatory organized at three different sites, with the headquarters being located in the United Kingdom and two telescope sites, in Africa and Australia. SKA will be composed of thousands of antennas distributed over distances of up to 3000 km and thus constitute the largest radio telescope on Earth. Australia will host 131,000 low-frequency dipoles (SKA-low), operating between 50 and 350 MHz and reaching maximum baselines of up to 65 km; South Africa will host 197 dishes (SKA-mid), covering the frequency range from 350 MHz to 15.4 GHz (with the goal of reaching 24 GHz), with baselines up to 150 km.[2]

The construction phase will take place from 2021 to 2029, with science commissioning starting around 2023. SKA is recognized as one of the "big data" challenges for the next decade. By the time the SKA is in steady-state operations, SKA-low and SKA-mid will produce, respectively, a raw data rate of ∼2 Pb/s (reduced to ∼7 Tb/s before arrival to the central signal

*Address all correspondence to Julián Garrido, jgarrido@iaa.es

processor) and 20 Tb/s. The SKA science data processor (SDP) will deliver up to 300 PB of calibrated data (called observatory data products, ODPs) from each telescope to a network of SKA regional centers (SRCs).[3] This model—a network of distributed facilities—has been used in the past by astronomy and particle physics communities to facilitate scientific analysis for state-of-the-art scientific instruments and infrastructures. The Large Hadron Collider (LHC) relies on a distributed computing infrastructure since its early days. Similarly, the Atacama Large Millimeter Array (ALMA) created a network of ALMA Regional Centers (ARCs) as an interface between the user community and the observatory. These centers are a key instrument to support scientific communities that carry out a variety of tasks, including preparation of observations, quality assurance of the data, archive operations, data reduction support, science operations, helpdesk, and commissioning.[4]

In the case of the SKAO, a combination of three main factors led to the decision of a distributed model: (1) the high data rates of the SKA telescopes, which impose extremely demanding computational requirements; (2) the data volumes, which are so large that direct delivery to users is not feasible; and (3) a globally distributed community that needs to work collaboratively.[3] The SRC network will provide direct access to ODPs for an international community of collaborating teams and organizations, as well as to the tools and processing power for their transformation into science-ready data (i.e., advanced data products, ADPs) and subsequent scientific exploitation. Preliminary studies analyzed different computing infrastructure to optimize their use,[5] and how they fulfill data processing requirements for the Low Frequency Array (LOFAR), one of the SKA-low pathfinders.[6]

The SRC network was conceived in September 2016, when the SKA organization established the SRC Coordination Group, with the mandate to define the requirements of the SRC network. This group (which includes LVM) identified an initial set of requirements associated with the six following areas: governance, science archive, accessibility and software tools, storage capacity, data processing capacity, and network connectivity.[7] Furthermore, international projects and initiatives started contributing to the design of the SRCs, analyzing the requirements from a regional point of view. The European project "Advanced European Network of E-infrastructures for Astronomy with the SKA" (AENEAS),[8] in which several of these authors participated, carried out, among other things, an estimation of the storage, computing, and data ingest capabilities required by a European SKA Data Centre. As well, they issued different reports with recommendations on the design of user interfaces for data discovery,[9] data processing,[10] and data transport.[11]

Other examples of SRC initiatives are the "Exascale Research Infrastructure for Data in Asia-Pacific astroNomy using the SKA" project (ERIDANUS),[12] and the Canadian Initiative for Radio Astronomy Data Analysis (CIRADA).[13] ERIDANUS is a design study focused on the prototyping of a data-intensive research infrastructure between China and Australia, and CIRADA is aimed at building the computing capacity required to scientifically exploit three radio telescopes [Very Large Array (VLA), Australian Square Kilometre Array Pathfinder (ASKAP), and Canadian Hydrogen Intensity mapping Experiment (CHIME)] and paving the way to a future Canadian SRC.

At the end of 2018, the SRC Steering Committee (SRCSC) took over from the SRC Coordination Group to take the SRC network from design to implementation on a global basis. This committee consists of one representative from each SKA member country that is carrying out activities to develop an SRC, as well as several representatives from the SKAO. At the time of writing, the countries participating in the steering committee are: Australia, Canada, China, France, Germany, India, Italy, Portugal, South Africa, Spain (represented by LVM), Sweden, Switzerland, The Netherlands, and United Kingdom. In addition, Japan and South Korea are observer countries in the committee. It is noteworthy that while the pace and the phase of the individual SRC developments in each country are different, in 2019, the Chinese SRC prototype was the first one to deploy dedicated infrastructure that was used to simulate SKA data transfer and processing.[14] The steering committee is coordinating these national initiatives to develop the first global network of SRC prototypes. To this end, seven working groups were established, focused on the following areas: SRC network architecture, data logistics, operations, federated computing and data software services, SKA science archive, compute, and science user engagement.

Currently, some SKA precursor and pathfinder telescopes are already producing such large data volumes that distributing them to individual users is unfeasible. Instead, users are required to access a common service that provides the necessary framework. The archives and science platforms of these pathfinders are thus valuable testbeds for the SRCs. The scientific exploitation of the data generated by these telescopes requires developing computing centers where data processing can be centralized, minimizing the transfer of large volumes of data around the world, and with sufficient capacities to store them and to provide archive services. Examples include Ilifu in South Africa,[15] which centralizes the processing of MeerKAT data in a cloud platform, and the Pawsey Supercomputing Centre in Australia,[16] which provides specialized services for the ASKAP. The Canadian Advanced Network for Astronomy Research has built a science platform[17] for data intensive astronomy that brings software to the data and provides user storage, user group management, or interactive and persistent virtual machines (VM), among other services. In Australia, the Commonwealth Scientific and Industrial Research Organization has built the ASKAP science data archive that provides access to data (i.e., images, image cubes, and catalogues) using both a virtual observatory and web services.[18]

The next few years are crucial to building expertise on processing and analyzing large and complex radio data to benefit scientifically from the SKAO telescopes. The community is working on it in different ways, e.g., via exploring synergies or participating in SKA science working groups. It is worth highlighting that the use of SKA precursors and pathfinder telescopes[1] is key for the community to get prepared for exploiting SKA and to optimize the scientific analysis of the SKAO data. In addition, a series of science data challenges are being released to the community by the SKAO[19] since November 2018, and this is planned to continue until the end of 2024. The challenges entail analysis of simulated SKA data products that resemble the type of data that the SKAO will produce, with each separate challenge exploring a particular aspect of the SKA data products. The second data challenge receives support from international processing facilities where each participant team needs to use one computational facility to access and process the data mimicking how the future SRCs will work with the SKA data.[20,21] The second data challenge also includes a reproducibility award for those who demonstrate reusable methods and reproducible results.

SKA will be a world-leading facility and as such it aims to follow and lead best practices in scientific integrity as those promoted by the open science movement. The concept of open science facilitates the reproducibility of scientific studies by enhancing the accessibility, understandability, and reusability of their data and methods. Furthermore, open science has impacts on areas related to some of the United Nations Sustainable Development Goals. In particular, the UNESCO recommendation on open science acknowledges how this initiative contributes to democratize information, by fostering enhanced sharing of scientific knowledge among scientific communities while promoting inclusion of underrepresented or excluded groups. Also, it contributes to reducing inequalities in the access to infrastructures.[22]

The adoption of open science values is rooted in SKA's foundational principles,[2] with the aim to promote collaborative, transparent, and accessible scientific research. SKA is the first facility including reproducibility as a scientific metric of success, and the SRC network will preserve ODPs, ADPs, and the workflows and tools used to generate them. We have contributed to the inclusion of open science in key documents for the SKAO (e.g., SKAO prospectus or the SKA construction proposal[2]) and to including reproducibility as a metric that will be monitored during SKAO operations.[23] The SKAO and the SRC network are working to enable best practices that make data and other digital research objects (e.g., algorithms, tools, workflows, protocols, or services) "findable, accessible, interoperable, and reusable" (FAIR). In particular, the SRC Coordination Group defined different requirements related to open science,[24] highlighting the requirement of "open access," which relates to the need for public links to SKA science data products, and the "reproducibility: provenance and workflow preservation" requirement, meaning that the SRCs must be capable of saving the provenance and the workflow associated with the data products generated at each SRC (see Sec. 5). Accordingly, open science introduces requirements into several SRCSC elements, and in particular to the SKA science archive. For example, the SKA Science Archive Working Group of the SRCSC is defining how SRCs will implement the FAIR principles and interoperate with the International Virtual Observatory Alliance (IVOA) ecosystem. The SKA-link project was led by the Instituto de Astrofísica de

Andalucía (IAA-CSIC) and gathered key members of the SKA SDP consortium and SRC initiatives, together with experts on cutting-edge e-Science. The goal of the SKA-link was to identify best practices and technologies to successfully exploit SKA data in SRCs, with an emphasis on tools that facilitate the reproducibility. Furthermore, the "European Science Cluster of Astronomy and Particle physics ESFRI research infrastructures" H2020 project (ESCAPE)[25] brings together the SKA community with others in the astronomy and particle physics fields, aiming to build a multidisciplinary open environment in which they can share data, tools, and methods according to the FAIR principles. We participate in providing different use cases for the development of the ESCAPE Science Analysis Platform[26] and prototyping the integration of its resources into the European Open Science Cloud.

In this paper, we present the development of the Spanish Prototype of an SRC (SPSRC). The IAA-CSIC is the Spanish institution that coordinates the scientific and technological activities related to SKA in Spain since 2011. It is in close contact with the science community, e.g., by supporting and facilitating the participation in the SKA science working groups or organizing SKA-related. The SPSRC initiative came as a natural consequence of these coordination activities, combined with several decades of experience in radio interferometry, e-Science, and open science applied to radio astronomy. The research areas of the IAA-CSIC astronomers overlap with many of the areas in which the SKA will be fundamental,[27] making the IAA-CSIC an ideal environment for building the SPSRC. The research topics range from star and planet formation, trans-neptunian objects and centaurs, galaxy evolution (including star formation or radio-jets), the Milky Way center, transient phenomena to cosmic explosions. The application of Very Long Baseline Interferometric (VLBI) techniques (which provide the high angular resolution SKA will achieve) is transversal to all these fields.

The IAA-CSIC participated in the design of the SDP until it passed the critical design review, contributing to the SKA preservation and delivery subsystems. Since then, IAA-CSIC has contributed to defining, designing, and implementing the SRC network and took part in several of the initiatives mentioned above (SRC Coordination Group, SRCSC, AENEAS and ESCAPE projects, etc.). Currently, the collaboration with other SRCs through the SRCSC and its working groups is key to ensuring interoperability and the success of the SRC network. Furthermore, the IAA-CSIC carries out research and development activities in other areas of relevance for the SKAO and SRC network, i.e., e-Science and open science. For example, we have actively participated in the IVOA, developing software for data analysis and contributing to standards for astronomy; we contributed to developing the Astronomy Edition of the Taverna workflow management system,[28,29] and we participated in building the research object concept,[30] a powerful tool for achieving open science.

The SPSRC aims to support preparatory scientific activities for SKA key science projects and SKA precursors/pathfinder science while promoting synergies and best practices on open science and FAIR principles. Open science at the SPSRC not only aims to enable scientific reproducibility but also to open up the facility to the whole astronomical community beyond that of radio astronomy alone, contributing to build a transversal facility that facilitates wavelength-agnostic analysis. In the next sections, we explain the deployment of the hardware (Sec. 2) and the cloud infrastructure providing computing and storage resources (Sec. 3). Sections 4 and 5 describe the SPSRC cloud services and the tools enabling collaborative analysis as well as open science practices, including the prototyping activities to develop the science archive. The user support services and training activities are presented in Sec. 6, and the operational and management aspects of the SPSRC are described in Sec. 7. Finally, in Sec. 8, we discuss the results achieved so far and describe the developments to be accomplished in the near future.

2 Hardware Platform

The technical specifications required for the SRCs have been estimated by the SRC Coordination Group as well as by the AENEAS project.[24,31] These specifications include aspects related to processing and storage capacity or network connectivity. They are further detailed in the SRC global size estimations document[32] for computing and storage. The AENEAS project conducted an analysis of computing load, data transfer, data storage, and network requirements for the

European SKA Science Data Centre,[33–36] including comparisons between different scenarios and technologies.

The hardware characteristics of the SPSRC are based on the previously mentioned reports. In addition, we surveyed 19 researchers representing seven IAA-CSIC research groups as a representative sample of the use cases of the IAA-CSIC (see Sec. 1). They represent potential future SPSRC users, and we gathered their current and future hardware and software requirements for data processing and analysis and were informed on their computational methodologies. In the second stage, we plan to investigate what requirements are needed for a wider community at a national level. We also conducted face-to-face meetings with the seven research groups to discuss the details of their particular science cases and data processing needs. In particular, data from the following topics were collected during these meetings:

- Data processing and software: Most relevant software tools used by the groups, as well as the degree of automatization of their workflows. This helped us to understand how the users manage their workflows and software. As well, the information was used to detect potential incompatibilities between software versions required by different groups.
- Data properties: Type and volume of data required to complete a project. We considered raw data, data generated during data processing, and final data products that will potentially need long-term storage.
- Computing infrastructures: Preferred and most common computing environments (e.g., desktop, clusters managed by the groups, supercomputers, or cloud platforms), including their general specifications and how they are accessed and used.

During the interviews, the degree of radio astronomy expertise was assessed to identify areas that would require science support from the SPSRC team. In addition, we identified which radio telescopes are commonly used by the groups to know the type of data analysis that these groups would conduct at the SPSRC.

Once we had gathered and analyzed the information from the survey and the interviews, we discussed potential implementations that would satisfy the hardware and software requirements of most of the groups. The emphasis was on those working with radio data and in particular with data from SKA pathfinders and precursors. We considered what options would be the least disruptive for their current methodologies to ensure a smooth transition. We also took long-term feasibility to expand and maintain the infrastructure into consideration. Our conclusions for the technical requirements are:

- Computing environment. The most requested software could be divided into two categories: radio astronomy software (CASA, AIPS, Miriad, Gildas, Difmap, among others) and visualization and analysis software (ds9, TOPCAT, Aladin, casaviewer, CARTA, VizieR, among others). In terms of workflow and software management, the responses included manual processing, execution of observatory pipelines, and development of own scripts and own pipelines. Finally, personal computers or dedicated single-node servers with a minority of high-performance computing or Cloud services were the preferences among the researchers. In summary, most of the groups needed to use specific software with specific versions, and they would feel more comfortable working with a self-managed resource. This could be achieved with a cloud infrastructure, such as OpenStack, Open Nebula, or VMware, where ad-hoc VM can be prepared with custom environments and operating systems for each project.
- Memory and CPU. A significant number of projects could be conducted in a reasonable amount of time with 16 core processors and 4 to 8 GB of memory per core. Some specific projects would benefit from up to 40 CPU cores. Specific memory-intensive tasks could be fulfilled in a fat node with 1 to 1.5 TB of RAM.
- Disk storage. Typical data sets range between 100 GB and 1 TB per project. We estimate that 500 TB would cover the data storage needs of the interviewed groups. Most of this disk space is used only temporarily for data processing for the duration of each project and only 10% to 30% of the total space requested is needed for long-term storage of final data

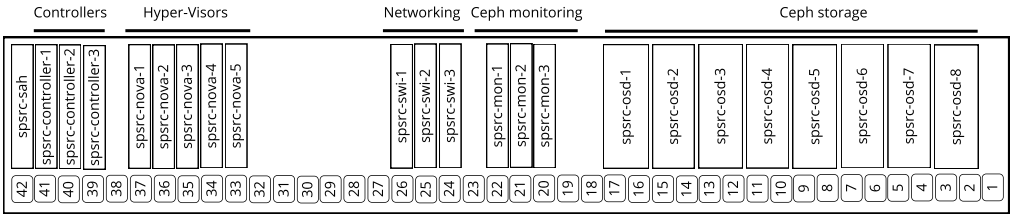

Fig. 1 Rack distribution (upright for space reasons).

products. Therefore, we estimate that the capacity suggested above would comfortably accommodate all the research done by all these groups during at least 3 to 5 years.

- Interface and interaction. A significant fraction of the projects needs continuous real-time interaction to tweak parameters and make decisions in the exploratory phase. Large unattended executions ranging from hours to weeks are not common but are expected in the near future due to new automatic pipelines becoming available. The most commonly desired options include a complete graphical desktop that can be remotely accessed with graphical interfaces such as x2go, Apache Guacamole, RDP, or VNC/vncserver and access through ssh, with X11 forwarding to open interactive windows.

This analysis of the technical requirements helped us make an informed decision on how to distribute the resources between CPU, memory, storage, and networking within the available budget. It was also considered the preliminary requirements in the SRC White Paper,[37] which was in preparation then. In particular, we decided to prioritize a large, high-performance storage capacity and to set up an infrastructure with high horizontal scalability (scale-out) that will allow to easily increase its computing capacity in future expansions.

From the survey, none of the ongoing or imminent projects explicitly required GPUs, so we did not prioritize including GPUs in the infrastructure in this first stage. However, given the relevance of the new data processing techniques that require GPUs (e.g., machine learning and deep learning techniques) and how they can be used to address the big data challenge in areas such as object detection and classification, we plan to include GPU nodes in the next hardware expansion.

After the requirements analysis and design phase in late 2019, we completed the procurement and deployment of the hardware in 2020. This was placed in a rack as shown in Fig. 1. The figure shows the distribution of the equipment and the organization of the deployment of cloud computing components, which will be detailed in Sec. 3.

2.1 *Storage*

The SPSRC hardware storage platform consists of eight dedicated nodes for storage and three nodes for monitoring. Each of the storage nodes has 32 solid state drive (SSD) disks, where one disk is set apart for the operating system and the rest (31) for storage. The overall storage is distributed as follows:

- Four storage nodes with 31 disks in each node. Each disk has a capacity of 7.6 TB, providing a total of 942.2 TB.
- Four storage nodes with 31 disks each, with 3.8 TB per disk, providing a total of 471.2 TB.
- Three monitoring nodes with two disks each of 7.6 TB per disk, configured with RAID1.

Altogether this provides a total raw storage of ∼1.4 PB. The hardware specifications for the storage nodes are listed in Table 1. To manage all storage in an efficient way, Ceph was deployed as the storage solution as described in detail in Sec. 3.3.

2.2 *Computing*

To deal with the demand for computing power, we deployed five compute nodes (to be managed and virtualized with OpenStack, see Sec. 3.1) at the SPSRC infrastructure, providing a total of

Table 1 Technical details describing the characteristics of the nodes used for mass storage.

Storage nodes for Ceph distributed storage	
4 Dell PowerEdge R740xd	CPU: 2 Intel Xeon 6230 (20C/40T)
	RAM: 16 × 16 GB RAM DDR4 2933 MT/s modules with a total of 256GB
	Disks: 32 × 7.68 TB SAS 12 Gbps SSDs
	Network: Broadcom 5720 with 4 1 GbE ports and Mellanox ConnectX-5 EX with 2 100 GbE QSFP28
4 Dell PowerEdge R740xd	CPU: 2× Intel Xeon 6230 (20C/40T)
	RAM: 16 × 16 GB RAM DDR4 2933 MT/s modules with a total of 256 GB RAM
	Disks: 32 × 3.84 TB SAS 12 Gbps SSDs
	Network: Broadcom 5720 with 4 × 1 GbE ports and Mellanox ConnectX-5 EX with 2 × 100 GbE QSFP28

200 CPUs, with 40 cores per node. Four of the nodes are identical and provide 384 GB of RAM memory, and the additional node has 1 TB of RAM memory. As a result, the most memory-intensive computations can be transferred to the 1 TB node and the rest can be distributed over the other nodes. The hardware specifications are listed in Table 2. All five nodes host two 7.68 TB SAS SSDs. As shown in the table, the two disks on each node are managed in RAID1 so that if one disk fails, there is another active disk ready for use at all times.

2.3 *Networking*

The demand for fast computational analyses on large amounts of data requires fast interconnection between computing and storage. The SPSRC hardware infrastructure combines 1 and 100 GbE ports for the internal network, storage, and computational nodes. The hardware specifications of the network switches are listed in Table 3. As shown in the table, we deployed a low-speed 1 GbE network switch to manage traffic (deployment and monitoring) and a high-speed 100 GbE network switch to communicate computing and storage internally as well as enable fast

Table 2 Features of the nodes for computing.

OpenStack hypervisors	
4 Dell PowerEdge R640	CPU: 2 Intel Xeon 6230 (20C/40T)
	RAM: 24× 16 GB RAM DDR4 2933MT/s moduleswith a total of 384 GB
	Disks: 2 × 7.68 TB SAS 12 Gbps SSDs
	Network: Broadcom 5720 with 4 1 GbE ports and Mellanox ConnectX-5 EX with 2 100 GbE QSFP28
1 Dell PowerEdge R640	CPU: 2 Intel Xeon 6230 (20C/40T)
	RAM: 16 × 64 GB RAM DDR4 2933MT/s modules with a total of 1024 GB
	Disks: 2 × 7.68 TB SAS 12 Gbps SSDs
	Network: Broadcom 5720 with 4 1 GbE ports and Mellanox ConnectX-5 EX with 2 100 GbE QSFP28

Table 3 Network configuration for management, data transfers, and storage.

Networking	
1 GbE networking for management	Dell S3048-ON with 48 1 GbE ports
100 GbE for OpenStack and Ceph networking	2 Dell Z9100-ON with 32 100 GbE QSFP28 ports each

data processing. To provide network redundancy, we duplicated 100 GbE switches with Multichassis Link Aggregation Group[38] technology to avoid single points of failure. Computing and storage nodes are redundantly connected by virtual link trunking as is the 1 Gbps management network.

In addition to the internal connection between nodes, it is essential to provide a high-capacity connection to and from the outside of the cluster, for example, to transfer large volumes of data or to interconnect the cluster with external data centers. Communication with the exterior currently goes through the RedIRIS Spanish national research and education network, with 10 Gbps fiber optics shared bandwidth. The IAA-CSIC network is planned to be upgraded to 100 Gbps proximately, which will allow us to have a 10 Gbps dedicated channel for the SPSRC.

We studied the performance of different network links, in particular between the SPSRC and the Chinese SRC prototype at the Shanghai Astronomical Observatory and the Italian National Institute for Astrophysics. We used iperf[39] to analyze and monitor the network quality and performance.[40]

3 Cloud Computing Infrastructure

In previous works, we have tested different cloud infrastructures, concluding that they are able to fulfil the requirements for processing data from the SKA pathfinder LOFAR. As well, they are offering interesting functionalities such as a straightforward and simple installation and maintenance of the software and the flexibility to adapt the infrastructure to the needs of the problem, among others.[6] These functionalities cover the needs of the potential SPSRC users as explained in Sec. 2. Thus, to govern all the SPSRC hardware, we deployed the most widely used enterprise/science platform for cloud computing, i.e., OpenStack.[41] The following sections detail how OpenStack has been deployed and configured in the SPSRC.

3.1 *OpenStack Deployment*

OpenStack is one of the most active open-source projects in the world, being one of the essential components for the creation of private and public Cloud infrastructure.[42] OpenStack has become the *de facto* standard for implementing cloud computing platforms through which to deliver infrastructure as a service (IaaS) that will, in its turn, finally deliver the entire stack of services related to platform as a service and software as a service in an extremely flexible and reliable way. With OpenStack, pools of computing, storage, and network resources are governed, with virtually infinite scaling capabilities. This is done via a command interface, API, and a control panel, which facilitate all its management.

We deployed the OpenStack version Train[43] within the computing, network, and storage infrastructure of the SPSRC described in Sec. 2. An important consideration when deploying this IaaS is the possibility of future scaling of resources and especially of managing them in a comfortable way, given the number of components. Thus, we use Kolla and Kayobe[44] as building and deployment tools for OpenStack.

These tools allow the deployment of OpenStack in containers on top of on-premise machines, as shown in Fig. 2. The advantage of this type of deployment is clear[44]: (a) containers (such as docker[45]) offer an attractive solution to isolate and distribute all OpenStack services throughout our SPSRC infrastructure efficiently, allowing to have production and test environments on the same infrastructure managing different versions of services without interfering with each other, (b) they enable a fast provision of nodes from a single startup host, in this way, the entire

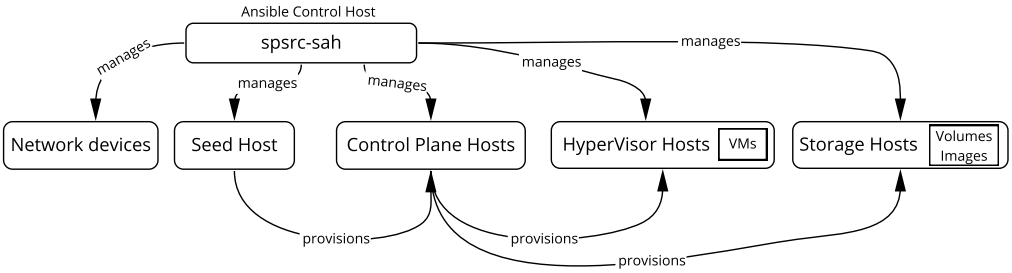

Fig. 2 Kayobe deployment for OpenStack.

orchestration of services and hosts is centralized, and (c) it is recommended to build systems designed around performance, sustainability, reliability and scalability, basic principles of robust container deployment, and key features for the design pillars of this SPSRC. Kayobe uses Ansible[46] to perform system configuration and then invokes Kolla–Ansible to perform the control plane service deployment. The latter also oversees the management of every level of the system stack, as well as the hardware life-cycle. With the Ansible playbooks and environment files in source control, a comprehensive infrastructure-as-code[47] system is delivered for a fast provisioning. The source code for the SPSRC deployment is hosted and maintained in a private Gitlab repository.

3.2 *OpenStack Architecture*

To tackle the design of the OpenStack deployment architecture, three types of requirements have been taken into account: project constraints (computing and storage), operational requirements (networks, service level agreements, and maintenance), and high availability (data, services, and network). Based on these requirements, a logical architecture was designed in which the following core services have been deployed within OpenStack:

(a) Computing:

- Nova: service that enables the provision of computing instances (virtual servers).
- Magnum: provides container orchestration engines for container deployment and management.

(b) Storage:

- Cinder: block storage service to deliver volumes to VM deployed with Nova.
- Glance: component to upload and discover data assets such as operating system images.
- Manila: shared file system service to provide a file sharing platform as a service.

(c) Network and authentication:

- Neutron: services and tools to provide network connectivity as a service between interface devices and the rest of the OpenStack services.
- Keystone: a component in charge of managing authentication and authorization of services and users.

(d) Service level agreements and maintenance

- Monasca: provides a highly scalable, performance, and fault tolerant monitoring-as-a-service solution. It also offers an extensible platform for advanced monitoring of all OpenStack services.
- Horizon: enables a web interface for the control panel of all user-manageable services, such as nova, neutron, cinder, glance, and others.

In addition to these core services being deployed, other services necessary to support and monitoring the architecture have been used:

- Core support services
  - MariaDB Galera Cluster: a database for high availability.
  - RabbitMQ: a message passing broker service for service intercommunication.
  - Memcached: data caching for distributed memory.
  - HAProxy: a tool for load balancing and high availability of services.
- Monitoring components
  - Prometheus: a monitoring and alerting toolkit.
  - Grafana: a query, visualization, and alert GUI to summarize infrastructure data.
  - Kibana: a search and data visualization tool for data indexed in Elasticsearch.

Figure 3 shows the set of OpenStack services deployed within the SPSRC infrastructure. Each of the blocks shown in the diagram are logical elements of OpenStack deployment. Each of these blocks (storage, monitoring, computing, and controllers) are a group of physical nodes with a set of replicated services (cinder, manila, nova, etc.), which supports high availability to the entire infrastructure, thus not having a single point of failure. The diagram also shows the two networks currently active: 1 GbE for infrastructure management and monitoring and 100 GbE for high speed communication between computing and storage. Figure 1 shows the layout of all these nodes inside the SPSRC rack.

3.3 *OpenStack Massive Storage Deployment with Ceph*

A main pillar of the SPSRC cloud computing infrastructure is storage and archive management. The explosive demand for data storage for the type of information processing that SKA will produce requires distributed storage infrastructures with near-Exascale[48] capacity, with agility in retrieving and archiving data and extreme reliability. The planning of mass storage in OpenStack is important because of its impact on the available budget. Consequently, it will put budgetary limits on the rest of the components in the infrastructure.

OpenStack with Ceph as the storage provider supports three types of continuous storage: block storage, object storage, and shared file systems. The SPSRC OpenStack setup supports block storage and shared file systems.

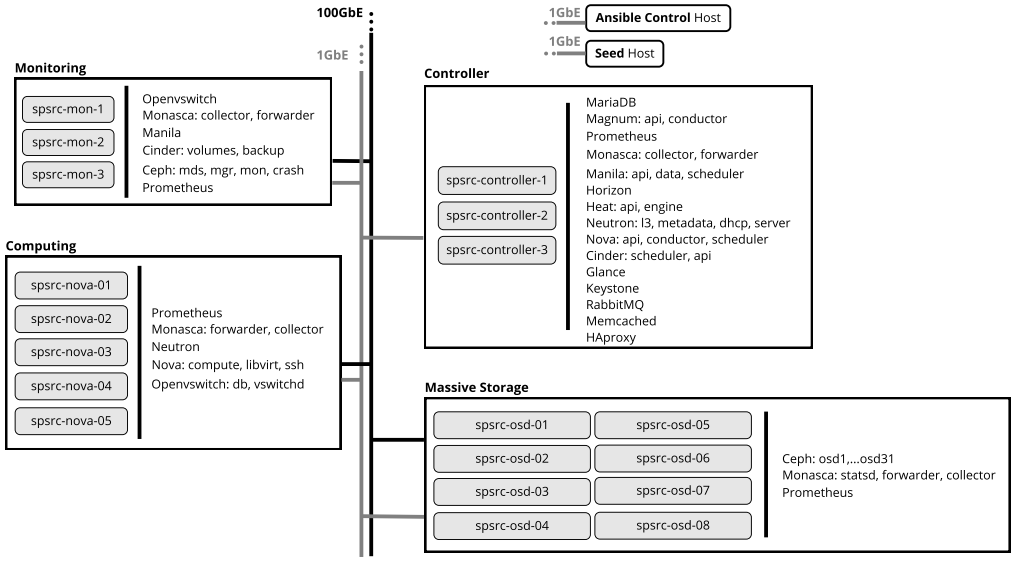

Fig. 3 OpenStack architecture, logical design, nodes, network, and services.

As discussed in the hardware section (see Sec. 2), the storage platform for OpenStack is deployed on Ceph, and the SPSRC is currently using version Octopus[49] (version 15.2.9). Ceph is a massively scalable, open, software-defined, storage system running on commodity hardware and focused on cloud storage and emerging data-intensive workloads. Ceph provides object, block, and file storage with a management console, a graphical user interface with cluster visualization, monitoring, and diagnostic information, as well as usage and performance statistics, and it is perfectly integrated with OpenStack.

The physical configuration of the storage platform with Ceph is composed of three Ceph monitors (mons) and eight Ceph storage nodes (also known as Object Storage Daemons, OSDs). First, we detail the Ceph OSD nodes and then the Ceph monitors. The Ceph OSD nodes are referred to as spsrc-osd-[00-08], as shown in Fig. 3. Thus, four of the nodes have 31 SSD disks, each with a capacity of 7.6 TB (1 drive writes per day (DWPD), i.e., many reads and few writes), and the remaining four nodes also have 31 SSD disks but with a capacity of 3.8 TB each (three DWPD, enabled for write durability), reaching a gross storage capacity of 1413.4 TB (1.4 PB). Each of these nodes includes a service called ceph-osd for each individual available SSD OSD, giving us a total of 248 OSD services (31 OSDs per node and eight nodes).

In terms of redundancy in Ceph, it was configured with a two-way replica option, which means that there are two copies of everything in the storage. This translates to a reduction of the total raw space of 1.4 PB to half, but on the other hand, we gain in reliability in terms of data availability.

To coordinate Ceph management and provide high availability for this storage, the infrastructure has three Ceph monitor nodes called spsrc-mon-[1-3] (Fig. 3, monitoring). These three nodes provide multiple services such as ceph-mgr, ceph-mds, ceph-mon, and ceph-crash to manage the cluster, to manage CephFS file systems, and to monitor and manage errors, respectively.

Ceph storage is perfectly integrated into OpenStack.[50] As a result, four pools (i.e., logical partitions for storage) have been created, taking advantage of the capacity and performance provided by the different types of disks. Currently, the system has the following pools:

- Manila: these pools have a 1DWPD-configuration (many reads and few writes). They provide share file systems type storage, which allows multiple VMs to mount a shared file system, similar to what would be NFS or HDFS file systems.
- Cinder-volumes: These have a 3DWPD configuration since the pools will be used to provide massive block storage space for VMs.
- Cinder-backups: Provides a storage service for block storage backups, with a 1DWPD configuration.
- Glance: This pool allows storing all the images of operating systems and assets with a 1DWPD configuration.

4 Cloud Services

4.1 *On-Demand Computing Resources*

The SPSRC cloud computing infrastructure allows researchers to use different configurations of VMs on demand. Several critical points define each configuration, in particular:

- Operating system: We have enabled a selection of the most popular Linux operating systems in the astronomy and astrophysics context. Via the OpenStack's Glance service, we offer the following Linux distributions to be used within VMs: CentOS 7/8, Ubuntu 18.04/20.04, and Debian 9/10.
- VM resources configuration: In OpenStack, flavors define the computing, memory, and storage capacity of the computing instances for the Nova service. A flavor is a hardware configuration available to a server that defines the size of a virtual server that can be launched. Figure 4 shows the set of configurations available to create VMs. As shown in the figure, three types of environments can be deployed, (a) general purpose VMs, with very flexible configuration possibilities ranging from light workload VMs to high load and heavy workload configurations, (b) VMs for SKA data challenge environments, with medium and high workloads, and (c) VMs for cluster compute nodes, such as Kubernetes

Name	CPU	RAM	Local HD	Target/Capabilities
spsrc.c2m4	2	4GB	50GB	Light load
spsrc.c4m8	4	8GB	50GB	
spsrc.c8m32	8	32GB	50GB	Multi-purpose
spsrc.c16m64	16	64GB	50GB	
spsrc.c24m194	24	194GB	50GB	
spsrc.c40m1000	40	1000GB	50GB	High load
sdc2.c16m64	16	64GB	100GB	Data challenges
sdc2.c32m128	32	128GB	100GB	
multihub.c10m24	10	24GB	50GB	Clusters
multihub.c16m32	16	32GB	50GB	

Fig. 4 List of flavors and features within OpenStack.

or queue management systems such as Slurm. The flavor with which a VM is created is not a limitation because, if required, the flavor can be changed easily at any time. That is, the resources available to a VM can be scaled up or down depending on changing project necessities and the server load.

- Mass storage: By default, all VMs have their root disk (i.e., the virtual hard drive where the operating system is installed) placed on the local SSD of a hypervisor (the VM monitor that runs the VM) to enable maximum performance. In addition to the default storage of each VM instance, it is possible to include two types of storage:
  - Block storage (cinder.volumes): This type of storage is the most common for downloading and working with large data sets. It allows users to create volumes of any size, from megabytes to terabytes, that are mounted on VMs. This type of storage can be detached from one instance and reattached to another while the data remain intact, although it cannot be shared by more than one instance. It does, however, allow the use of extralarge volumes to store data and offers an excellent read/write performance.
  - Shared file system storage (manila.ceph): This storage model is available for use in VMs that want to access a shared data or software repository or a service marketplace that can be shared by different users and services. The main advantage is that this type of storage can be shared by many VMs, which gives a great flexibility in working collaboratively. All VMs that are created can include a number of shared file system spaces for different purposes. In particular, we have deployed a shared file system to store a catalog of container images with software ready to use by any of the projects hosted in the SPSRC (Sec. 4.4).
- Networking: All VMs have a public IP and their own "fully qualified domain name (FQDN)", so they are directly accessible from the internet through a domain, following this pattern: vm_name.iaa.csic.es, for example, spsrc01.iaa.csic.es (Fig. 5). VM users belonging to the group spsrc-group can set up services within a predefined and strict range of ports, where only a subset of ports are open to outside.
- Access mode to VM resources: Each VM can be accessed through two different protocols: SSH and RDP. By default, all VMs provide SSH access to users through public/private keys. In addition, Apache Guacamole is installed on the VMs. This tool provides a clientless remote desktop gateway. That is, it offers streamlined connection to VMs from any browser, supporting standard protocols, such as VNC, RDP, and SSH.

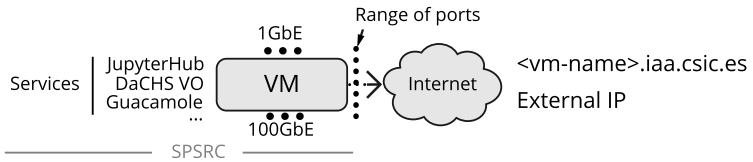

Fig. 5 Network and external connectivity diagram.

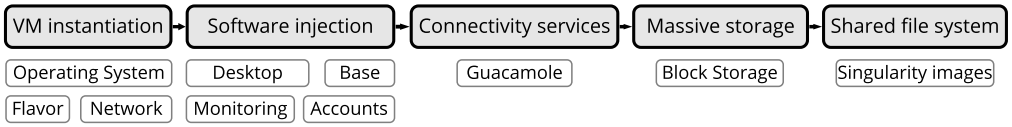

Fig. 6 The process of instantiating a VM in the SPSRC infrastructure.

- Core software and science services: Two sets of software are loaded onto each VM during creation. One is the software of the selected operating system, which includes the basic services, access/management accounts, monitoring services, and a desktop that are installed using Ansible. The other set consists of services for science such as JupyterHub and a container hub of containerized astrophysics and astronomy software as well as services that are installed on demand by the user. These last two service types are detailed in the following subsections.

Figure 6 shows the general process of creating a VM, including the stages previously detailed.

4.2 *On-Demand Elastic Clusters*

The SPSRC infrastructure is a flexible resource, ready to deliver multiple configurations to support terabytes of storage or offer data analysis tools and workflow services. To this end, OpenStack allows elastic clusters, i.e., multiple independent clusters to be deployed on-demand. Currently, the infrastructure hosts a Kubernetes cluster with a configuration consisting of 10 cores and 24 GB of RAM per node (from 2 up to 20 nodes) and supports several tens of JupyterHub instances running at the same time. Similarly, with this configuration, it would be possible to enable the OpenStack infrastructure with on-demand and flexible clusters[51,52] with Slurm or HTCondor, for science-intensive jobs.

4.3 *JupyterHub*

A JupyterHub[53] service makes it possible for users to publish and share their codes, with rich annotations, and execute them interactively. Two types of JupyterHub setups are available in the SPSRC. The first option allows any VM user to launch their individual copy of JupyterHub for all available resources, such as mass storage, data sharing between users, and integration with other container images to run within the JupyterHub interactive process. Alternatively, users can request an on-demand cluster on Kubernetes (see Sec. 4.2) that initiates JupyterHub services. Currently, we have a cluster supporting three types of users (see more details in Sec. 7.1): SPSRC/IAA-CSIC (internal, institute, and projects users), European Science Cluster of Astronomy (ESCAPE), and participants in schools, such as SOMACHINE2020 and SOMACHINE2021, where features and performance of the profiles are customizable to adapt to the demand for computing.

4.4 *ContainerHub*

The ContainerHub is a service available for all created VMs and clusters. The ContainerHub is a shared catalog of applications and services for radio/astronomy that currently provides more than 30 ready-to-use images from Singularity,[54] a container service focused on reproducibility and execution of scientific work. With the ContainerHub service, VM instances can execute this software without having to install virtual environments or modify its session while respecting the configuration of libraries and software versions. One of the advantages of providing a centralized and containerized software catalog is that from the SPSRC management, it is possible to include and update images in all running instances, for all the users simultaneously and in real time. The catalog available at the SPSRC currently contains updated images built for specific projects. Among these images are stimella,[55] LOFAR pipelines, the set of radio astronomical software packages KERN suite, and different versions of CASA. The hub will grow with new

software and updated versions as required by the hosted projects, being immediately available to the whole community of users. This single place for images also optimizes disk space, avoiding duplicates of software in different VMs.

5 Science Archive

As stated in Sec. 1, the SKA will generate the largest amount of radio-astronomical data to date and the data products coming out of the observatory will be ingested into the SRC network at a rate of 600 PB/year. The SKA science archive is currently being designed as a distributed but logically centralized system that aims to provide fully functional and continuous capabilities.[37] The size and complexity of the SKAO data will necessarily impact the design of the archive and its connection to other subsystems that enable science analysis. The archive shall provide enough capabilities to search and explore data directly, to minimize data movement through the SRC network. It is primordial fulfilling the FAIR[56] principles and being compliant with the virtual observatory recommendations. The SKA science archive working group of the SRCSC is responsible for the definition of the archive requirements as well as for the design and prototyping of its main components. At the time of writing, this group is led by the Spanish and Canadian members of the SRCSC, and it consists of 35 people, which also includes a core group of 18 experts from Australia, France, Italy, The Netherlands, Portugal, SKAO, Spain, Sweden, and the United Kingdom.

In this section, we focus on the activities related to the SPSRC science archive, which comprises the storage space to preserve raw and processed data and the subsystems for searching and accessing the data. Considering that the SKA Observatory is starting the construction phase this year (2021), the SPSRC archive is initially being designed to satisfy the requirements of SKA precursors/pathfinder science cases, with full ability to scale up for SKA. The handling and sharing of the expected amount of data with the (authorized) scientific community demands careful planning on hardware characteristics and which services will be implemented to access archived resources.

As described in Sec. 2, we conducted a series of interviews to define the detailed requirements for this phase. With a capacity of 400 TB of net file storage dedicated to the archive, the hardware characteristics described in Table 1 will allow for the required I/O rates to access and process the data from the SPSRC computational environment. High I/O rate is one of the most important requirements, due to the large size of the data, which consequently impacts data processing times and the feasibility of deploying analysis services on top of the archive. The SPSRC archive will include advanced services on top of the infrastructure, such as user-archive interface services (supporting authentication, authorization, and querying), data and metadata visualization, cut-out services, a web portal, access from virtual observatory services, APIs (i.e., via scripts), etc. Some of these services require a significant amount of computational resources due to the nature of the data from the SKA precursors. The resilience of the subsystem is another key aspect that the SPSRC has addressed by including different levels of redundancy, i.e., at hardware and storage levels.

The subsystems to allow searching and accessing the data are an ongoing development. After testing existing virtual observatory libraries and software, we are currently deploying DaCHS in the infrastructure. DaCHS is an integrated framework for building web services connected to the Virtual Observatory ecosystem while supporting the entire workflow from ingestion to data mapping to service definition.[57] The SPSRC will host a series of services that will be registered at the European Virtual Observatory (EURO-VO) registry to ensure that they are discoverable from the astronomy software ecosystem. (IVOA Registries have information of all Virtual Observatory known web services. See Ref. 58.)

In particular, the SPSRC will provide access to catalogs using web services that enable SQL-like queries[59] to tabular data. These kinds of services can be queried from Virtual Observatory desktop software[60,61] and from libraries accessing Virtual Observatory services.[62,63] Images and data cubes will be findable from services compliant with the IVOA Simple Image Access recommendation.[64] In addition, we are developing specific services to provide access to derived radio data such as cut-outs (subsets of full data sets), by following the IVOA Server-side

Operations for Data Access (SODA) recommendation,[65] a standard to develop services that perform operations on the data. Catalogs and image services will be implemented on the DaCHS framework whereas the SODA services will be developed separately. Interoperability with the rest of astronomy archives is fundamental to create a transversal facility that aims to enhance research and facilitate multimessenger science based on Open Science/FAIR practices.

As part of the SPSRC Science Archive developments, we are contributing to the development of the archive for the ASKAP WALLABY Hi survey.[66] This work is undertaken by its technical working group 7, and several copies of this archive are planned to be hosted at the SPSRC, Australian SRC, and in Canada. The WALLABY team has extended the capability of SoFiA (source finding software for three-dimensional spectral line data) by developing an automated, parallel HI source-finding pipeline for the WALLABY survey (SoFiA 2).[67] SoFiAX executes SoFiA-2 and automatically merges, resolves, and inserts extracted sources into a database for further inspection and analysis. The resulting cubelets, moment maps, or spectral data are accessible via a table access protocol (TAP) interface. At the time of writing, we are experimenting with a containerized DaCHS deployment.[57]

Considering that SKA will not generally preserve raw data due to size issues, it is especially relevant to understand the processes and data transformations happening in the SKA SDP. With the aim to provide archive services that contribute to a better understanding of the data, we have studied how provenance data models can be applied to the SKA science data products and to the derived products that will be created at the SRCs. The term provenance refers to capturing the whole process involved to obtain the ADPs, such as actions, pipelines, versions, and parameters. This action of preserving the workflows used to generate the SKA data products and their provenance is an important step toward SKA science reproducibility. In collaboration with GMV Aerospace and Defence S.A.U, we have compared the ProvONE data model,[68] the W3C data model recommendation, and the IVOA data model recommendation.[69] ProvONE and the IVOA data models are both based on the W3C standard. The ProvONE extension was designed for a better description of scientific workflows and in particular its computational processes, whereas the IVOA recommendation is aiming to bring the provenance data model to the astronomy domain by adding new elements that are domain-specific. As a result of this study, we have concluded that the IVOA provenance data model can be applied to the SKA [70] in a similar way to its application to a Cherenkov Telescope Array provenance implementation.[71] However, the SKA pipelines and computational workflows are so complex that there may be a benefit using a combination of two models or a modification of the IVOA recommendation.[70]

6 Support and Training

Our aim is to support preparatory activities for the SKA key science projects and the science enabled by the SKA precursors and pathfinders. Therefore, the SPSRC has allocated human resources to support the projects hosted in the platform. Support and training are key to optimize the throughput of the platform by helping users to make efficient use of the available software and hardware resources, to gather new community requirements and to train the community on how to manage the data in the new data processing paradigm required by the SKAO. The SPSRC provides support and training on radio interferometry, state-of-the-art software tools, and automation of data processing workflows, all to enhance the scientific throughput of the platform.

The user support and helpdesk service are delivered through different channels to ensure flexible and effective communication. First, we have developed documentation providing information for users and a detailed description of the most useful procedures.[72] Second, we use a dedicated SPSRC email address to deliver announcements, as well as to receive and solve questions from the users. This is complemented with a Slack workspace for the SPSRC in which we give technical and scientific support, including specific support in private channels for each project we host. When required, we also organize videocons and face-to-face meetings to discuss particular issues that users want to clarify. Last, we created a repository of singularity images with radio astronomical software to be used from the different VMs.

The most common support questions are on technical issues related to configuration and usage of the platform. We also help users to get familiar with and use the services provided,

for example, how to use Guacamole as an access point, how to use JupyterHub to execute notebooks interactively, and how to access our SPSRC repository of singularity images. From the science point of view, we help users to access data from different observatories, find resources describing data processing pipelines for different instruments, and clarify general questions on radio interferometry, calibration, or potential use of instruments. Examples of these are discussions on best practices to access and retrieve MeerKAT data or discussions on how the calibration procedures of the MeerKAT pipeline oxkat[73] work, among other topics. To motivate the use of open science tools, we also give support and provide examples on tools, such as git, Github, conda software management, Jupyter notebooks, Binder, and Zenodo.

Apart from direct support using the communication channels described above, we also organize training events to promote the use of these tools and good practices related to them for a wider community. For example, we have conducted a series of short hands-on training events called Open Science Droplets,[74] describing and showing the usage of open science tools. We are also currently preparing two training schools: one focused on open science and another one on radio interferometry. The goal of these programs is to motivate and facilitate the community to adopt more flexible, efficient, collaborative, open tools. Apart from being a useful asset for conducting their own research, these practices will enhance the quality of the future SKA science.

7 SPSRC Operations and Management

The SPSRC infrastructure is actively managed by a team with different purposes. First, to host a series of development and scientific projects that make use of those resources and services. Second, to enable the community to access the platform and get the support and training needed. Third, to ensure the daily operations of the available services. Lastly, to upgrade the software resources and to test and deploy new functionalities and services. Here, we summarize the current usage of the platform by development, scientific, and training programs as well as describing how the SPSRC is managed and organized.

7.1 *Usage of the SPSRC Services*

The ultimate purpose of setting up all the hardware, software, services, and support described above is to provide a platform to conduct research projects. The SPSRC started to host the first science projects at the end of 2020. Once the deployment of the OpenStack server was finalized, by September 2020, a commissioning period started, until the end of 2020, to benchmark the disk and CPU usage of the VMs, deploy access services, and to verify all the functionalities of those services.

During the lifetime of the platform thus far, we have hosted 20 projects: 16 requiring VMs and four requiring JupyterHub deployments. We classify the hosted projects in three categories (see Fig. 7): (1) research projects, led by the IAA-CSIC researchers to conduct data processing and analysis on different astrophysical fields; (2) development projects, which have the goal of creating or enhancing services needed in an SRC and improving the platform; (3) training projects that include data processing or analysis schools and also temporary machines to be used during other observatory schools or to verify the capability of the platform for specific purposes (see Sec. 6)

The research projects currently hosted at the SPSRC include analysis of data from the SKA precursor MeerKAT: observations of radio emission induced by star-planet interaction. We also host an e-MERLIN (SKA pathfinder) legacy project on resolved radio emission of ultraluminous infrared galaxies,[75] and a project on star-planet interaction based on LOFAR (SKA pathfinder) observations. Other research projects include, for example, the analysis of galactic radial profiles.

As some examples of development projects, we have deployed an online file repository and are developing an archive platform compatible with the virtual observatory (see Sec. 5), and we host a project to test the installation and execution of data processing pipelines for SKA precursors and pathfinders, among other projects. As part of a master's thesis, we conducted a study of the data transfer quality between different international research centers using network tests.

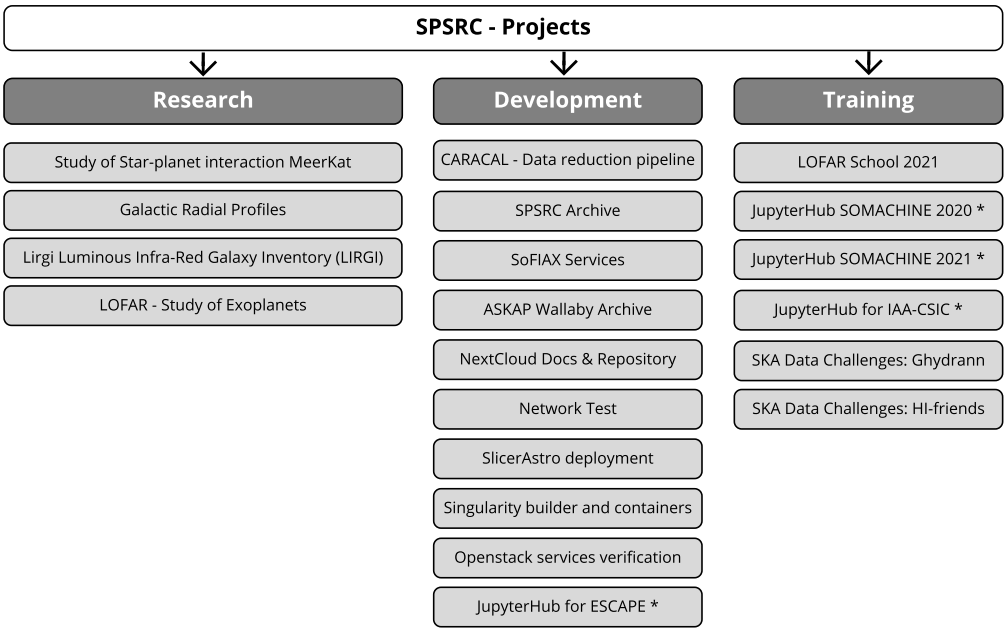

Fig. 7 Overall view of the hosted SPSRC projects.

The training projects at the SPSRC cover a wide range of services. We have hosted a JupyterHub service for two editions of SOMACHINE, which is an international school on machine learning, big data, and deep learning in astronomy. Likewise, we deployed temporary VMs to allow users to participate and execute the tutorials of the sixth LOFAR Data School. In addition, we are one of eight facilities providing infrastructure for the second SKA Data Challenge (see Sec. 1) issued by the SKA Observatory, hosting two teams.

Currently about 80% of the CPUs and memory are assigned to active projects. The usage has been continuously increasing since the beginning of the year. In terms of storage, until the archive services will become available (see Sec. 5), the main use is processing storage via block storage, with a mean assignation of 12 TB per month, on average. Current projects have required transferring and storing 110 TB of raw data. We have received requests to transfer a volume of 338 TB of raw data in the near future. Subsequent data processing usually requires disk space between one and five times the raw data volume. Apart from the block storage assigned to individual projects, we have deployed a shared storage unit that hosts a central service of software containers, currently including a catalog of singularity images with radio astronomical software ready to be used by any user from the different VMs.

7.2 *Management and Organization*

The SRC White Paper[37] provides a guideline on how to scale up the SRC capacity and functionality, and it also estimates the associated costs. It explains that the SRC network needs to achieve most of the functionality by 2025 while only 10% of the required storage/processing capacity, and it should achieve full capacity and functionality by 2030. The SPSRC is being developed considering these guidelines. Its development was fostered in July 2018 since the IAA-CSIC was accredited as a "Severo Ochoa" Center of Excellence by the Spanish Ministry of Science, Universities and Research, after an evaluation by an international committee of its research activity and the presentation of a strategic plan for the period 2018 to 2021. This plan included the creation of an SRC prototype. This marked a major milestone in the roadmap to develop the SPSRC, as well as the SKA Office in Spain. The SPSRC has already been presented at national and international level including contributions in the "SKA in Spain!" meeting (2019), the "SEA 2020 meeting" (2020), the "ADASS XXX" (2020), "2021 SKA Science Conference A precursor view of the SKA Sky" (2021), among others.

At the time of writing, the SPSRC is independently funded by the Severo Ochoa program and complementary competitive regional and national funds. In particular, the existing infrastructure (see Sec. 2) was acquired using competitive funding (∼700 k), and we are currently planning to scale up the platform, if the additional funding already requested is secured.

The development of the SPSRC is led by the IAA-CSIC, and the team is coordinated by the Spanish representative at the SRCSC. In close cooperation with the IAA-CSIC team responsible of the coordination of the Spanish participation in SKA, the SPSRC team disseminates the SPSRC capabilities, gathers community requirements, manages the platform usage by establishing scientific priorities and human resources availability, organizing the support astronomers work and providing a link with the technical personnel. The technical coordinator is responsible for supervising the computational platform installation and maintenance, defining the services offered in coordination with the SRC Network, the SKAO, as well as with other related initiatives (e.g., European projects, SKA precursors collaborations, etc). The SPSRC infrastructure was deployed in the Computing Center at the IAA-CSIC with the support of the head of the facility. In addition, the External Advisory Board of the IAA-CSIC, formed by international experts who review IAA-CSIC activities and provide advice for future actions, informed very positively on the development of the SPSRC in its last report (2020).

The current team is composed of eight FTEs combining radio astronomers, astroinformatics, and software engineers with expertise in open science, system administration, OpenStack management, radio interferometry, machine learning, artificial intelligence, and outreach.

8 Conclusions and Future Work

Preparatory activities for the SRCs started in 2016 with the SRC Coordination Group, and they increased with the AENEAS and ERIDANUS projects. In Spain, the development of a prototype is led by the IAA-CSIC. One of the first activities was to gather requirements from a variety of science cases (especially those related to SKA precursors and pathfinders). In 2019, we interviewed seven groups as a representative sample of use cases. Then, we designed the SPSRC computing infrastructure to satisfy the requirements identified in those interviews, together with the SRC Coordination Group and AENEAS requirements. As a result, the SPSRC infrastructure is flexible and can redistribute resources. In addition, it can be adapted to heterogeneous projects and will support the science with SKA precursors and pathfinders. These abilities are key to identify (and solve) potential bottlenecks in the future SKA data processing. Moreover, the scalability offered by this infrastructure will allow increasing the SPSRC resources steadily until fulfilling the computing requirements for the next phases of the project and those that a fully operational SKA will introduce by (around) the year 2029.

The set-up of the SPSRC computing infrastructure was finalized in July 2020, followed by a commissioning period that lasted until the end of that year. During 2021, we gave progressively access to different research groups to deploy their own software and at the same time services like we deployed JupyterHub on the infrastructure. In the lifetime of the platform, it has hosted 20 projects (including research, development, and training projects) that have required 16 VMs and four JupyterHub services.

The services offered by the SPSRC are oriented to facilitate collaboration with a globally distributed community and to enable the best possible scientific practices. The SPSRC team is making a special effort to develop a platform that enables scientists to follow open science practices. This effort is channeled through three main areas:

- Engaging with initiatives focused on investigating ways for supporting open science, such as the ESCAPE project, the collaboration with GMV Aerospace and Defence S.A.U, and the participation in the IVOA.
- Populating the SPSRC platform with tools able to support open science. In particular, the SPSRC platform includes a JupyterHub service as well as a container image repository. In addition, a science archive is being implemented taking into account the standards and models published by IVOA.

- Training in open science practices. In addition to the open science droplets training events, given that the SPSRC platform is focused on supporting open science, the training on SPSRC tools will contribute to empower the community with the skills required to analyze SKA data following the FAIR principles.

The establishment of a Spanish SRC will play an essential role in providing the scientific community with access to SKAO data, as well as processing capacity and tools to generate and analyze ADPs. This will be done in collaboration with other SRCs and as a consequence, relevant topics for future work and research can be derived from the collaboration to be established within the framework of the SRC network. This collaboration, in the coming years, will make it possible to interconnect and interoperate all the individual SRCs, offering a global distributed and fully resilient service. The SPSRC team will carry out open science and radio astronomy training activities to bring best practices to the community. As well, we will offer support to precursor and pathfinder projects as primary activities during the interim. A key improvement of the SPSRC infrastructure will be the inclusion of a data transfer node, which will enhance data transfers and connectivity to the SKAO and SRC network. Another key objective will be to complete the science archive development, with core services that will allow access to the data through virtual observatory protocols. Apart from these, other improvements will be considered, such as the provision of a marketplace service for the creation of VMs, the allocation of storage space, or the installation of software and services on demand, as well as the use of workflow management services such as Slurm.

Acknowledgments

We acknowledge financial support from the State Agency for Research of the Spanish Ministry of Science, Innovation and Universities through the "Center of Excellence Severo Ochoa" awarded to the Instituto de Astrofísica de Andalucía (SEV-2017-0709) and from the Grant no. RTI2018-096228-B-C31 (Ministry of Science, Innovation and Universities/State Agency for Research/European Regional Development Funds, European Union). In addition, we acknowledge financial support from the Ministry of Science, Innovation and Universities and the European Regional Development Funds (EQC2019-005707-P) and the Regional Government of Andalusia (SOMM17-5208-IAA-2017). LVM, JG, SSE and SLV acknowledge The European Science Cluster of Astronomy and Particle Physics ESFRI Research Infrastructures project that has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 824064. We would like to explicitly acknowledge Dr. José Ruedas, head of the Computer Centre at IAA-CSIC, for his technical assistance. We acknowledge the contribution from Theresa Wiegert. LVM, SSE, and SLV acknowledge financial support from the Grant No. COOPB20448 (Spanish National Research Council Program of Scientific Cooperation for Development i-COOP+2019). LVM, JG, and JM acknowledge financial support from the Grant No. RED2018-102587-T (Spanish Ministry of Science, Innovation and Universities/State Agency for Research). LVM, JG, SSE, JM acknowledge financial support from the Grant No. IAA4SKA P18-RT-3082 (Regional Government of Andalusia). LVM acknowledges financial support from the Ministry of Science and Innovation, from the budgetary line 28.06.000x.430.09 of the General State Budgets of 2021, for the coordination of the participation in SKA-SPAIN. LD acknowledges financial support from the Grant No. PTA2018-015980-I (Ministry of Science, Innovation and Universities and the Spanish National Research Council). MP acknowledges financial support from the Grant No. 54A Scientific Research and Innovation Program (Regional Council of Economy, Knowledge, Business and Universities, Regional Government of Andalusia and the European Regional Development Funds 2014-2020, program D1113102E3).

References

1. R. Braun et al., "Advancing Astrophysics with the Square Kilometre Array, Vol1-2," PoS AASKA14 (2015).

2. SKA Organization, "SKA Phase 1 Construction proposal," 2020, http://www.skatelescope.org/wp-content/uploads/2021/02/22380_Construction-Proposal_DIGITAL_v3.pdf
3. A. Chrysostomou et al., "Operating the Square Kilometre Array: the world's most data intensive telescope," *Proc. SPIE* **11449**, 114490X (2020).
4. E. Hatziminaoglou et al., "The European ALMA regional centre network: a geographically distributed user support model," *Messenger* **162**, 24–29 (2015).
5. S. Sánchez-Expósito et al., "Web services as building blocks for science gateways in astrophysics," *J. Grid Comput.* **14**(4), 673–685 (2016).
6. J. Sabater et al., "Calibration of LOFAR data on the cloud," *Astron. Comput.* **19**, 75–89 (2017).
7. R. Bolton et al., "SKA Regional Centre requirements," SKA Organisation SKA-TEL-SKO-0000735 rev3 (2019).
8. AENEAS, "Advanced European Network of E-infrastructures for Astronomy with SKA," 2019, https://www.aeneas2020.eu.
9. V. Galluzzi et al., "*Recommendations on the design of user interfaces for data discovery, access, and retrieval for the ESDC*," AENEAS H2020 Project-731016, Deliverable 5.2 (2018).
10. A. Costa et al., "*Recommendations on the design of user interfaces for data processing, reprocessing, analysis and visualization for the ESDC*," AENEAS H2020 Project-731016, Deliverable 5.4 (2018).
11. R. Hughes-Jones and D. Vicinanza, "*Report on data transport tests and recommendations*," AENEAS H2020 Project-731016, Deliverable 4.5 (2019).
12. ERIDANUS, "Exascale Research Infrastructure for Data in Asia-Pacific astroNomy Using the SKA," 2021, https://eridanus.net.au.
13. CIRADA, "The Canadian Initiative for Radio Astronomy Data Analysis," 2020, https://cirada.ca/.
14. T. An, X.-P. Wu, and X. Hong, "SKA data take centre stage in China," *Nat. Astron.* **3**, 1030–1030 (2019).
15. ILIFU, "Cloud Computing for Data-Intensive Research," 2021, http://www.ilifu.ac.za.
16. Pawsey Computing Centre, "The Pawsey Supercomputing Centre," 2021, https://pawsey.org.au/.
17. J. J. Kavelaars et al., "Digital research infrastructure in astronomy," White paper identifier W062 (2019).
18. M. Huynh et al., "The CSIRO ASKAP science data archive," *Astron. Soc. Pac. Conf. Ser.* **522**, 263 (2020).
19. A. Bonaldi et al., "Square kilometre array science data challenge 1: analysis and results," *Mon. Not. R. Astron. Soc.* **500**, 3821–3837 (2020).
20. SKA-Observatory, "SKA science data challenges," 2020, https://sdc2.astronomers.skatelescope.org/about-the-challenges.
21. SKA-Observatory, "Teams ready for SKA science data challenge 2," SKAO Magazine Contact 6 (2020).
22. UNESCO, "Draft text of the UNESCO recommendation on open science," SC-PCB-SPP/2021/OS-IGM/WD3, 2021, https://unesdoc.unesco.org/ark:/48223/pf0000376893.
23. SKA Organization, "SKAO establishment and delivery plan," 2020, https://www.skatelescope.org/wp-content/uploads/2021/02/22380_SKA_Est-Delivery-Plan_DIGITAL_v3.pdf.
24. R. C. Bolton and the SRCCG, "SKA Regional Centre requirements," SKA-TEL-SKO-0000735-v1, 2017, https://astronomers.skatelescope.org/wp-content/uploads/2017/10/SKA-TEL-SKO-0000735_01-SKA_Regional_Centre_Requirements.pdf.
25. ESCAPE, "European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures," 2021, https://projectescape.eu/.
26. Z. Meyer-Zhao et al., "ESCAPE Deliverable 5.1. Preliminary report on requirements for ESFRI Science analysis use cases," 2019, https://projectescape.eu/deliverables-and-reports/d51-preliminary-report-requirements-esfri-science-analysis-use-cases.
27. M. Pérez-Torres et al., Eds., *The Spanish Square Kilometre Array White Book*, Sociedad Española de Astronomía (2015).

28. J. Garrido et al., "AstroTaverna: tool for scientific workflows in astronomy," 2013, https://ui.adsabs.harvard.edu/#abs/2013ascl.soft07007G.
29. J. Ruiz et al., "AstroTaverna-building workflows with virtual observatory services," *Astron. Comput.* **7**, 3–11 (2014).
30. K. M. Hettne et al., "Structuring research methods and data with the research object model: genomics workflows as a case study," *J. Biomed. Semant.* **5**, 41 (2014).
31. M. Van Haarlem, "European SKA Regional Centre requirements," AENEAS H2020 Project-731016, Deliverable 2.2 (2019).
32. R. Bolton and A. Chrysostomou, and the SRCCG, "A model for SKA Regional Centre global size estimation, SKA Organisation SKA-TEL-SKO-0001000," 2018, https://astronomers.skatelescope.org/wp-content/uploads/2019/05/SKA-TEL-SKO-0001000-01_AModelforSKARegionalCentreGlobalSizeEstimation-signed.pdf.
33. A. M. M. Scaife et al., "*Analysis of compute load, data transfer and data storage anticipated as required for SKA Key science*," AENEAS H2020 Project-731016, Deliverable 3.1 (2018).
34. A. M. M. Scaife et al., "*Initial system sizing*," AENEAS H2020 Project-731016, Deliverable 3.3 (2018).
35. A. M. M. Scaife and M. Ashdown, and WP3 Team, "Report on design & costing for ESDC," AENEAS H2020 Project-731016, Deliverable 3.4 (2019).
36. R. Hughes-Jones and D. Vicinanza, "*Architecture and cost model for the European ESRC network*," AENEAS H2020 Project-731016, Deliverable 4.3 (2019).
37. P. Quinn et al., "SKA regional centres white paper," SKA Regional Centre Steering Committee (2020).
38. E. S. Reddy, K. P. Kumar, and H. Shahnasser, "A TestBed for deployment of datacenter switches for training purposes," *J. Comput. Commun.* **5**(03), 129 (2017).
39. J. Dugan et al., "iPerf – speed test tool for TCP, UDP and SCTP," 2020 https://iperf.fr/.
40. A. Mesa, "Global data transfer survey for SKA Interferometer data processing nodes," Master's thesis, University of Granada (2020).
41. O. Sefraoui, M. Aissaoui, and M. Eleuldj, "OpenStack: toward an open-source solution for cloud computing," *Int. J. Comput. Appl.* **55**, 38–42 (2012).
42. J. B. Raja and K. V. Rabinson, "IAAS for private and public cloud using Openstack," *Int. J. Eng.* **5**(4), 99–103 (2016).
43. OpenStack project, "OpenStack train: OpenStack open source cloud computing software," 2020, https://www.openstack.org/software/train/.
44. Kayobe and OpenStack, "openstack/kayobe," 2021, https://github.com/openstack/kayobe.
45. D. Merkel, "Docker: lightweight Linux containers for consistent development and deployment," *Linux J.* **2014**(239), 2 (2014).
46. W. Bentley, *OpenStack Administration with Ansible*, Packt Publishing Ltd. (2016).
47. M. Artac et al., "DevOps: introducing infrastructure-as-code," in *IEEE/ACM 39th Int. Conf. Software Eng. Companion (ICSE-C)*, IEEE, pp. 497–498 (2017).
48. A. M. M. Scaife, "Big telescope, big data: towards exascale with the Square Kilometre Array," *Philos. Trans. R. Soc. A* **378**(2166), 20190060 (2020).
49. "Ceph octopus," 2020, https://docs.ceph.com/en/latest/releases/octopus/.
50. X. Zhang, S. Gaddam, and A. T. Chronopoulos, "Ceph distributed file system benchmarks on an openstack cloud," in *IEEE Int. Conf. Cloud Comput. in Emerg. Markets (CCEM)*, IEEE, pp. 113–120 (2015).
51. B. Bockelman, J. C. Bejar, and J. Hover, "Interfacing HTcondor-CE with Openstack," *J. Phys.: Conf. Ser.* **898**(9), 092021 (2017).
52. P. Llopis et al., "Integrating HPC into an agile and cloud-focused environment at CERN," in *EPJ Web Conf.*, EDP Sciences, Vol. 214, p. 07025 (2019).
53. T. Kluyver et al., "Jupyter notebooks: a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds., pp. 87–90, IOS Press (2016).
54. G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: scientific containers for mobility of compute," *PLoS One* **12**(5), e0177459 (2017).

55. S. Makhathini, "Advanced radio interferometric simulation and data reduction techniques," PhD thesis, Rhodes University (2018).
56. M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data* **3**, 160018 (2016).
57. M. Demleitner et al., "Virtual observatory publishing with DaCHS," *Astron. Comput.* **7-8**, 27–36 (2014).
58. IVOA, "IVOA Registry of Registries," 2021, http://rofr.ivoa.net.
59. P. Dowler et al., "Table access protocol Version 1.1," IVOA Recommendation 27 September 2019 (2019).
60. M. B. Taylor, "TOPCAT & STIL: Starlink Table/VOTable processing software," *Astron. Soc. Pac. Conf. Ser.* **347**, 29 (2005).
61. F. Bonnarel et al., "The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources," *Astron. Astrophys. Suppl. Ser.* **143**, 33–40 (2000).
62. A. Ginsburg et al., "Astroquery: an astronomical web-querying package in Python," *Astron. J.* **157**(3), 98 (2019).
63. S. Becker and M. Demleitner, "TAP support in PyVO," *Astron. Soc. Pac. Conf. Ser.* **521**, 483 (2019).
64. P. Dowler, F. Bonnarel, and D. Tody, "IVOA simple image access Version 2.0," IVOA Recommendation 23 December 2015 (2015).
65. F. Bonnarel et al., "IVOA server-side operations for data access Version 1.0," IVOA Recommendation 17 May 2017 (2017).
66. B. S. Koribalski et al., "WALLABY: an SKA pathfinder H i survey," *Astrophys. Space Sci.* **365**(7), 118 (2020).
67. T. Westmeier et al., "SoFiA 2: an automated, parallel HI source finding pipeline for the WALLABY survey," *Mon. Not. R. Astron. Soc.* **506**, stab1881 (2021).
68. A. S. Butt and P. Fitch, "Provone+: a provenance model for scientific workflows," *Lect. Notes Comput. Sci.* **12343**, 431–444 (2020).
69. M. Servillat et al., "IVOA provenance data model Version 1.0," IVOA Recommendation 11 April 2020 (2020).
70. J. Garrido-Sánchez and S. Sánchez-Expósito, "Study of provenance data for the SKA science dataprocessor," in *IVOA May 2021 Interoperability Meeting (IVOA interop 2021)* (2021).
71. M. Sanguillon et al., "IVOA provenance data model: hints from the CTA provenance prototype," *Astron. Soc. Pac. Conf. Ser.* **512**, 581 (2017).
72. L. Darriba et al., "SKA Regional Centre Prototype at the IAA-CSIC User Documentation," 2021, https://spsrc-user-docs.readthedocs.io.
73. I. Heywood, "oxkat: semi-automated imaging of MeerKAT observations," 2020, https://ui.adsabs.harvard.edu/abs/2020ascl.soft09003H.
74. J. Moldón et al., "Open science droplets," 2021, https://droplets-spsrc.readthedocs.io/
75. LIRGI, "Luminous Infra-Red Galaxy Inventory," 2021, http://lirgi.iaa.es.

Julián Garrido works as a project manager of the Spanish participation in the SKA. He contributed to developing standards and tools to support open science, reproducibility, and interoperability in astronomy. Since 2017, he has been involved in the activities aimed to design and implement the SKA Regional Centres (SRC). His research interests include open science, artificial intelligence, and astronomy.

Laura Darriba received her PhD in physics in the field of simulations of compact groups of galaxies from the University of Barcelona. She has worked as a data scientist in industry in the UK using machine learning models to predict market behavior. She has been a member of the team developing the SPSRC since 2019.

Susana Sánchez Expósito has a degree in computer engineering and has been working for the IAA-CSIC as a R&D engineer since 2007. She has had the technical responsibility for the projects e-CA and AMIGA4GAS, investigating e-Science and distributed computing infrastructures. Since 2015, she has been involved in the design of platforms for processing SKA data. Currently, she is the technical lead of the development of the SPSRC.

Manuel Parra-Royón is a postdoctoral researcher within the development group of the SPSRC. He has contributed to the design of ML-based Brokers for workflow efficiency at CERN and in the early stages of analysis of the Cherenkov Telescope Array. Since 2019, he works with data processing and data workflows for astrophysics and astroparticles. Currently, he is working at the SPSRC, being responsible for the infrastructure, architecture, and computing services.

Javier Moldón is a postdoctoral researcher specialized in radio astronomical observations in the fields of nuclear activity in ultra-luminous IR galaxies, star-forming history of the universe, and galactic and extragalactic transient events. He has participated in the commissioning and development of the radio interferometers LOFAR and e-MERLIN, and he currently works at the SPSRC, being responsible for the scientific operations, user support and training.

María Á. Mendoza is a science-domain services engineer in the development of SPSRC. She has a degree in physics and PhD in computer science and artificial intelligence. She has participated as a software developer in the projects Calar alto legacy integral field area (CALIFA) survey (IAA-CSIC), preparing for Euclid mission (IEEC-CSIC) and UGR contribution to PLATO2.0 (University of Granada).

Sebastian Luna-Valero received his computer science degree from the University of Malaga in Spain. He has been working in research computing in academia both in Spain and the United Kingdom since 2013. He was affiliated to IAA-CSIC from 2018 to 2021 working as a system administrator and helping with the deployment of the SPSRC infrastructure. As of May 2021, he has moved to the EGI Foundation to work as a cloud community support specialist.

Antxon Alberdi is currently a director at the Instituto de Astrofísica de Andalucía (IAA-CSIC). His research focuses on the study of the (sub-)parsec-scale relativistic jets in AGNs, radio supernovae and supernova factories in Starburst Galaxies, using high angular resolution and high sensitivity interferometric observations. He belongs to the SEAC (SKA External Advisory Committee) and is a core member of the SKA Radio Continuum Working Group, being responsible of the Focus Group on Nearby Galaxies. He is also member of the EHT Scientific Collaboration.

Isabel Márquez is the scientific director of the Severo Ochoa program at the IAA-CSIC. Her research mostly relates to nuclear activity (AGN) in galaxies, in particular the extreme of low luminosity AGN, devoted to the analysis of the origin and properties of LINERs from a multiwavelength perspective. Her main achievements include the study of the mechanisms related to the fueling processes in the circumnuclear regions of active galaxies, and how they compare to those present in nonactive galaxies. Some of her work focuses on the understanding of the differential properties from isolation to galaxies in clusters.

Lourdes Verdes-Montenegro is a scientific researcher at IAA-CSIC. She is an expert in multifrequency analysis of galaxy evolution, with emphasis in the study of atomic gas (HI). She leads the AMIGA team, a multidisciplinary group pioneering in applying eScience and Open Science to radio astronomy research. She coordinates the Spanish participation in the SKA and the development of the SPSRC.