8.6 Radio astronomy (CC)

- Ambition
- User stories
- Use cases
- Architecture & EOSC-hub technologies considered /assessed
- Requirements for EOSC-hub
  - Technical Requirements
  - Capacity Requirements
- Validation plan

Short description	Radio astronomy CC task
Type of community	Competence Center
Community contact	Hanno Holties, Rob van der Meer
Meetings	
Supporters	

Ambition

The Radio Astronomy CC will support researchers to find, access, manage, and process data produced by the International LOFAR Telescope. It aims to lower the technology threshold for the Radio Astronomical community in exploiting resources and services provided by the EOSC. Particular aspects that will be addressed are federated single sign-on access to services in a distributed environment and support for data-intensive processing workflows on EOSC infrastructure, notably having access to user workspace connected to high-throughput processing systems, offer portable application deployment, and provide integrated access to a FAIR science data repository. The community is to be empowered to optimally profit from these and increase the science output from multi-petabyte radio astronomical data archives of current and future instruments. The RACC will achieve this by undertaking activities including integration with available federation and data discovery services.

Users will be provided with access to large-scale workspace storage facilities within the EOSC-hub to store and share temporary data and products from pipelines. RACC will empower science groups to deploy their own processing workflows. The RACC lessons learned will serve as input for the design and construction of a European Science Data Center for the Square Kilometre Array (SKA), e.g. via the complementary AENEAS and ESCAPE projects.

User stories

No.	User stories
US1	As an Observatory, we want to offer Single Sign On to our users and manage community access through a central federated collaboration management service, such as COmanage, to improve user experience and consolidate user administration for services.
US2	As a scientific user, I want to perform LOFAR data analysis on archived data-products using available scalable compute infrastructure, allowing for long-term storage and inspection of results. It must be possible to automate initiation and monitoring of processing workflows including data staging and storage. I want to use portable software deployment such that time spent on porting applications is minimized, an integrated workflow management framework, and user workspace to store data products that are to be evaluated or further processed.
US3	As a scientific user, I want to enter science-grade data products in a science data repository that supports the FAIR principles to ensure long-term data preservation and attribution of effort. This will further improve sharing of data with colleagues and access to data from other science domains. It should be possible to access data in the science data repository using direct links to individual data objects via an anonymously accessible public URL such that other services, e.g. those provided by the Virtual Observatory, can be built to provide access to the data.

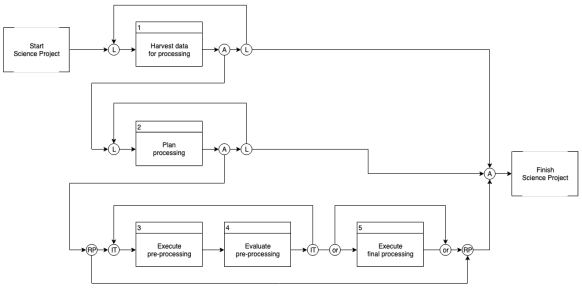
Use cases

Step	Description of action	Dependency on 3rd party services (EOSC-hub or other)

UC1	Observatory registers service with federation	Federated AAI (Comanage, e.g. EGI Check-In)
UC2	User registers for LOFAR collaboration(s)	Federated AAI (Comanage, e.g. EGI Check-In)
UC3	Observatory manages collaboration(s)	Federated AAI (Comanage, e.g. EGI Check-In)
UC4	Collaboration administration provisions non-web service(s)	Federated AAI (Comanage, e.g. EGI Check-In)
UC5	User authenticates to federation	Federated AAI (Comanage, e.g. EGI Check-In)
UC6	Observatory releases data analysis software in repository	-
UC7	User enters data analysis software in repository	-
UC8	User selects data for retrieval /processing	-
UC9	User requests/stages data-products	-
UC10	User retrieves data-products	dCache WEBDAV/Macaroons
UC11	User initiates data analysis workflow	HTC processing cluster supporting Singularity, CWL workflows, CVMFS (application distribution)
UC12	User inspects output data	dCache WEBDAV/Macaroons
UC13	User registers output data	B2SHARE, PID service (EPIC)
UC14	Observatory registers archived data	B2FIND (option), PID service (EPIC)

The first five use cases are related to providing federated authentication and authorization to LOFAR and EOSC services for the LOFAR community, addressing user story US1.

Use cases UC6 - UC12 are related to user data processing for science projects and address user story US2. UC6 - UC9 are included for completeness but are handled by services provided by the LOFAR Observatory and, other than integration in a federated AAI, do not depend on 3rd party services. The diagram below is a control flow diagram of the user processing workflow that will build on the use cases in this document.



Legend for control flow diagrams

(A) Potentially concurrent flows

(or) Mutually exclusive flows

(L) Loop until exit condition satisfied

(RP) Perform in multiple instances (potentially concurrent)

(IT) Perform number of times

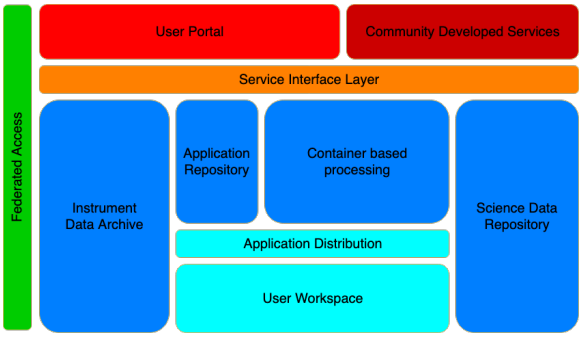
The main user processing workflow.

At the moment, the functions in the diagram above are handled by community-developed services and can be viewed as representations of the ‘User’ role in the use cases provided here (1: UC8; 2: -; 3: UC9 - UC11; 4: UC12; 5: UC10, UC11, UC13). Existing functions have been developed by one of the key science projects but are not available as a service to the wider ‘long-tail’ community. A stepwise implementation of these functions in a central User Portal will be undertaken to support the wider community in the future.

Use cases UC13 and UC14 address user story US3. UC13 is considered critical. UC14 may provide added value for data-linking purposes and to increase use of data from the LOFAR archive outside of the radio astronomical community.

Note that in the use cases the roles ‘Observatory’ and ‘User’ can represent a person, but also an automated process that is performed on behalf of a person.

Architecture & EOSC-hub technologies considered /assessed



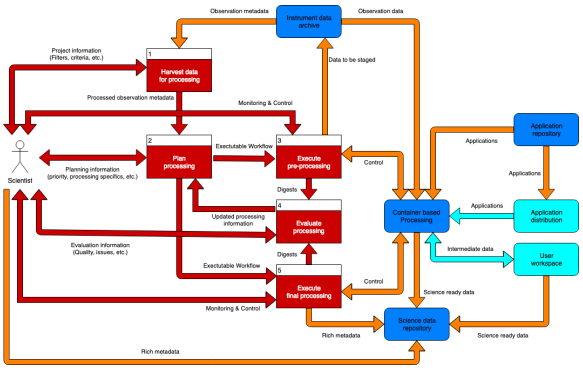
High level of the architecture considered for the Radio Astronomy Competence Center.

The high level architecture for the RACC incorporates the suite of services that together address the user stories and is to support all of the defined use cases. The component ‘Federated Access’ interfaces with all components that require user level authentication and authorization support.

At the top of the diagram are two components representing a central LOFAR User Portal, supported by the LOFAR Observatory, and Community Developed Services that use the same service interfaces to build custom functionality.

Within EOSC-hub, the focus will be on the integration of Federated Access and on integration of the blue-coloured components in the lower part of the diagram (dark blue for services directed at users, light blue for integration supporting services). Steps towards implementation of a central User Portal will be undertaken but full implementation depends on work carried out in other projects as well (e.g. ESCAPE). The components ‘Instrument Data Archive’ end ‘Application Repository’ are considered to be provided and supported by the LOFAR Observatory. The other blue-coloured components will be built as much as possible on available EOSC services.

The diagram below shows how the functions in the user processing workflow interact with the services in the RACC architecture. Note that all these functions are to be provided by the components at the top of the RACC Architecture diagram.



Interaction of the user processing workflow with RACC components.

The following table maps components of the RACC architecture to the service provided by EOSC:

Component	EOSC Service	Description
Federated Access	EGI Check-In, using COmanage for collaboration management.	COmanage is the federated collaboration management system of choice. A pilot is underway to federate using EGI Check-In. Depending on outcome, alternative federation providers (notable Geant or SURFnet) may be considered.
Container Based Processing	EGI/Grid and High Throughput processing clusters	It is envisioned that the processing systems support CWL based workflows using Singularity containerized applications. RACC would be interested in a workflow management service provided by EOSC for this type of processing capability. Alternatively, a workflow management service would need to be deployed and maintained as an RACC specific service.
Application Distribution	CVMFS	The main objective is to provide transparent local access to selected application containers and software installations on processing clusters.
User Workspace	dCache WEBDAV & Macaroons	The user workspace is envisioned to use the same storage infrastructure as the instrument data archive. The objective is to move away from personal X509 based access over GridFTP toward Macaroon based access over WEBDAV.
Science Data Repository	B2SHARE, EPIC/B2HANDLE, B2FIND	User generated science-level data is to be entered in B2SHARE. Optionally, data in the LOFAR instrument data archive will be registered in B2FIND. A LOFAR-scoped handle generating service is to be provided by EPIC/B2HANLDE.

Requirements for EOSC-hub

Technical Requirements

Requirement number	Requirement title	Link to Requirement JIRA ticket	Source Use Case

RQ1	EOSC-hub to support integration of LOFAR services in EGI Check-In federated AAI	 EOS CWP 10-80 - Jira .	UC1
RQ2	EOSC-hub to support customization of the COmanage Community Organisation for the LOFAR.	 EOS CWP 10-80 - Jira .	UC3
RQ3	EGI Check-In to allow any IdP in EduGAIN, plus social identity providers Google, Orcid, etc., to be allowed for SSO access to federated LOFAR services.	 EOS CWP 10-80 - Jira .	UC2
RQ4	EOSC-hub to provide a LOFAR scoped Persistent ID generation service.	Assume this to be provided by SURFsara provided EPIC service	UC13, UC14
RQ5	EOSC-hub to provide a Science Data Repository (B2Share) that supports LOFAR data sizes and a radio astronomy oriented metadata model.	 EOS CWP 10-81 - Jira .	UC13
RQ6	B2Share to provide direct data access URL's	Already possible? (TBC)	UC13

RQ7	EOSC-hub to provide a B2FIND service for harvested registration of archive data using LOFAR PID's and a LOFAR metadata model (option)	 EOS CWP 10-82 - Jira .	UC14
RQ8	EOSC-hub to support CVMFS mounting on processing clusters	 EOS CWP 10-83 - Jira .	UC6, UC11
RQ9	EOSC-hub to support Singularity containerized applications on processing clusters.	 EOS CWP 10-84 - Jira .	UC11
RQ10	EOSC-hub to support CWL based workflow management framework(s)	 EOS CWP 10-85 - Jira .	UC11
RQ11	EOSC-hub infrastructure partners for LOFAR to support required dCache features (WEBDAV, Macaroons, User workspace, staging API/GFAL)	To be dealt with by CC partners	UC9, UC10, UC12
RQ12	EOSC-hub infrastructure partners for LOFAR to integrate storage and compute systems in accordance with requirements for LOFAR processing workflows	To be dealt with by CC partners	UC11

Capacity Requirements

The table below provides an overview of capacity requirements that will enable the main use cases at a scale that is considered minimal to support science projects carried out by the full LOFAR community. It should be noted that a proper access model is to be developed. The current practice of individual scientists securing processing and storage capabilities for their projects is considered inefficient and ineffective. A centrally managed allocation process would be preferable, where resources needed for completion of data processing and for storing data in a science data repository are requested and allocated in a single process together with the allocation of observing time and storage capacity in the LOFAR instrument data archive. Organization and funding of the resources required for science data generation and publication are however unclear. It is our ambition to gain access to these resources via the EOSC.

EOSC-hub services	Amount of requested resources	Time period
dCache storage	Archive: 7PB growth per year; Workspace: 1-2 PB	Expand existing at LTA sites
HTC processing system	10+ Million Core hours	Fraction (10%) to support pilot & development activities from July 2019; Bulk in 2020
B2SHARE	50 – 100 TB growth per month. Consider pilot at some hundreds of TB level before deciding on growth path.	Pilot from July 2019
B2HANDLE /EPIC, (B2FIND option)	Supporting 10+ million data products	From June 2019
Comanage	Supporting 1000+ ID’s, 100+ COU’s	From April 2019 (output EOSCpilot)

Validation plan

Functionality will be validated at user story level. For the Radio Astronomy Competence Center activities undertaken within the EOSC-hub project will initially be made available to a small number of pilot users for validation. Validation of production services available to the full LOFAR community is considered an important end-goal but depends on demonstrated service maturity and available resource capacity and is not necessarily in scope for this project.

US1: The LOFAR Observatory will duplicate the administration for a small set of projects to the federated collaboration management service. Associated users will be invited to register and provided with links to services that have been integrated with the federation (these may be duplicates of a production services or a separate authentication entry point, depending on the capabilities of the service). Support for the User Story will be considered a success when the users have been able to access the services using the federated AAI and were authorized access in accordance with their collaboration membership and role(s) therein. User feedback will be collected as input for further development. The federated collaboration is required to support a community of 2500+ members, and provide functionality and effective user interfaces for 1000+ collaborative teams (scale of existing user and scientific project databases).

US2: A representative of the community will be asked to perform LOFAR data analysis for each integrated workflow. Validation is considered a success if the user is able to successfully conclude the data analysis workflow and access the results using the RACC provided services. Upon success, scaling up to supporting the required processing for a full science project will be considered. A single workflow, for processing a single LOFAR observation, must be allowed to take 20 TB of data as input, use a workspace of 40 TB and generate up to 1 TB of output data. It uses up to 20,000 core hours for it's processing and jobs can run for up to one week. The workflow is typically divided over up to 250, mostly independent, processing 'threads' over which the total capacity is distributed. A typical science project (not considering the Key Science Projects) will process up to 10 observations.

US3: A representative of the community will be asked to register science-level data generated by a LOFAR processing workflow in the science data repository and use the associated PID as a reference in publications and/or sharing results with colleagues. Validation is considered a success if the representative succeeds. A next step in the validation is automatic ingest of data in the science data repository. The repository must support datasets at LOFAR scale: Science-level data generated from a single LOFAR observation will typically consist of 50 files of 10GB each. Files up to 100GB are possible. There are up to 100 observations to be processed each month.