Ref. Ares(2019)6730358 - 30/10/2019

Project Title	European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructure
Project Acronym	ESCAPE
Grant Agreement No	824064
Instrument	Research and Innovation Action (RIA)
Topic	Connecting ESFRI infrastructures through Cluster projects (INFRA-EOSC-4-2018)
Start Date of Project	01.02.2019
Duration of Project	42 Months
Project Website	www.projectescape.eu

D5.2 - DETAILED PROJECT PLAN

Work Package	WP5, ESAP
Lead Author (Org)	Zheng Meyer-Zhao (ASTRON)
Contributing Author(s) (Org)	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON), Michiel van Haarlem (ASTRON), Arpad Szomoru (JIVE), Stelios Voutsinas (UEDIN), Giuliano Taffoni (INAF), Susana Sánchez Expósito (IAA-CSIC), Pierre Chanial (EGO), Hugh Dickinson (OU), Matthias Füßling (CTAO)
Due Date	31.10.2019, M9
Date	30.10.2019
Version	1.0

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Version	Date	Authors	Notes
0.1	14.10.2019	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON), Michiel van Haarlem (ASTRON), Arpad Szomoru (JIVE), Stelios Voutsinas (UEDIN), Giuliano Taffoni (INAF), Susana Sánchez Expósito (IAA-CSIC), Pierre Chanial (EGO)	First draft. Ready for review within work package 5.
0.2	21.10.2019	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON),Michiel van Haarlem (ASTRON),Hugh Dickinson (OU),Arpad Szomoru (JIVE), Matthias Füßling (CTAO)	Processed all comments from member institutes of work package 5.
1.0	30.10.2019	Zheng Meyer-Zhao (ASTRON)	Submitted version.

Disclaimer

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Table of content

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Executive Summary

This document describes the project plan for ESCAPE Work Package 5, ESAP - ESFRI Science Analysis Platform, detailing its goals, objectives and structure. It will serve as the work plan at the beginning of the project and will be updated throughout the project.

The ESAP main objectives are to define and implement a flexible science platform for the analysis of open access data available through the EOSC environment that will allow EOSC researchers to identify and stage existing data collections for analysis, tap into a wide-range of software tools and packages developed by the ESFRIs, bring their own custom workflows to the platform, and take advantage of the underlying HPC and HTC computing infrastructure to execute those workflows.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Project Summary

ESCAPE (European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures) addresses the Open Science challenges shared by ESFRI facilities (SKA, CTA, KM3NeT, EST, ELT, HL-LHC, FAIR) as well as other pan-European research infrastructures (CERN, ESO, JIVE) in astronomy and particle physics. ESCAPE actions are focused on developing solutions for the large data sets handled by the ESFRI facilities. These solutions shall: i) connect ESFRI projects to EOSC ensuring integration of data and tools; ii) foster common approaches to implement open-data stewardship; iii) establish interoperability within EOSC as an integrated multi-messenger facility for fundamental science. To accomplish these objectives, ESCAPE aims to unite astrophysics and particle physics communities with proven expertise in computing and data management by setting up a data infrastructure beyond the current state-of-the-art in support of the FAIR principles. These joint efforts are expected result into a data-lake infrastructure as cloud open-science analysis facility linked with the EOSC. ESCAPE supports already existing infrastructure such as astronomy Virtual Observatory to connect with the EOSC. With the commitment from various ESFRI projects in the cluster, ESCAPE will develop and integrate the EOSC catalogue with a dedicated catalogue of open-source analysis software. This catalogue will provide researchers across the disciplines with new software tools and services developed by astronomy and particle physics community. Through this catalogue ESCAPE will strive to cater researchers with consistent access to an integrated open-science platform for data-analysis workflows. As a result, a large community “foundation” approach for cross-fertilisation and continuous development will be strengthened. ESCAPE has the ambition to be a flagship for scientific and societal impact that the EOSC can deliver.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Acronym List

Partners in WP5

CERN	European Organization for Nuclear Research
CNRS-LAPP	Laboratoire d'Annecy de Physique des Particules (CNRS)
CSIC	Consejo Superior de Investigaciones Científicas
CTAO	Cherenkov Telescope Array Observatory
FAIR GMBH	Facility for Antroproton and Ion Research
EGO	European Gravitational Observatory
FAU	Friedrich-Alexander University Erlangen-Nuremberg
INAF	Istituto Nazionale di Astrofisica
IFAE	Institut de Física d'Altes Energies
JIVE	Joint Institute for VLBI ERIC
KIS	Leibniz-Institut fuer Sonnenphysik
NWO-I-ASTRON	Netherlands Institute for Radio Astronomy (NWO-I)
NWO-I-Nikhef	Nationaal instituut voor subatomaire fysica (NWO-I)
RUG	Rijksuniversiteit Groningen
SKAO	Square Kilometre Array Organisation
UCM	Universidad Complutense de Madrid
UEDIN	The University of Edinburgh

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

General

API	Application Programming Interface
ASTERICS	Astronomy ESFRI & Research Infrastructure Cluster
CS	Citizen Science
DIOS	Data Infrastructure for Open Science
DO	Digital Object
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
ESAP	ESFRI Science Analysis Platform
ESCAPE	European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures
ESFRI	European Strategy Forum on Research Infrastructures
ESF/RI	ESFRIs and major RIs as projects within ESCAPE
FAIR	Findable, Accessible, Interoperable, Reusable
MVP	Minimum Viable Product
RA	Right ascension
REST	Representational State Transfer
RI	Research Infrastructure
RO	Research Object
SAP	Science Analysis Platform
VO	Virtual Observatory
WP	Work Package

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

1. Introduction and Goals

The ultimate goal of the European Open Science Cloud (EOSC) is to empower the scientific community to effectively extract high-impact results from the incredible amounts of data generated by current and future facilities and research infrastructures. The goal of ESCAPE WP5 is to implement a flexible science analysis platform for the analysis of (open access) data available through archives which are part of, or connected to the EOSC environment. This platform will support researchers in creating and executing reproducible scientific workflows and methods, enhancing the sharing not only of data but also of the scientific workflows, following the *FAIR* principles.

ESCAPE WP5 will define and implement a platform that will allow EOSC researchers to identify and stage existing data collections for analysis, tap into a wide-range of software tools, packages and workflows developed by the ESFRIs, bring their own custom workflows to the platform, and take advantage of the available HPC and HTC computing infrastructure to access the data and execute those workflows.

Our approach is to provide a set of functionalities from which various communities and ESFRIs can assemble a science analysis platform geared to their specific needs, rather than to attempt providing a single, integrated platform to which all researchers must adapt. Deploying an EOSC-based science platform provides a natural opportunity to integrate with the data and computing fabric this environment encompasses while simultaneously accessing the tools, techniques, and expertise other research domains bring to that environment.

The ESFRI Science Analysis Platform (ESAP) developed through ESCAPE WP5 will provide a flexible and expandable analysis environment for the astronomy and physics community and constitute an absolutely essential resource for the big data challenges of the next generation of ESF/RIs.

In the ESCAPE project proposal, several tasks are defined along with the contributing partners. In section 5, we will describe the tasks within this work package. Based on the description of the task given in the proposal and on the use cases from the ESFRIs, we identified a list of service components in the first deliverable D5.1 - Preliminary report on requirements for ESFRI science analysis use cases. In order to define the functionalities of these services, we apply the methods of collecting user stories, which are in turn summarized into a list of requirements. Section 6 gives a more detailed description of these user stories and of the requirements derived from them. In section 7, we list the service components of the ESAP, identify their functionalities and map the requirements derived from section 6 into these service functionalities. The high-level architectural

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

design of ESAP is presented in Section 8. Section 9 describes the implementation plan of ESAP, and we make the risk analysis in Section 10.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

2. WP5 structures

The work in ESAP is organised in 5 tasks:

- Task 5.1: Data Aggregation and Staging
- Task 5.2: Software Deployment and Virtualization
- Task 5.3: Analysis Interface, Workflows and Reproducibility
- Task 5.4: Integration with HPC and HTC Infrastructures
- Task 5.5: WP5 Management

2.1. Task 5.1: Data aggregation and staging

This task will provide users of the science platform with the capability to access and combine data from multiple collections and stage that data for subsequent analysis. It will draw heavily on the work in WP4 based on the IVOA interoperability protocols to enable data discovery, but also require extension by relevant work in WP2 to handle the staging of potentially large and distributed data collections. Generalization or adaptation of the IVOA software stack and protocols to non-astronomical use cases may be required.

Staging of the data for analysis will require the ability to dynamically allocate user workspace across a distributed data infrastructure including storage and database resources.

2.2. Task 5.2: Software deployment and virtualization

A key component of the EOSC science platform will be the availability of readily accessible versions of the software, tools, scripts, and packages developed by the various ESF/RIs, and their communities. This task will incorporate the work on the software repository described in WP3 and focus on tools and services to support the virtualization of relevant software packages and pipelines. Principle extensions to the work in WP3 will likely include additional work on the containerization of software, provenance and versioning metadata, and specifically the deployment of packages available in the ESCAPE-EOSC software repository to user workspaces. Similarly, the ability to capture user information about the provided software (usage information of bugs for example) and to transmit that information back to the WP3 repository will also be supported.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

2.3. Task 5.3: Analysis interface, workflows, and reproducibility

The analysis interface task combines a number of elements to form the working surface for the user of the EOSC science platform. It integrates the data access and staging element of Task 5.1 along with the access to the EOSC software repository and containers described in Task 5.2 into an interactive workspace such as a Jupyter notebook or similar technologies. In addition, this task will provide tools and services that simplify the porting of customized analysis and processing workflows to the science platform environment. For example, this task will support the user in mapping their individual workflows into a common deployment language such as the Common Workflow Language (CWL) in order to more easily deploy them across the heterogeneous and distributed computing infrastructure underpinning the EOSC. These workflows, along with their logs and user annotations, can be stored in the user workspace to be retrieved, improved, re-run, or shared with others. The ability to preserve and share the workflows and processing history from previous analysis sessions will enhance the reusability of EOSC data collections and improve the reproducibility of community results.

2.4. Task 5.4: Integration with HPC and HTC infrastructures

Once data for analysis has been located and staged, and workflows have been defined, either by accessing the EOSC software repository or by the user directly, the next step is to deploy those workflows on the underlying processing infrastructure. For many of the involved ESF/RIs, the data scales involved require significant computational resources (storage and compute) to support additional processing and analysis. The EOSC-ESFRI science platform therefore must interface to an underlying HPC or HTC infrastructure. Consequently, it is important to make efficient use of the full performance potential of the HPC centres, e.g. by optimizing the access to file systems from the data processing layer and by ensuring the portability of science applications with container solutions. As with the data, this infrastructure is likely to be large, widely distributed geographically, and definitely heterogeneous. Deploying user-initiated processing and analysis tasks on this HPC infrastructure while simultaneously maintaining interactivity and responsiveness in the analysis system will be a challenge and requires a mixture of dynamic resource allocation and optimization. This task will draw upon the infrastructure work outlined in WP2 on the “datalake” design and implementation.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

2.5. Task 5.5: WP5 Management

This task is led by NWO-I-ASTRON and will provide the coordination of the various technical tasks in WP5, as well as overall coordination with the other work packages in ESCAPE.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

3. ESAP use cases

The development of the ESAP will be guided by the requirements posed by the use cases that the ESCAPE communities presented during the ESFRI Use Case Requirements workshop (M5.1) on 16-17 April 2019. While in deliverable ESCAPE-D5.1 we provided a preliminary report on these requirements, here we analyse in detail the ESCAPE use cases by means of the so-called “User Stories”. They are a short and natural language description of features expressed by the person (the user) who needs the given capability of the system.

3.1. User Stories

User stories are artifacts used in software development to facilitate the communication between developers and users and to support the identification of the requirements to be fulfilled by the system. Next, in table 1 we list a set of user stories in the form of:

“As a <type of user> I want to <perform some task>, so that I can <achieve some goal>”.

The field <type of user> indicates the role of the user, such as generic scientist, ESFRI user, software developer etc., <perform some task>, a capability or feature for the platform and <achieve some goal>, the benefit of having this capability. The first column of the table shows an identifier for each user story. In the last column we provide a value indicating the degree of interest (importance) of the story from the point of view of the user. A higher value means greater importance.

The user stories in Table 1 have been gathered from participating ESF/RIs, with the goal of deriving common requirements that are interesting for users across the ESCAPE communities. Following an Agile software development methodology, this first version of user stories will evolve with new user stories or modifications of the existing ones, thus generating new requirements that will guide an incremental implementation of the ESAP.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

USER STORY ID	As a <type of user>	I want to <perform some task>	so that I can <achieve some goal>	Importance Points (Importance, 1-5)
U-1	Scientist	For a given survey, find all Objects that I'm interested in a certain region of the sky, given an RA, declination & a radius	See if there is overlap with another survey I am using, and compare the metadata of the two.	5
U-2	Anonymous User	Find what type of data is available	See if the type of science I'm interested is served through the platform.	3
U-3	Publisher	Publish my data through the Virtual Observatory	Reach a larger target audience.	2
U-4	Lofar user	Query LOFAR LTA database for public data	Find the location of the datasets, stage the data, process the data after they are staged at computing facilities that have direct access to the datasets.	5
U-5	SKA large programme user	Process large amount of data using pipelines that I can use from the Science Analysis Platform	Produce Advance Data Products	5
U-6	SKA large programme user	Include the Advanced Data Products that I have produced in the SAP catalogue	Assign a DOI to the data products, so they can be sharable/findable/citable.	5
U-7	SKA large programme user	Deploy a new pipeline for processing SKA Observatory Data to the ESCAPE computing infrastructure	Execute this pipeline later on in order to produce Advanced Data Products	5
U-8	SKA large programme user	Create/manage a group of collaborators that can access to the programme data	Ensure (only) users allowed to work on the programme data have access permissions to	5

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

			these data	
U-9	SKA Regional Centre	Ensure that the data access policies of SKA data products are adhered to	Provide appropriate access to public data products whilst respecting the specific (dynamic) data access right to individual data collections	5
U-10	Anonymous User	Find examples and notebooks for a given dataset that I'm interested in	Use existing workflows as a starting point to explore the data	3
U-11	ESFRI user	Provide data and analysis tools used in an article	Enable readers to repeat that analysis.	5
U-12	ESFRI manager	Distribute a set of low level ESFRI data	Look for potential interest of other communities	4
U-13	ESFRI manager	Distribute a set of low level ESFRI data	Organize a data challenge	3
U-14	Scientist	Find a combination of data, software and services.	Users should be able to query the platform to identify the combination of data, software, and computing resource to repeat analysis.	3
U-15	Scientist	Select the SAP framework more suitable to their analysis	SAP provides different data processing (e.g. interactive, batch, desktop like). User should be able to select the one (s)he prefers	3
U-16	CS[1] Project Developer (Scientist)	Search and retrieve a subset of public experimental data	Process the data to generate subjects that are appropriate for a CS experiment.	3
U-17	CS Project Developer (Scientist)	Deploy services that interface directly with an external CS platform e.g. Zooniverse.	Make use of bespoke, project specific algorithms to accelerate aggregation and volunteer consensus for subjects in a CS project.	3
U-18	Citizen Scientist	Search and retrieve public data provided multiple	Obtain data that are pertinent to a particular subject or set of	3

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

		ESFRIs that are related in some specific way e.g. they relate to the same coordinate on the sky	subjects that I have encountered while participating in a citizen science project.	
U-19	Citizen Scientist	Perform a subset of simple analyses and reduction steps using public data as inputs	Explore data that are pertinent to a particular subject or set of subjects that I have encountered while participating in a citizen science project.	3

Table 1: User stories gathered from the participating ESFRIs.

Notes:
1 Citizen Science

3.2. Requirements

In addition to the "User Stories" table we have a "Requirements" table, which lists a number of requirements that were produced based on the user stories.

For each entry in this table we have a "Requirement ID (REQ. ID)", which is a unique identifier for each table entry, which is useful in order to be able to refer to each requirement without using the full description. The description of each requirement can be found under "Requirement Description", where we identify what needs to be done in order to satisfy the requirements of a given user story, which is then referenced using the field "Use Case". We also provide a field named "Story point" which we use to calculate the complexity of requirements.

In order to estimate the complexity of a requirement, we use the Fibonacci sequence which is commonly used in Agile project planning. The advantage of using this type of system for calculating complexity of tasks, is that it helps better reflect the uncertainty of estimating large tasks, which often means that the task may need to be broken down into smaller tasks. The values produced represent the difficulty of implementing each requirement, which depends on a number of factors. The point calculations take into account the amount of work that needs to be done for each task, the complexity, the uncertainty of the work, as well as the duration of the tasks in total.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

REQ. ID	Requirement Description	Use Case	Story point (Complexity 1, 2, 3, 5, 8..)
R-1	Users should be able to get a list of available data, searchable by different criteria, including keyword, science domain, institute, datatype etc.	U-1, U-2, U-16, U-17, U18	13
R-2	Users should be able to get a list of known (VO & Other) tools and software for users & publishers.	U-3, U-9, U12, U13	2
R-3	Users should be able to, for a given project & dataset, query for metadata and aggregate information (i.e. find location of data).	U-4, U11	3
R-4	Users should be able to stage a given dataset at the appropriate facility.	U-4	8
R-5	Users should be able to execute a job on a given dataset, including but not limited to: batch or real-time queries & pipelines, depending on the capabilities of the facility, which need to be made clear to the user.	U-4, U-5, U-7	21
R-6	Platform needs to accommodate restricted data access, so that groups of authorised users are the only ones that are able to access a given private data set, shared to them via the platform.	U-8, U-9	13
R-7	Users should be able to select from an existing list of Workflows (Notebooks) and either download, or deploy on available facilities.	U-10, U-19	13
R-8	Users should be able to assign PIDs to every digital object that is part of a Research Object.	U-6	5
R-9	User generated data needs to be queryable via ESAP.	U-9	5
R-10	Users should be able to ingest advance data products generated from data processing and/or data analysis back to the project data archive.	U-6	21

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

R-11	User should be able to select computing facilities on the basis of their capacity. E.g. She needs an HPC resource with a specific acceleration (GPU) because the software to be run requires it.	U-15, U-14, U-8	13
R-12	Less experienced users (e.g. citizen scientists) should be able to filter the list of available software tools to include only those deemed pertinent to the data that they have selected.	U-19	5
R-13	Users should be able to schedule computational tasks at regular intervals e.g. to periodically retrieve new classification data from a CS experiment.	U-18	13

Table 2: List of requirements based on the user stories in Table 1.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

4. Functional description of the ESAP services

In ESCAPE deliverable D5.1 - Preliminary report on requirements for ESFRI science analysis use cases, we identified service components that should be available on ESAP based on the use case requirements gathered from participating ESF/RIs (Figure 1). In this section, we will describe in detail what functionalities each service component needs to have, and which requirements listed in the Requirements Table (Table 2) fall under each service component.

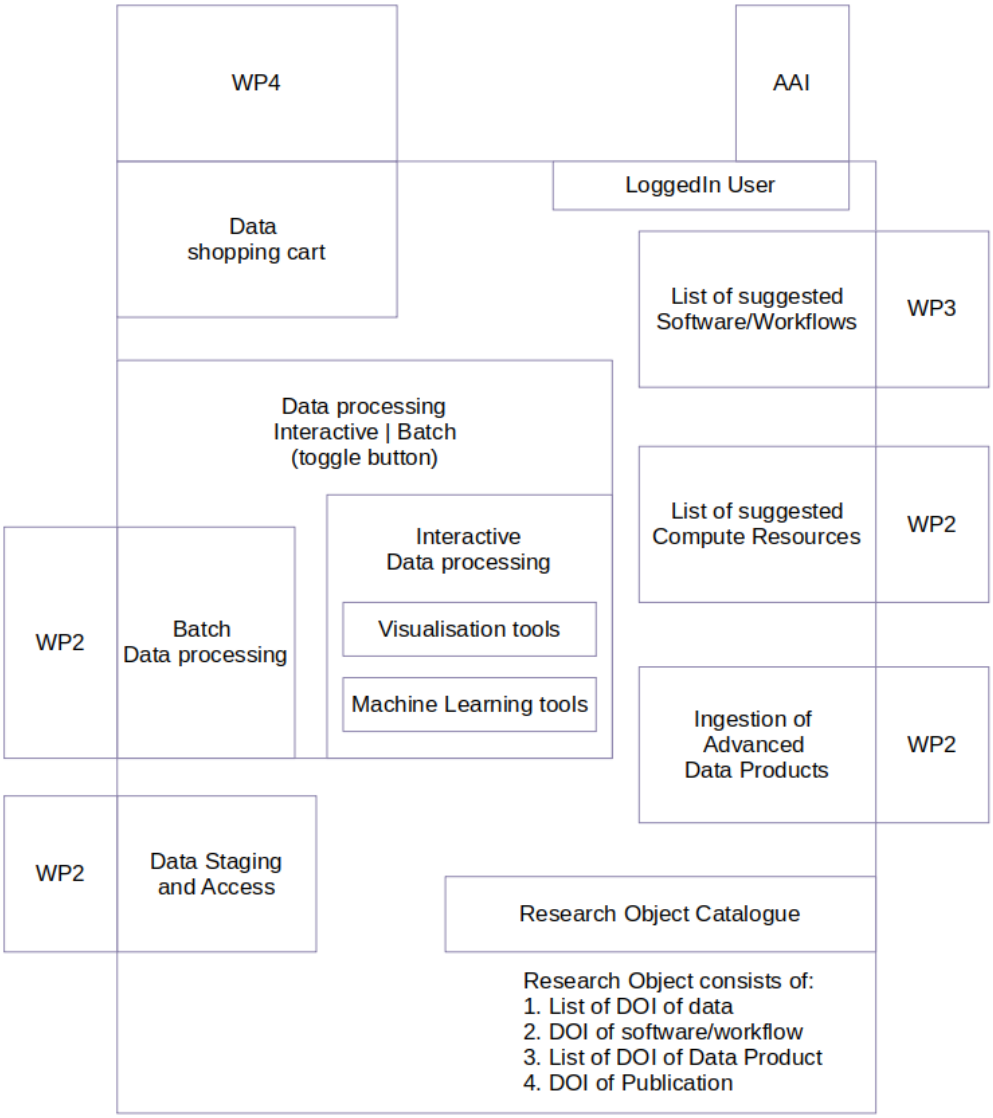

Figure 1: Service components based on input by ESFRIs

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

4.1. Data selection shopping cart

All ESFRI use cases (will) have their own data collections, where data and metadata are hosted outside of the scope of the ESAP. This means that the ESAP needs to connect to external databases hosted by different ESFRIs (e.g. the Lofar LTA catalogue) or libraries (e.g. astronquery.utils.tap) for data query. The ESAP will provide corresponding interfaces that will profit of the CEVO WP work on data access and benefit from the Virtual Observatory standards when possible.

Data selection on the ESAP UI needs to have a tick box next to the data (actually the metadata) returned from data query and a “add to basket” button. (would rephrase to something like: ESAP should be able to ‘remember’/locally store data selections)

The following requirements from the Requirements Table are under this service:

R-1	Users should be able to get a list of available data, searchable by different criteria, including keyword, science domain, institute, datatype etc.	U-2	13
R-3	Users should be able to, for a given project & dataset, query for metadata and aggregate information (i.e. find location of data)	U-4	3

4.2. AAI

The ESFRI science analysis platform will be openly accessible for all users to look for public data in astronomy, astrophysics and particle physics. However, proprietary project data and computing/storage resources will be protected by authentication, authorization and accounting mechanisms.

Users should be able to authenticate themselves with their host institution, private email or social media credentials on the ESAP. Whether a user is authorized to view certain data or to use certain computing/storage resources on a given infrastructure is dependent on the policy of the ESFRI the user belongs to. This means each ESFRI will need to have the information available and allow the ESAP to query the information.

The AAI impacts almost all the Requirements in Table 2, here we list the most affected ones :

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

R-6	Platform needs to accommodate restricted data access, so that groups of authorised users are the only ones that are able to access a given private data set, shared to them via the platform	U-8, U-9	13
R-10	Users should be able to ingest the advanced data products generated from data processing and/or data analysis back to the project data archive	U-6	21
R-4	Users should be able to stage a given dataset at the appropriate facility	U-4	8
R-3	Users should be able to, for a given project & dataset, query for metadata and aggregate information (i.e. find location of data)	U-4	3

4.3. Data Staging and Access

The data staging and access service is directly connected to the Data Lake infrastructure being developed by WP2 DIOS. For data staging, ESAP will forward the users' staging request and negotiate authentication and authorisation with the Data Lake infrastructure delivered by WP2, collect the response from the Data Lake infrastructure, and direct the response back to the user.

Data access for interactive analysis or download may be in the form of Structured Query Language requests to a database through standard Virtual Observatory APIs or non-standard access points depending on the specific data publisher, or through data access libraries via a Jupyter Notebook (i.e. astropy). In the second case the data access library abstracts the user from the specific API that they are connecting to, but will in most cases be through HTTP requests through which they will be able to select a subset of the data and then visualize it in the notebook, possibly repeating the process while they navigate through the data. In any case the goal is to provide a plethora of examples for data access possibilities depending on the data, and to provide abstraction layers as much as possible so that the complexity of accessing heterogeneous datasets is minimized.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Data access for batch data processing will be through the batch computing nodes. Given that a user is authorized to access certain data, an access token will be returned by the AAI service used in the Data Lake Infrastructure, such that this access token can be passed onto the computing nodes to directly access the data during batch job execution.

The following requirements from the Requirement Table are under this service:

R-4	Users should be able to stage a given dataset at the appropriate facility	U-4	8

4.4. Batch Data Processing

Data processing through ESAP will be available in two modes, namely batch data processing and interactive data analysis. This doesn't mean that ESAP provides the computing resources needed for data processing, but rather connects to infrastructure providers where the user has access to. This requires ESAP to be able to interface with different data processing infrastructures, i.e. HTC, HPC or Cloud platforms. Many existing workload management systems are proven to be capable of dealing with this, which makes the development of ESAP in terms of connecting to different compute infrastructures easier. ESAP will investigate the existing workload management systems and select one to interface with. In this way, batch job submission will be handled by the selected workload management system that in turn connects to the underlying HTC, HPC or Cloud platforms.

The following requirements from the Requirement Table are under this service:

R-5	Users should be able to execute a job on a given dataset, including but not limited to: batch or real-time queries & pipelines, depending on the capabilities of the facility, which need to be made clear to the user	U-4, U-5, U-7	21
R-11	User should be able to select computing facilities on the basis of their capacity. E.g. She needs an HPC resource with a specific acceleration (GPU) because my software requires it.	U-15, U-14, U-8	13

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

R-13	Users should be able to schedule computational tasks at regular intervals e.g. to periodically retrieve new classification data from a CS experiment.	U-18	13

4.5. Interactive Data Analysis

Data analysis is commonly a process that produces science ready data from raw data acquired by the experimental facilities. The processes established by scientists for data analysis are quite different depending on the different communities (e.g. the way SKA data will be processed is different from the way LHC data are processed). However, usually data analysis is an interactive process. This means that ESAP should implement a set of tools that allow interactivity and visualization.

4.5.1. Data analysis with Visualisation tools

Data Visualization is one of the core services to be provided. There are two ways a scientists use the visualization features: interactive plots (real-time information about the specific points or areas currently explored, plus possibility of highlighting/hiding specific content) and deep content interaction (this extends beyond graphical plots, to additional type of media and content, and generally requires one or more complex UI elements for collecting external inputs).

A large number of data visualization tools and services are available (e.g. ipywidgets), some of which are developed and maintained by ESCAPE communities (e.g. VISiVO). Tools and services for visualization will be part of the ESCAPE and EOSC marketplace. A framework to find, deploy and configure those tools on ESAP is necessary. Data exploration of large data sets requires interaction with large computing facilities. Jupyter is a possible framework to offer interactive visualization capabilities, together with remote graphical access services.

4.5.2. Data analysis with Machine Learning tools

Machine learning tools are either developed by ESCAPE partners or third parties. Given a trained model, ESAP needs to provide the ability for user to select models and apply the models on selected datasets. These machine learning tools will be integrated either by WP3 into the software repository or can be uploaded by individual users to the ESAP. These tools will then be made available through the interactive data analysis interface.

The following requirements from the Requirement Table are under this service:

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

R-7	Users should be able to select from an existing list of Workflows (Notebooks) and either download, or deploy on available facilities	U-10	13

4.6. List of suggested Software/Workflows

This ESAP service component is the link to the software repository being developed by WP3: OSSR - Open-source scientific Software and Service Repository. The functionality of this service is to filter the software repository based on the data selected by the user, such that only software/workflow that is relevant to the selected data/datatype will be listed. This will save users lots of time in looking through a number of irrelevant software/workflows, therefore, allowing users to focus on data processing and/or analysis, instead of drowning in the lake of various software.

The following requirements from the Requirement Table are under this service:

R-2	Users should be able to get a list of known (VO & Other) tools and software for users & publishers.	U-3, U-9	2

R-7	Users should be able to select from an existing list of Workflows (Notebooks) and either download, or deploy on available facilities	U-10	13

R-12	Less experienced users (e.g. citizen scientists) should be able to filter the list of available software tools to include only those deemed pertinent to the data that they have selected.	U-19	

4.7. List of suggested Compute Resources

In order to avoid large volumes of data transfer, it is generally preferable to bring the compute to data. The recommended list of compute resources will be based on the location of the datasets selected by the user, given that the user is authorized in using the recommended compute resource.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

For small datasets, the choice of recommendation will purely be based on which compute resources a user is authorized to use.

The following requirements from the Requirement Table are under this service:

R-11	User should be able to select computing facilities on the basis of their capacity. E.g. She needs an HPC resource with a specific acceleration (GPU) because my software requires it.	U-15, U-14, U-8	13

R-13	Users should be able to schedule computational tasks at regular intervals e.g. to periodically retrieve new classification data from a CS experiment.	U-18	13

4.8. Ingestion of Advanced Data Products

Users may want to ingest the advanced data products generated from data processing and/or data analysis back to the project data archive. The data ingest policy and the necessary tools used to verify the quality of the advanced data products need to be defined and developed by the research infrastructure. The ESAP will need to interface with the quality verification tools provided by the ESF/RI in order to verify whether an advanced data product meets the quality requirements of the ESF/RI for data ingest. A data ingest request will be sent to the Data Lake Infrastructure API only when the data product meets the quality requirements of the ESF/RI for data ingest.

The following requirements from the Requirement Table are under this service:

R-10	Users should be able to ingest the advanced data products generated from data processing and/or data analysis back to the project data archive	U-6	21

4.9. Research Object Catalogue

As described in ESCAPE deliverable D5.1, Research Objects (RO) are aggregations of Digital Objects involved in a scientific experiment such as datasets, analysis codes,

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

provenance information, annotations and other related information. Current RO models make use of Persistent Identifiers (PIDs) to aggregate and interlink the experiment elements, which may be distributed in different repositories.

The Research Object Catalogue service component of the ESAP does not only need to provide query capability for existing research object, based on its datasets, analysis codes etc., but also need to deliver a mechanism to create new research object, including the service of getting PIDs for Digital Objects by connecting to existing services that generate PIDs.

The following requirements from the Requirement Table are under this service:

R-8	Users should be able to assign DOIs to Research Objects	U-6	5
R-9	User generated data needs to be queryable via ESAP	U-9	5

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

5. ESAP Architectural Design

The ESAP platform will be built following the microservice architectural design: each functionality will be implemented as an independent interconnected service. Independence means that if one service is removed, the others will still be able to process API requests, and the platform will still continue to work although with reduced functionality. The services are interconnected through the API Gateway and communicate between each other through exposed API endpoints to create, read, process and delete resources.

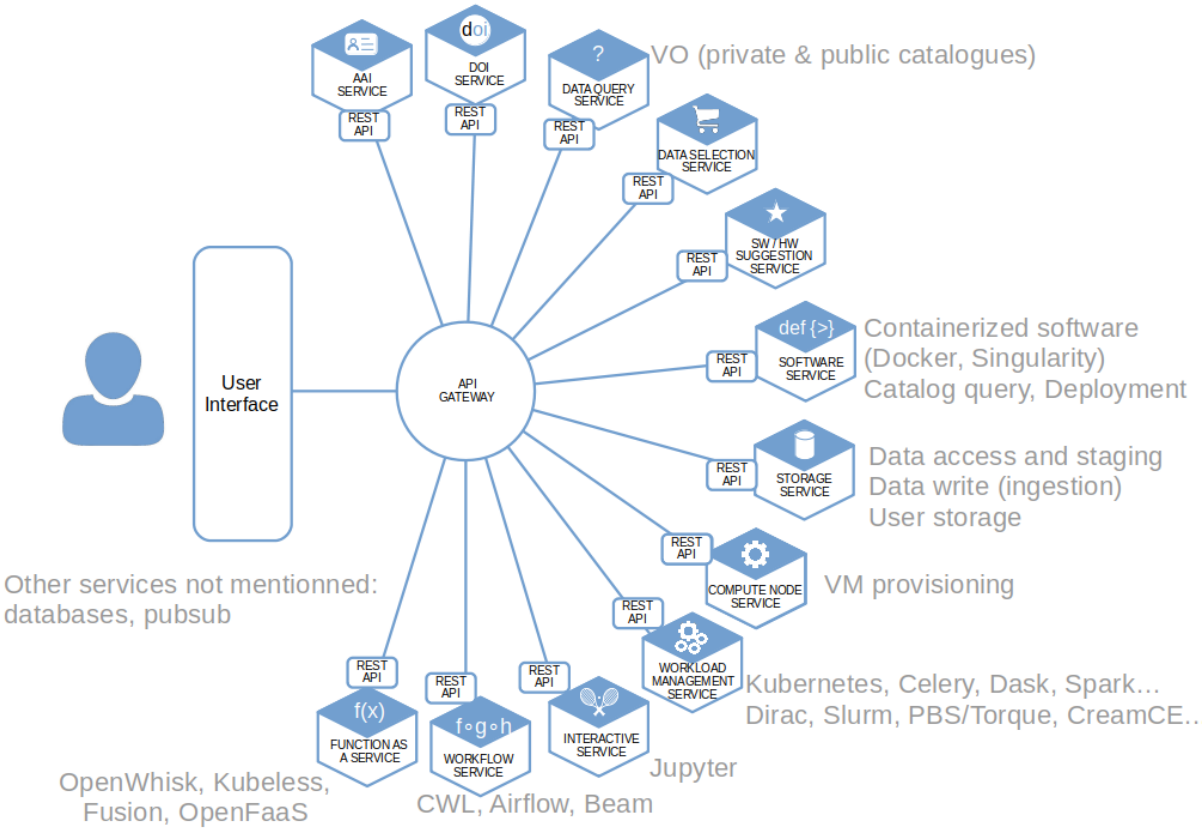

Figure 2: ESAP global architectural design

The motivation of this design is to

- reduce the overall complexity of the platform
- increase robustness
- enforce clear interfaces through the service API definitions
- decouple the development of the different services
- easily add or replace functionalities

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

The global architectural design of ESAP shown in Figure 2 gives an overview of the microservices and API gateway architecture ESAP plans to use. We will use this design as a starting point and work out a detailed architectural design, where each service component, its APIs and the routing among the services will be presented. This will be the first task to be fulfilled before any implementation.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

6. Implementation Plan

6.1. Description of the Implementation Approach

We plan on following an agile approach during the development of the project. This type of project management follows an iterative approach, where we define a “Minimum Viable Product” as a goal, and then iteratively add features in small increments. One of the advantages of using such a methodology for this project is that it allows us to build a prototype, and depending on user feedback, adjust the targets and potential user features that we add to the platform. Additionally, it guarantees that we will have a working platform with at the very least a basic usable science platform based on our MVP definition, minimising the risk of uncertainty that comes with designing with a top-down approach.

As mentioned previously in this document the requirements and thus the design of the platform will be closely tied with the user stories, which ensures that by evaluating the complexity of tasks in combination of the impact of each user story, we are incrementally meeting attainable targets that produce useful features.

With this approach a focus will be solid continuous communication, which will be achieved via frequent telecons, real-time chat applications and face-to-face interactions.

6.2. ESAP Implementation Plan

In the ESCAPE project proposal, a list of milestones and deliverables are defined. Work package 5 will follow the timeline of the defined milestones and deliverables, and make sure that we can reach the milestones and deliver the ESAP prototype on time. Table 3 and table 4 gives an overview of the upcoming milestones and deliverables in work package 5.

Number	Milestone Name	Month	Due
M5.3	Initial science platform prototype with discovery and data staging	M18	July 2020
M5.4	Deployment of initial set of ESFRI software on prototype platform	M20	September 2020
M5.5	Second WP5 workshop to analyse prototype performance	M22	November 2020

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

M5.6	Integration of Science Platform with OSSR repository	M28	May 2021
M5.7	Integration of Science Platform with Data Lake expanded prototype	M30	July 2021
M5.8	Delivery and integration of new ESFRI visualization and analysis tools	M36	January 2022
M5.9	Final WP5 ESFRI user training workshop on the Science Platform	M38	March 2022

Table 3: List of milestones

Number	Deliverable name	Month	Due
D5.3	Performance assessment of initial Science Platform prototype	M24	January 2021
D5.4	Final assessment of the performance of the Science Platform prototype and plan for deployment of production version within the EOSC	M42	July 2022

Table 4: List of deliverables

6.2.1. Overview of available resources

In order to reach the milestones and deliver the prototype on time, we will engage each partner in work package 5 contributing to the service components development. The service components identified fall into the different tasks described in the project proposal. Therefore, we categorize the service components under each task and list how each partner is going to contribute to the tasks.

ESAP service components under Task 5.1: Data aggregation and staging

- Data selection shopping cart
- Data staging and access
- AAI

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Contributing partners:

CERN, CNRS LAPP, EGO, FAIR GMBH, KIS, NWO-I-ASTRON, NWO-I-Nikhef, SKAO, UCM

Individual contributions:

CERN - *12 PM, 0.46 FTE in the period of 01-10-2019 - 01-01-2022.*
Provision of user tools to estimate and report data availability and latency. The solution can be based on the Rucio WEB-UI.

CNRS LAPP - *12 PM*
Deploy various components of the LSST Science Platform and test their usability for realistic science analysis use cases.

EGO - *6 PM*
Make Virgo data available through ESAP.

FAIR GMBH - *6 PM*
Integration of the analysis platform with the Data Lake.

KIS - *6 PM*
No information available yet.

NWO-I-ASTRON - *6 PM, 0.5 FTE in the period of 01-01-2020 - 31-12-2020*.
Study of the interface of existing tools such as RUCIO and pyvo, and implement the integration in collaboration with WP2.

NWO-I-Nikhef - *6 PM throughout the project*.
Compute and data services integration, specifically for HTC infrastructures. Study into leveraging AAI developments in ESCAPE and the global R&E community to access data. Consider access management to staged data.

SKAO - *6 PM throughout the project*.
Understanding Rucio and contributing to Rucio development. Understanding how it can be integrated with other technologies.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

UCM - *9 PM, 0.6 FTE in the period 2020 - 2022.*
Help will be provided to make Cherenkov Telescope Array data sets available.

ESAP service components under Task 5.2: software deployment and virtualisation

- List of suggested Software/Workflows
- Data analysis with visualisation tools
- Data analysis with machine learning tools

Contributing partners:

CTAO, CSIC, EGO, FAIR GMBH, JIVE, NWO-I-ASTRON

Individual contributions:

CTAO - *6 PM throughout the project*.
New analysis techniques (alert correlation) with CTA

CSIC - *6 PM throughout the project*.
Provision of use cases to demonstrate ESAP functionality. Deployment/ingestion of the science domain software involved in CSIC use cases to the platform in order to test it.

EGO - *12 PM throughout the project*.
Contribute to development of ESAP service components

FAIR GMBH - *6 PM throughout the project*.
Build containers (Docker, Singularity) for specific use cases (HEP)

JIVE - *18 PM, 0.43 FTE throughout the project*.
Simplify re-running pipelines by the use of Jupyter. Creation of a model of the provenance of radio data at JIVE.

NWO-I-ASTRON - *12 PM, 0.5 FTE in the period of 01-01-2020 - 31-12-2021*. Implement in ESAP support for selecting containerized software and executing them on a testbed cluster.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

SKAO - *6 PM throughout the project*. **in addition to project plan**
Deploying docker containers that package astronomy applications including enabling machine learning use cases. Using CVMFS as a software distribution tool to enable integration and visibility at multiple compute sites.

ESAP service components under Task 5.3: Analysis interface, workflows, and reproducibility

ESAP services under this task:

- Interactive Data analysis
- Research object catalogue
- AAI

Contributing partners:

CSIC, CTAO, EGO, FAIR GMBH, FAU, IFAE, JIVE, KIS, NWO-I-ASTRON, RUG, SKAO, UEDIN

Individual contributions:

CSIC - *12 PM throughout the project*.
Perform a continuous evaluation of the level of reproducibility and achievement of the FAIR principles supported by the ESAP platform .

CTAO - *12 PM throughout the project*.
Implement a multi-messenger application into the ESAP, serve as the link to WP4 (definition of data model, VOevent etc) and WP3 (Services and SW).

EGO - *12 PM throughout the project*.
Contribute in integrating ESAP with existing tools.

FAIR GMBH - *6 PM throughout the project*.
Practice with analysis platform tools (JupyterHub and Swan)

FAU - *12 PM, 0.5 FTE from 01-01-2020 - 01-01-2022*.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Porting of KM3NeT developments on an analysis interface, workflow and reproducibility into the ESAP environment.

IFAE - *13 PM, 0,33 FTE in the period 01-07-2019 - 01-07-2020.*
Workflow implementation for the MAGIC-DL3 files production, using the staging layer provided by WP2 and the Data Lakes. Use REANA for further reproductions, batch jobs and CWL as a standard description language. Creation of Advanced data products for ingestion in GammaHub (DL3 to tables). Implementation in ESAP support for Apache Hadoop Stack. Study the implementation of a full virtualized cluster (GammaHub) to easy increase/decrease the number of nodes. Explore the integration of the GammaHub platform with IVOA standards.

JIVE - *18 PM, 0.46 FTE throughout the project.*
Analysis of the functionality of the JIVE data archive. Enable the re-running of pipelines with different parameters. Enable feedback from users to the archive.

KIS - *18 PM throughout the project*.
Contribute to visualisation tools integration from Q3 2020.

NWO-I-ASTRON - *24 PM, 0.5 FTE in the period 01-09-2019 - 30-04-2022*
Implement the ESAP interface which provides users ability to do interactive processing, and provide tools for the ingest of the derived data products back into the Data Lake.

RUG - *6 PM throughout the project*.
Practice with analysis platform tools (JupyterHub and Swan)

SKAO - *12 PM throughout the project*.
Exploring the current technologies and their suitability including running DIRAC workflows and JupyterHub instances.

UEDIN - *12 PM, 0.20 FTE in the period 01-04-2019 - 31-07-2022.*
Provide effort in building an analysis platform (JupyterHub/Swan/other) in the form of documentation, technical reports, scripts & automated configuration & deployment. Work on reproducible workflows for discovering, accessing and visualising any dataset using Virtual Observatory standards. Provide examples and documentation on running batch processing workflows in an Analysis Platform (i.e. use PySpark to run a job on a Spark cluster, in a Jupyter Notebook).

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

ESAP service components under Task 5.4: Integration with HPC and HTC infrastructures

- Batch data processing
- Data staging and access
- List of suggested Compute Resources
- Ingestion of advanced data products
- AAI

Contributing partners:

CERN, INAF, NWO-I-ASTRON, NWO-I-Nikhef, RUG, SKAO

Individual contributions:

CERN - *6 PM, 0.25 FTE in the period 01-01-2020 - 01-01-2022.*
Integrate the Science Platform prototype with the Data Lake prototype.

INAF - *12PM in the period 01.01.2020 - 01.01.2022.*
Work on the integration of the Science Platform prototype with the Data Lake, HPC and HTC resources (local and large computing centers). Perform benchmarks and evaluations. The LOFAR Italian infrastructure for INAF astronomers will be leveraged as use case. A Science Platform on HPC resources will be tested and provided for the Italian LOFAR community.

NWO-I-ASTRON - *6 PM, 0.5 FTE in the period 01-07-2020 - 30-06-2021.*
Integrate ESAP with workload management tools and the WP2 Data Lake.

NWO-I-Nikhef - *12 PM throughout the project.*
Consider the ability to use the services. Support software engineering practices.

RUG - *12 PM throughout the project*.
Integrate the Science Platform prototype with the Data Lake prototype deployed in WP2.

SKAO - *6 PM throughout the project*.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Exploring integration required for existing workload management tools that may provide the base layer of the ESAP.

6.2.2. ESAP Minimum Viable Product core team

In order to make the work of each WP5 partners visible and accessible from the science analysis platform, we use the micro-service architecture and API gateways to design the ESAP. A high-level architectural design is described in Section 5.

Work package 5 formed an MVP core team, the responsibility of which is to:

- Study existing technologies, tools, services and select an appropriate service to serve as the MVP of ESAP
- Figure out what issues need to be addressed before implementing/integrating each service
- Map the requirements translated from user stories to architectural design, and adjust the architectural design if needed.
- Iterate the process till the architectural design matches the requirements of user stories
- Engage partners to work out the detailed architectural design together with the MVP core team
- Engage partners to try out the basic implementation to evaluate API design
- Engage partners to use service mockup to test API design
- Keep track of every components and the interaction between them.

Having the Minimum Viable Product and the detailed architectural design in place, the core team will be able to engage partners to make their contributions available through ESAP APIs, such that service components can be added to the MVP step by step to produce the ESAP prototype.

6.3. Connection with other WPs

Many service components of ESAP will be provided by or connect to the work from other work packages, therefore, it's essential for work package 5 to follow the development within other work packages. Work package 5 doesn't only receive information from partners that participate in multiple work packages, we also actively communicate with other work package leads, participate in other work package meetings, and organise joint work packages event. In this way, we will make sure that the products from other packages can be smoothly integrated into ESAP prototype.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

7. Risk Analysis

In this section we describe the identified risks to the project, evaluate the probability of them occurring, their potential impact and finally what measures the work package partners need to take in order to avoid them.

Risk Description	Probability	Impact	Risk Management Measures
Retention of key staff, with domain knowledge	Low	High	Partners will be responsible to hire and retain the appropriate staff, and inform the WP leaders if they cannot.
Delays on dependant deliverables from other Work Packages	Medium	Medium	Ensure that there is continuous communication with other WP leaders, and that any delays in their deliverables can be accounted for in the planning of this WP.
Hardware availability	Low	High	The Work Package leaders will ensure that hardware for development & production will be planned for and available when needed.
Expertise on key technologies & tools (Jupyter, Machine learning, AAI etc)	Low	Medium	Proper planning of the required services that need to be implemented, along with frequent tech reviews where the team will discuss and overview these technologies and tools
Delays in development of identified services	Medium	Medium	Agile approach with a focus on a bottom-up approach, along with continuous meetings where progress on the different services is discussed

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Missing user requirements	Low	Low	An iterative bottom-up approach, where user feedback and requirements are reviewed and updated at each cycle

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.