



Project Title	European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures
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	2019-02-04
Duration of Project	42 Months
Project Website	www.projectescape.eu

D5.3 – Performance Assessment of Initial Science Platform Prototype

Work Package	WP5, ESFRI Science Analysis Platform
Lead Author (Org)	John D. Swinbank (ASTRON)
Contributing Author(s) (Org)	Ian Bird (CNRS-LAPP), Hugh Dickinson (Open University), Rafael Garrido (IAA-CSIC), Gareth Hughes (CTAO), José Ramón Rodón (IAA-CSIC), Susana Sánchez Expósito (IAA-CSIC), Giuliano Taffoni (INAF), Lourdes Verdes-Montenegro (IAA-CSIC), Harro Verkouter (JIVE), Nico Vermaas (ASTRON), Stelios Voutsinas (U. Edinburgh)
Due Date	2021-08-31 (M31)
Date	2021-08-26
Version	1.0

Dissemination Level

- | | |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | PU: Public |
| <input type="checkbox"/> | PP: Restricted to other programme participants (including the Commission) |
| <input type="checkbox"/> | RE: Restricted to a group specified by the consortium (including the Commission) |
| <input type="checkbox"/> | CO: Confidential, only for members of the consortium (including the Commission) |

Versioning and contribution history

Revision	Date	Description
1.0	2021-08-26	Version as submitted
0.4	2021-08-26	Incorporate feedback from Dickinson & Taffoni; improve formatting
0.3	2021-08-25	Incorporate feedback from WP5 members; add Executive Summary
0.2	2021-08-20	Use ESCAPE document format
0.1	2021-08-17	Initial draft for internal distribution to ESCAPE WP5

Disclaimer

ESCAPE – 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 reports on the current status of ESAP, the ESFRI Science Analysis Platform. ESAP is a key part of the interface between the services delivered by the ESCAPE project and the scientific community: it will provide a unified mechanism by which users can discover and interact with the data products, software tools, workflows, and services that are made available through ESCAPE, and it is designed to be extensible and flexible to adapt to the emergent requirements of future projects. The development of ESAP is the fundamental activity of ESCAPE Work Package 5.

ESAP, in and of itself, does not provide any compute or analysis capabilities. Rather, it acts as a broker between users and the various services which are available to them. For example, ESAP will help users identify datasets which are of interest to them (perhaps by interrogating the ESCAPE “data lake”, or an ESFRI-specific archive), to locate software and workflows which can help them analyze that data, and connect them to services which can execute analyses codes on their behalf (perhaps interactively, such as in a Jupyter notebook, or by scheduling jobs on a batch processing system).

ESAP abstracts the details of the various heterogeneous underlying systems from the user, so that they can use a unified, coherent interface to access all of the various services they need. It does this by adopting a modular, flexible architecture. The user will connect to a service-independent web-based *user interface*, which in turn communicates with the *API Gateway*. By adopting a set of standard programming interfaces and conventions, the Gateway can easily be extended to address whatever current or future capabilities are exposed through the EOSC.

We have assessed the performance of ESAP by comparing both its current, prototype, implementation, and the ultimate design vision, against community needs and expectations. These needs and expectations were derived in three ways:

- from the initial requirements developed in the early stages of the ESCAPE project;
- from the use cases that ESCAPE ESFRIs have been developing over the course of the project to date;
- by hosting a workshop, at which ESAP was presented to the community and their feedback was solicited.

Based on this evaluation, we conclude that ESAP is making good progress towards an ultimate vision that is well aligned with the needs of the scientific community. No major changes to the ESAP plans are expected as a result of this analysis, but the feedback received will play an important role in improving and refining plans for further development of ESAP over the remaining project duration.

This document is submitted as ESCAPE project deliverable D5.3, *Performance Assessment of Initial Science Platform Prototype*.

Contents

1	Introduction	8
2	The ESAP Vision	9
2.1	High-Level Summary	9
2.2	Conceptual Model	9
2.3	Major Capabilities	10
2.4	Extensibility and Supported Services	14
3	Current Capabilities	16
3.1	User Interface	16
3.2	Authentication and Authorization	16
3.3	Data Orchestration within ESAP	16
3.4	Data Discovery	18
3.5	SAMP	18
3.6	Interactive Data Analysis	19
3.7	Managed Database	20
4	Requirements	21
5	Use Cases	23
5.1	Cherenkov Telescope Array	23
5.2	JIVE	27
5.3	KM3NeT	28
5.4	LOFAR	29
5.5	SKA	30
6	Workshop	32
6.1	Logistics	32
6.2	Meeting Summary	32
6.3	Questionnaire	34
6.4	Outcomes	35
6.5	Lessons Learned for Future Workshops	35
7	Conclusions	37
A	Report on Project Milestone MS31	38
B	Post-Workshop Questionnaire	39

List of Figures

1	ESAP in its environment.	11
2	The high-level architecture of ESAP.	11
3	The front page of the ESAP test deployment at ASTRON.	17
4	Using ESCAPE IAM to authenticate with ESAP.	17
5	The “shopping basket” viewed through the ESAP web interface.	18
6	Query results displayed though ESAP.	19
7	The ESAP IDA workflow.	19

List of Abbreviations

- AAAI** Authentication, Authorization, and Accounting Infrastructure.
- ANTARES** Astronomy with a Neutrino Telescope and Abyss environmental RESearch.
- API** Application Programming Interface.
- CASA** Common Astronomy Software Applications.
- CEVO** Connecting ESFRI Projects to EOSC through VO framework.
- CTA** Cherenkov Telescope Array.
- DIOS** Data Infrastructure for Open Science.
- EOSC** European Open Science Cloud.
- ERIC** European Research Infrastructure Consortium.
- ESAP** ESFRI Science Analysis Platform.
- ESCAPE** European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures.
- ESFRI** European Strategy Forum on Research Infrastructures.
- ESO** European Southern Observatory.
- EVN** European VLBI Network.
- FAIR** Findable, Accessible, Interoperable, Reusable.
- GB** Gigabyte.
- GPU** Graphics Processing Unit.
- GUI** Graphical User Interface.
- HPC** High-Performance Computing.
- HTC** High-Throughput Computing.
- HTTP** Hypertext Transfer Protocol.
- IAM** Identity and Access Management.
- IDA** Interactive Data Analysis.
- IVOA** International Virtual Observatory Alliance.
- JIVE** Joint Institute for VLBI ERIC.
- JSON** JavaScript Object Notation.
- LHC** Large Hadron Collider.
- LOFAR** LOw Frequency ARray.
- OSSR** Open-source scientific Software and Service Repository.
- PID** Persistent Identifier.

QoS Quality of Service.

REST Representational State Transfer.

SAMP Simple Application Messaging Protocol.

SKA Square Kilometre Array.

SQL Structured Query Language.

UWS Universal Worker Service.

VLBI Very Long Baseline Interferometry.

VO Virtual Observatory.

WP Work Package.

1 Introduction

This document is European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE) project deliverable D5.3: “Performance Assessment of Initial Science Platform Prototype” [4].

The ESFRI Science Analysis Platform (ESAP) is the primary deliverable of ESCAPE Work Package 5 (WP5). A primarily web-based application, it serves a key part of the interface between the services delivered by ESCAPE and the scientific community, providing mechanisms by which users can discover, access, and analyze the data, workflows, and services which the European Strategy Forum on Research Infrastructures (ESFRIs) affiliated with ESCAPE, and other projects, publish to the European Open Science Cloud (EOSC).

At the time of writing, the ESCAPE project is approximately at its mid-point. A substantial amount of planning and development has been undertaken, and the project is now looking to consolidate on that start, by demonstrating increasing integration of the various services that have been developed, and by showing how they can be used to address the various use cases which are advanced by the associated ESFRIs. This makes it an opportune moment to step back and assess the current state and future plans for ESAP development. It is that analysis which is undertaken in this document.

This document assess the performance of ESAP by considering:

1. To what extent the prototype meets the requirements which were specified in *D5.2: Detailed Project Plan* [10]?
2. To what extent the prototype meets the needs currently being expressed by the ESCAPE-affiliated ESFRIs, as expressed through the use cases currently being collated by the project?
3. What other goals or desires for ESAP might the community express through a workshop and accompanying call for feedback?

Note that none of these topics directly address “performance” as narrowly defined in terms of raw throughput or latency. In practice, ESAP is unlikely ever to be a limiting factor here: it acts merely as a conduit to help users interact with other systems. Those other systems handle bulk data storage, transport and compute; they do not place substantial load on ESAP itself. Furthermore, while the ESAP system itself has been designed to be scalable where appropriate, at this stage in development we have focused on prototyping and exploring capabilities, rather than raw throughput. In short, therefore, a raw computational performance measurement is not regarded as being significant for the terms of this report.

This report is structured as follows. It starts by describing the vision for ESAP in §2. This outlines how the ultimate ESAP deliverable is foreseen: what will it do, and how will it be deployed? In §3, it moves on to describe the current state of ESAP development.

Those first two sections are the essential background to the second part of the report, which directly attempts to answer the questions outlined above. In particular, §4 tabulates the documented requirements and compares them to current development; §5 discusses how ESAP relates to all the use cases collected in the ESCAPE project platform; and §6 describes a workshop that was called to solicit community feedback. Finally, a summary of the results of this analysis is presented in §7.

2 The ESAP Vision

This section presents a brief overview of the vision for and current design of ESAP. It supplements and expands upon earlier discussions [4, 10] to describe current thinking about how ESAP can best meet its goals.

2.1 High-Level Summary

ESAP will provide a flexible system for analysing data made available through EOSC. It will assist users in engaging with the services provided in the other ESCAPE work packages by:

- providing a flexible interface for querying and retrieving data from a variety of archives and data repositories, with particular emphasis on those which are stored in or accessible through the services provided by ESCAPE WPs 2 (DIOS: Data Infrastructure for Open Science) and 4 (CEVO: Connecting ESFRI Projects to EOSC through VO framework), as well as the citizen science platforms addressed through WP6;
- enabling users to explore the software repositories, like the WP3 Open-source scientific Software and Service Repository (OSSR), to identify and select analysis tools and workflows which are appropriate to their needs;
- helping users to identify interactive data analysis and batch computing facilities which are accessible to them;
- facilitating the staging of data, software, and workflows to compute facilities, providing access to those facilities for end users, and subsequently retrieving the results of processing.

ESAP will be, by design, extensible: rather than attempting to anticipate every possible type of data repository, software, compute system, or other service provider, the platform will provide generic interfaces through which it can be extended to encompass new functionality.

In short, our approach is not to attempt to provide a single, integrated platform to which all researchers must adapt, but rather a set of functionalities from which various communities and research infrastructures can assemble an analysis platform geared to their specific needs. 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. At the same time, we expect that instances of ESAP may usefully be deployed in other contexts, from providing services to just a few users within a small project, to supporting major pieces of infrastructure; it must therefore be capable of operating effectively at a range of scales.

2.2 Conceptual Model

ESAP, in and of itself, provides no compute or analysis capabilities (beyond a simple ability to view tabular data and preview images). Rather, it acts as a broker between users and the various query and analysis services which are available to them. These might include, for example:

- bulk data query systems, which can help the user locate and access data files (images, visibility data, etc) in archives, data lakes, or similar bulk storage systems;
- tabular data query systems, which can help the user find relevant entries in source catalogues and similar relational systems;

- Interactive Data Analysis (IDA) systems, which provide the user compute and visualization tools in a convenient environment with access to relevant datasets (for example, a Jupyter [7] notebook, or containerized analysis application);
- bulk data processing systems, which provide batch (non-interactive) processing of data at-scale in HPC or HTC environments;
- scientific software repositories, which provide access to specialist analysis tools and workflows;

A given instance of ESAP is configured with information about available services¹. When a user connects, the ESAP instance should:

- help the user select services which are relevant to them (for example, by clearly presenting the available services; by making clear what science cases those services support, by taking account of the user's access privileges, etc);
- facilitate authentication and authorization with the various services, as necessary;
- provide a consistent and convenient way for the user to access services (for example, by providing the user with a single way to enter a particular query, and then automatically translating that to the requirements of each individual service);
- mediate data flow between services (for example, by enabling the user to locate data with an archive query, dispatch the data to the processing facility, and schedule processing of the data on a bulk data processing system).

This relationship is illustrated schematically in Fig. 1: this shows the end user communicating directly with ESAP, which mediates their interactions with a range of other services, deployed across a variety of different infrastructures.

Note that the user communicates with a single ESAP instance, while that instance mediates interactions with a range of different services from a variety of infrastructure providers.

2.3 Major Capabilities

2.3.1 User Interface

ESAP is primarily a web application: the central hub (the “API Gateway”, or “back-end”) runs on one or more servers, and users interact with it by making HTTP requests. The work package will provide a customizable front-end application (“ESAP-GUI”) which runs in the browser and communicates with the back-end. This separation of concerns is illustrated in Fig. 2. In principle, it may be possible to support alternative GUIs which communicate with the same back-end. Providing such alternatives is out of scope for this work package, but provides scope for future extension of the work if appropriate.

2.3.2 Authentication and Authorization

Users may be asked to log in to access ESAP itself, or to use some or all of the services mediated by a given ESAP instance.

¹This configuration is instance-specific: for example, a central EOSC installation of ESAP might provide access to a wide range of services, spanning the entire EOSC, while an institutional or project-level system may only be configured with information about local resources.

D5.3 – Performance Assessment of Initial Science Platform Prototype

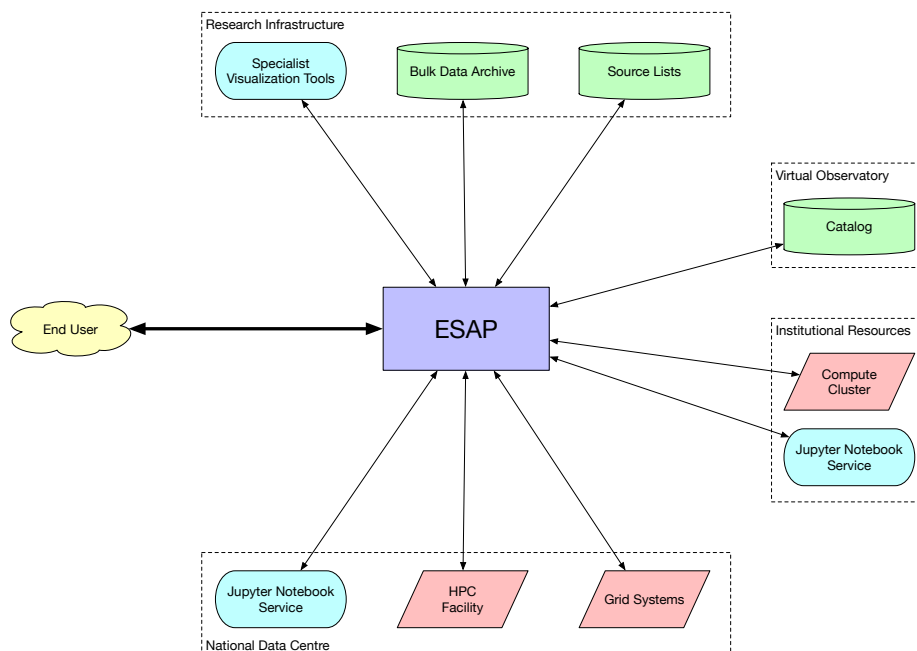


Figure 1: ESAP in its environment.

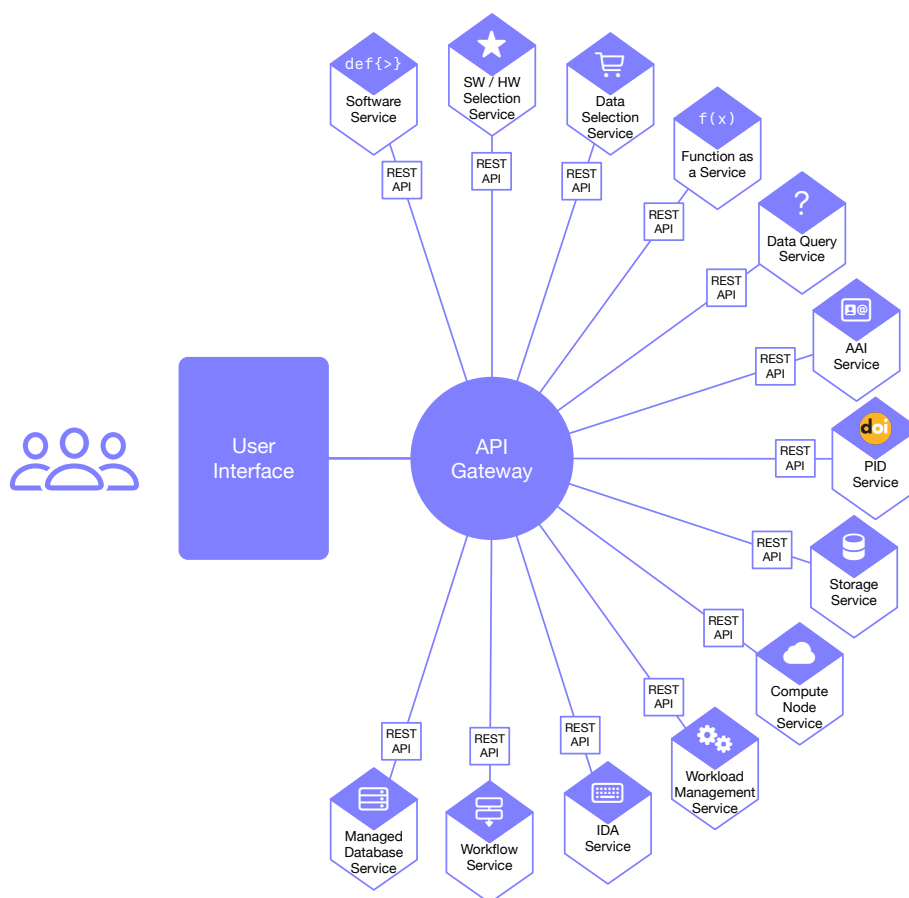


Figure 2: The high-level architecture of ESAP.

This step is not required: if both the owner of the ESAP instance and the owner of any services being accessed make them available to the general public, then ESAP need not force the user to log in. In general, however, users are expected to log in before using the data management services (§2.3.3).

ESAP as delivered by this work package will provide for user authentication through the ESCAPE Identity and Access Management (IAM) service². Where possible, ESAP is designed to be flexible and adaptable to other systems, but explicit support for other systems is outside the scope of this work package.

2.3.3 Data Orchestration within ESAP

The fundamental — if simplified — workflow envisioned for ESAP is that the user will query one or more archives to identify data of interest, then dispatch that data to IDA or bulk processing systems for processing.

To support this model, ESAP will maintain a per-user list of active data items: the “shopping basket”. This basket is persistent: (a representation of) the data the user has selected is serialized as JSON, and the results are stored in a database. Note that the basket is not generally expected to contain a complete representation of the data in question (it will not store multi-GB images or query results), but rather it will contain sufficient metadata that the data can be fetched and manipulated on demand (for example, it will store the query which produces the result in question, or a path or other identifier which enables data to be fetched from the “data lake” or other storage).

Services integrated with the ESAP system will be able to edit, augment, and update the contents of the users’ shopping basket.

The shopping basket metaphor will be extended to include services — such as IDA or batch compute facilities — and workflows from the OSSR and other repositories.

2.3.4 Data Discovery and Staging

ESAP will provide a uniform interface which enables users to dispatch queries to a multiplicity of archive services. These will include both federated, multi-facility systems such as the Virtual Observatory (VO) and facility- or ESFRI-specific archives. It also includes the “data lake” being developed as part of the Data Infrastructure for Open Science (DIOS) system in ESCAPE WP2.

The data discovery system will adapt itself dynamically to the type of archive being queried. For example, it will be possible to query astronomical archives by using astronomy-specific parameters such as the celestial position where appropriate.

When data of interest to the user has been located, if appropriate it will be possible to arrange for the data to be “staged” — that is, to be moved from the archive to storage which is available with low-latency from an appropriate analysis system.

2.3.5 SAMP

ESAP aims to provide full support for the International Virtual Observatory Alliance (IVOA) Simple Application Messaging Protocol (SAMP) [14]. This makes it possible for users of other SAMP-compliant tools — including TOPCAT [15], Aladin [2] and Astropy [16] — as well as archive interfaces like ESASky [9] to exchange data with

²<https://iam-escape.cloud.cnaf.infn.it/login>

ESAP. This means that users can take advantage of the advanced querying and data manipulation capabilities provided by these tools and facilities in conjunction with the possibilities offered by ESAP, maximizing interoperability and avoiding duplication of effort.

2.3.6 Interactive Data Analysis

IDA describes a scientist interacting with a dataset in real time to perform their analyses. That is, they type commands or manipulate controls, and observe the results that are produced or the figures that are displayed. Contrast this with batch processing, discussed in §2.3.7.

The processes and tools required for IDA differ substantially from field to field and from facility to facility. For example, the way that data from the Square Kilometre Array (SKA) will be analyzed is very different to the processes applied to data from the Large Hadron Collider (LHC). It is therefore essential that ESAP implement a flexible capability for interfacing with a variety of IDA services.

The architecture described in §2.3.1, together with the data management system described in §2.3.3, are designed to make this possible. Specifically, this will be implemented by developing APIs through which ESAP can provide elements of the “shopping basket” to the IDA system — including both data and software specifications — and then by accepting appropriately authenticated updates from the IDA system as the user saves their analysis. The expectation is that the IDA system will write substantial data products (such as output images) to bulk storage (such as the DIOS data lake), and return references to them to ESAP for further analysis.

2.3.7 Batch Data Processing

Batch data processing describes a situation which is in many ways similar to IDA (§2.3.6), but with a number of significant differences:

- the work is carried out asynchronously: the user submits a job, and then returns some time later to examine the results;
- the user does not interact with the computing systems while processing takes place;
- processing generally happens at scale, perhaps being distributed over multiple computing systems.

ESAP will support this by:

- providing a generic API for interacting with batch compute systems, combined with one or more adaptations of this interface to specific systems;
- providing a user interface for asynchronous processing, where ESAP tracks the progress of user jobs, and notifies the submitter when they are complete.

2.3.8 Service and Software Discovery

ESAP will provide deep integration with the OSSR, and other repositories of software and services if appropriate. This will make it possible for users to discover capabilities which are of relevance to them. In particular, ESAP will help users discover software workflows and compute and storage infrastructure which can be used to execute both IDA and batch processing tasks (as described in §§2.3.6 & 2.3.7).

The user should be provided with a range of help in identifying software and services which are of relevance to their needs. That is, based on metadata sourced from the OSSR, ESAP should help the user make informed

decisions based on criteria such as (but not limited to):

- software which is capable of processing the types of data stored in their shopping basket (§2.3.3);
- software which is appropriate for the type of analysis they wish to perform (addressing particular science goals, capable of being executed in batch or interactive mode, etc);
- services which are capable of executing the workflow or software package which the user has selected;
- services which are local to the storage location of bulk data, or which can instantiate efficient bulk data movement.

2.3.9 Managed Database

The ESAP Managed Database service is a new capability, first publicly proposed and discussed at the Second WP5 Workshop (§6). The Managed Database service provides users with the capability to define and use their own relational databases directly within the ESAP system. It is possible to directly load the results of queries against external archives into the user's database space, and then to submit complex Structured Query Language (SQL) queries to the database system. This provides the user with advanced data analysis capabilities — for example, the ability to perform complex catalogue cross-matching — without requiring that they set up and administer their own database system. Further, it opens the prospect of integrating ESAP with external SQL federation services such as Trino³ or openLookeng⁴.

2.3.10 Provenance and PIDs

Processing, controlled and mediated through ESAP, will result in *advanced* data products: refined, augmented, or reduced versions of the input data. These data products, taken together with the workflows that have been used to produce them and resulting scientific publications, form the *research objects* which are the fundamental outputs of the scientific community. In order to facilitate Findable, Accessible, Interoperable, Reusable (FAIR) access to data, ESAP will provide mechanisms for tracking the provenance of these research objects and will assist users in providing them with Persistent Identifiers (PIDs) [3].

2.4 Extensibility and Supported Services

As described in §§2.2 & 2.3.1 above, the ESAP system is designed to be intrinsically extensible: the core API Gateway provides generic interfaces into which additional services can be integrated with minimal effort. The recent addition of the (prototype) Managed Database service (§2.3.9) validates this concept: this service was integrated with ESAP over the course of no more than a few weeks.

However easy it is to integrate services with ESAP, it is clearly impossible for the ESCAPE team to integrate *all possible* services: there are simply too many domain-specific tools in use in the scientific community for this to be practical. Instead, we focus on:

- providing a number of service integrations which demonstrate key capabilities and facilitate the expressed science goals and use cases of ESCAPE-affiliated ESFRIs;
- providing documentation and examples to make it possible for new services to be quickly and easily integrated with ESAP without direct intervention from the ESCAPE team.

³<https://trino.io>

⁴<https://openlookeng.io>

D5.3 – Performance Assessment of Initial Science Platform Prototype

The detailed list of service integrations which will be supplied in the core ESAP delivery by ESCAPE WP5 is still under development. However, we expect to provide at least:

- data query and data discovery based on major ESFRI archives;
- integration with Rucio-based [1] data lake systems;
- VO query capabilities;
- SAMP integration;
- integration with Jupyter-based IDA facilities, probably based around BinderHub [6] and/or Rosetta⁵;
- integration with at least one batch computing service, probably through DIRAC [13].

⁵<https://rosetta.oats.inaf.it/main/>; <https://github.com/sarusso/Rosetta>

3 Current Capabilities

This section provides an overview of current ESAP capabilities and describes how they relate to the vision described in §2.

The capabilities described here are available through a test deployment of ESAP at ASTRON. This system may be accessed at <https://sdc-dev.astron.nl/esap-gui/>. Note that this system is provided without support to facilitate development and demonstration; the service is not expected to be reliable, nor to be available indefinitely.

3.1 User Interface

The front page of the ESAP test deployment at ASTRON is shown in Fig. 3. It presents an attractive and usable web-based interface that enables the user to rapidly discover which services are configured in this ESAP instance (in the top bar; here, Archives, Interactive Analysis and IVOA-SAMP). The interface highlights the available archives: in this case, corresponding to Apertif [12], the ASTRON VO system⁶, and the Zooniverse citizen science system⁷.

As discussed in §2.3.1, the user interface shown is a separate process from the core “API Gateway” back-end system for ESAP. The interface is a cross-platform web application implemented in React⁸, which communicates with the API Gateway over a REST interface. The API Gateway itself is written in Python, using Django⁹ and its companion REST framework¹⁰. In principle, it would be possible for alternative or specialist user interfaces to communicate with the API Gateway using the same interface; as of now, however, the authors are aware of no other ESAP interfaces.

3.2 Authentication and Authorization

ESAP is fully integrated with the ESCAPE project’s IAM service. Fig. 4 shows an example of a user authenticating with the ESAP test system using ESCAPE IAM.

After the user has been authenticated, ESAP should automatically be able to forward their credentials to downstream services, to prevent the user from having to re-authenticate multiple times. Final integration of this capability is still ongoing.

3.3 Data Orchestration within ESAP

Shopping basket capabilities, as described in §2.3.3, are available in the current version of ESAP to users who have authenticated through ESCAPE IAM (§3.2).

Fig. 5 shows an example of the shopping basket visualized through the web interface. Note that the formatting of the contents of the basket varies depending on the source of the data.

Future work will provide additional flexibility for managing the contents of the shopping basket.

⁶<https://vo.astron.nl/>

⁷<https://www.zooniverse.org/>

⁸<https://reactjs.org/>

⁹<https://www.djangoproject.com/>

¹⁰<https://www.django-rest-framework.org/>

D5.3 – Performance Assessment of Initial Science Platform Prototype

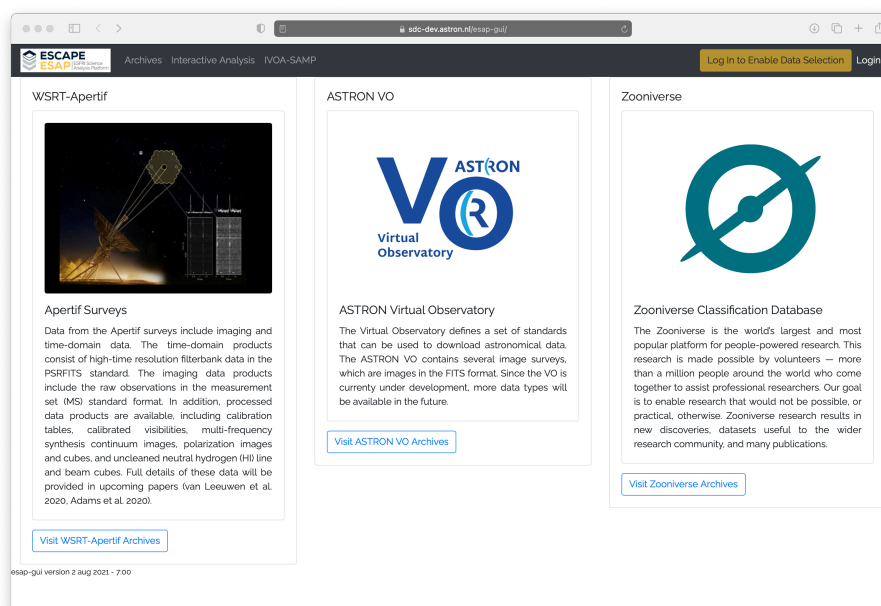


Figure 3: The front page of the ESAP test deployment at ASTRON.

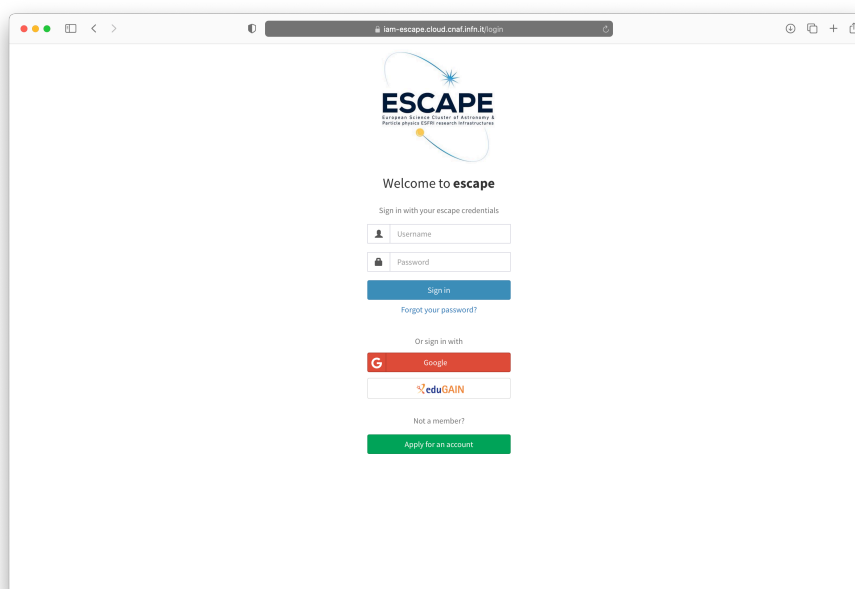
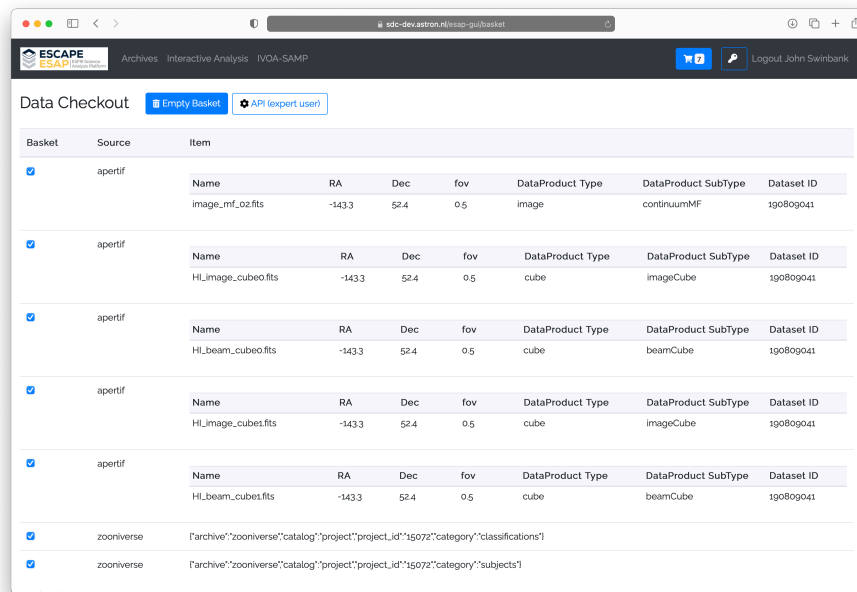


Figure 4: Using ESCAPE IAM to authenticate with ESAP.



Basket	Source	Item	Name	RA	Dec	fov	DataProduct Type	DataProduct SubType	Dataset ID
<input checked="" type="checkbox"/>	apertif		Name	RA	Dec	fov	DataProduct Type	DataProduct SubType	Dataset ID
			image_mf_02.fits	-143.3	52.4	0.5	image	continuumMF	190809041
<input checked="" type="checkbox"/>	apertif		Name	RA	Dec	fov	DataProduct Type	DataProduct SubType	Dataset ID
			HI_image_cube0.fits	-143.3	52.4	0.5	cube	imageCube	190809041
<input checked="" type="checkbox"/>	apertif		Name	RA	Dec	fov	DataProduct Type	DataProduct SubType	Dataset ID
			HI_beam_cube0.fits	-143.3	52.4	0.5	cube	beamCube	190809041
<input checked="" type="checkbox"/>	apertif		Name	RA	Dec	fov	DataProduct Type	DataProduct SubType	Dataset ID
			HI_image_cube1.fits	-143.3	52.4	0.5	cube	imageCube	190809041
<input checked="" type="checkbox"/>	apertif		Name	RA	Dec	fov	DataProduct Type	DataProduct SubType	Dataset ID
			HI_beam_cube1.fits	-143.3	52.4	0.5	cube	beamCube	190809041
<input checked="" type="checkbox"/>	zooniverse	{"archive":"zooniverse","catalog":"project","project_id":"15072","category":"classifications"}							
<input checked="" type="checkbox"/>	zooniverse	{"archive":"zooniverse","catalog":"project","project_id":"15072","category":"subjects"}							

Figure 5: The “shopping basket” viewed through the ESAP web interface.

3.4 Data Discovery

Interfaces between ESAP and a variety of archive systems have been implemented, to varying degrees of polish and reliability. At time of writing, the Apertif, ASTRON VO and Zooniverse interfaces are the most reliable and functional. Examples are shown in Fig. 6. Note that the query interface and the form of the results returned adapt appropriately depending on the nature of the archive in question, so that — for example — the user is given the opportunity to specify celestial coordinates when querying the Apertif archive, but not when querying the Zooniverse system.

The leftmost columns in the result interface give the user the opportunity to select individual rows from the results to add them to their shopping basket (§3.3). Future versions of ESAP will offer more flexibility here; this might include, for example, bulk addition of many rows to the basket without selecting each individually.

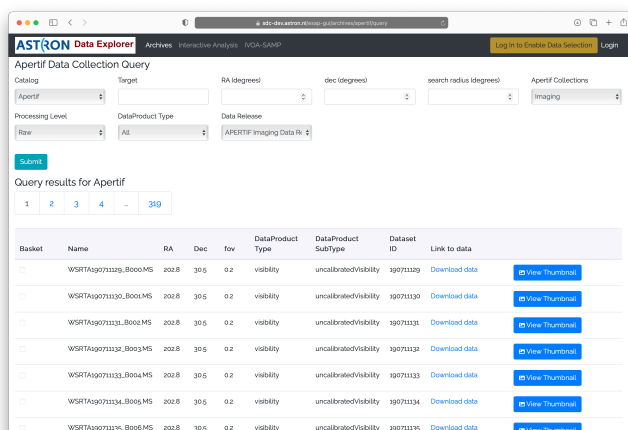
Currently, archives are queried individually and synchronously. That is, the query is targeted towards one archive, and the user interface blocks until a response is received. This may not be appropriate for all archives, in particular those which offer access to extremely large catalogues. Future upgrades are expected to include dispatching a query to multiple archives (with appropriate conversion from the form in which it is specified by the user to service-specific interfaces), and asynchronous queries (in which the user can log off and return later to inspect the query results), perhaps based on the IVOA Universal Worker Service (UWS) pattern [5].

3.5 SAMP

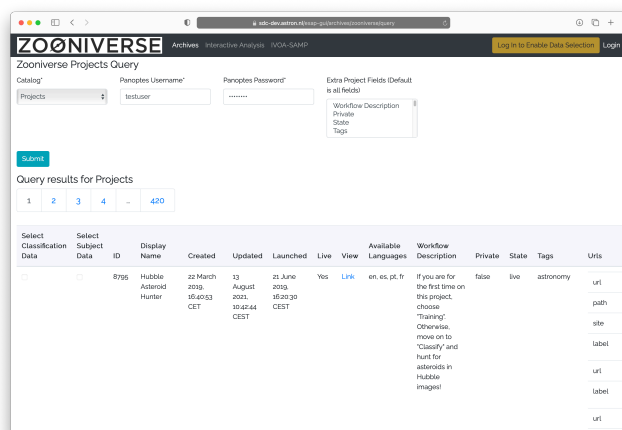
ESAP incorporates support for SAMP using the sampjs library¹¹. Users can transmit data to the ESAP web interface from other SAMP-enabled applications, and from there add it to their shopping basket. Transmission in the other direction — from ESAP to other applications — is not yet available.

¹¹<https://github.com/astrojs/sampjs>

D5.3 – Performance Assessment of Initial Science Platform Prototype

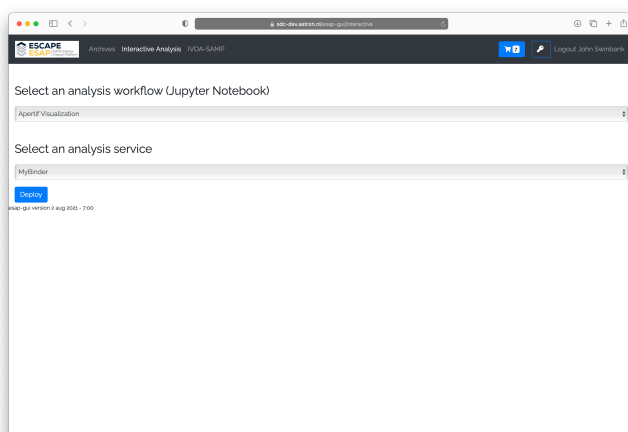


(a) Apertif.

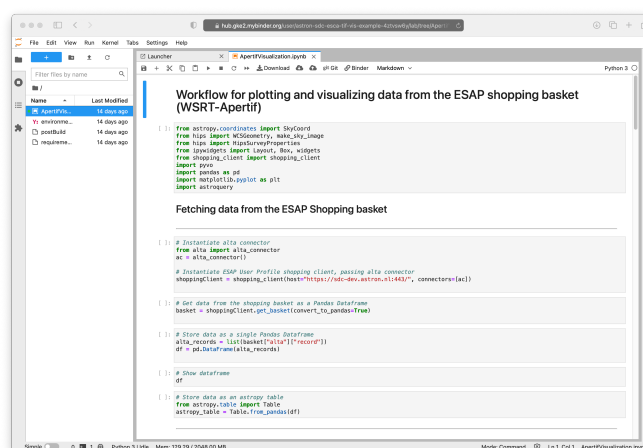


(b) Zooniverse.

Figure 6: Query results displayed though ESAP.



(a) Workflow selection.



(b) Jupyter notebook.

Figure 7: The ESAP IDA workflow.

3.6 Interactive Data Analysis

The current version of ESAP directly embeds information which enables it to execute a limited number of workflows, described in the form of Jupyter notebooks. A user interface exists to make it possible to execute these notebooks on a number of different analysis systems, but the existing test system only provides access to MyBinder¹².

Upon choosing an appropriate workflow and analysis service, the user is redirected to the notebook environment. In that environment, a Python library — initially developed specifically to address Zooniverse classification data, but now adapted to a wide range of data types — makes it possible for them to access their ESAP shopping basket, and hence to download or otherwise manipulate the data that they have selected.

Selecting the workflow and execution service, then working in the notebook, are shown in Fig. 7.

¹²<https://www.mybinder.org/>

There some usability and polish issues remain outstanding regarding shopping basket integration in the notebook: currently, the user must manually use the Python API to access data, including supplying an appropriate authentication token (copied-and-pasted from the ESAP interface). Ultimately, this process should be improved: data should be automatically staged to the notebook system, and no additional authentication should be required.

This system is not yet integrated with the WP3 OSSR, or any other external software or service collections. Discussions are currently underway between WP3 and WP5 to clarify these interfaces. When complete, this integration will provide a much richer experience for the user to discover and schedule workflows.

Although the current IDA system focuses on Jupyter notebooks, we expect to extend this service to address other forms of interaction in future work, likely making use of the Rosetta system (§2.4).

3.7 Managed Database

Development of the Managed Database service, described in §2.3.9 started in summer 2021. A functional prototype is now available, and development is progressing rapidly. As yet, however, the service is not fully integrated with the ESAP deployment at ASTRON.

4 Requirements

ESCAPE deliverable *D5.2: Detailed Project Plan* [10] defines a series of functional requirements on the ESAP system. At this stage in development, it is not expected that all of the requirements have been met. However, this is an opportune moment to consider the how far the current capabilities, as described in §3, show progress towards meeting those requirements. This is summarized in the table below.

ID	Description	Current Status
R-1	Users should be able to get a list of available data, searchable by different criteria, including keyword, science domain, institute, datatype etc	Basic query interfaces which meet these requirements are available, as shown in §3. Development and refinement of sophisticated query interfaces is expected to continue throughout the project duration.
R-2	Users should be able to get a list of known (VO & other) tools and software for users & publishers.	This requires integration with the ESCAPE WP3 OSSR. At the time of writing, work on that integration is ongoing but incomplete; it is not currently possible to publish the results of an OSSR query in ESAP, but a (brief) list of workflows is available, as shown in §3.
R-3	Users should be able to, for a given project & dataset, query for metadata and aggregate information (i.e. find location of data).	Included in the query system shown in §3.
R-4	Users should be able to stage a given dataset at the appropriate facility.	Planned, subject to ongoing integration with WP2 (DIOS).
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.	Basic capabilities are currently available for IDA jobs, and the same capabilities will ultimately be made available for batch computing when it is integrated. Advanced capabilities will be based on deeper integration with DIOS and OSSR, which will be forthcoming.
R-6	The 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.	This requirement is handled through access restrictions at the service level.
R-7	Users should be able to select from an existing list of Workflows (Notebooks) and either download, or deploy on available facilities.	This functionality is currently available, as shown in §3
R-8	Users should be able to assign PIDs to every digital object that is part of a Research Object.	This capability is planned, but not yet available.

D5.3 – Performance Assessment of Initial Science Platform Prototype

ID	Description	Current Status
R-9	User generated data needs to be queryable via ESAP.	ESAP provides access to all data which is available through configured archives, including user-generated data.
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.	This is an interaction between the analysis service and the archive, which is not mediated by ESAP in the current design.
R-11	Users should be able to select computing facilities on the basis of their capacity. E.g. an HPC resource with a specific acceleration (such as a GPU) might be needed because the software to be run requires it.	This capability is planned, but not yet available.
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.	This capability is planned, but not yet available.
R-13	Users should be able to schedule computational tasks at regular intervals e.g. to periodically retrieve new classification data from a citizen science experiment.	This capability is planned, but not yet available.

In summary, of the thirteen requirements described, five have been met (R-1, R-3, R-5, R-7, R-9), although enhancements to the implementation are ongoing, six are currently the subject of ongoing work (R-2, R-4, R-8, R-11, R-12, R-13), and two are satisfied by the integration of ESAP with other ESCAPE infrastructure (R-6, R-10).

5 Use Cases

The ESCAPE project has formalized its collection of use cases on the project platform¹³. Each of the ESFRIs involved in the project has been updating and modernizing its use cases and incorporating them into this platform, a task that is still ongoing at time of writing. This section focuses on those use cases which have been included in the platform to date, considering how each one can be addressed by the current or envisioned capabilities of ESAP. Additions and refinements to the material presented here are expected as new use cases are added; these updates will be collated within the platform itself, rather than by updates to this document. A full description of exactly how each use case might be executed is out of scope for this document; instead, we focus on highlighting those areas of the use cases where ESAP may be expected to contribute, and how that relates to the ESAP vision described in §2.

Use cases are organized by the originating ESFRI.

5.1 Cherenkov Telescope Array

5.1.1 CTA001: Long haul ingestion and replication

This use case has no direct impact on ESAP development.

5.1.2 CTA002: Data Reprocessing

Each of the stages of the use case may be addressed by ESAP as follows:

1. *Raw data is identified on tape via metadata e.g. using `getMetaData` method.*
 - ESAP provides flexible routines for searching archives via metadata, as described in §2.3.4.
 - Currently, ESAP does not integrate with the Cherenkov Telescope Array (CTA) archive, but this functionality is planned.
2. *Data volume is calculated.*
 - ESAP does not directly provide mechanisms for calculating aggregates over selected data, but extension to address this part of the use case is likely straightforward.
 - Arbitrary calculations based on metadata are possible by routing the workflow through a Jupyter notebook using the functionality described in §2.3.6.
3. *Data is staged from tape storage to temporary disk.*
 - Data staging functionality is planned for ESAP, as described in §2.3.4, but has not yet been implemented.
4. *Data is reprocessed using CTA pipeline software via the workload management system using a cache area for on-the fly, transient data products.*
 - Batch processing functionality is planned for ESAP, as described in §2.3.7, but has not yet been implemented.
5. *Final data products are verified.*

¹³<https://project.escape2020.de>

- Details of the verification process are not specified in this use case, but it seems like that ESAP's IDA functionality, described in §2.3.6, would be appropriate.

6. *Cache and temporary data is cleared.*

- No direct impact on ESAP, although potentially an interface could be added to ESAP to enable convenient management of the relevant data.

7. *Ingest the resulting new data into the data lake.*

- Data can be transmitted from analysis systems to the data lake without the direct involvement of ESAP. However, ESAP may contribute to providing it with an appropriate PID, as described in §2.3.10.

8. *Update the corresponding metadata.*

- No direct connection to ESAP.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the project's vision (§2), although further work is required to complete delivery of these capabilities.

5.1.3 CTA003: Generation of Instrument Response Function

Each of the stages of the use case may be addressed by ESAP as follows:

1. *User input of parameters. (Time period, model, systematic uncertainty)*

- This would take place in an IDA environment provisioned through ESAP.

2. *Validation of parameters.*

- This would take place in an IDA environment provisioned through ESAP.
- Appropriate software and services would be sourced from the OSSR and included in the environment by ESAP.

3. *Confirm valid simulation does not already exist.*

- This would take place in an IDA environment provisioned through ESAP.
- Appropriate software and services would be sourced from the OSSR and included in the environment by ESAP.

4. *Calculate number of required events.*

- This would take place in an IDA environment provisioned through ESAP.
- Appropriate software and services would be sourced from the OSSR and included in the environment by ESAP.

5. *Search for compute resources.*

- This will be addressed through ESAP's service discovery functionality (§2.3.8).

6. *Job submission via workflow management system.*

- This will be addressed through ESAP's batch computing functionality (§2.3.7).

7. *Validation of simulated instrument response function.*

- This will take place in an IDA environment provisioned through ESAP.
- ESAP's data discovery functionality (§2.3.4) can be used to identify the appropriate simulated data for use in the IDA environment.

8. *Ingest instrument response function in the data lake with appropriate metadata.*

- Data can be transmitted from analysis systems to the data lake without the direct involvement of ESAP. However, ESAP may contribute to providing it with an appropriate PID, as described in §2.3.10.

9. *Remove simulated data.*

- No direct impact on ESAP, although potentially an interface could be added to ESAP to enable convenient management of the relevant data.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the project's vision (§2), although further work is required to complete delivery of these capabilities.

5.1.4 CTA004a: Interactive analysis of (simulated) CTA science data by a principal investigator

Each of the stages of the use case may be addressed by ESAP as follows:

1. *User logs in to ESAP and is identified as a CTA project principal investigator.*

- ESAP is integrated with ESCAPE project Authentication, Authorization, and Accounting Infrastructure (AAAI), as described in §2.3.2.
- ESAP itself does not track project affiliation or access rights, but rather delegates this to the services which directly manage data access.

2. *Search for (simulated) CTA high-level data by project identifier.*

- Addressed by §2.3.4.

3. *Search for corresponding instrument response function for the data selected.*

- Addressed by §2.3.4.
- Further integration work is required to make it straightforward to use the outputs of one ESAP query as the inputs to a further query.

4. *Search for corresponding metadata, log files etc.*

- Addressed by §2.3.4.

5. *The data can now be analysed in interactive mode using Jupyter.*

- Addressed by §§2.3.6 & 2.3.8.
- ESAP will assist by identifying appropriate analysis workflows, identifying IDA services capable of executing those workflows, and making it straightforward for the user to access them together with the data.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the project's vision (§2), although further work is required to complete delivery of these capabilities.

5.1.5 CTA004b: Batch analysis of (simulated) CTA science data by a principal investigator

This use case proceeds as CTA004a (§5.1.4) until step 6, where analysis is performed by a batch processing system rather than interactively. The same comments apply to ESAP's role in this workflow, with the exception that batch processing is described by §2.3.7 rather than §2.3.6.

5.1.6 CTA005: Analysis of a (simulated) AGN using a combined workflow (gammapy & AGNpy)

Each of the stages of the use case may be addressed by ESAP as follows:

- *User logs in to ESAP and discovers data based on their own data rights.*
 - Addressed as per CTA004a (§5.1.4).
- *Search for and select simulated CTA data in the data lake.*
 - Addressed as per CTA004a (§5.1.4).
- *Search for the corresponding instrument response function for the data selected.*
 - Addressed as per CTA004a (§5.1.4).
- *Search for the corresponding metadata, log files etc.*
 - Addressed as per CTA004a (§5.1.4).
- *Analyze the data in an interactive session, including identifying an appropriate workflow and execution service, making data available in the interactive environment, and storing the results.*
 - Addressed as per CTA004a (§5.1.4).

In short, all aspects of this use case which are of relevance to ESAP are addressed in the project's vision (§2), although further work is required to complete delivery of these capabilities.

5.1.7 CTA006: Combined CTA + KM3NeT Analysis

The workflow for this use case is the same as that for CTA005 (§5.1.6).

5.1.8 CTA007: Search for public CTA high-level data via the VO

Each of the stages of the use case may be addressed by ESAP as follows:

- *User navigates to ESAP.*
 - ESAP is conveniently accessible through its web user interface; §2.3.1.
- *User finds CTA data via the ESAP VO service.*
 - The services accessible through a given instance of ESAP are not fixed, as described in §2.2; not every instance of ESAP will necessarily provide a VO service.

- However, the ESAP data discovery system (§2.3.4) is explicitly being designed with the ability to query the VO in mind.
- It is expected that ESAP will ship with VO support in the core deliverable (§2.4).
- *The data is then combined with other data using VO tools (e.g. Aladin lite)*
 - ESAP’s “shopping basket” (§2.3.3) is explicitly designed to make it straightforward to collect data from multiple sources and dispatch it to e.g. IDA environments.
 - ESAP’s SAMP capability, §2.3.5, is designed to enable interoperability with VO tools.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the project’s vision (§2), although service availability may vary with ESAP configuration.

5.2 JIVE

The stated goal of this use case is that *the scientist can find and read European VLBI Network (EVN) data through ESAP and perform analyses on (public) data sets*. Specific items listed are:

- *Provide EVN data sets in the VO.*
- *Offer a Jupyter notebook for standard calibration and data reduction in the OSSR.*
- *Offer a radio-astronomy-specific Jupyter kernel based on the Common Astronomy Software Applications (CASA) [8] project libraries and tools for handling of radio astronomy data.*
- *Provide a JupyterHub service at JIVE where users can execute the notebook close to the EVN archive, hosted at JIVE*

The following functionality is required from ESAP to address these goals:

- ESAP must be able to discover EVN data.
 - Data discovery is explicitly part of ESAP, as discussed in §2.3.4.
 - It is expected that ESAP will ship with VO support in the core deliverable (§2.4).
- ESAP must be able to access Jupyter notebooks made available through the OSSR.
 - This is an explicit goal of ESAP’s IDA system, as described in §2.3.6.
 - The current prototype is able to start the execution of Jupyter notebooks, but OSSR integration is still a work-in-progress, as described in §3.6; a preliminary version of that functionality is expected in the third quarter of 2021.
- ESAP must be able to access the Jupyter-CASA kernel made available through the OSSR.
 - Software and services would be sourced from the OSSR and included in the environment by ESAP.
- ESAP must be able to access the JupyterHub service made available through the OSSR.
 - Software and services would be sourced from the OSSR and included in the environment by ESAP.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the project’s vision (§2), although service availability may vary with ESAP configuration.

5.3 KM3NeT

5.3.1 ANTARES data set analysis

The stated goal of this use case is that *legacy data from ANTARES is provided by KM3NeT and used as KM3NeT full analysis example*. Specific items listed are:

- Offer ANTARES data through VO.
- Create and offer ANTARES example notebook.
- Offer data services for data interpretation.

The following functionality is required from ESAP to facilitate these goals:

- ESAP should be able to discover data in the VO.
 - Data discovery is explicitly part of ESAP, as discussed in §2.3.4.
 - Support for the VO is explicitly planned (§2.4).
 - The current prototype integrates with ASTRON VO services only (§3.4), but this will be extended with future development.
- ESAP should facilitate an interactive analysis service, including Jupyter notebooks.
 - This is an explicit goal of ESAP's IDA system, as described in §2.3.6.
 - The current prototype is able to start the execution of Jupyter notebooks; work on deeper integration is ongoing.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the current ESAP vision; development to provide full functionality is ongoing.

5.3.2 KM3NeT data releases

The stated goal of this use case is that *the scientist can find and read KM3NeT event data through ESAP and perform analyses on public data sets*. Specific items listed are:

- Provide data set in data lake.
- Offer Jupyter notebook for dummy analysis in the OSSR.
- Offer KM3NeT-specific libraries for handling of data.
- Check use of Rucio in KM3NeT.

The following functionality is required from ESAP to address these goals:

- ESAP must be able to discover and stage data in the (Rucio-based) data lake.
 - Data discovery is explicitly part of ESAP, as discussed in §2.3.4.
 - Support for Rucio is explicitly planned (§2.4); an advanced prototype exists, but it is not currently deployed for use (§3.4).
 - Further Rucio integration is expected in the second half of 2021.

- ESAP must be able to access Jupyter notebooks made available through the OSSR.
 - This is an explicit goal of ESAP's IDA system, as described in §2.3.6.
 - The current prototype is able to start the execution of Jupyter notebooks, but OSSR integration is still a work-in-progress, as described in §3.6; a preliminary version of that functionality is expected in the third quarter of 2021.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the current ESAP vision, and prototype versions of currently-unavailable functionality are expected shortly.

5.4 LOFAR

Note that LOFAR is not an ESFRI, but nevertheless provides a number of important use cases that are tracked through the ESCAPE project platform, and it serves as an important precursor to the SKA.

5.4.1 LOFAR0001: Long haul ingestion and replication

This use case has no direct impact on ESAP development.

5.4.2 LOFAR002: Data Processing

The stated goal of this use case is that *the ability to process data that is in the data lake at an external location*.

The following ESAP functionality is relevant:

- ESAP must be able to discover and stage data in the (Rucio-based) data lake.
 - Data discovery is explicitly part of ESAP, as discussed in §2.3.4.
 - Support for Rucio is explicitly planned (§2.4); an advanced prototype exists, but it is not currently deployed for use (§3.4).
 - Further Rucio integration is expected in the second half of 2021.
- ESAP must be able to schedule interactive or batch processing workflows on data discovered in the data lake.
 - These are explicit goals of ESAP development, as described in §§2.3.6 & 2.3.7.
 - Current IDA capabilities are described in §3.6.
 - Batch computing capabilities are not currently available in ESAP, but their development is expected during late 2021.
- ESAP must be able to (visually) inspect relevant intermediate and final results.
 - These are explicit goals of ESAP development, as described in §2.3.6.
 - Current IDA capabilities are described in §3.6.

In summary, all aspects of this use case which are of relevance to ESAP are addressed in the current ESAP vision, and prototype versions of currently-unavailable functionality are expected shortly.

5.4.3 LOFAR003: Integration of the existing (LOFAR) LTA in the data lake

This use case has no direct impact on ESAP development.

5.5 SKA

5.5.1 SKA global data lake proof-of-concept for astronomy-scale data products

This use case has no direct impact on ESAP development.

5.5.2 JHub notebook access to data lake data products

The goals for this use case may be addressed as follows:

- *Store SKA data in data lake managed by Rucio.*
 - This has no direct impact on ESAP.
- *Retrieve this data from data lake into a Notebook environment.*
 - ESAP will provide data-discovery tools for use the data lake, as described in §§2.3.4 & 2.4.
 - ESAP's Jupyter integration will make it possible to seamlessly access that data from a notebook environment, as described in §2.3.6.
- *User logs in to JupyterHub via ESCAPE IAM, and is then able to interact with the data lake seamlessly without re-uploading credentials.*
 - As described in §3.6, in the current prototype implementation it is necessary for the user to supply credentials, in the form of a token, directly into the notebook environment. This limitation will be addressed in a future development.

In short, all aspects of this use case which are of relevance to ESAP are addressed in the current ESAP vision, and prototype versions of currently-unavailable functionality are expected shortly.

5.5.3 Data product replica prepared for compute on request, interactive session started

The goals for this use case may be addressed as follows:

- *Identify Rucio data via ESAP, take the data identifier to a JupyterHub server running the Rucio-JupyterLab extension environment, download it, calculate checksum.*
 - Identifying data in Rucio is covered in §§2.3.4 & 2.4.
 - Integration of ESAP with Jupyter systems is addressed in §2.3.6.
 - Analysis within the notebook (e.g. calculation of checksums) falls outside the scope of ESAP itself, but should certainly be possible.
- *Do above but with a custom Docker image for the user's environment (via BinderHub).*
 - As above, integration with Jupyter notebook systems, including flexible and comprehensive treatment of the user environment, will be supported by ESAP.

D5.3 – Performance Assessment of Initial Science Platform Prototype

- Implementation details are not yet finalized; the details of the technology stack (BinderHub, Docker, etc) may vary from those requested, while still satisfying the scientific goals of the use case.
- *Compute to data model: interactive service (JupyterHub) is dynamically launched at data location.*
 - ESAP itself will use standardized APIs to provide access to services which have been configured at remote sites. The detailed mechanism by which the service is instantiated at the remote site is outside ESAP's scope.
- *Rucio data identified via ESAP that is not close to any JupyterHub service. ESAP creates Rucio rule to move data as Quality of Service (QoS) transition and user is sent to Jupyter server as before.*
 - As described in §2.3.4, it is expected that ESAP will provide a capability for “staging” data to locations which are appropriate for processing.
 - Development of this capability is still being planned, and the implementation may differ in technical details to that which is proposed in this use case.

This use case proposes a number of detailed technical approaches. While these are, in general, aligned with current ESAP plans, details of the final implementation may vary somewhat from the details expressed in the use case, without compromising the scientific usefulness of the solutions.

Members of the ESAP development team will continue to engage with colleagues from WP2/DIOS and the SKA to ensure that a properly-functional, fully-integrated solution to address this use case is delivered.

6 Workshop

The *Second WP5 Workshop to Analyse Prototype Performance* was held on 5 August 2021. This workshop was convened to satisfy ESCAPE project milestone MS31. The report submitted to mark the completion of this milestone is provided in Appendix A.

At this workshop, participants were presented with a summary of the current status of ESAP, of future development plans, and of the status of integration with other ESCAPE work packages. Their feedback was directly solicited in the workshop, and they were given the opportunity to follow up after the workshop either through a structured questionnaire or by e-mail directly to the WP5 Coordinator.

This section provides an overview of the workshop, a summary of the material presented, and notes on relevant discussions and feedback. It concludes, in §6.5, with some reflections on workshop organization and improvements which might be implemented for future workshops in a similar vein.

6.1 Logistics

The workshop was intended to solicit opinions and feedback from across the ESCAPE project, with a particular focus on the various ESFRI stakeholders in ESAP development. There was no formal advertisement or registration procedure; instead, announcements of the workshop were provided to:

- The ESCAPE Executive Board;
- The ESCAPE Technical Coordination Team, which includes representatives from the ESFRIs and all of the ESCAPE work packages;
- The ESCAPE Work Package 5 membership.

Announcements made the scope and intended audience of the workshop clear, and encouraged recipients to invite others (in particular, members of their ESFRI or work package) who might be interested to attend the workshop.

The workshop took place on the morning of 5 August 2021. Given the ongoing Covid-19 pandemic, the workshop took place on Zoom¹⁴; no “hybrid” or in-person option was available.

The workshop was scheduled for three hours, including discussion time. In practice, substantially more time was needed; this is discussed in §6.5.2. The INDICO page at <http://indico.in2p3.fr/e/SecondWP5Workshop> provides the complete agenda for the workshop, as well as slides and other supporting material for the various presentations.

The number of participants connected to the workshop fluctuated with time, with a maximum of 47 people connected. These were drawn from across the various ESCAPE project work packages and the associated ESFRIs. Section 6.5.1 briefly reflects on the level and type of participation in the workshop.

Following the workshop, participants were asked to complete a questionnaire with their feedback. This is discussed in §6.3.

6.2 Meeting Summary

The agenda of the meeting was broadly divided into three parts, which are discussed separately here.

¹⁴<https://www.zoom.us>

6.2.1 Project Overview

ESAP Overview (Swinbank) provided an overview of the vision and current state of development of ESAP; the material broadly paralleled that presented in §2 & 3 of this document. This included an extensive live demonstration of the current capabilities of ESAP.

The response from the audience was broadly positive. Key outcomes of the discussion included:

- The project should be careful not to redevelop user interfaces which already exist, especially where those extant interfaces have been carefully developed to facilitate particular science cases or user needs. This was especially emphasized for working with ESO archives: while it is important to facilitate working with them through ESAP, it is necessary to avoid investing development effort which broadly duplicates the work which ESO has already done. The discussion focused around how ESAP can use VO technologies, notably SAMP, in facilitating interactions between services wherever possible.
- The project should carefully track how ESAP development is related to the use cases expressed by the ESFRIs. WP5 acknowledges this input; see, for example, §5 in this document.
- The project should consider the future of this work in the context of the EOSC-Future project. WP5 acknowledges this input, and will continue to work with ESCAPE leadership on this point.
- It was noted that the description of IDA services (§2.3.6 & 3.6 focuses on Jupyter-based systems. This is primarily driven by the use cases that have been expressed by the various project stakeholders (§5). However, as described in §2, ESAP is designed to be extensible to other service types as the need arises. Similarly, the OSSR expects to register a range of different software and service types.

ESAP Architecture (Vermaas) provided a more technical deep-dive into how services can be integrated with the ESAP system. This included essential background material for the overview presented earlier.

6.2.2 Future Development Plans

Batch Processing (Hughes) gave an overview of the current status of, and plans for further developing, batch processing capabilities within ESAP. Refer to §§2.3.7 & 2.4 of this document for an overview of these plans.

A key aim of this talk was to solicit guidance from the community about which batch systems they regard as a priority, as well as developer effort to help integrate ESAP with their own systems. The WP5 leadership adopts the position that selection of a batch system or systems should be largely community driven: rather than WP5 choosing a standardized system, WP5 will work with the community to enable whatever batch systems they require (within the limits of available resources).

The primary advocates of ESAP interfaces with batch computing come from CTA, who require interfaces with DIRAC. We agreed that development would prioritize this work; members of the community were encouraged to work with WP5 developers to push this effort forward.

Managed Database (Chanial) discussed the Managed Database service being proposed for ESAP (§2.3.9), including an impressive live demo. The discussion focused around technical aspects of the system: the technologies used, and how they could be integrated with the rest of the ESAP system.

The suggestion of the Managed Database service being used as the basis for potential improvements to the shopping basket (§2.3.3) was discussed. This could unlock a range of new possibilities. However, the Managed

Database is currently an immature prototype, and the work package is cognizant of its commitments to make robust deliveries of promised functionality on time. The current plan is therefore to continue development of the existing shopping basket system, while working on the Managed Database service in parallel and remaining alert to opportunities for future enhancements.

Interested members of the community were encouraged to connect directly with the WP5 developers to provide additional use cases and guide the development effort.

6.2.3 Interactions with Other Work Packages

DIOS / Work Package 2 (Grange, Di Maria, Hilmy) demonstrated the “data lake as a service”: a powerful interface between a Jupyter notebook service and the Rucio-based data lake. There was general agreement that this represented a key capability, and that plans should proceed for integrating it with the overall ESAP service offering. Additional discussion focused on technical aspects of the data management.

OSSR / Work Package 3 (Hughes, Voutsinas, Graf) provided an extensive introduction to how software can be registered with the OSSR (“onboarding”), then moved on to discuss the current state of and future plans for integration between the OSSR and ESAP, as discussed in §§2 & 3 of this document. A number of important questions for future development were raised, and this workshop resolved to schedule further close discussion between WP3 and WP5 members to clarify interfaces and goals. However, there was general agreement about the overall direction of travel and goals for this work.

VO / Work Package 4 (Grange) provided a brief overview of ESAP integration with the VO. The discussion linked closely back to the earlier discussion regarding interface duplication: the consensus was that VO technologies like SAMP and VOTable [11] offer important opportunities for ESAP to interface between various different services without simply duplicating their existing interfaces. However, effectively taking advantage of VO technologies requires careful attention to metadata management: ESAP developers must be cognizant of this and integration must be treated with care.

6.3 Questionnaire

At the workshop, participants were directed to a questionnaire which was hosted on Google Forms¹⁵. The questionnaire was designed to provide a structured way to collect feedback from the workshop participants about what they heard at the meeting: after a period of reflection and time to experiment with a “live” ESAP test system, they would be able to record their considered opinion.

The complete questionnaire as sent to users is shown in Appendix B of this document.

In practice, only two users completed the questionnaire. This result is obviously somewhat disappointing; possible reasons and alternative approaches which make be taken at future workshops are discussed in §6.5.4.

Highlights of the responses received were:

- One (anonymous) response which indicated an intention to deploy ESAP on their own local infrastructure;
- An emphasis on the necessity to define an API for connecting ESAP with batch compute systems;

¹⁵<https://docs.google.com/forms>

- A desire to enable computing without duplicating or staging data when dealing with extremely large datasets;
- Suggestions for streamlining some aspects of the ESAP interface.

Given the small number of responses received, the questionnaire is clearly of limited value in understanding the response of the community to ESAP. However, the overall level of participation in the workshop and the quality of the resulting discussion provide a high level of confidence in its validation of the ESAP platform, so this is not regarded as a major concern. The WP5 team will continue to use a variety of different methodologies to solicit input from stakeholders over the remainder of the project duration.

6.4 Outcomes

As recorded above, there were a number of insightful points raised during discussion at the workshop. In particular, the emphasis on avoiding simple reimplementation or duplication of existing user interfaces was well taken, as were the thoughts on how best to make use of VO technologies. The WP5 team will schedule a number of follow-up discussions — internally, cross-work-package, and with other stakeholders — to develop plans in response to the feedback received.

Overall, though, the workshop provides strong validation of the current plans and implementation of ESAP: while there were many suggestions for improvement and minor course updates, there were no substantial objections raised to any of the plans discussed, and nor were there major gaps identified that the ESAP vision does not adequately account for.

6.5 Lessons Learned for Future Workshops

6.5.1 Zoom

At this stage in the global Covid-19 pandemic, workshop participants are familiar with the Zoom platform; few technical issues were reported, and the workshop organizers are not aware of anybody who was unable to participate due to issues with the technology.

It is likely that holding the workshop on Zoom increased the number of participants: it is unlikely that as many people would have been able to join the workshop if it involved travel to meet in person.

On the other hand, this also means that many of the participants seemed disengaged. A relatively small fraction of the participants actively took part in the debate; many others likely connected for a few minutes without investing substantial attention

Future workshops could seek to mitigate this by more actively soliciting participation in the discussion: rather than simply calling for questions or comments at the end of talks, an active chair could request feedback from participants, or otherwise structure the discussion to ensure that all participants can fully participate.

6.5.2 Discussion Time

The workshop was scheduled for a total of three hours, which included a number of presentations and time set aside for discussion (see the agenda in §6.1). In practice, this was not adequate to the amount of discussion which arose. Effectively all of the scheduled talks engendered discussion which over-ran the slot allocated to the talk. The final talk did not finish until the nominal end time of the workshop, leaving no (scheduled) time for the final discussion session. In practice, discussion continued for 45 minutes after the workshop among

those who did not have to leave for other commitments. While this additional discussion time was valuable, it meant that the views of those who did have to leave were excluded from full consideration.

6.5.3 Live Demonstrations and Test Systems

Three live demonstrations were run during the workshop. These worked surprisingly well: there were no significant technical problems, and holding the workshop on Zoom meant that everybody had a clear, full screen view of what was being displayed. Feedback on all of these sessions was very positive, and nobody reported any difficulties in viewing or understanding the material.

A test system was also made available which participants could connect to and experiment for themselves. Little obvious engagement was seen with this system: no user feedback was received based on their experiences with it and no bugs or problems were reported.

Combined with other lessons (notably §6.5.4), this is indicative of the challenges involved in maintaining user engagement after the formal end of the workshop programme: participants are eager to participate in discussions during the scheduled workshop, but will then move on to other activities rather than follow-up on workshop activities.

6.5.4 Questionnaire

Only two responses were received to the questionnaire that workshop participants were asked to complete (§6.3), and no other feedback was sent to the WP5 Coordinator.

This speaks to the same difficulties in maintaining user engagement as seen in §6.5.3. It may also indicate that the questionnaire was too detailed: rather than participants being able to quickly submit their thoughts, they were forced to click through multiple pages, and may have moved on to other high-priority tasks before finishing.

One immediate lesson learned is that the workshop organizers should be more proactive about reminding the audience of the existence of the questionnaire. Although it was repeatedly mentioned during the meeting itself, there was no follow-up after the meeting. The response rate would likely have been improved by sending reminders directly to workshop participants a few days after the meeting.

Taken in concert with §6.5.1 and §6.5.2, this suggests that efforts should be made to build structured audience feedback into the agenda of the workshop. Rather than facilitating free-form discussion and then relying on post-workshop input, like a questionnaire, for fully-considered audience feedback, the focus should be on giving the audience time and opportunity to record mature opinions during the workshop itself.

7 Conclusions

This document has presented an evaluation of the current status of, and future plans for, ESAP, the ESFRI Science Analysis Platform. It has summarized the long-term vision for the project (§2) and taken stock of the current state of development (§3). It has compared these against the documented project requirements (§4) and the use cases which have been collected in the ESCAPE project platform (§5). This analysis has highlighted a number of areas in which further work is required to fully deliver the promise of the ESAP system, which — at the current stage in development, roughly halfway through the project — is to be expected. In addition, a number of minor “course corrections” have been identified. However, the overall vision and direction of development for ESAP have been validated by these comparisons.

To complete the analysis, a workshop was held at which ESAP was presented to members of the ESCAPE project and the ESFRI community and their opinions and inputs were solicited (§6). This workshop resulted in valuable and insightful discussion, which will inform future plans for ESAP development. However, no major areas were identified in which ESAP’s capabilities diverged from community expectations, or in which the ESAP vision was unable to service the community’s needs.

In short, therefore, this analysis concludes that the performance of the ESAP is in-line with the expectations of the community and appropriate for the current stage in its development. Substantially more hard work will be required for it to fully achieve its potential, but the progress to date and plans for the future have been successfully validated.

A Report on Project Milestone MS31

The following text was submitted to the Sigma EU Portal on 6 August 2021.

The Second ESCAPE WP5 Workshop took place on 5 August 2021. This workshop brought together members of the ESCAPE ESFRI community, contributors to other ESCAPE work packages, and members of the WP5 team, to evaluate the progress being made on the development of ESAP, the ESCAPE ESFRI Science Analysis Platform. The twin goals of the workshop were to ensure that ESAP capabilities remain closely matched to the ESFRI's needs, while also being well integrated with the work being carried out in other work packages. This workshop was intended to satisfy ESCAPE project milestone MS31.

Due to ongoing restrictions on travel, the workshop took place virtually, using Zoom. While total participation varied slightly through the event, a peak of 47 participants was recorded. The INDICO platform was used for organizing the agenda and collecting presentation materials; these materials remain online and accessible for future reference on INDICO¹⁶. The majority of the meeting was not recorded, but two “live demos” were captured for posterity and are also available through INDICO.

The first part of the workshop presented an overview of the ESAP system. This included discussion of the overall vision for ESAP, accompanied by a demonstration of its current capabilities. A prototype version of the ESAP system was made available to participants, and they were invited to experiment with it. This was followed by an explanation of how service and infrastructure providers can integrate their offerings with ESAP, and then a discussion of future development plans, including integration with batch processing systems and a “managed database” capability.

The second part of the workshop focused on cross-work package activities. This included an impressive demonstration of integration of a Jupyter-based interactive data analysis system with the “data lake” being developed in ESCAPE WP2, together with an explanation of how software and services can be registered with the Open-source scientific Software and Service Repository (OSSR) developed by WP3, and then made available through ESAP and an illuminating discussion of how ESAP capabilities can best integrate with the ongoing Virtual Observatory developments that WP4 is contributing to.

Throughout, the meeting provoked lively and insightful discussion, and suggested a number of promising avenues for future development and collaboration. At the end of the meeting, participants were asked to complete a questionnaire, reflecting on their experiences at the meeting and their goals for ESAP. The results of the discussion and questionnaire will be summarized in ESCAPE deliverable D5.3, which will be forthcoming.

¹⁶<http://indico.in2p3.fr/e/SecondWP5Workshop>

B Post-Workshop Questionnaire

3. To date, have you been directly involved in ESAP development?
Direct involvement includes tasks such as writing code or documentation which form part of the ESAP core (API Gateway, GUI or Managed Database); it doesn't include development of infrastructure which you might one day hope to integrate with ESAP (like scientific analysis code, or stand-alone JupyterLab systems).

Mark only one oval.

☐ Yes ☐ No

4. Before participating in the workshop, did you have a good understanding of what the goals of ESAP development are?

Mark only one oval.

☐ Yes ☐ No

5. After participating in the workshop, do you have a good understanding of what the goals of ESAP development are?

Mark only one oval.

☐ Yes ☐ No

Making Resources Available Through ESAP

6. Does your project have services or data which you would like to see made accessible through some centrally provided ESAP instance?
"centrally provided" could mean hosted by WP5, by the ESCAPE Project more generally, or by EOSC.

Mark only one oval.

☐ Yes ☐ No *Skip to question 9*

Second ESCAPE WP5 Workshop
Thank you for participating in the Second ESCAPE WP5 Workshop.

The aims of this workshop were to describe the vision behind ESAP – the ESFRI Science Analysis Platform – and to collect feedback from project stakeholders about how future developments can best meet their needs. We appreciate you taking the time to fill in this form to let us know how we're doing!

To ensure that your answers are fully accounted for in our post-workshop planning, please submit your thoughts by 12 August 2021. However, we welcome feedback at any time: please feel free to reach out to swinbank@astron.nl at your convenience. You may also wish to get in touch by e-mail if you are uncomfortable about providing details through this Google-managed survey.

1. Please enter your e-mail address
This is entirely optional; we'll only use it to get in touch with you if we need clarification or more information about your answers. If you'd prefer to not to give your address to Google, feel free to send feedback directly to swinbank@astron.nl.

2. Do you work for other otherwise represent an ESFRI or other major research infrastructure involved with the ESCAPE Project?

Mark only one oval.

☐ Yes, and I can speak on behalf of that project
☐ Yes, but I'm answering this questionnaire on a purely personal basis, rather than representing the needs of my infrastructure
☐ No

D5.3 – Performance Assessment of Initial Science Platform Prototype

<p>7. If you answered "yes" to the above, please provide a brief description of the services or data.</p> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <p>8. If you answered "yes" to the above, what support, documentation, or other material would best help you get started?</p> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <p style="text-align: center;">Hosting ESAP Instances</p> <p>9. Are you interested in deploying ESAP within your own infrastructure? <small>That is, to run your own local "science platform", either accessible to the world or just to your own userbase.</small></p> <p>Mark only one oval.</p> <p style="text-align: center;"> <input type="radio"/> Yes <input type="radio"/> No </p>	<p>10. If you answered "yes" to the above, what sort of services would you be providing access to? <small>Be as specific or otherwise as you like — we'd be interested in generic statements like "batch compute", but also descriptions of specific workflow management systems, for example.</small></p> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <p>11. If you answered "yes" to the above, what support, documentation, or other material would best help you get started?</p> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <p style="text-align: center;">Addressing Use Cases</p> <p>12. What types of service would you be interested in accessing through ESAP?</p> <p><i>Tick all that apply.</i></p> <div style="display: flex; flex-direction: column; gap: 5px;"> <input type="checkbox"/> Data Discovery & Query <input type="checkbox"/> Interactive Data Analysis specifically based on Jupyter notebooks <input type="checkbox"/> Interactive Data Analysis not based on Jupyter notebooks <input type="checkbox"/> Batch Computing based on DIRAC <input type="checkbox"/> Batch Computing not based on DIRAC <input type="checkbox"/> Managed Database </div> <p>Other: <input type="checkbox"/> _____</p>
--	---

D5.3 – Performance Assessment of Initial Science Platform Prototype

<p>13. Others, namely...</p> <hr/> <hr/> <hr/> <hr/> <hr/>	
<p>14. Please describe any outstanding use cases which you don't think can be effectively addressed through ESAP as it is currently envisioned. Do you have suggestions for enhancements or improvements that would help?</p> <hr/> <hr/> <hr/> <hr/> <hr/>	<div data-bbox="762 1180 874 1686"> <p>ESAP Demo</p> <p>There is a demo version of ESAP at https://ack-dev.astron.nl/esap-gui/ for you to try our and experiment with. Be aware that this is very much a work-in-progress system; there are known bugs and glitches, and uptime is not guaranteed. Nevertheless, please give it a try; we welcome feedback.</p> </div>
<p>15. Bug and problem reports</p> <p>If you found any obvious errors with the ESAP demo system (pages which didn't load, dialogue boxes reporting problems, etc), please describe them here.</p> <hr/> <hr/> <hr/> <hr/> <hr/>	<p>16. Suggestions</p> <p>Please leave any suggestions for functionality or usability enhancements here.</p> <hr/> <hr/> <hr/> <hr/> <hr/>



References

- [1] Martin Barisits et al. “Rucio: Scientific Data Management”. In: *Computing and Software for Big Science* 3.1 (Aug. 2019), p. 11. ISSN: 2510-2044. DOI: 10.1007/s41781-019-0026-3. URL: <https://doi.org/10.1007/s41781-019-0026-3>.
- [2] F. Bonnarel et al. “The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources”. In: *Astronomy and Astrophysics Supplement* 143 (Apr. 2000), pp. 33–40. DOI: 10.1051/aas:2000331.
- [3] European Commission Expert Group on FAIR Data. *Turning FAIR into Reality*. Tech. rep. European Commission. URL: <https://op.europa.eu/s/pCqN>.
- [4] *Grant Agreement Number 824064 — ESCAPE*. Nov. 2018.
- [5] P. A. Harrison and G. Rixon. *Universal Worker Service Pattern Version 1.1*. IVOA Recommendation 24 October 2016. Oct. 2016. DOI: 10.5479/ADS/bib/2016ivoa.spec.1024H.
- [6] Project Jupyter et al. “Binder 2.0 - Reproducible, interactive, sharable environments for science at scale”. In: *Proceedings of the 17th Python in Science Conference*. Ed. by Fatih Akici et al. 2018, pp. 113–120. DOI: 10.25080/Majora-4af1f417-011.
- [7] Thomas Kluyver et al. “Jupyter Notebooks - a publishing format for reproducible computational workflows”. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by Fernando Loizides and Birgit Schmidt. Netherlands: IOS Press, 2016, pp. 87–90. URL: <https://eprints.soton.ac.uk/403913/>.
- [8] J. P. McMullin et al. “CASA Architecture and Applications”. In: *Astronomical Data Analysis Software and Systems XVI*. Ed. by R. A. Shaw, F. Hill, and D. J. Bell. Vol. 376. Astronomical Society of the Pacific Conference Series. Oct. 2007, p. 127.
- [9] Bruno Merín et al. “ESASky v. 2.0: All the Skies in your Browser”. In: *Astronomical Data Analysis Software and Systems XXVII*. Ed. by Pascal Ballester et al. Vol. 522. Astronomical Society of the Pacific Conference Series. Apr. 2020, p. 89.
- [10] Zheng Meyer-Zhao (ASTRON) et al. *D5.2: Detailed Project Plan*. Tech. rep. ESCAPE, Oct. 2019. URL: <https://projectescape.eu/deliverables-and-reports/d52-detailed-project-plan>.
- [11] Francois Ochsenbein et al. *VOTable Format Definition Version 1.4*. IVOA Recommendation 21 October 2019. Oct. 2019.
- [12] T. Oosterloo et al. “Apertif; the next stage”. In: *Westerbork Telescope 50th Anniversary*. Vol. 361. Sept. 2018, 16, p. 16.
- [13] Federico Stagni et al. *DIRACGrid/DIRAC: v6r20p15*. Version v6r20p15. Oct. 2018. DOI: 10.5281/zenodo.1451647. URL: <https://doi.org/10.5281/zenodo.1451647>.
- [14] M. Taylor et al. *Simple Application Messaging Protocol Version 1.3*. IVOA Recommendation 11 April 2012. Apr. 2012. DOI: 10.5479/ADS/bib/2012ivoa.spec.1104T.
- [15] M. B. Taylor. “TOPCAT & STIL: Starlink Table/VOTable Processing Software”. In: *Astronomical Data Analysis Software and Systems XIV*. Ed. by P. Shopbell, M. Britton, and R. Ebert. Vol. 347. Astronomical Society of the Pacific Conference Series. Dec. 2005, p. 29.
- [16] The Astropy Collaboration et al. “The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package”. In: *Astronomical Journal* 156, 123 (Sept. 2018), p. 123. DOI: 10.3847/1538-3881/aabc4f. arXiv: 1801.02634 [astro-ph. IM].