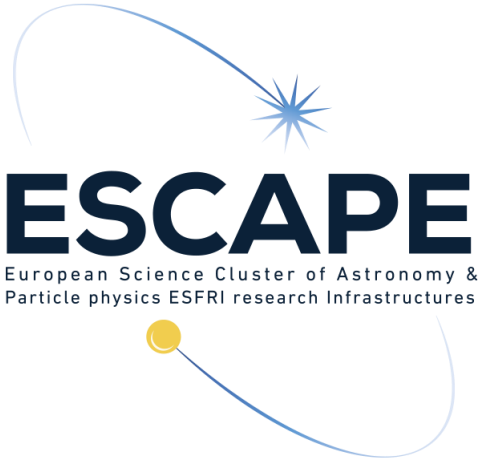

Project Title	European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructure
Project Acronym	ESCAPE
Grant Agreement No	824064
Instrument	Research and Innovation Action (RIA)
Topic	Connecting ESFRI infrastructures through Cluster projects (INFRA-EOSC-4-2018)
Start Date of Project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

D5.1 PRELIMINARY REPORT ON REQUIREMENTS FOR ESFRI SCIENCE ANALYSIS USE CASES

Work Package	WP5, ESFRI Science Analysis Platform
Lead Author (Org)	Zheng Meyer-Zhao (ASTRON)
Contributing Author(s) (Org)	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON), Michiel van Haarlem (ASTRON), David Groep (Nikhef), Rosie Bolton (SKAO), Hugh Dickinson (Open University), Susana Sánchez-Expósito (IAA-CSIC), José Ramón Rodón (IAA-CSIC)
Due Date	31.07.2019, M6
Date	30.07.2019
Version	1.1

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)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	12.06.2019	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON), Michiel van Haarlem (ASTRON)	First version
0.2	27.06.2019	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON), Michiel van Haarlem (ASTRON)	Processed comments/suggestions from ESFRIs and workshop participants
0.3	22.07.2019	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON), Michiel van Haarlem (ASTRON), David Groep (Nikhef), Rosie Bolton (SKAO), Hugh Dickinson (Open University), Susana Sánchez-Expósito (IAA-CSIC), José Ramón Rodón (IAA-CSIC)	Integrated suggestions/comments from WP5 F2F meeting
1.0	30.07.2019	Zheng Meyer-Zhao (ASTRON), Yan Grange (ASTRON), Michiel van Haarlem (ASTRON)	Submitted version
1.1	21.08.2019	Jayesh Wagh (CNRS)	Final deliverable

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° 210506816.

Table of contents

Table of Figures

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

AENEAS: Advanced European Network of E-infrastructures for Astronomy with the SKA

API: Application programming interface

AWS S3: Amazon Simple Storage Service (Amazon S3)

BRITE: BRIght Target Explorer

CERN: European Organization for Nuclear Research

CEVO: Connecting ESFRI projects to EOSC through the Virtual Observatory framework

CNRS: Centre National de la Recherche Scientifique

CoRoT: Convection, Rotation and planetary Transits

CTA: Cherenkov Telescope Array

CNAF: INFN National Centre for Research and Development in Information Technology

CC-IN2P3: IN2P3 Computing Centre

CASA: Common Astronomy Software Applications

DESY: Deutsches Elektronen-Synchrotron

DIOS: Data Infrastructure for Open Science

DOI: Digital object identifier

ECO: Engagement and Communication

EGO-Virgo: European Gravitational-Wave Observatory

eIDAS: Electronic IDentification, Authentication and trust Services

ELT: Extremely Large Telescope (was E-ELT)

EOSC: European Open Science Cloud

EOSC-Hub: Integrating and managing services for the European Open Science Cloud

ERIC: European Research Infrastructure Consortium

ESA: European Space Agency

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

EST: European Solar Telescope

EVN: The European VLBI Network

FAIR: Findable, Accessible, Interoperable, Reusable *or* Facility for Antiproton and Ion Research

FITS: Flexible Image Transport System

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.

GCN: Gamma-ray Coordinates Network

HEP: High Energy Physics

IVOA: International Virtual Observatory Alliance

KM3NeT: Cubic Kilometre Neutrino Telescope

LOFAR: Low-Frequency Array

LSST: Large Synoptic Survey Telescope

ORCID: Open Researcher & Contributor ID

OSSR: Open-source scientific Software and Service Repository

PLATO: PLAnetary Transits and Oscillations of stars

SKA: Square Kilometer Array

TESS: Transiting Exoplanet Survey Satellite

TOUCAN: The VO gateway for asteroseismic models

UEDIN: University of Edinburgh

VO: Virtual Observatory

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-fertilization and continuous development will be strengthened. ESCAPE has the ambition to be a flagship for scientific and societal impact that the EOSC can deliver.

Executive summary

ESCAPE WP5 ESFRI Science Analysis Platform (ESAP) organised an ESFRI Use Case Requirements workshop (M5.1) on 16-17 April 2019 in Groningen, The Netherlands, in order to identify the requirements for the ESFRI Science Analysis Platform from ESCAPE ESFRI partners. Before the workshop, ESCAPE ESFRI partners were asked to fill in a survey form which was created by WP5. During the workshop, several discussion sessions were held to discuss identified ESAP components and the priorities of implementing each of the services.

In this document, we analyse the answers of the ESFRI Survey in terms of data properties, computing system properties, and software properties. We also summarise the ESFRI use case requirements by identifying the required ESAP services and the priorities of implementing these services. Finally, we show a list of questions that have been raised during the workshop which need to be resolved in the future. Answers to the ESFRI Survey can be found in Appendix A.

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

ESCAPE WP5 organised an ESFRI Use Case Requirements workshop (M5.1) on 16-17 April 2019 in Groningen, The Netherlands. The workshop was held at the Centre for Information Technology (CIT) of the University of Groningen.

Before the workshop, ESCAPE ESFRI partners were asked to fill in a survey form which was created by WP5. During the workshop, ESFRI partners gave short presentations regarding their research facilities and their computing/storage/network requirements. The results of the ESFRI survey were shown to the participants during the workshop. We also invited three speakers to talk about Virtual Observatory (VO) (UEDIN), EOSC service Functions-as-a-Service (Faas) (DESY), and Jupyter Notebook as an interface (JIVE). Several sessions were held to discuss identified ESAP components and the priorities of implementing each of the services.

In this report, we will show the analysis of the ESFRI survey results, topics of the workshop discussions, and the summary of ESFRI use case requirements.

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. ESFRI survey

To obtain input for the requirements of the ESCAPE Science Analysis Platform (ESAP) and to gain more insight into ESFRI's data products and their processing/storage requirements, we distributed a survey among the ESFRIs and other research project partners in the ESCAPE project. In total 11 different responses were submitted by the ESFRIs. In this section, we give an analysis of the results. The list of questions and raw output can be found in Appendix A. The main outcomes of the survey are summarised in the sections below.

2.1. Participants

Of the eleven respondents, six are astronomical observatories (EGO - Virgo, ESO - Paranal, EST, LOFAR, VLA, CTA), two are particle colliders (FAIR, HL-LHC) and one, KM3NeT, is an astroparticle physics instrument. JIVE represents the use case of correlating data from multiple observatories. The other two use cases represent scientific analysis of scientific data from multiple observations (Asteroseismology) and citizen science (Zooniverse).

2.1.1. SKA

The SKA is an ESFRI Landmark project under development now and expected to be generating scientifically useful data products at scale by the late 2020s. It will be the world's largest Radio Observatory, supporting a large community of individual users (>1000) across hundreds of institutes, working on (eventually) thousands of unique astronomy projects.

SKA represents a great leap in the scale of data products compared to the rates at which historical facilities have generated data, and so even though the vast majority of the data processing needed to generate science-ready data products will take place within the confines of the SKA Observatory, using dedicated hardware, the user interaction with data products – i.e. the scientific analysis of the observatory data products, will need to take place within a platform that connects user workflows with relevant data products and compute services.

However, although the scale of SKA data products generated will be a new challenge in the astronomy arena (comparable in data rate only with HL-LHC, but with much larger file sizes), the scope of user interactions needed can be well exercised by considering existing astronomy/ astronomer work flows on current facilities, and so we now consider examples that can be used for testing the functionality of a SAP.

2.1.2. VLA use case

The Very Large Array (VLA) is a well-established radio telescope facility located in New Mexico and one of the SKA pathfinder telescopes. The use case based on VLA data is intended to represent how SKA user would interact with the ESA platform in order to study how this platform could support SKA community to achieve the reproducibility and FAIR standards of the data and methods. This use case is a real scientific study (M. Jones et al. in prep.) of the evolution of the HI (neutral hydrogen gas) content of galaxy groups. This study has been performed with particular attention to reproducibility, and for carrying it out, different tools from the EOSC-Hub has been used/tested, what will provide

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.

some information about the already existing EOSC tools which may be considered to be integrated in the ESAP platform.

Notice that computing/storage requirements from the VLA use case are not representative for those from SKA. A detailed description of the expected compute load and data transfer/storage requirements from SKA can be found in the AENEAS D3.1 Deliverable[1].

2.1.3. LOFAR

The International LOFAR telescope is an interferometric array based in the Netherlands but with stations of antennes spread over several countries. LOFAR has been operational since 2010 and can therefore be seen as a smaller-scale demonstrator for future SKA operations. Also, the data distribution and access from the LOFAR telescope will be part of the data served and distributed by the ASTRON Science Data Centre, making technology developed in ESCAPE potentially relevant to the instrument. The questionnaire was answered from the point of view of the observatory without a specific processing use case in mind and is representative of current daily operations.

2.1.4. CTA

The Cherenkov Telescope Array (CTA) is a global project to build the world's largest and most sensitive ground-based observatory for gamma-ray astronomy at very-high energies. It will also be the first observatory at this energy regime open to the world-wide astronomy and physics communities, providing a unique resource of data products and tools.

2.1.5. JIVE

The Joint Institute for VLBI ERIC (JIVE) is the central node of the European VLBI Network (EVN), a distributed array of radio telescopes, in and outside of Europe, offering astronomers the highest resolution view of radio sources. To address the ever growing size of data sets being produced by current (and future) instruments, JIVE will bring the compute to the data by integrating radio-astronomical data reduction techniques into Jupyter-style notebooks, and offering these in containerized form to the user community. Data reduction pipelines will be constructed featuring minimal recomputation techniques. The EVN archive will be modernised and enhanced with mechanisms to feedback user data and publications, and to archive re-processed data. Data reduction tools will be further developed and modified to make them ready for the large instruments of the future. Methods to store the raw VLBI data and offer re-correlation, or other types of processing, as a service will be investigated.

2.1.6. HL-LHC

The Large Hadron Collider (LHC) at CERN will enter its High Luminosity phase (HL-LHC) in the mid 2020s. The challenge will consist in storing ten times more data of higher complexity with respect to the previous years. For this reason, HL-LHC foresees a considerable amount of organized processing of data resident in high latency and less expensive media (today based on tape technologies). It will be important in the context of ESAP to demonstrate therefore the capability to stage and process data from tape archives in orchestrated manner. From the data processing perspective, the HL-LHC

[1] See D3.1 Deliverable "Analysis of compute load, data transfer and data storage anticipated as required for SKA Key science" at https://www.aeneas2020.eu/project-deliverables/

experiments developed an ecosystem of tools and services to run workflows on distributed batch resources, for data reduction. The ESAP developments would complement the existing tools and focus on the end user analysis, offering a notebook-based solution for interactive analysis.

2.1.7. KM3NeT

The KM3NeT observatory has started operation with the first detection units for optical detection of high-energy cosmic neutrinos in the Mediterranean Sea. As the experiment's building phase and development of high-volume data processing and management coincide with the ESAP efforts, KM3NeT can both contribute the perspective of an event-based experiment with a wider particle physics agenda to ESAP as well as serve as a test bed for candidate technologies in the ESAP development process. The most prominent use case for KM3NeT lies in making low-statistics data sets available for multi-messenger analyses.

2.1.8. Asteroseismology

The Asteroseismology use case is based on the analysis of time series of primary and secondary targets. Development of specific numerical codes for the treatment and analysis of "big data" generated by photometric space missions CoRoT and Kepler now in phase of exploitation and TESS and BRITE in phase of observation. This study includes standard harmonic analyses and non-standard fractal analyses, also applicable to spatial series. Currently they are strongly involved in the preparation of the PLATO (ESA) space mission. PLATO2.0 is the ESA M3 mission to fully characterize exoplanets around nearby stars. This use case involves asteroseismic tools, compliant with IVOA standards (e.g. TOUCAN).

Together with the VLA (section 2.1.2), this use case aims to support the continuous evaluation of the level of reproducibility and achievement of the FAIR principles supported by the platform. These two use cases complement each other and go beyond the two selected subdisciplines, as representatives of some of the most complex datasets that the astronomical community is using today, respectively three-dimensional data, and time series together. This will contribute to test the cross-disciplinarity of the ESAP platform.

2.1.9. FAIR

FAIR is the international accelerator Facility for Antiproton and Ion Research under construction in Darmstadt. It will use the upgraded adjacent GSI accelerators as injector chain. While the first processing of the RAW data of FAIR experiments needs to be done on-site and online in order to reduce the size of the data before storing it, FAIR user communities will need to further reduce and analyse the data interactively and by means of batch processing. Therefore FAIR is highly interested in contributing to and taking advantage of the design and construction of the ESAP analysis platform, where a user can have an organised view of the available data and applicable software, apply different filters for selecting them, add his own scripts and transparently have this translated into consistent processing distributed close to the data with the produced output made available within the same platform also for example for being interactively displayed or further reprocessed.

2.1.10. EST

The European Solar Telescope (EST) is a next generation large-aperture solar telescope. This 4-metre telescope will be optimised for studies of the magnetic coupling between the deep photosphere and

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.

upper chromosphere. This will require diagnostics of the thermal, dynamic and magnetic properties of the plasma over many scale heights, by using multiple wavelength imaging, spectroscopy and spectropolarimetry. To achieve these goals, the EST will specialize in high spatial and temporal resolution using various instruments simultaneously that can efficiently produce 2D spectral information. EST will be located in Canary Islands, one of the first-class locations for astronomical observations.

2.1.11. EGO-Virgo

The European infrastructure EGO operates the Virgo detector, which is an interferometric gravitational wave antenna operational since 2017, that produces data made available to the Virgo and LIGO consortia. The data are processed through two distinct channels. The first one is crucial for multi-messenger astronomy: the low-latency analysis detects gravitational wave events that are advertised through GCN circulars for follow-ups by electro-magnetic telescopes. The second one is for offline analysis and a variety of goals are pursued such as detection of continuous gravitational waves or coalescence of compact binaries.

2.1.12. ESO- La Silla Paranal

ESO operates the La Silla Paranal Observatory, providing some of the world's largest and most advanced observational facilities at three sites in Northern Chile: La Silla, Paranal and Chajnantor. The Very Large Telescope (VLT) at Cerro Paranal is ESO's premier site for observations in the visible and infrared light. All four unit telescopes of 8.2m diameter are individually in operation with a large collection of instruments. On La Silla, ESO operates two major telescopes (3.6-m telescope, New Technology Telescope (NTT). They are equipped with state of the art instruments either built completely by ESO or by external consortia, with substantial contribution by ESO.

ESO is currently compiling requirements for a replacement of its existing data reduction infrastructure, and is interested in exploiting synergies with the ESAP effort. The new data reduction system will provide a platform for interactive data reduction for its community of about 5000 observers. It will also be used for non-interactive in-house processing of data for quality control, and for the production of science ready data products for ESO's Science Archive Facility.

2.1.13. Zooniverse

The Zooniverse is the world's largest and most popular platform for citizen science and people-powered research. This research is made possible by volunteers — hundreds of thousands of people around the world who come together to assist professional researchers. The citizen science approach has proven utility for complex analysis of large and intermediate-sized datasets. In the future it will also provide a viable mechanism for generating the training data required by deep learning algorithms that will be essential to process PB-scale datasets that the ESCAPE partner instruments will generate.

Zooniverse data analyses are performed interactively using a general-purpose web interface with broad but limited functionality. An interactive project builder allows researchers to construct specific workflows for data annotation and classification and analysis by combining elements from a library of predefined marking and text-entry tools. Data are typically presented as preprocessed static images, but low volume video and audio assets are also supported. Functionality to render numerical

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 graphically is in the early stages of development. Data resources are typically uploaded to and served from Zooniverse-managed cloud storage, but presentation of web-accessible data hosted by research institutes is also supported. Volunteer annotations are persisted to a relational database and can be retrieved programmatically or via the project builder.

In this international collaboration, the required storage and computing resources are geographically distributed in many different computing centers. The most representative ESAP use case would consist in seamlessly running the offline analysis pipelines on different e-infrastructures.

2.2. Data properties

The first question we asked on the data is what the data type is. The three particle physics experiments (HL-LHC, PANDA/FAIR and KM3NeT) describe the data as event-based, while the astronomy experiments pick observation based. The Asteroseismology and FAIR projects answered their data consists of time series only. EGO-Virgo data consists of timelines from which they extract astrophysical events. LOFAR and the EST also provide that they have time series data on top of it being event- or observation-based. The Citizen science project is rather different since the data it generates is produced by citizen scientists submitting classifications of data that have typically undergone substantial reduction and preprocessing relative to their raw state.

2.2.1. Data size and type

On the question what the data size is or will be, the replies were very diverse. Some projects give the amount of data per time unit, others per file, and others the total size. The size of single observations or events varies between tens of KB to tens of TB. The data production varies from a few TB to a few EB per year.

From the answers we can see that astronomy projects (with “observation-based” data products) have pre-defined targets on the sky, with data products planned specifically with science goals for that target in mind. Particle and astroparticle scientists work on “experiment” data - where conditions are created within an experiment that will trigger data capture and the generation of data products when a particular event (e.g particle interaction / decay shower) occurs. Citizen science also tends to see results as “experiment” based, since here too conditions are set up which will lead to the generation of results but with no direct control over precisely when (or from which citizen contributors) these results will happen. However, even in astronomy, we can easily imagine use cases where we need to consider an “experiment” approach - where the collection of data products presented to the users enables statistical analyses that go far beyond any individual “observation” or single project.

2.2.2. Data accessibility

All projects that generate data which can be mapped to a point on the sky either support or want to support the use of the Virtual Observatory standard. In Figure 1, we show how the readiness to the VO is distributed among projects.

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.

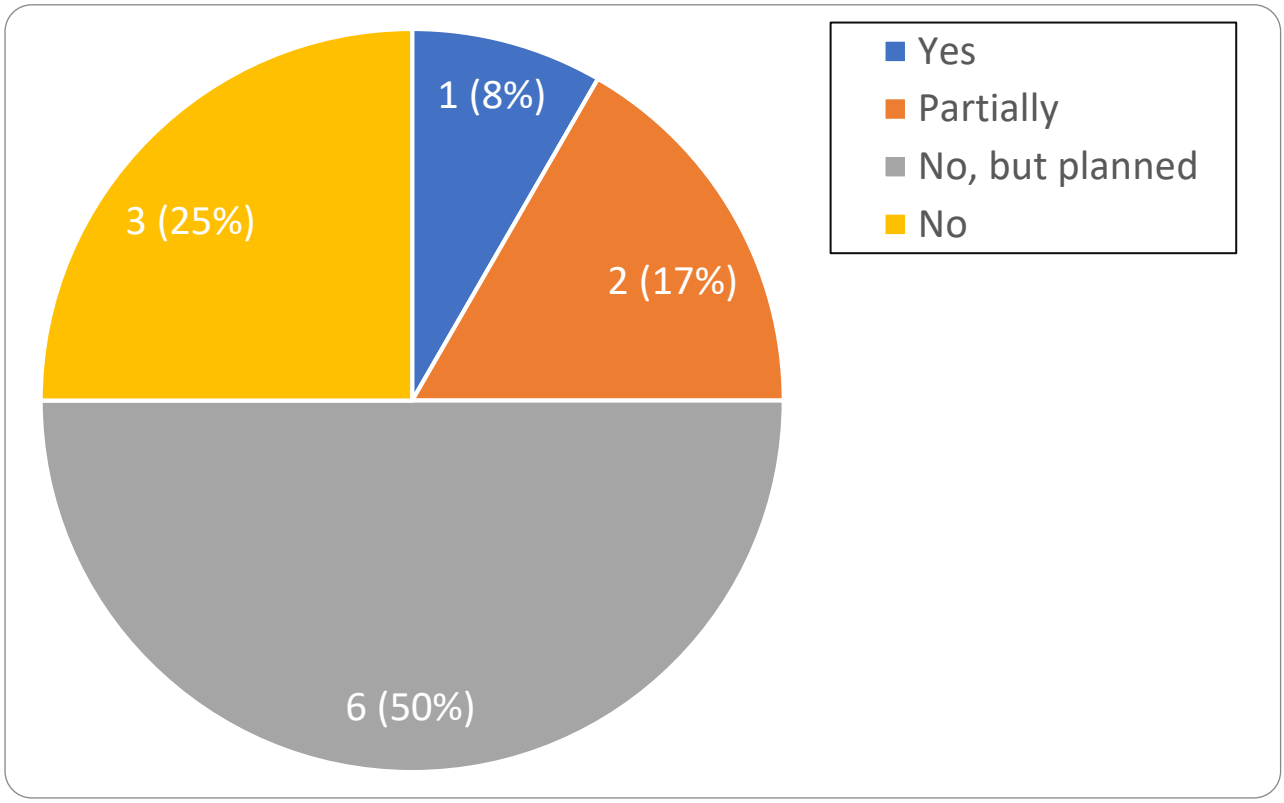

Figure 1 Breakdown of whether the projects want to provide access to their data using the Virtual Observatory protocols.

Most projects are making, or want to make, the data public after a proprietary period. Zooniverse uses public data as input for citizen science projects. VLA (SKA pathfinder) and Asteroseismology use cases offer fully public data in addition to having a proprietary period.

2.2.3. Other data properties

More details of the other data properties specified can be found in Appendix A. Here, we summarise a few relevant points made in the space provided to answer the question about this topic. The HL-LHC mentions the amount of metadata to be stored to be significant as well. EST provides data cubes in FITS format.

2.3. Computing system properties

Two experiments, EST and JIVE, do not store the data in a geographically distributed manner. Out of the ones that do store data in a geographically distributed way, LOFAR and the Asteroseismology experiment do not duplicate stored data. Most Zooniverse subject data are stored in AWS s3 buckets (i.e. cloud storage) but can be served from locations managed by the science project (which may be useful in the context of ESCAPE). The classification database to store volunteer annotations is also hosted on AWS. On the question whether processing is geographically distributed, the answers are

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 same as to the question whether the storage is for all but JIVE, which has central storage but distributed processing.

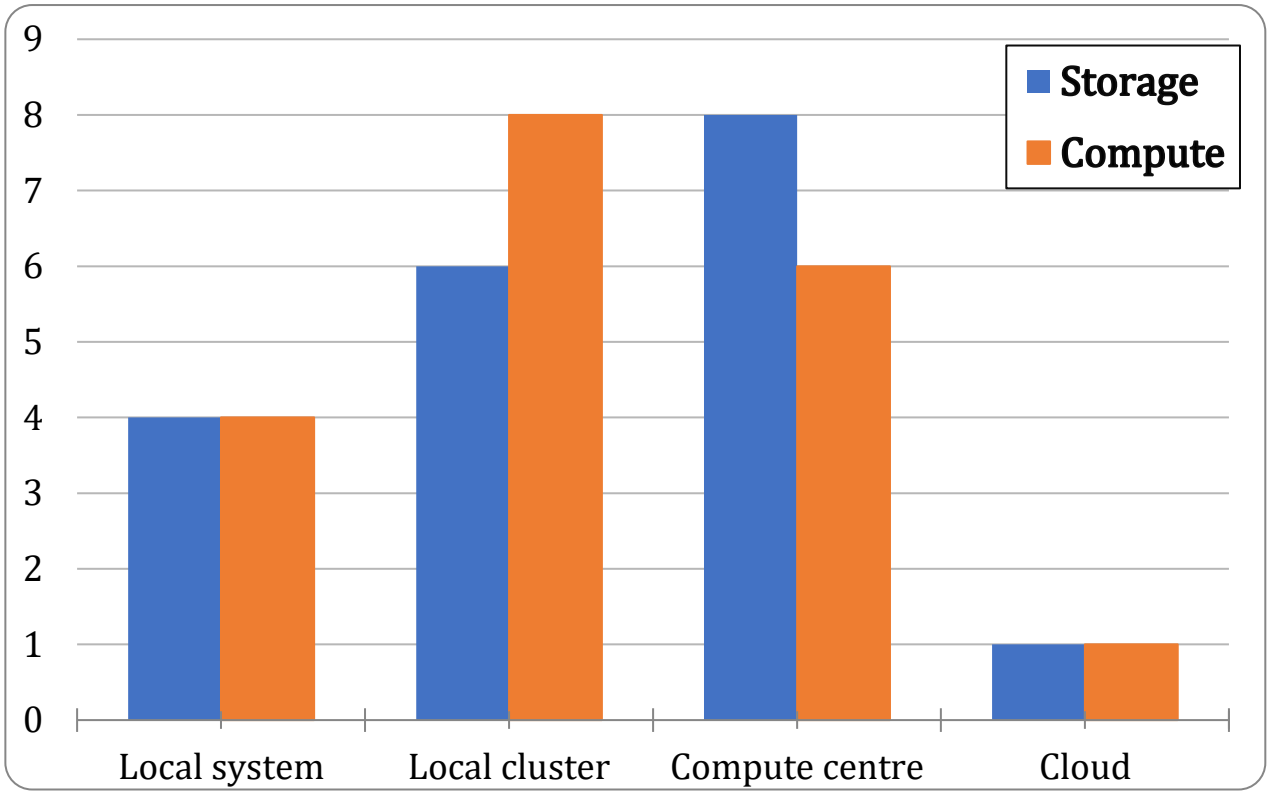

Figure 2 Overview of locations where projects store data and perform computations. Local systems cover personal laptops, desktops, and servers. Compute centres also include the Grid, national e-infrastructure, etc.

Figure 2 shows the distribution of processing and storage locations as reported by the project representatives. Most types of system are used by the communities represented in ESCAPE. Also, two projects reported that possible future extensions of the processing infrastructure could be cloud systems (Asteroseismology and JIVE) and clusters (JIVE) or at the compute centre where the data is stored (LOFAR).

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.

2.4. Software properties

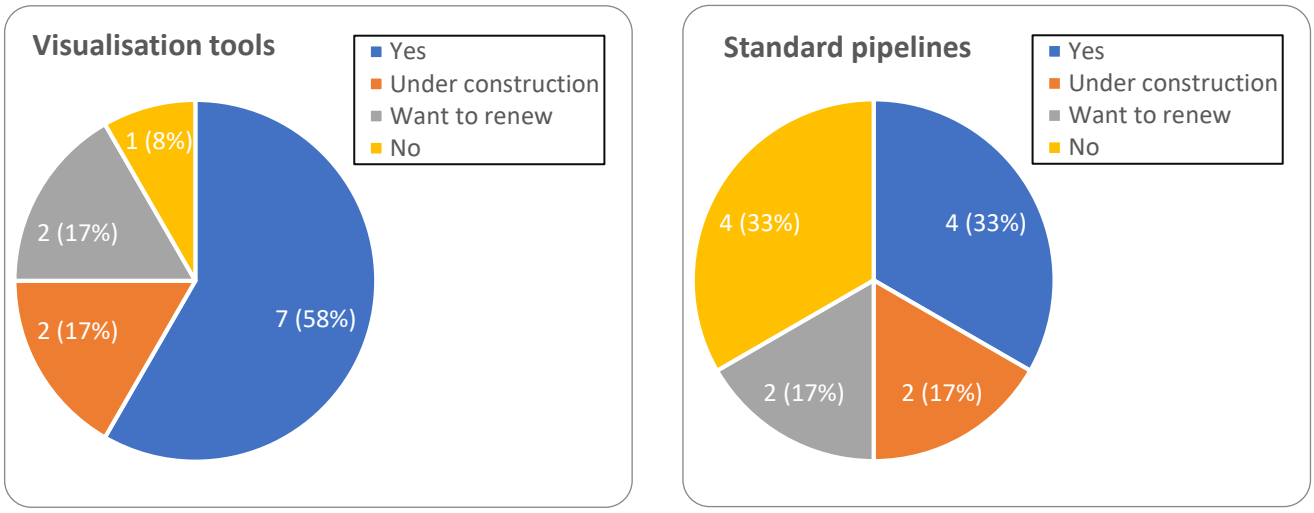

Figure 3 Development status of visualisation tools (left) and pipelines (right).

In the questionnaire, software has been split into two types. The first one being the tools needed to visualise the results from an experiment. Those should be tightly coupled to the user interface. The second one is the pipeline processing software which generates the data products. In Figure 3, we show the status of the development of visualisation tools and processing pipelines. Most of the projects have visualisation tools available. Several are under construction or the wish exists to renew them. For a small percentage no visualisation tools are available and need to be developed. In the case of pipelines, the readiness is lower and much more development is to be expected.

The general comments on the software status can be summarised as the software being very diverse and under constant development.

2.5. Analysis of the answers

From the information filled in by (ESFRI) use cases in the survey form, we draw the following conclusions:

- ESAP needs to support project data with a variety of sizes.
- ESAP needs to support both data which is VO compliant such as data from astrophysics observatories as well as experiment/simulation data which are not mapped to the VO yet. The VO standards might be extended to handle new types of data.
- The amount of processing required should be different for data coming out of the instrument/observatory and the data generated by users. However, both types of data should be supported by the platform anyway.
- Querying duplicated data needs to be supported as roughly half of the experiments duplicate their data storage.
- Visualisation tools and pipelines need to be integrated into ESAP.
- For experiments that store simulations, ESAP needs to support searching these simulations.

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. Requirements for ESFRI Science Analysis Platform

During the discussion sessions of the ESFRI Use Case Requirements workshop, participants from ESFRI partners and WP5 have identified the required services of the ESFRI Science Analysis Platform (ESAP) and discussed the implementation priorities of these services.

3.1 User story

In this section, we start by looking at the identified services of ESAP from a user's perspective. We describe how data flows on the platform after a user gets onto the platform.

A user coming to the ESAP services can be (and in many cases on initial access will be) anonymous. At this stage, the user can be presented with services that permit anonymous (if potentially rate limited or otherwise controlled) access, such as searching for public data and services provided by the public Virtual Observatory registry (as discussed in WP4) or any other access service that follows the FAIR principle as e.g. services from Zooniverse.

For the ability to use proprietary data (both the ability to search for this data through meta-data, as well as access to the data itself) the user will have to login to the ESAP platform services. In the research infrastructures that participate in the ESAP, users are granted access based on community membership, following the model defined in the Authentication and Authorization mechanisms (AAI) that also supports the Data-Lake vision, and provides group management and heterogeneous authentication capabilities, following the AARC Blueprint Architecture [AARC-BPA]. Authenticated users, based on their community and group membership attributes, access the services in the ESAP and can be presented with their private data sets, and with resources and services to which they have been granted access.

The user is able to use single sign-on (SSO), and this SSO should preferably work seamlessly across services both within the ESAP as well as across other community services (e.g. an Observatory portal). This SSO capability can be provided at two levels: the first option (which will allow the provisioning of collective services across the ESAP and the other service providers) implies that the services also connect to the Community AAI Proxy in which membership and user identifiers are coherently managed. For many of the RIs, this can be a preferential model as it allows services to collectively provide a capability – they all recognize the user as the same person. It also allows integration with generic e-Infrastructure service providers in the EOSC based on agreements between the community and the e-Infrastructure federations. Alternatively, the SSO capability is provided at the authentication layer (e.g. at the user's home organization in eduGAIN, or a generic identity source such as ORCID or eIDAS). This provides an SSO experience, but no intrinsic collective service capability. In order to allow the user to 'recognise' the login capability and make that as seamless as possible (by-passing the perennial 'WAYF' ('Where Are You From') or identity provider discovery service) problem of finding out what preferred authentication source the user would be using), the recommendations by RA21 [RA21.org] would be a good basis to present the login capability also on the ESAP portal services. The AAI capability and its SSO properties should also work in non-web scenarios, and in particular permit access control to the (network-accessible) API of the services in 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.

After logging onto the platform, the user will be able to add the list of data queried from VO or other data query interface into the data “shopping cart”.

The user is presented the upper level of data hierarchy, which includes the RIs offering data to that user. If the user is anonymous, (s)he is presented open data only, while an authenticated user is presented data of the authenticating RI in addition. The user can interactively browse a data hierarchy: for example the user selects 'raw' or 'simulation', then (s)he can choose among different years of production, then (s)he can choose among production cycles within that year, than (s)he can choose a specific time interval, etc. Metadata describing the datasets can be displayed when hovering over it. The resulting list can be additionally filtered by other characteristics (data of production, type of calibration, etc., depending on the labels carried by the data). The user selects the data and is possibly shown the size of the selected data.

Based on the data format, the platform will suggest a list of compatible software or workflows, which will be able to process the selected data. The software and workflows mentioned above are provided either by WP3 via ESAP interface or by other ESAP users.

ESAP will accommodate two processing modes, i.e. interactive data processing and batch data processing. If interactive mode is selected, users can choose from a list of containers with jupyter notebook environment and science domain specific software stack installed. Users can then play with the datasets (should be small in size and quantity) with the selected software, and possibly add his/her own processing scripts into his/her processing space. Users can also upload their own containers with specific domain software installed on them. These scripts/containers might be hosted somewhere else, e.g. Github/Gitlab/dockerHUB. After experimenting with the datasets and the selected/developed software, users may choose to stage bigger/more datasets, and execute the workflows they’ve been trying out/testing interactively on these datasets on potential (i.e. list of recommended) HPC/HTC systems on large scale in a batch processing mode . If possible the user can see the progress of the processing or is notified when the job is done or when the job is terminated.

The processing will result in advanced data products. The ESAP should then support the user in submitting those products to the data archive or to publish the data products. Both the data and workflows that have been used to produce the advanced data products, including the data products itself, should persist a unique identifier, i.e. DOI (Digital Object Identifier). If the produced advanced data products are published in any form of publications, the publication should also have a DOI. All the above mentioned attributes, i.e. data, workflows, advanced data products, and publications form a research object, which is an entity that packs the data and workflows used to produce the advanced data products and the resulting scientific publications together, that will be searchable through a research object catalogue hosted on ESAP. Access rights to advanced data products need to be set by the users/user group (public, private, project, etc). Research objects (RO) are always public, since the result has been published in the scientific literature. Those data should be accessible in such a way that the archive of user-generated data can be queried through ESAP.

The whole set of values for the query is saved as a template so that it can easily be reused with or without modifications. Users should be able to query both observatory-provided data or data

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.

provided via any other access service that follows the FAIR principle as well as user processed data and it should be clear what data falls in which category. Also the provenance metadata of the post processed data should be accessible by the user so that they can find what observation or experiment a post-processed data set has been generated from.

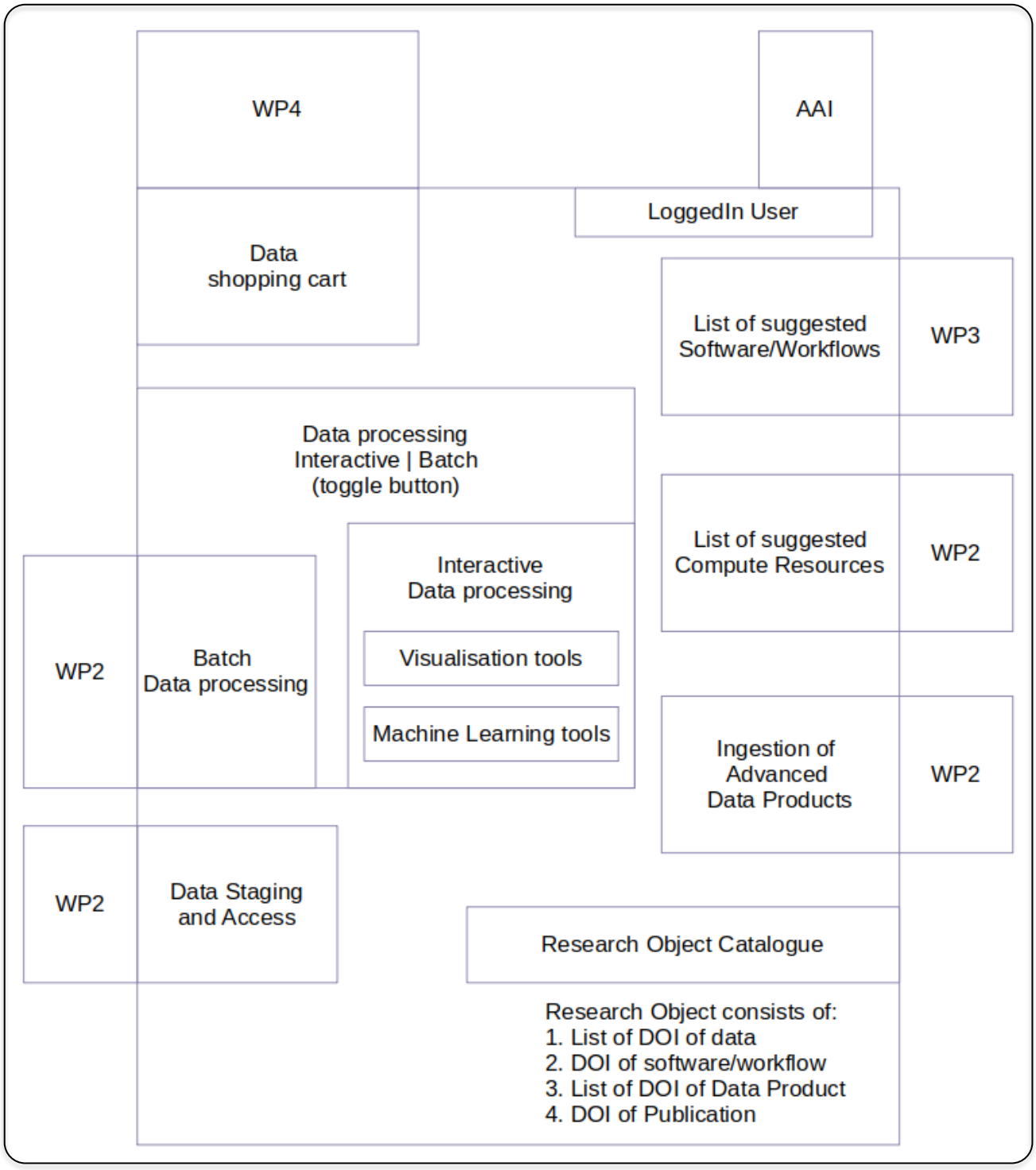

Figure 4 Schematic overview of the functionality of ESAP and its link to the other work packages within the ESCAPE project

3.1. Identified services

In this section, we give a summary of the service components that have been identified during the ESFRI Use Case Requirements workshop. In the following subsections we group the services in chronological order of the user story: data finding, staging and access, processing, and finally ingesting the result. The subsections thereafter will give an overview of the overarching topics of Authentication and Authorisation Infrastructure (AAI), provenance and research object catalogue.

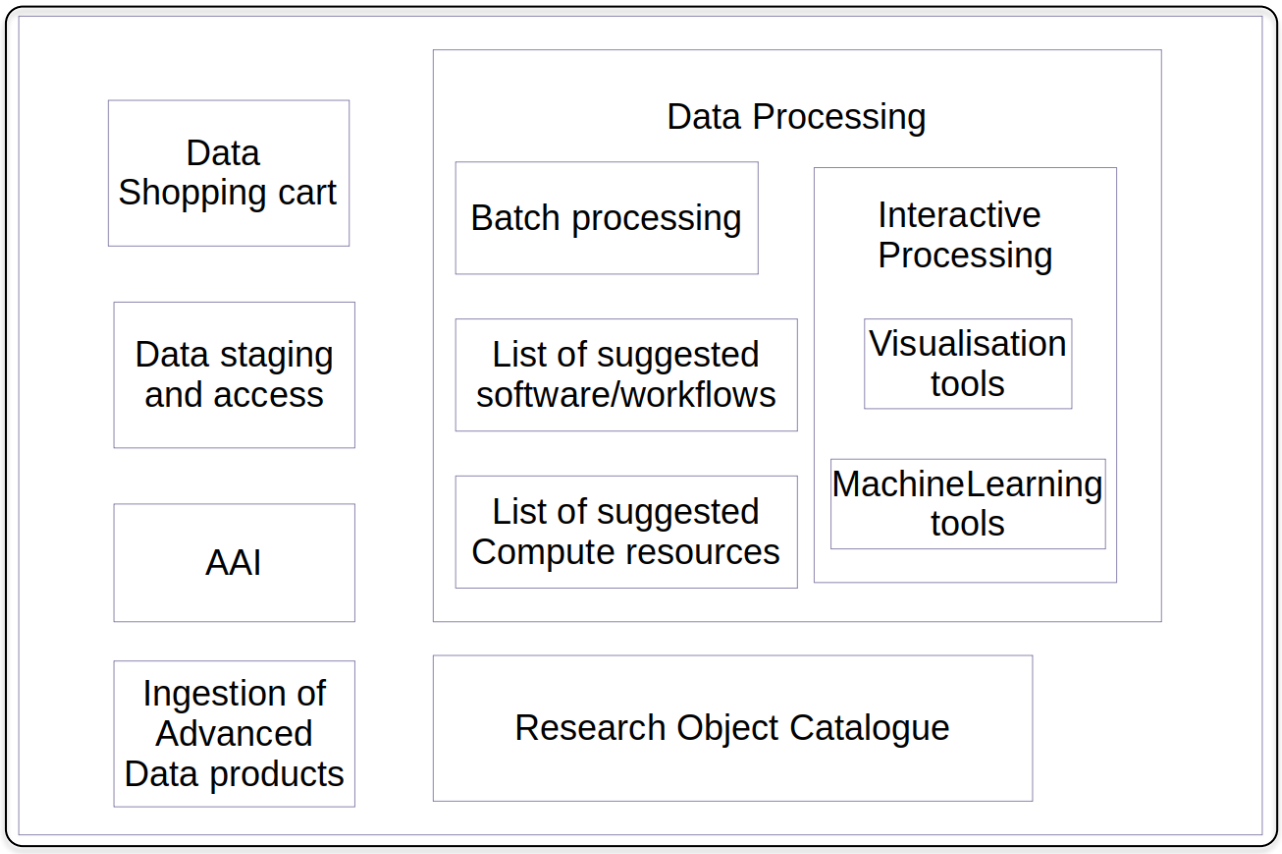

Figure 5 Overview of the services identified during the workshop.

3.1.1. Data finding

Most data that can be mapped on a position in the sky can be accessed using the Virtual Observatory (VO)[2] protocols. A significant fraction of the work in work package 4 is to make sky-based data that is not yet in a VO repository accessible through the VO (e.g. visibility data from radio astronomy). However, it might not be possible to access data from HEP experiments and simulations via the

[2] Note that the acronym VO is also used for "virtual organisation", which can lead to some misunderstanding in discussions on the topic. In this document, however, we will only use it as shorthand for "virtual observatory".

existing VO protocols in a straightforward way. Therefore, we will need to discuss how to make this data available through ESAP using different protocols. For Zooniverse classifications, the applicability of VO protocols is typically inherited from the data that were classified.

This means that the ESAP interface should provide access to VO services and any other FAIR access services needed to make the data findable. Also, the interface should support showing the difference between data that has been created by an observatory, experiment or simulation and data that has been further processed by users (including linking to publications through research objects if applicable).

3.1.2. Data staging and access

When data has been found, it should be made accessible for processing. We will refer to this action as staging. The exact mechanism behind this will need to be offered by the framework developed in work package 2. The staging operation and, if applicable, specification of boundary conditions (e.g. availability of specific resources, proximity of other data sets for creating collections) should be accessible to the user, through the ESAP interface.

ESAP should also provide data access to the data for analysis purposes. The user needs to be able to have both access using an interactive interface. e.g. providing visualisation tools or machine learning libraries, and a non-interactive interface, i.e. for extending the analysis to larger data sets, the interface should allow for executing batch jobs to obtain results from large or complex data sets.

3.1.3. Data Processing

For data processing, ESAP should be able to recommend a user what processing tools are available for a given data type, e.g., based on file type, based on the information provided by the repository that is being developed in work package 3. Also, ESAP should be able to provide options for performing the compute jobs, as also referred to in section 3.1.2.

The standard workflows that can be obtained from the work package 3 repository can be adapted by users. ESAP should also be able to make adapted or, more generally, user-defined workflows accessible to other users to use for their processing needs. In parallel to what is mentioned in section 3.1 on data access, the provenance of a workflow (i.e. standard or defined by a specific user) should be clear when searching for one that fits a user's needs.

Processing of data should be offered in both interactive and batch processing modes. The interactive mode will typically be used to tweak what pipeline components and parameters to use to perform optimal processing using only a small subset of the data, after which batch processing can be executed using the optimal settings. The method to switch between both should be as simple as possible, preferably by using a single toggle switch.

An alternative method of processing data that should also be supported is setting up a pipeline which will be triggered by the arrival of a data set at a specific location. Again, this should be made available to the user in a fashion that is as simple and accessible as possible.

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.

3.1.4. Ingestion of advanced data products

The ESAP needs to provide an easily accessible way to make data shareable to a wider community, for example, through EOSC service B2Share. Moreover, ESAP should provide a mechanism to rate the quality of the data shared by ESAP users, such that the community will contribute to the quality control of the advanced data products. Published data should be findable through research object catalogue, which means the ESAP research object catalogue should link to the services that host published/shared research data.

3.1.5. AAI

For Authentication and Authorisation (AAI) to ESAP, work package 5 would like to use a federated AAI system. Such solutions have already been developed in external projects and hopefully will be present in the EOSC service catalogue, such that ESAP can integrate with the service directly.

3.1.6. Research Object Catalogue

The Research Objects are aggregations of Digital Objects involved in a scientific experiment such as datasets, analysis codes, provenance information, annotations and other related information. An RO provides an effective way to share, publish and link every piece of an experiment, showing the relationships among them. Current RO models make use of Persistent Identifiers (PIDs) to aggregate and interlink the experiment elements, which may be distributed in different repositories. As stated in the *Turning FAIR into reality*[3] report, "data need to be accompanied by PIDs and metadata rich enough to enable them to be reliably found, used and cited". Therefore, in order to achieve the FAIR-ability of the digital objects hosted in the ESCAPE framework, ESAP should support users in assigning PIDs to their digital objects and aggregating them into ROs.

[3] https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283

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.

3.2. Priority of service implementations

During the WP5 workshop, the participants were asked to vote for what services they think need to have the highest priority of being implemented. The results of that vote are shown in Figure 6.

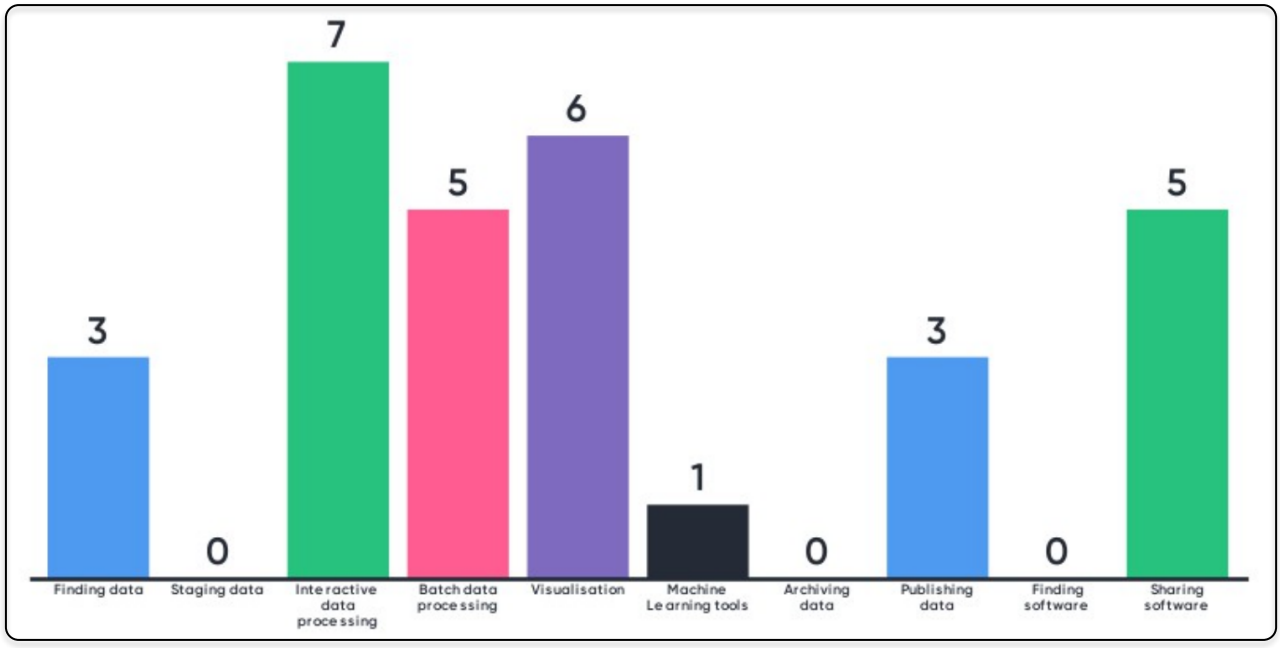

Figure 6 Prioritisation of services by the workshop participants.

The services that the participants of the workshop believed have the highest priority of being implemented are:

1. Interactive data processing
2. Visualisation
3. Batch data processing / sharing software

Finding data and publishing data share the fourth place, while machine learning tools got one vote. Interestingly staging and archiving data is not considered a priority by the participants and even though sharing software is considered important, finding software also did not get any votes. Of course, the result shown above is a reflection on the people participated in the vote during the workshop, not all ESFRI partners and ESCAPE WP5 member institutes joined the voting.

The goal of prioritisation is to define a Minimal Viable Product (MVP) that contains the most required functionality while still covering a significant part of the processing chain.

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.

3.3. Summary of ESAP requirements

The ultimate goal of the ESFRI Science Analysis Platform is to support scientists using data from ESCAPE ESFRI partners to have easy access to their research data, the available compute infrastructure and storage, to seamlessly publish their advanced data products and software/workflows, to be able to easily share/query published data, and to obtain the provenance of the published data.

To achieve these, ESAP needs to support users of the platform to search for and select data[4], select available software/workflows[5], select available compute resources[6], process data either interactively or in batch mode, publish/share data, software, and/or research objects as described in section 3.1.To enable this a single sign-on mechanism, giving seamless access to all integrated services, will be desirable.

Whilst there are clearly common building blocks needed for this platform that will serve users from across all communities, the user stories from different thematic science areas (e.g. astronomy vs. collider physics vs. astroparticle physics) are quite different. We will identify platform elements that are common to all but develop separate SAPs for the different themes.

Identified ESAP services	Implementation priority	Links to Other Wps
Finding data	3	WP4
Staging data	0	WP2
Interactive data processing	7	WP2
Batch data processing	5	WP2
Visualisation	6	
Machine Learning tools	1	
Archiving data	0	WP2
Publishing data	3	WP4
Finding software	0	WP3
Sharing software	5	WP3
List of suggested software/workflows		WP3
List of suggested compte resources		WP2
AAI		WP2
Research Object Catalogue		WP4

Figure 7 Overview of the identified services with their priorities and links to the other ESCAPE work packages.

The identified ESAP services, their implementation priorities, and links to other work packages are listed in Figure 7. Note that, some identified ESAP services in Figure 7 don't have any number

[4] Link to work package 4

[5] Link to work package 3

[6] Link to work package 2

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.

associated with its implementation priority. This is because these services are identified during the discussion of the use case requirements workshop. The more detailed implementation plan of these services will be provided in the next deliverable “Detailed Project Plan for WP5”.

Appendix A: survey

In this appendix, we show the raw results of the survey answered by the ESFRI partners. In the notes (at the end of this appendix), we explain where and why changes were applied to the answers for the analysis in section 1 of this document.

Data properties

Name of (ESFRI) RI / project	**Data Format**	**Data size and size distribution per experiment/observation (e.g. for LOFAR images this could be 0.5-10TB total per observation, consisting of 244 equally-sized subbands as tar):**
Zooniverse	Citizen Science classifications	Small, eg ~1e6 images with ~1e6 binary classifications, ~1e6 annotations to images, etc. Zooniverse currently hosts about 1e9 images in total.
LOFAR	Time series, Observation based	Varies considerably per experiment. Size raw data: several tens of TB, size processed data prior imaging: a few TB for a few hours of observation. The typical imaging data set is distributed over 244 or 488 frequency bands (subbands). Each frequency band is stored in the measurement format which is a directory containing both the metadata and data in separate files.
PANDA/FAIR[c]	Event-based, Time series	For PANDA (as benchmark for FAIR experiment): "event" size is typically ~30 kBytes with a total data rate of 200 Gbytes/sec.
HL-LHC	Event-based	Order of few Exabyte/year from 2026 in files of ~10GB
VLA use case (SKA pathfinder)	Observation based, Observation data from the VLA, a SKA pathfinder	300MB per observation (2 Observations)

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.

JIVE-ERIC	Observation based	0.01-5 TB
ESO - La Silla Paranal Observatory (optical near IR)	Observation based	A bit more than 1 PB, growing by 15-20% a year
European Solar Telescope (EST)	Time series, Observation based	
EGO-Virgo	Time series	~500 KB/sec per observatory (2 Ligo, 1 Virgo) of scientific data (Hoft + state vector, etc.), ~40MB/sec of raw data. This has to be multiplied by the science run duration: For one year run and for one observatory we reach ~1PB of raw data.
KM3NeT	Event-based	Maximum requirement * full set of events O(10)GB/day * simulations for reference purposes of about the same order Minimum requirement * preselected events of a few GB/month * simulation as reference tables or by providing software (O(10) GB)
CTA	Event-based (DL0, DL3), Observation-based (DL5)	O(1 TB/observation-hour) raw data (DL0) collected during observations for internal processing, O(GB/observation-hour) reduced data (DL3) for Science Users, O(MB/observation-hour) science-ready data products (DL5) for Science Users. For a full year, this leads to O(10 PB/yr) raw data (DL0) for internal processing, O(TB/yr) reduced data and science-ready data products (DL3, DL5) for Science Users
FAIR	Time series	1TB/s into online farm, 10GB/s saved on disk (CBM: Data rate to online farm: 400 GB/s, Simulated

		events: 250 kB/event)
Asteroseismology (No ESFRI project)	Time series	For time series data this could be from 1 GB to 50 GB per time series, consisting of one exposition sampling in 32 points per second. The files are ASCII files, FITS files and HDF5 file formats

Name of (ESFRI) RI / project	**Source of data**	**Science-relevant data generated**
Zooniverse	Simulated, From Experiment (Simulated data have utility for calibration of volunteer response). These represent the input data to be classified. The output data products are sets of classifications or data annotations.	Science-ready data (e.g. data cubes). Data presented in suitable form for citizen science.
LOFAR	From Observation	Raw data (e.g. visibilities), Reduced data (e.g. calibrated uv data), There should be a distinction between what the Radio Observatory provides to the users and what the users can produce themselves on their own processing clusters
PANDA/FAIR	Simulated, From Experiment	
HL-LHC	Simulated, From Experiment	RAW data + Analysis Formats (from x10 to x1000 times smaller than RAW, but in multiple versions)
VLA use case (SKA pathfinder)	From Observation	Raw data (e.g. visibilities), Reduced data (e.g. calibrated uv data), Science-ready data (e.g. data cubes)
JIVE-ERIC	From Observation	Raw data (e.g. visibilities), calibration data

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.

ESO - La Silla Paranal Observatory (optical near IR)	From Observation	Raw data (e.g. visibilities), Reduced data (e.g. calibrated uv data), Science-ready data (e.g. data cubes)
European Solar Telescope (EST)	From Observation	Raw data (e.g. visibilities), Reduced data (e.g. calibrated uv data), Science-ready data (e.g. data cubes)
EGO-Virgo	From Observation	Science-ready data (e.g. data cubes)
KM3NeT	Simulated, From Experiment	Reduced data (e.g. calibrated uv data), Science-ready data (e.g. data cubes)
CTA	Simulated, From Observation	Raw data (DL0, e.g. telescope image cubes), Reduced data (DL3, e.g. event lists), Science-ready data products (DL5, e.g. sky maps, spectra)
FAIR	Simulated, From Experiment	
Asteroseismology (No ESFRI project)	Simulated, From Observation	Raw data (e.g. visibilities), Reduced data (e.g. calibrated uv data)

Name of (ESFRI) RI / project	**Is the metadata VO compliant?**	**What access rights apply to the data?**
Zooniverse	Likely to depend on compliance of input data. Classification output data can probably be made to conform with the VO DataSet Characterisation model.[a]	Proprietary period after which public, Zooniverse-badged experiments must be based on public data but the classifications can be proprietary for a period.
LOFAR	Only some final data products generated by the users are VO compliant.	Proprietary period after which public
PANDA/FAIR	No	Proprietary period after which public

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.

HL-LHC	No	Proprietary period after which public, the "embargo time" and the type of data made public varies for different experiments
VLA use case (SKA pathfinder)	No, but planned to make it so	Public
JIVE-ERIC	No, but planned to make it so	Proprietary period after which public
ESO - La Silla Paranal Observatory (optical near IR)	Metadata of processed data exposed through the archive are VO compliant. It is planned to do so for raw data as well, to the extent possible.	Proprietary period after which public
European Solar Telescope (EST)	No, but planned to make it so	Proprietary period after which public
EGO-Virgo	No, but planned to make it so	Proprietary period after which public
KM3NeT	No, but planned to make it so	Proprietary period after which public
CTA	No, but planned to make it so	Proprietary period after which public
FAIR	No	Proprietary period after which public
Asteroseismology (No ESFRI project)	No, but planned to make it so	Public

Name of (ESFRI) RI / project	**Are there any other aspects relevant to describe the data generated?**	**Is the data (or will the data be) geographically distributed?**
Zooniverse		Input data (subjects for classification) can be stored using AWS s3 protocols or served directly from locations managed by the research team. Output data (classification results) are

		stored in cloud-hosted relational database and can be downloaded by science teams to multiple locations.
LOFAR		Yes
PANDA/FAIR		Yes
HL-LHC	~10 "system metadata" (size, checksum, creation time) and ~30 "user metadata"	Yes
VLA use case (SKA pathfinder)[b]	Currently data is public and can be accessed through the VLA archive interface	We have download the raw data from the VLA archive to the EUDAT infrastructure. Then we plan to use the VO to publish final data
JIVE-ERIC		No
ESO - La Silla Paranal Observatory (optical near IR)		It currently isn't, but it might be in the future, even though there is presently no need to do so.
European Solar Telescope (EST)	2-D images (FITS format) of a sequence of spectra at different spatial positions or a sequence of 2D images of the solar surface at various wavelength positions. Additionally, calibration data for polarimetry, focus, dark current, flatflieding.	Yes
EGO-Virgo		Yes
KM3NeT		Yes
CTA	The reduced data sets come with an appropriate Instrument Response Function per good time interval in the observation.	Yes (raw data), Maybe (reduced data, science-ready data products)
FAIR		Yes
Asteroseismology (No ESFRI project)	The data is generated by different satellites: CoRot, Keppler, Brite and TESS. In the future they will also be generated through the Plato and Cheos satellites	Yes

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.

Compute and storage properties

Name of (ESFRI) RI / project	Is the data (or will the data be) duplicated?	Is data processing (or will data processing be) geographically distributed?
Zooniverse	Not explicitly, although AWS cloud storage implies redundant data backup.	Use case dependent. For example the use of classification data as labels to train machine learning algorithms may entail cloud computation.
LOFAR	No	Yes
PANDA/FAIR	Yes	Yes
HL-LHC	Yes	Yes
VLA use case (SKA pathfinder)	Yes	We are using the European Federated Cloud for the data processing and we plan to use SciServer.
JIVE-ERIC	No	Yes
ESO - La Silla Paranal Observatory (optical near IR)	No	It currently isn't, but it might be in the future, even though there is presently no need to do so.
European Solar Telescope (EST)	Yes	Yes
EGO-Virgo	Yes	Yes
KM3NeT	Yes	Yes
CTA	Yes	Yes
FAIR	Yes	Yes
Asteroseismology (No ESFRI project)	Yes	Yes

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.

Name of (ESFRI) RI / project	How/where is the data stored in the community at the moment (e.g. local private cluster, commercial cloud, national e-infra, scientists' laptop/desktop) and will this change in the future? [h]	How/where is the data processed in the community at the moment (e.g. local private cluster, commercial cloud, national e-infra, scientists' laptop/desktop) and will this change in the future?
Zooniverse	Currently AWS S3 and AWS Relational Database Service.	Local private clusters and using cloud computation e.g. AWS EC2 instances.
LOFAR	Data are distributed through a long-term archive. Processing continues wither at the archive or, more often, at external computing facilities at the location of the user - no desktop, but compute clusters	See above
PANDA/FAIR	The low-level data will be stored centrally at FAIR computing center, and high-level filtered data will be distributed on private clusters, scientists laptops/desktops of institutes involved in the experiment.	only simulated data are presently processed. The processing of this data is primarily done at large data centers and private clusters. The aim is to streamline this in accordance to the data flow of experimental data eventually.
HL-LHC	Worldwide LHC Computing Grid (~200 centers)	Mostly in WLCG resources + HPS resources + Cloud resources + private clusters (mostly analysis) ... the all lot.
JIVE-ERIC	Archive at JIVE-ERIC and on scientists' laptops, server	Current: scientists' laptop/desktop, future maybe on cluster+cloud
ESO - La Silla Paranal Observatory (optical near IR)	Local private closest, mostly.	Local private closest, mostly.
European Solar Telescope (EST)	Institute cluster/scientists' laptop/desktop --> national e-infra	Institute cluster/scientists' laptop/desktop -> national e-infra

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.

EGO-Virgo	Both at the Observatories sites and on supporting Computing Centers (CNAF, CC-IN3P3, etc.)	Data are processes at the Observatories sites and on supporting Computing Centers (CNAF, CC-IN3P3, etc.)
KM3NeT	Temporary on-site filtering and storage of data, permanent storage at CCLyon and CNAF	Mainly processed (offline calibration, reconstruction etc.) at CCLyon or at participating institutions.
VLA use case (SKA pathfinder)	VLA archive is implemented in a local facility (local private cluster)	Local private cluster or even laptops/desktops
CTA	Only simulation data available at the moment that is stored on data center infrastructures on the EGI GRID. Future CTA raw data (DL0) will be permanently stored in the CTA bulk archive on CTA data center infrastructures, Reduced data and Science-ready data products (DL3, DL5) will be permanently stored in the CTA science archive on CTA data center infrastructures.	Simulated data is produced on data center infrastructures on the EGI GRID, simulated reduced data and simulated science-ready data products processed off-line on local private clusters or scientists' laptop/desktop. During operations, reduced data and science-ready data products (DL3, DL5) will be accessible to the community and can be processed off-line on local private cluster or scientists' laptop/desktop. Limited resources may be available for interactive data exploration on CTA infrastructure.
FAIR	At the moment the data is stored in a dedicated cluster at GSI. Replication and distribution of the data will depend on the computing models of the FAIR experiments.	At the moment the data is processed on the cluster at GSI.
Asteroseismology (No ESFRI project)	Dedicated servers and VO Services	Dedicated servers

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.

Name of (ESFRI) RI / project	Are there any other aspects relevant to describe the compute and storage systems used?[d]
LOFAR	Ambition is to move to a setup where data is directly processed at the LTA by the Radio Observatory and all users
FAIR	Specific aspects will be decided by the computing models of the FAIR experiments.

Software properties

Name of (ESFRI) RI / project	Visualization tools available?	Standard processing Pipeline available?
Zooniverse	Yes, in the form of standalone Python scripts available on Github. This functionality should be integrated in the science platform.	No, needs to be developed and this is funded in WP6[b]
LOFAR	limited visualisation tools in the software currently in production in the RO environment. More options have been developed by the community of users.[d]	there are some processing pipeline available in the production environment, but more advanced ones need to be implemented. They are not available through a 'science platform' though, as such a science platform for LOFAR does not exist yet[d]
PANDA/FAIR	Yes, but it should be integrated in the science platform	Yes, but it should be integrated in the science platform
HL-LHC	Yes, but we would like to renew it	Yes, but we would like to renew it
VLA use case (SKA pathfinder)	Yes, but it should be integrated in the science platform	No, needs to be developed
JIVE-ERIC	Yes, but it should be integrated in the science platform	No, needs to be developed

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.

ESO - La Silla Paranal Observatory (optical near IR)	Given the long history of optical/near-IR astronomy, there are plenty of options out there. Some level of central integration would certainly be beneficial, if it doesn't come at the expense of flexibility and usability.[a]	Same as above.[a]
European Solar Telescope (EST)	No, needs to be developed	No, needs to be developed
EGO-Virgo	Yes, but it should be integrated in the science platform	Yes, but it should be integrated in the science platform
KM3NeT	Yes, but under construction	Yes, but under construction
CTA	Yes, but under construction	Yes, but under construction
FAIR	Yes, but it should be integrated in the science platform	The building blocks are available but no standard processing pipeline is yet available.[e]
Asteroseismology (No ESFRI project)	Yes, but it should be integrated in the science platform	Yes, but it should be integrated in the science platform

Name of (ESFRI) RI / project	**Are there any other aspects relevant to describe the software used?**
Zooniverse	Current analysis software: https://github.com/zooniverse/swap https://github.com/zooniverse/Data-digging https://github.com/zooniverse/aggregation-for-caesar
LOFAR	The concept of science platform still needs to be developed for LOFAR
VLA use case (SKA pathfinder)	There different software tools involved in this use case: CASA libraries, CASA plotms, SoFiA, python notebooks. We have also encapsulated in docker containers the tools involved in each of the steps of the workflow
ESO - La Silla Paranal Observatory (optical near IR)	Very very diverse...

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.

FAIR	The software building blocks for data processing are based on the FairROOT analysis framework

Notes for the table

[a] We used “Yes” in the analysis

[b] We used “No” in the analysis

[c] We tried to group the answers in “Local system” (dedicated servers, laptops, pcs, etc), “Local cluster”, “Compute centre” (including grid, national infrastructure) and “Cloud” to make Figure 2.

[d] We used “Want to renew” in the analysis.

[e] We used “Under construction” in the analysis

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.