



ESCAPE

European Science Cluster of Astronomy &
Particle physics ESFRI research Infrastructures

ESCAPE Legacy Booklet

Author list

Giovanni Lamanna - LAPP (CNRS-IN2P3)

Ian Bird - LAPP (CNRS-IN2P3)

Mark Allen - CDS (CNRS-INSU)

Xavier Espinal - CERN

Kay Graf - FAU

Stephen Serjeant - OU

John Swinbank - ASTRON

Disclaimer

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein.

The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.



Publication date: October 2022



Table of contents

A word from the ESCAPE Coordinator	5
A word from the ESCAPE External Expert Advisory Board	7
ESCAPE - A preamble	8
ESCAPE DIOS - Data Infrastructure for Open Science	9
A Scientific-Data Lake for Open Science	10
The ESCAPE Data Lake implementation	11
Workload orchestration in a Data Lake.....	11
Understanding and addressing the scientific community needs	12
The Data Lake model at work: continuous assessment and Data Challenges	13
Outlook.....	14
ESCAPE OSSR - Open-Source Scientific Software and Service	17
The OSSR Vision.....	18
Achievements to Date.....	18
Future Goals	21
ESCAPE VO - The Virtual Observatory in ESCAPE	22
The vision of the astronomical VO in ESCAPE and EOSC	23
Achievements to date	24
Future Goals	27
ESCAPE ESAP - ESFRI Science Analysis Platform	29
The ESAP vision.....	30
Achievements to date	31
Future goals.....	32
ESCAPE ECO - Engagement, Communication and Citizen Science	35
The ECO Vision.....	36
Achievements to date	38
Future Goals	42
ESCAPE to the FUTURE	44

List of figures

Figure 1: ESCAPE “EOSC Cell”	8
Figure 2: The ESCAPE Data Lake Model. An overarching Data Management and Orchestration layer to deliver data to a different type of computing infrastructures and resources.	14
Figure 3: A real data workflow use case at the MAGIC telescope. The remote storage in La Palma island is Data Lake aware and acts as a buffer injecting fresh data to the ESCAPE Data Lake. Once the files are consolidated and with the right replication level, the files are deleted from the local buffer allowing new data to be stored.....	15
Figure 4: Data injected to the DL from three radio source observations in external locations. Users in an external location download the data, process and store results back to the DL. Users interested in combining results stored with other public data to also cover the visible spectrum. Combined optical data from the Hubble located via the VO (WP4). Optical and radio data aggregate via the ESAP (WP5), combined analysis done. Results uploaded back to the DL.....	15
Figure 5: Extreme I data management tests by SKAO to certify “End-to-end data lifecycle” from radio telescopes in South-Africa and Australia.....	16
Figure 6: OSSR Architecture.....	18
Figure 7: Deep Learning applied into ESO archive	26
Figure 8: Image from the IVOA.....	27
Figure 9: ESAP provides a single, consistent, interface and point of access to a variety of services drawn from a range of providers. Links to the other work packages ESCAPE are indicated.	30
Figure 10: ECO work package tasks, adapted and updated from the figure in the Description of Work.....	36
Figure 11: Schematic representation of experts leading the way for the non-specialist interactions with the EOSC.	38
Figure 12: Schematic model of the connections between ESCAPE services, in their relations to citizen science. Adapted from a diagram by G. Lamanna and I. Bird.....	38

List of tables

Table 1: Main Results of each ESFRI/RIs	25
Table 2: Communications key performance indicators.....	42



List of acronyms

Acronym	Full text
CERN	Conseil Européen pour la Recherche Nucléaire
CPU	Central Processing Unit
CS	Citizen Science
CTAO	Cherenkov Telescope Array Observatory
DAC21	Data and Analysis Challenge
DIOS	Data Infrastructure for Open Science
DLaaS	Data Lake as a Service
ECO	Engagement and COmmunication
EEAB	External Expert Advisory Board
EGO	European Gravitational Observatory
ELT	Extremely Large Telescope
ENVRI-FAIR	European Environmental and Earth System Research Infrastructures- Findable, Accessible, Interoperable, Re-usable
EOSC	European Open Science Cloud
ESA	European Space Agency
ESAP	ESFRIs Science Analysis Platform
ESFRI	European Strategy Forum on Research Infrastructures
ESO	European Southern Observatory
EST	European Solar Telescope
FAIR	Facility for Antiproton and Ion Research
FAIR	Findable, Accessible, Interoperable, Re-usable
FTS	File Transfer mechanism
GPU	Graphics Processing Unit
HL-LHC	High-Luminosity Large Hadron Collider
HPC	High Performance Computing
IAM	Identity and Access Management
ICT	Information and Communication Technologies
IVOA	International Virtual Observatory Alliance
IWAPP	Innovative Workflows in Astro & Particle Physics
JIV-ERIC	Joint Institute for VLBI ERIC

LOFAR	Low Frequency Array
OSSR	Open-source Scientific Software and Service Repository
OU	Open University
PANOSC	Photon and Neutron Open Science Cloud
PID	Persistent Identifiers
RI	Research Infrastructures
SKAO	Square Kilometre Array Observatory
SRIA	Strategic Research and Innovation Agenda
SSHOC	Social Sciences & Humanities Open Cloud
VO	Virtual Observatory
WP	Work Package



A word from the ESCAPE Coordinator

Science is a source of progress for humanity and is constantly evolving to understand, decipher, decode, and push back the limits of knowledge. More inclusively Science, Technology and Innovation have vital importance to respond to major human and societal challenges.

Scientific Research is progressing towards the new paradigm of “Open Science” for more open, transparent, collaborative and inclusive scientific practices to enhance the impact of science in our society, fostered by the expansion of information and communication technologies (ICT).



This is the fundamental motivation of the ESCAPE scientific community and it is also the challenge shared by pan-European Research Infrastructures (RIs) that are members of the ESCAPE science cluster that are ESFRI projects and landmarks such as CTAO, ELT, EST, FAIR, HL-LHC, KM3NeT and SKAO as well as other pan-European research infrastructures such as CERN, ESO, JIV-ERIC and EGO.

More broadly, the destination of open science has led EU Member States to launch the European Open Science Cloud (EOSC) initiative. A unique cloud that allows universal access to research data via a single platform.

The ambitions of EOSC and open science are about a change in the way citizens might perceive research and public investment in research, a chance to allow everyone to participate in the scientific process and research practice by accessing research data, and a hope to accelerate discoveries and increase scientific value by sharing data and transferring knowledge across scientific communities.

There is no openness in the broadest sense without a concerted programme of work that: aims to overcome sociological barriers between scientific communities; challenges standards and practices to share any digital object beyond any scientific result; supports shared developments and cross-fertilisation actions to target the interoperability of data from multiple sources. In summary, openness is not possible without an operational implementation of the FAIR (findable, accessible, interoperable and reusable) practice for data, and ESCAPE's mission is to achieve this.

ESCAPE is one of the five Science Clusters that resulted from the 2018 H2020 topic initiative aimed at “Connecting ESFRI infrastructures through Cluster projects” to advance in the FAIRness data stewardship and to link to EOSC. Other Science Clusters are ENVRI-FAIR (Environment and Earth Sciences), EOSC-LIFE (Biomedical Science), PANOSC (Neutron and light sources facilities) and SSHOC (Social Science and Humanities).

What is relevant and specific in ESCAPE is that it represents a unique cross-fertilization opportunity for the concerned scientific community that was fostered by two complementary excellences: the data stewardship of large astronomical archives through the Virtual Observatory (VO) infrastructure and the exabyte-scale data management and large-scale distributed computing in particle physics (such as WLCG). Furthermore, astrophysics and particle/nuclear physics communities are generators and consumers of large volume of complex data and at the same time early adopters of ICT and data management innovative solutions pushing the state-of-the-art. The partners in ESCAPE recognise the strong synergies and potential commonalities which are there at several levels: the research collaborations themselves are often synergistic and overlapping, with cross-over between all parts of the community, and within their national research institutes; there are often common funding agencies for astronomy and particle/nuclear physics in many countries; the data and computing facilities that all of these ESCAPE partners use are often host to both astronomy and particle or nuclear physics experiments.

Thus, the natural synergies of the science domains are also reinforced by these factors and have anticipated naturally a vision towards the architectural implementation of EOSC for our community. Indeed, the ESCAPE work structure aims at deploying a domain-based “EOSC cell” composed of a federated data infrastructure (DIOS), a catalogue of co-created high-quality, open-source software (OSSR), the integration of (VO) services for multi-messenger astronomical archived data, the science analysis platform instances customized to the needs of RIs and user communities (ESAP) and services to improve access to data through citizen science crowdsourcing experiments for most of the ESCAPE facilities (ECO).

The successful work programme, the achievements and the ability of ESCAPE to build a global picture of open science and to implement it properly are widely recognised. The top-down approach of the RIs willing to continue cooperative actions by joining their efforts is confirmed. The bottom-up demands of the scientists involved not to interrupt but to continue the cross-fertilisation in science and innovation that ESCAPE has been able to build, are strongly considered. Horizontally, the leverage arm that the cluster work programme provides to universities and institutes is an essential added value.

For these reasons, although the H2020 Science Cluster grant is coming to an end, a new ESCAPE open collaboration agreement is established. This will maintain the exceptional collaborative and human experience represented by the Science Cluster and strengthen the role and impact of astronomy and nuclear/particle physics in the field of open science and, more broadly, in the European Research Area.

Giovanni Lamanna, Director of the CNRS laboratory LAPP and ESCAPE Coordinator





A word from the ESCAPE External Expert Advisory Board

During the entire funding period, the ESCAPE External Expert Advisory Board (EEAB) provided advice and evaluated the achievements of the project. The EEAB conveyed the vision from thematic international and national research institutes and universities, their encompassed communities as well as from the pan-European research fields to orientate activities towards a full achievement of the goals of ESCAPE.

The communities represented by the EEAB are very eager to have a dedicated infrastructure to address the Open Science challenges shared by ESFRI facilities (CTA, ELT, EST, FAIR, HL-LHC, KM3NeT and SKAO), other pan-European research institutions (CERN, EGO, ESO and JIVE) as well as smaller and national initiatives in astronomy, astroparticle physics, nuclear physics and particle physics. The EEAB is particularly pleased with the progress made by ESCAPE in developing solutions for the FAIRness of large data sets handled by the research infrastructures to the benefit of the broad community of users. The EEAB acknowledges the considerable progress in the project with respect to building the astronomy and particle physics cell of the EOSC. Based on the major ESFRI projects, the aim is the ambitious digitisation of the entire research field, which is very broad in scope. In parallel, there is a growing awareness that much can be gained by sharing “Big Data” and best practices between experiments and communities. We consider it important that ESCAPE promotes the use of data format standards to facilitate data access between experiments and supports the transition to open access publishing strategies as well as encourages making data publicly available (as “Open Data” and for “Citizen Science”).

In all these aspects, ESCAPE is at the forefront of the global developments and has defined future standards that are highly appreciated by the communities. It is important that synergies and similarities as well as different levels in the tools of the individual research areas were identified as well as further developed by taking into account the strengths of each individual community and making them available to the others.

We recognize that the interfaces between the ESCAPE Work Packages have become more and more important for the success and for the sustainability of the whole project. It is essential to continue to promote and establish coordination across adjacent disciplines that share similar data-related challenges and that have overlapping scientific ambitions.

The EEAB is very impressed with what has been achieved under ESCAPE and will monitor how the tools and concepts are applied and further developed, which is particularly important for science in the overlapping area of the individual research fields. Therefore, this report cannot be the end of the mentioned activities, but the necessary beginning of a new era of research data management in sustainable basic research in Europe and beyond. This report is a milestone and the EEAB supports the efforts to maintain and further develop ESCAPE as a cross-field infrastructure.

Christophe Arviset, ESA – European Space Agency

Andreas Haungs, Chair of APPEC - Astroparticle Physics European Coordination committee

Karl Jakobs, Chair of ECFA - European Committee for Future Accelerators

Marek Lewitowicz, Chair of NuPECC - Nuclear Physics European Collaboration Committee

Colin Vincent, Chair of ASTRONET Board - Astronomy European Collaboration



ESCAPE - A preamble

ESCAPE brings together a large fraction of the European RIs in Astronomy, Astrophysics, Particle and Nuclear Physics. The ESCAPE project is supported and supervised by national universities and institutes of the European Union member states that are organized in thematic consortia such as APPEC, ASTRONET, ECFA and NuPECC. The large professional community engaged in ESCAPE-related science extends to tens of thousands of scientists.

ESCAPE aims to address the Open Science challenges shared by its partners and the community. These challenges are technical, operational, sociological and scientific. Open Science allows scientific information, data and outputs to be more widely accessible and harnessed. In order to achieve the inclusiveness of the ESCAPE scientific domains in the global implementation of EOSC, a work program is structured to deliver an “EOSC cell” where five technical Work Packages are in charge of the development and deployment of five corresponding pillars of such a cell for open data research.

- 🔗 **DIOS**: Data Infrastructure for Open Science (WP2);
- 🔗 **OSSR**: Open-source scientific Software and Service Repository (WP3);
- 🔗 **VO**: The Virtual Observatory in ESCAPE (WP4);
- 🔗 **ESAP**: ESFRI Science Analysis Platform (WP5);
- 🔗 **ECO**: Engagement, Communication and Citizen Science (WP6).



Figure 1: ESCAPE “EOSC Cell”



ESCAPE DIOS - Data Infrastructure for Open Science



The ESCAPE Data Infrastructure for Open Science (DIOS) is a federated data infrastructure able to cater for the multi- Exabyte needs of current and future ESFRIs and other RIs. It provides FAIR data management principles, serving global and varied scientific communities in a scalable and performing fashion. The DIOS Data Lake model integrates a common set of tools to orchestrate the computing resources provided by computing centres to one or many RIs. This common set of tools facilitate data management, data transfer, data access, an information system and a common identity management framework. The Data Lake model implements the vision of a single data entity, hiding the inherent complexity and enabling policies, rules and data life-cycles to be applied to on the Data Lake infrastructure as a whole.

A Scientific-Data Lake for Open Science

A Data lake is a form of storage consolidation offering a structure where data is concentrated in selected sites, and globally orchestrated as a single entity. This offers the ability to implement common policies, rules and data life-cycles acting and operating on the Data Lake as a whole, e.g. data is stored and accessed from the Data Lake as an entity, not as a specific site. The Data Lake service can hence be used both in a very simple manner: *a la cloud* push/pull experience ironing out all the complexity for end-users¹, but also as an extremely capable fine-grained system that can be exploited by data management experts² to implement access control, fine-tuned data replication, data pre-placement, storage quality of service, and leverage content delivery services, etc.

The ESCAPE Data Lake is built from a set of sites providing storage services to specific ESFRIs/RIs and user communities, with the goal to carry out independently well-defined tasks. This requires that their combined storage capacity and bandwidth can meet the demands of the designated task and that usage of the different sites is transparent to the users. This means that a form of trust relationship exists (through a common AAI), a way to locate data is in place as a high level data management system (RUCIO) and a File Transfer mechanism (FTS) guarantee a high-level transport layer with the required protocols and interfaces with the storage systems at the sites. The Data Lake model also opens the door to storage consolidation by simplifying the scope of storage used for data processing oriented facilities. Unmanaged transient storage such staging areas, a streaming-cache layer or buffers (data transport layers from now on) are used to transition files between the Data Lake and the different types of sites providing computing resources: standard grid-like sites, HPCs, commercial/private/hybrid clouds, or other short term resources, etc. These data transport layers would enable file re-usability and latency hiding mechanisms that would eventually facilitate data access to workloads that are active or expected to become active shortly. Data transport layers are typically located at the edge of the sites, where data is moved/removed by data management

-
- 1 Lightweight data management requirements: user upload store data to a URL-like endpoint, the data get FAIR-ed behind the scenes and readily available for the user and his collaborators also on an URL-like style
 - 2 Sciences with Peta/Exa-byte scale Data Management needs, harnessing different storage, data life-cycles and large user communities.

systems or triggered by the application access, in case the data is not present the system will fetch the file from the Data Lake contacting the data management system. The implementation of these data transport layers is transparent to the users and potentially breaks the co-location paradigm of CPU's and data.

The ESCAPE Data Lake implementation

The ESCAPE project aims to address Open Data and Open Science challenges in Astronomy and Particle Physics for the coming Exa-scale era. Through ensuring FAIR data principles, helping to connect RIs to the EOSC and fostering common integration of tools among the RIs and the resource providers. Different components constitute the architecture of the Data Lake, which is designed to provide a flexible and scalable infrastructure in terms of data organisation, management, and access. The storage services hosted at different facilities are the core of the infrastructure, exposing different storage systems to the Data Lake. These systems are seamlessly integrated to work as a whole via well-known protocols: gridFTP, http/webdav, and xroot; offering download, upload, and data streaming functionalities, third-party transfer capabilities and data deletion on a unified namespace.

The open-source data management system Rucio is the enabling technology that implements data orchestration in the Data Lake enabling: data policies, replication rules, file layout transitions, data life-cycles, distributed redundancy, etc. The data transfer technology stack is implemented through GFAL, FTS and Rucio. GFAL is a multi-protocol data management library providing an abstraction layer of storage systems complexity. FTS provides reliable data transfer at large scale between the storage systems enabling a Third Party Copy ability to transfer data directly between arbitrary storages end-points. Rucio uses GFAL for upload/download operations, FTS employs GFAL³ in order to perform the actual transfer. GFAL also works on the storage file system level with the supported protocols. Rucio provides a common namespace for the users to interact with the data.

The authentication and authorization schema used to manage access control relies on the use of X.509 certificates and the VOMS stack while the transition towards token based authorization is ongoing in parallel. This transition will be achieved by using the OpenID Connect (OIDC) identity layer on top of the OAuth2 protocol. Rucio supports OIDC authentication, this is being tested at the moment. These capabilities are implemented through the INDIGO IAM (Identity and Access Management) service and accommodates the needs of RIs to deal with embargoed/restricted-access data. Interoperability of the Data Lake services is ensured by the Compute Resources Information Catalogue (CRIC) a central catalogue containing services information and configuration for Rucio, such as. supported protocols at sites and their prioritisation.

Workload orchestration in a Data Lake

This ecosystem of tools and services have been integrated and commissioned to build the Data lake. This effectively keeps complexity to a minimum, but when needed by the RI's data management requirements, allows access to the underlying services for more fine-grained control. The Data Lake addresses FAIR principles by construction and facilitates the transition to Open Data and Open Access in Open Science frameworks. The users perceive one single data container which effectively federates underlying distributed resources, providing a common global namespace and allowing to define, policies, quotas, access rights, data life-cycles, replication rules, file redundancy levels and more. Data is stored in the Data Lake with the defined file/dataset attributes: replication, accessibility, lifecycle, etc. and it is immediately ready to be consumed by the user community.

The way data gets delivered to the user community can be varied, from end-user laptop and notebooks or to the user-neighbourhood infrastructures: large storage at a university or caching layers. Therefore, the processing facility can range from a very simple resource: a laptop of a single user downloading a file via http, or a notebook service on a cloud accessing data via the Rucio-CLI; or on the other hand the resources can be large structures: heavy-duty batch systems, a large cloud provisioning or a punctual but massive CPU/GPU allocation on an High Performance Computing (HPC) centre. This reflects the fact that computing resources are getting very heterogeneous, the standard one-size-fits-all grid-site model is vanishing. Also hinting that data locality relying on a fully fledged storage system is not always needed nor effective to exploit such a variety of compute resources.

A sufficiently flexible de-centralised Data Lake can be leveraged to efficiently serve data for a variety of wide-ranging use cases. It is a quite different task to address a data processing campaign where the same job types might run on hundreds of Petabytes, compared to the case of hundreds of user-analysis jobs running on a few files of few Gigabytes each. Both extremes can be managed by the Data Lake; while the first heavy-duty example might be better orchestrated with data pre-placement policies set by the user on selected sites providing a large number of cores, the user analysis use-case might run on any place reading from an arbitrary storage location in the Data Lake at anytime. These two extreme examples illustrate there is space for anything in between them and emphasises once more the fact that the level of complexity is a choice that needs to be adapted to the user/RI needs.

Understanding and addressing the scientific community needs

Some specific topics have been identified and addressed to make the Data Lake model flexible and widely adoptable:

- ☞ Token-based authentication integration across the several layers of the Data Lake infrastructure: Rucio, FTS, storage systems and potential integration with other AAI providers. end-to-end workflows run integrally using token-based authentication, including the data access and processing from analysis platforms;
- ☞ Full data life-cycle accommodation with the ability to define replication rules and policies: data redundancy, location and an eventual mapping to storage quality-of-service provided at the sites to adjust data popularity needs and optimise the use of available storage types and cost;
- ☞ Webdav/HTTP ability has been promoted to be the de-facto standard transfer protocol in the Data Lake. The widespread acceptance of HTTP protocols also provides a flexible way to interact and integrate storage resources;
- ☞ Fostering Rucio evolution by gathering and providing feedback from new scientific communities using Rucio. This interest and feedback is being taken into consideration by the Rucio core team who are evolving the service to cater for these new scopes and needs. There is a special interest from astroparticle, astronomy and cosmology communities to enlarge and expand the Rucio metadata capabilities;
- ☞ Expanded Data Lake monitoring capabilities. The monitoring platform provides a good overview of the ongoing activities in the infrastructure, such as in-flight transfers and scheduled data movements, but also actual usage of resources in terms of storage by service provider, and by experiment;
- ☞ Active Deployment and Operations team (DepOps). Early in the project we identified the need to share experience with the tools and the operations. The DepOps team is formed by experiments and site representatives and coordinated via a well-established meeting, and is demonstrated to be instrumental in fostering knowledge transfer and expertise sharing. The DepOps teamwork was crucial to prepare and drive the Data Challenges;

- ☞ Demonstrate that the Data Lake “big-data” scope could be integrated also into a much simpler “just-a-bunch-of-files” user-perspective. This aimed to develop a Rucio extension allowing to connect the Data Lake infrastructure with the various analysis machines a user might be using. In this spirit, we developed a Rucio extension that is pluggable to vanilla Jupyter notebooks, giving the ability to browse and download data in the Data Lake from the web browser/notebook. New improvements and ideas resulted in further development and integration of the Data Lake capabilities with user environments, and produced the integrated “Data Lake as a Service” (DLaaS) offering, providing increased browse/download/upload data capabilities, token integration, data movement within the notebook; permitting also integration of local storage, and leveraging content delivery and caches, etc., allowing and extending the integration possibilities with Analysis Platforms, computing infrastructures, and resources local to the services provider;
- ☞ Integration of heterogeneous resources has been demonstrated. The Data Lake flexibility enables heterogeneous data processing and compute facilities to be used in conjunction with the Data Lake. For example, the use of commercial clouds for compute and for storage through the implementation of Swift/S3 storage endpoints in the Data Lake. Integration with HPC facilities has been enabled through activities with CINECA/HPC, and a collaboration with the FENIX/HPC project.

The Data Lake model at work: continuous assessment and Data Challenges

There has been a continuous assessment and evolution of the Data Lake model. The early pilot Data Lake assessment culminated in a joint exercise - the “Full Dress Rehearsal” (FDR20). The several layers of the pilot Infrastructure were exercised during a 24h production-like window where experiments executed relevant workloads covering a wide range of activities, from data recording from the experiment detectors/sources to data browsing and access via notebooks for user analysis purposes.

The FDR20 served to confirm that the infrastructure was solid and the concept successfully addressed the actual needs of the experiments present in ESCAPE. Nevertheless, the FDR20 exercise allowed us to pinpoint areas for improvement and where to focus subsequent efforts.

The next big challenge was the “Data and Analysis Challenge” (DAC21), where during 10 days the RIs in ESCAPE ran production-like Data Management, Processing and Analysis workloads. This included data acquisition activities from data sources, policy-driven data replication and data lifecycle implementation. Data processing was a fundamental target of the DAC21, therefore emphasis was put on a variety of use-cases of processing activities including interplay possibilities using large scale resources (batch systems and clouds) and user-analysis oriented platforms (online notebooks and analysis platforms).

During DAC21 several experiments deployed and used for the first time private installations of the Data Lake Data Management and File transfer tools, thus demonstrating its seamless integration into a wide common Data Lake Storage infrastructure. Hence demonstrating good potential for sustainability and reinforcing the synergies with the ramping-up activity in the EOSC Future.

Incorporation of third party centres was proven. One of the examples is the MAGIC telescope (CTA precursor) where an external storage was integrated in the Data Lake. This external storage is the telescope’s data acquisition (DAQ) buffer space, in charge of recording data from the telescope. After the data is stored locally at the telescope, the Data Lake Data Management tools take care of injecting the data into the Data Lake and once the data is consolidated, and with the adequate level of replication, it gets deleted from the DAQ buffer to free space and have room for new data acquisition (Fig.2).

Another successful example of integration is the workload run by LOFAR (Fig.3), where data from three

different radio sources is injected into the Data Lake, subsequently distributed and then accessed via a notebook for Analysis. This data then is combined at the notebook level with other data from the visible wavelengths coming from the Virtual Observatory catalogue. As a final product a combined wavelength analysis was delivered.

The extreme long haul data management tests by SKAO certified end-to-end data lifecycles (Fig. 4). Injecting regular data products at each of the two “telescope” storage endpoints (IDIA in Cape Town and AARNET in Perth), incorporating the following aspects, to mimic telescope data product placement into a staging area: Upload the data to a non-deterministic source storage endpoint, register the data in place in Rucio, Subscription-based movement, Long haul transfer, Life-time based QoS transitions (including start and end dates), short lifetime (1hr) at source. This test was fully successful. Then transfer data from IDIA and AARNET to Manchester and Lancaster (both UK sites).

Outlook

The Data Lake model has been successfully proven as a capable working system able to address the scientific computing needs of the various scientific communities in ESCAPE. The ESRIs/RIs could implement, in a production-like scenario, entire data life cycles ranging from raw data recording to data distribution for user access. Some of the tools and services used in the ESCAPE Data Lake have been already picked up by some ESRIs/RIs for testing, and are currently being evaluated as a possible solution to address the upcoming challenges in their computing models. The ESCAPE Data Lake model has been certified as a potential solution to address RI’s Exa-Byte scale computing needs, and able to deliver data and provide access to/from a varied type of processing facilities: large batch systems, computing clouds and HPCs. On the other hand the ESCAPE Data Lake also proved its usefulness to address simple data needs with its ability to hide complexity and providing simple interfaces for data access.

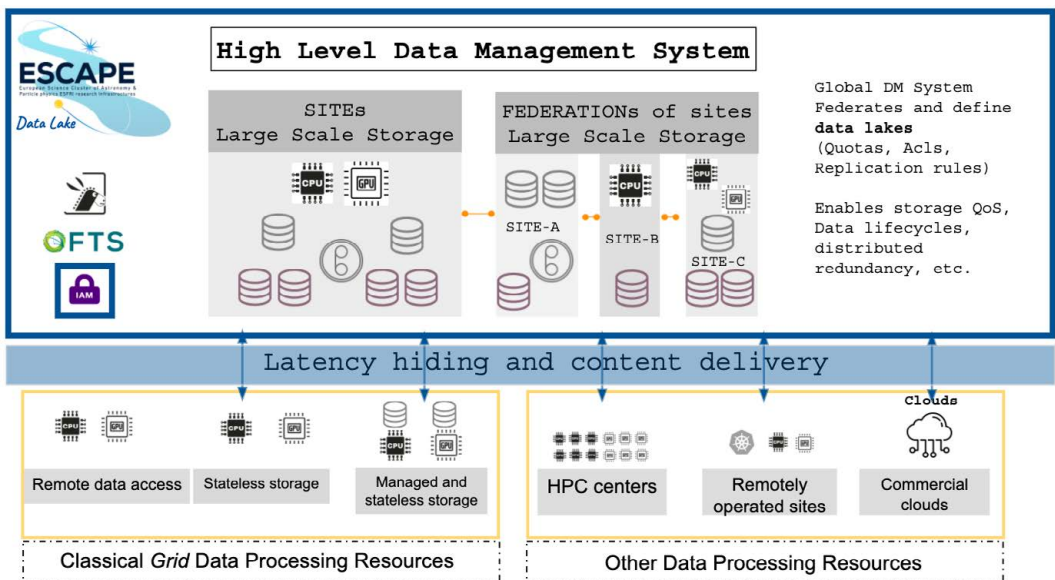


Figure 2: The ESCAPE Data Lake Model. An overarching Data Management and Orchestration layer to deliver data to a different type of computing infrastructures and resources.



Figure 3: A real data workflow use case at the MAGIC telescope. The remote storage in La Palma island is Data Lake aware and acts as a buffer injecting fresh data to the ESCAPE Data Lake. Once the files are consolidated and with the right replication level, the files are deleted from the local buffer allowing new data to be stored.

Example: multi-wavelength analysis



1. Data injected to the DL from **three** radio source observations in external locations
2. User in external location download the data, process and store results back to the DL
3. User interested in combining results stored with other public data to cover also visible spectrum
4. Combined optical data from the Hubble located via the VO (WP4)
5. Optical and radio data aggregate in via the **ESAP (WP5), combined analysis done**. Results uploaded back to the DL.

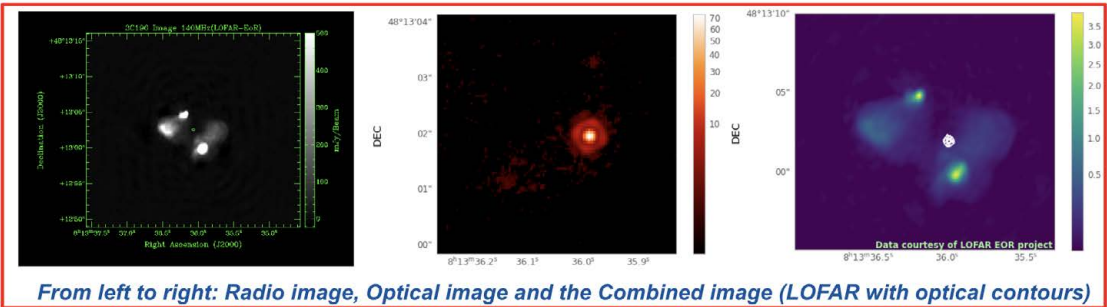


Figure 4: Data injected to the DL from three radio source observations in external locations. Users in an external location download the data, process and store results back to the DL. Users interested in combining results stored with other public data to also cover the visible spectrum. Combined optical data from the Hubble located via the VO (WP4). Optical and radio data aggregate via the ESAP (WP5), combined analysis done. Results uploaded back to the DL.



Storage Endpoints

European, South African and Australian sites
Best efforts basis currently

Mix of deterministic and non-deterministic RSEs (**non-deterministic to mimic SKA data staging source storage**)

Instance will continue to grow over coming year, as SKA Regional Centre partners join and help assess functionality. Anticipate Spain, China, France, and maybe Canada



Long-haul transfers from Australia and South Africa to European locations during DAC21

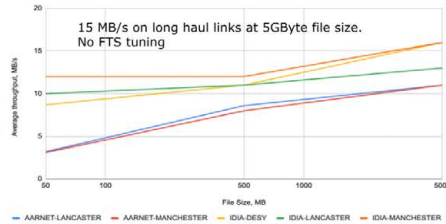


Figure 5: Extreme I data management tests by SKAO to certify “End-to-end data lifecycle” from radio telescopes in South-Africa and Australia.



ESCAPE OSSR - Open-Source Scientific Software and Service

Activities within OSSR are broadly divided into three major areas:

- 🔗 Support a community-based approach for continuous development, deployment, exposure and preservation of domain-specific open-source scientific software and services in the global context of the EOSC catalogue of services - the OSSR itself;
- 🔗 Enable open science interoperability and software re-use for the data analysis of the ESCAPE ESFRI projects based on FAIR principles;
- 🔗 Create an open innovation environment for establishing open standards, common regulations and shared software libraries for multi-messenger/multi-probe data;
- 🔗 Educate stewards for FAIR software by knowledge transfer, collection of best practices and software schools.

The OSSR Vision

The OSSR vision is to establish a trustable, sustainable repository for software and services and to foster collaboration on the co-creation of high-quality, open-source software for open science. Contributing to the OSSR enhances software quality through guidelines and recommendations for software contributions, enables the development of interoperable use of the software through extended metadata and increases findability of software through OSSR integration in the EOSC and other research environments. It thus makes software a first class citizen of open science and the EOSC.

Achievements to Date

Architecture

The OSSR is based on collecting software in the form of a code repository or container image in a Zenodo community⁴, which is connected to a specialized landing page and searchable via a python client library. The landing page⁵ is the entry point of users to the OSSR products, as well as to other services within the ESCAPE EOSC cell. It also contains links to documentation and training materials.

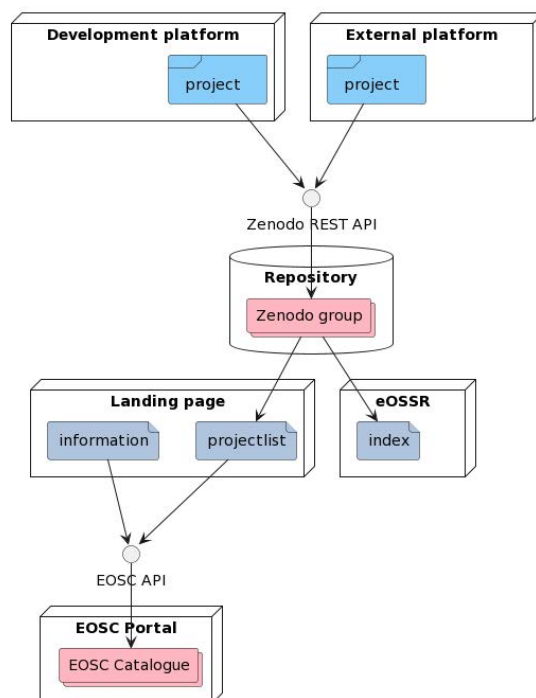


Fig. 6: OSSR Architecture

4 <https://zenodo.org/communities/escape2020/>

5 <http://purl.org/escape/ossr>

The development platform⁶ provides a common place to gather the common developments, ideas, guidelines and templates for the community, as well as a platform for new developments if required by an institution/group without access to another solution. It showcases the full software lifecycle up to the publication in the repository. The development platform is not to substitute the development platforms already used by each institution/group, the technical enhancements necessary to create project links to the OSSR can be equally well applied on the respective native development platforms.

The repository backend of OSSR is the ESCAPE2020 Zenodo community⁷. It holds the long-term archived open science projects developed in the platforms. Each OSSR record is required to have an additional codemeta.json-file to hold the extended metadata that describes the software.

To facilitate easy access, the eOSSR Python library⁸ gathers all the developments made for the OSSR. In particular, it includes an API to programmatically access the OSSR, retrieve records and publish content, functions to map and crosswalk metadata between the CodeMeta schema adopted for the OSSR and Zenodo internal schema and functions to help developers automatically contribute to the OSSR, in particular using their continuous integration.

Integration of the OSSR in aggregating portals will increase the visibility and findability of the OSSR entries. To this end, the OSSR will be integrated in the EOSC portal. Additional portals will be considered, and already the OSSR has been integrated to re3data⁹, the registry of research data repositories.

Onboarding

Partners in the ESCAPE project and the wider science community are encouraged to onboard their scientific software, public datasets (limited in size), container images, or repositories with full analyses environments to the OSSR¹⁰. They are requested to complete an onboarding process which involves the curated presentation of their project and upload of the contribution to the Zenodo community, triggering a short review process. Although currently focusing on ESCAPE-related projects, contributions are encouraged from any related research field, as is already exemplified by a cooperation with the DMA project in the Helmholtz research program Matter.

Content

The OSSR is designed to be flexible through the customized adaptation of the metadata. Perspective entries to the repository have to be stored as software or dataset in Zenodo, but could range from installable software packages to containerized images of software to full analysis environments including data, or smaller data bases. Extended metadata for installation or integration of the content in a larger research environment lies within the responsibility of the content provider, however, metadata required for interoperability and findability of specific entries can be integrated in the OSSR schema definition and eOSSR library functions on request. Thus, the OSSR functionality can be adapted to the requirements depending on the contributors' use cases.

6 <https://gitlab.in2p3.fr/escape2020/wp3>

7 <https://zenodo.org/communities/escape2020>

8 <https://escape2020.pages.in2p3.fr/wp3/eosr/>

9 <https://www.re3data.org/repository/r3d100013827>

10 <https://escape2020.pages.in2p3.fr/wp3/ossr-pages/page/contribute/onboarding/>

Best practices

In order to collect best software practices, a workshop was organized with contributions from a wide field of scientific software experts¹¹. The outcome of the workshop is made available in the OSSR and used to inform contributors about recommendations on software development. Focus topics include software interoperability, development strategies, licensing practices and software solutions. In the guidelines and rules of participation to the ESCAPE OSSR¹², this led, amongst others, to recommendation of an open source license, the requirement to provide documentation alongside the software, and ensure software provenance through adequate version control.

Transmitting best practices and know-how

Producing FAIR software in practice requires know-how. The ESCAPE Data Science schools organised by OSSR transmit such knowledge and educate software stewards. During these schools, scientists in the field of astronomy, astro-particle and particle physics are taught the necessary ingredients for their software to become a part of open science by experienced code custodians. The 2021 special online edition welcomed more than 1000 registered participants. Following the FAIR paradigm and as an example of good practices in code development, all the school material is openly available online, including scientific programme, agenda and links to all contributions (software repository, notebooks, presentations and recordings).¹³

Cooperation

An exemplary effort to foster cooperation is given in the ConCORDIA project to produce CORSIKA turnkey containers for various use cases in astroparticle physics and linked research fields shared for research conducted with primary and secondary detectors underwater or in ice and for muons in the low atmosphere. It involves setting up the CORSIKA containers and running test productions in various computing centres and assessing and certifying the quality and physics relevance domain of the simulation.

Cross-fertilisation

Although cross-fertilization occurs on various levels in the work package, a special focus was put on the establishment of an Innovation Competence Group which was finalized by the organisation of an all-hands meeting, implemented as an online workshop IWAPP - Innovative Workflows in Astro & Particle Physics¹⁴.

The construction of the innovation group started with a series of 12 dedicated meetings of the Focus Group 3 and Task 3.4 of OSSR, where the different ESCAPE partners presented their activities in terms of innovation in data management, software and data analysis, in particular regarding the use of artificial intelligence and especially in the form of deep-learning techniques for querying large data archives, pre-processing data, object classification and parameter inference. The group continues to be active on specific topics, especially in regard of establishing a common project in the context of EOSC-future.

11 <https://escape2020.pages.in2p3.fr/wp3/woss/>

12 https://escape2020.pages.in2p3.fr/wp3/ossr-pages/page/contribute/guidelines_ossr/

13 <https://doi.org/10.5281/zenodo.5838436>

14 <https://indico.in2p3.fr/event/20424/>

Future Goals

This section reflects the future work within the OSSR context.

Technical Developments

The current OSSR architecture, provides a robust and flexible repository implementation. However, there are opportunities for further enhancements that cannot be completed within the scope of the ESCAPE project, as:

- 🔗 Extending the CodeMeta schema beyond the current implementation to keep track of further developments within the repository (see also 4.2);
- 🔗 Extending the eOSSR library allowing for enhanced searches, additional development platforms and archives;
- 🔗 Support for a seamless integration with the ESAP and Virtual Research Environment and the full support of the outcomes of the ESCAPE/EOSC-Future Test Science Projects.

Sustainability

The long-term sustainability of the OSSR service was a design requirement from the beginning. Several choices have been made so that the individual components have an impact also in the future even if OSSR as a service itself should not be maintained, this is e.g. for the simple interfaces between the different components, the metadata standard and the archive in Zenodo.

Several actors in the current OSSR focus group on software and service collection have expressed their interest to join a *Curation interest group* that continues the onboarding - on a lower level - also after the ESCAPE project. More importantly, the maintenance of OSSR is also one of the goals of the recently funded ESCAPE collaboration.

Future integration and cooperation

One major future goal is that a repository in the form of the OSSR - or future developments of it - become a natural part of the EOSC Exchange Layer or even the EOSC core services and the interoperability between repositories part of the EOSC Interoperability Framework. This will be a medium-term goal for the future integration of OSSR.

For this, the EOSC architecture integration should follow the recommendations of the EOSC working group on *Scholarly infrastructures for research software*¹⁵ and the current *EOSC Task Force on Infrastructure for Research Quality Software*.

Repositories as the OSSR should be sustained as essential part of the infrastructure for scientific software. The collaboration also with other communities should be further pursued and the architecture harmonised. This should be a mid-term future effort between the different EOSC Clusters or their follow-up organisation. In a long-term vision, an open-science software repository could be lead as a foundation.

15 European Commission, Directorate-General for Research and Innovation, Scholarly infrastructures for research software: Report from the EOSC Executive Board Working Group (WG) Architecture Task Force (TF) SIRS, Publications Office, 2020, <https://data.europa.eu/doi/10.2777/28598>



ESCAPE VO - The Virtual Observatory in ESCAPE



The activities of VO (Virtual Observatory), are organized around three main areas. Firstly, the integration of astronomical VO services into the EOSC. Secondly, the implementation of FAIR principles for ESFRI data via the development and use of common interoperability standards. Thirdly, the application of new techniques to data in astronomy archives to add value to the scientific content. These activities are integrated with the other areas of ESCAPE so that VO services are progressively made available in ESAP, VO software libraries and applications can be on-boarded to OSSR, and data from VO services can be used in citizen science projects.

The vision of the astronomical VO in ESCAPE and EOSC

The VO is a framework of open standards for making astronomy data FAIR. It is an established and operational framework that has proven to be a great success for many aspects of the interoperability and FAIRness of astronomy data. It is an essential component of the astronomy data landscape, as has been strongly stressed in the ASTRONET Infrastructure Roadmap (2008, 2014, 2022). The standards are defined by the community driven International Virtual Observatory Alliance (IVOA). Astronomy data providers, in particular ground- and space-based observatories publish their data using the IVOA standards, and compliant scientific tools and services enable the discovery, access and (re-)use of the data by the whole astronomy research community.

The vision of the VO is that astronomical datasets and other resources should operate as a seamless whole. It can be best thought of as an ecosystem of tools and services that discover and access astronomy data sets distributed worldwide. ESCAPE WP4 brings together partners with scientific and technical expertise in the VO framework together with partners who are connected to the astronomy ESFRI to pursue the definition and adoption of these open IVOA standards based on the requirements and priorities of the ESFRI so that they fully participate in the global VO framework. This is supported by a comprehensive set of training activities for researchers and data providers, as well the development of innovative new software tools and libraries to enable new scientific capabilities.

The vision of the VO in EOSC is that the growing and operational VO interoperability framework becomes integrated into the EOSC, facilitating VO services to benefit from EOSC, and supporting interdisciplinary science making astrophysics resources discoverable in a wider scientific environment. As such VO is part of the ESCAPE thematic cell that will become a part of EOSC. Furthermore, it will serve as an example of how the catalogue of an entire discipline, the VO registry of Astronomy, can be connected to EOSC as a consistent whole that is supported by open community based standards.

Achievements to date

Connection of the astronomical VO to EOSC

The main result to date of the work toward connecting the VO to EOSC is the mapping of VO Resource metadata against the EUDAT B2FIND metadata. This mapping of VO resources to DataCite-based¹⁶ EUDAT catalogue service has been validated as working, and it can be accessed through the EUDAT B2FIND service itself¹⁷. The EUDAT B2FIND service is already onboarded to the EOSC Portal and the metadata is harvested by OpenAIRE which demonstrates the interoperable propagation of metadata as a first step toward having all of the VO resources visible in EOSC portal.

Interoperability Standards to address ESFRI/RI needs

Over the course of the ESCAPE project there has been much progress on the activities for implementation of FAIR principles for ESFRI data through the VO. The WP4 work on the definition and implementation of standards is done by using, and building on, the IVOA framework of standards. WP4 has contributed to the update of the underlying architecture of these standards (their organization, functions and inter-relationships) with the release of the IVOA Architecture Version 2.0¹⁸ in 2021.

The ESCAPE ESFRI requirements for standards have been prepared via specific meetings and the annual WP4 Technology Forum event where all partners develop the requirements and concepts for standards including a significant amount of technical detail, and scientific motivation. These requirements are then brought to the IVOA meetings as contributed presentations and demonstrations as input for the standardization process. The results of the ESCAPE input at the IVOA meetings are collected and assessed in Milestone reports.

The table below lists the main results that have been achieved toward the development of interoperability standards and their implementation in tools and services related to the various ESFRI/RIs. These standards are applicable to a wide range of astrophysical data and services and the table also indicates the ESCAPE partners that have contributed to the results. An important achievement is that through this process the ESFRI/RIs have built their capacity for implementation of open standards, and become actors in the development of the standards. ESCAPE has enabled the sharing of expertise and experience, for example, the ESO archive has been used as an exemplary case of operational implementation of the IVOA standards¹⁹, sharing their experience on implementing the standards²⁰.

16 DataCite Metadata Schema; <https://schema.datacite.org/>

17 <http://b2find.eudat.eu/group?q=ivoa&sort=title+asc>

18 <https://ivoa.net/documents/IVOAArchitecture/20211101/index.html>

19 ADQL - Aladin Lite - DataLink - HiPS - ObsCore - SAMP - SODA - SSA - STC-S (point, circle, polygon, multi-polygon) - TAP (DALI, VOSI, UWS, UCD, UTYPE, ...) – VOTable.

20 <https://indico.in2p3.fr/event/23481/>









ESFRI/RIs	Partners	Results toward interoperability standards and tools
ESO-ELT. 	CNRS-ObAS, ESO, HITS INAF, INTA, UEDIN, UHEI.	<ul style="list-style-type: none"> 🔗 IVOA standards for data access and visualisation 🔗 DataLink v1.1²¹, MOC2.0 🔗 Support of VO standards in ESO archive services used as exemplary case to demonstrate the use of interoperability standards in a large ESFRI archive 🔗 Standards relevant to Optical/IR/survey astronomy 🔗 Tools for visualisation 🔗 Aladin Lite v3 prototype
EGO/VIRGO. 	CNRS-ObAS, EGO (INFN).	<ul style="list-style-type: none"> 🔗 Development of MOC2.0²² standard (<i>approved March 2022</i>) 🔗 Major update of mocpy as a reference implementation of the standard 🔗 Tools & libraries integrated into GW community software (<i>e.g. ligo.skymap</i>) 🔗 Paper accepted in Astronomy & Computing journal 🔗 Including python notebook and tutorial
SKAO, JIVE, ALMA, (LOFAR).   	ASTRON, CNRS-ObAS ESO, INAF, JIVE, SKAO, UHEI.	<ul style="list-style-type: none"> 🔗 Creation and support of the IVOA Radio Astronomy Interest Group 🔗 Publication of an IVOA note: <i>Radio astronomy in the VO: services implementation review v1.1</i>²³ 🔗 Proposal for an extension of the IVOA ObsCore Data Model for radio visibility data 🔗 Example TAP services developed by ESCAPE partners accessible in VO tools and in the ESCAPE platform
CTA, KM3NeT.  	CTAO, CNRS-ObAS, CNRS-CPPM, Obs-Paris, UHEI.	<ul style="list-style-type: none"> 🔗 Data Provenance standard (ProvDM) approved by IVOA 🔗 Many activities for adoption and implementation - Workshop²⁴ 🔗 Reference paper published on a: Management System for Provenance Information
EST 	CNRS-ObAS, INTA, KIS, ORB, UHEI.	<ul style="list-style-type: none"> 🔗 VO metadata developed for Solar Physics. Preparation and submission of formal proposal for additional semantic metadata to be included in IVOA UCDS 🔗 Prototype TAP services for solar data implemented at ROB using EPN-TAP

Table 1: Main Results of each ESFRI/RIs

21 DataLink v1.1 (in review) <https://ivoa.net/documents/DataLink/20211115/index.html>

22 MOC v2.0 <https://ivoa.net/documents/MOC/20220317/index.html>

23 IVOA Note publication: <https://ivoa.net/documents/Notes/RadioVOImp/index.html>

24 <https://indico.in2p3.fr/event/21913/>

One illustrative example of the results is the development of sky-spatial and temporal coverage systems for indexing of astronomy data, designed to support fast data access and management of complex sky regions, for example for gravitational wave follow-up campaigns connected to EGO/Virgo. The IVOA MOC 2.0 standard which provides a “Multi-Order Coverage” map based on the HEALPix tessellation, was led by ESCAPE and become an IVOA standard in March 2022. This provides a general capability for being able to search and cross-match astronomy data sets based on their sky coverage and their temporal coverage. The new capabilities were applied by EGO-Virgo and resulted in a journal paper “*Multi Order Coverage data structure to plan multi-messenger observations*”²⁵. The paper includes an interactive python notebook and on-line video tutorial materials, both of which were used in the WP4 training event the ‘2nd Science with interoperable data school’²⁶ in February 2022. Other notable results are the new VO services implemented by the ESCAPE partners in the areas of radio astronomy, high energy astronomy and solar physics. And furthermore the maturation of the standards and tools for data provenance, in particular for high energy use cases of CTA and KM3Net.

Application of new techniques to data in astronomy archives.

ESCAPE WP4 has made important progress on the visualisation of multi-scale astronomy data, with the development of a new WebGL enabled version of the Aladin Lite application which serves as the visualisation component of a number of Astronomy archives. This hierarchical visualisation system is based on the IVOA HiPS and MOC standards as has grown into a distributed network of HiPS nodes around the globe. The new version enables the use of multiple sky projections, and importantly enables use of the FITS format version of the HiPS standardised surveys. The figure below show the new version visualisations of the ESO VVV survey and the MeerKAT radio image of the centre of the galaxy, as well as an all-sky view using the Mellinger optical data.

Achievements have also been made for the application of Deep Learning to the content of the ESO archive, with a number of demonstrations being presented to the astronomy community, and also scientific results published²⁷.

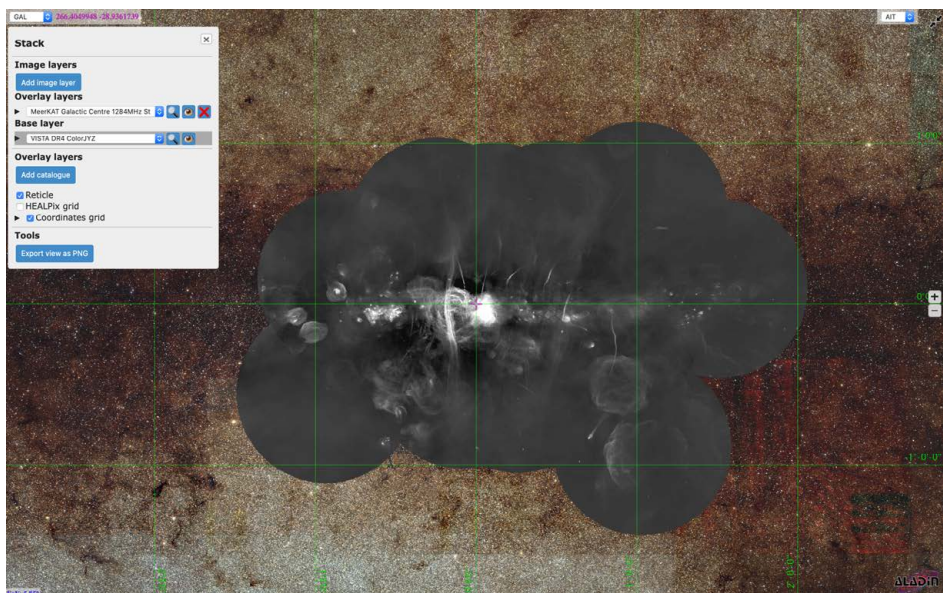


Figure 7: Deep Learning applied into ESO archive

25 Greco et al. 2022, accepted in Astronomy and Computing

26 <https://indico.in2p3.fr/event/25225/>

27 MNRAS paper : N. Sedaghat et al. (2021), Open Access in [arXiv](https://arxiv.org/)

Future Goals

The VO is established as the interoperability framework for astronomy with the IVOA actively developing the interoperability standards in response to the scientific and technical needs of the astronomy infrastructures. This is a rapidly evolving field in particular due to the emergence of multi-messenger and time-domain astrophysics, as well as the significant increase in the volume of the data from current and future instruments. The SKAO for example will bring astronomy into a new era of big data, and like with the major observatories and space agencies, the SKAO plans to use VO standards and has very recently become a member of the IVOA.

Coordination of the European representation in the IVOA for the development of the VO is crucial for its sustainability. ESCAPE has extended the network of cooperating infrastructures who use and contribute to the VO, and the achievements described above will be a long lasting legacy of ESCAPE provided that the framework can be well maintained. As such an essential future goal is to maintain the coordination of VO activities in Europe to meet the scientific and technical needs.

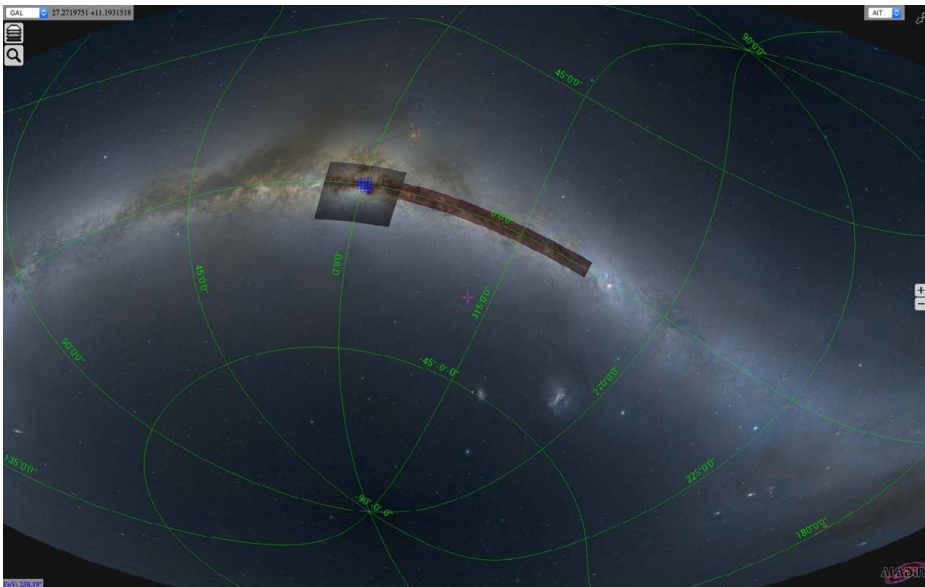


Figure 8: Image from the IVOA

The immediate activities that follow from the success of the ESCAPE project include the continued implementation of the standards and tools. The work on spatial and temporal indexing of data has applicability across all of the astronomy infrastructures, and the goal is to foster innovative use of this system for multi-messenger and time-domain astronomy. Implementation of the visualization tools built in ESCAPE are expected in the ESO and ESA archives, and a wide range of data access interfaces and science platforms. In terms of standards, one of the clear next steps is to define the standards necessary for science platforms and their implementation is various kinds of virtual research environments.

In the ESCAPE project we have taken the first steps of integrating the VO into EOSC. This has been done while EOSC itself was in a very early stage, well before the SRIA and the formation of the EOSC Association. The next clear steps are to make the astronomy resources more visible in EOSC Portal via involvement in EOSC Future and other projects. This will also involve a necessary assessment and test of the new 'enhanced EOSC Resource catalogue' for on-boarding of 'data sources' which is expected to be a significant evolution of the current on-boarding process.

The ESCAPE project has enabled the integration of the astronomical VO activities into a wider scope of addressing the Open Science challenges of astronomy astroparticle and particle physics infrastructures. The VO activities are part of the ESCAPE thematic cell including the Data Lake, Software Repository and Science Platform, and next steps are clear in terms of enabling VO interoperability so that VO data may be accessed in the Science Platform, with VO data software in the repository and using the Data Lake as potential back-end to VO services and also for staging of VO data.



ESCAPE ESAP - ESFRI Science Analysis Platform

ESAP is the ESFRI Science Analysis Platform: 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. Preparing other services such as data products, and tools for integration with ESAP and their subsequent use have also been part of the team activity within ESCAPE and across EOSC.

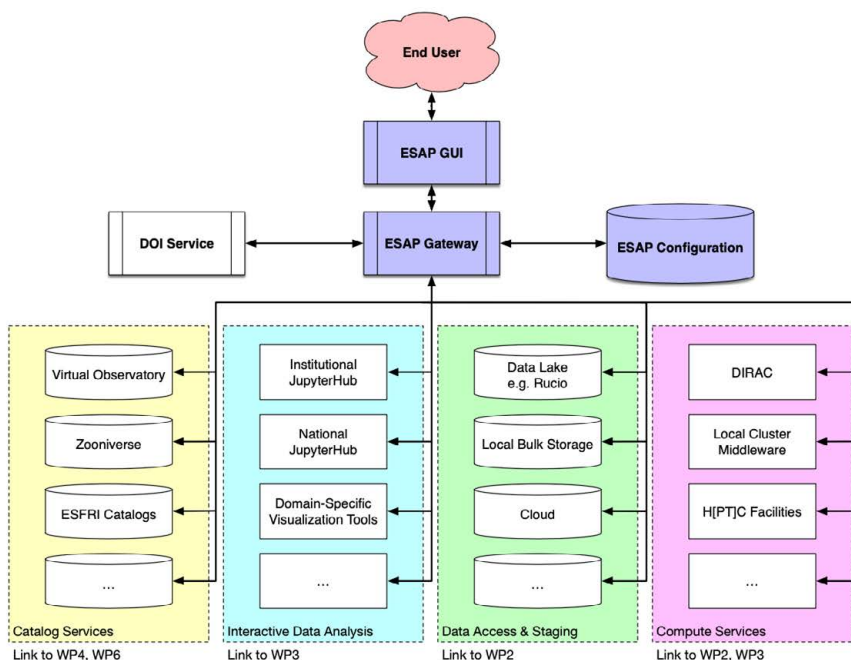


Figure 9: ESAP provides a single, consistent, interface and point of access to a variety of services drawn from a range of providers. Links to the other work packages ESCAPE are indicated.

The ESAP vision

ESAP is a science platform toolkit: an integrated set of software components which ESFRIs, ESCAPE project partners, and other groups can use to rapidly assemble and deploy platforms that are customized to the needs of their particular user communities and which integrate their existing service portfolios. These various deployed instances of ESAP then provide the key interfaces between the services delivered by the ESCAPE project and the wider scientific community.

In order to meet this goal, ESAP has been designed to be flexible and adaptable to the particular needs of each user community. This is achieved by abstracting the details of heterogeneous underlying infrastructures away from users, a task that is achieved by ESAP's modular, plugin-driven architecture: an

ESAP instance is integrated with its surrounding environment by enabling and configuring an appropriate selection of plugins, and new capabilities are easily added by writing new plugins.

While this plugin based system makes ESAP infinitely re-configurable, its basic functionality includes tools for discovering and accessing data, services, and software from ESCAPE project repositories; access to a range of computing and analysis services; and orchestration of data, services, and software to help users create and access research environments that meet particular needs. This model is shown schematically in Figure 9, which illustrates the range of services that ESAP can help the user access. The flexibility of this approach means that instances of ESAP can be deployed at a variety of scales, from providing services to just a few users within a small project, to supporting major pieces of infrastructure.

Achievements to date

Platform capabilities

The core ESAP system – the user interface, the API Gateway, and a selection of plugins connecting it to key services – are now mature and well tested. This includes:

- 🔗 **Core data management capabilities:** As the user accumulates and manipulates datasets, they accumulate results and carry them with them through the system. Subsequent analysis tasks augment this data collection. Ultimately, it can be persisted for future references and, if appropriate, published for wider use;
- 🔗 **Data archive interfaces:** The user is able to use ESAP to search for and discover data in a wide range of archives. While this certainly includes archives which are built upon open standards – and, in particular, VO interfaces as described below – it is possible to incorporate bespoke and non-standard data sources by writing appropriate plugins. For example, plugins are currently available that provide access to data collections from Apertif²⁸ and Zooniverse²⁹, amongst others. Note that the ESAP interface adapts appropriately depending on the data collection being queried, and even makes it possible to search multiple semantically-related archives simultaneously;
- 🔗 **Interactive data analysis facilities:** The user can spawn interactive analysis jobs based on Jupyter³⁰ running at a range of different computing facilities. These jobs can integrate with the other ESCAPE work packages, as described below;
- 🔗 **Data Lake integration (DIOS):** ESAP provides direct access to query the Data Lake, discovering data and working with it directly within the core ESAP interface. Further, the Data Lake as a Service system provides direct integration of the Data Lake with the interactive analysis environments available through ESAP;
- 🔗 **Software Repository integration (OSSR):** The ESCAPE OSSR collects software from across the various ESFRIs associated with the project. Deep integration between ESAP and the OSSR makes it possible for ESAP users to discover this software and dispatch it – together with their chosen datasets – directly to selected analysis environments;
- 🔗 **Virtual Observatory protocol support (CEVO):** ESAP provides pervasive support for VO standards, developed in conjunction with ESCAPE WP4. Users can locate and access data for processing via ESAP using VO systems, and even leverage the IVOA SAMP standard to have ESAP communicate with applications running on their personal computer;
- 🔗 **Managed database support:** ESAP's (prototype) managed database support makes it possible to fetch

28 <https://alta.astron.nl>

29 <https://www.zooniverse.org> 3 <https://jupyter.org>

30 <https://jupyter.org>

results from multiple remote catalogues into a local database that can be used for advanced analytical functionality like cross-matching diverse catalogues.

Software packaging and support

The ESAP team have packaged and integrated a wide variety of scientific analysis software, covering a variety of the ESCAPE ESFRIs, in forms that are designed for re-use and ultimately for integration with the OSSR. Space constraints make it impossible to list all of the various initiatives in this document, but illustrative examples include:

- 🔗 Publication of solutions to the SKAO Science Data Challenge to the OSSR;
- 🔗 Packaging and containerization of the R3BRoot and CmbRoot³¹ tools;
- 🔗 A reproduction package for the scientific analysis of Hickson Compact Group 16 described in Jones et al. (2021)³².

Infrastructure deployment

Deployment and provisioning of infrastructure for the long term is not within the scope of ESAP team. However, it does support a number of software installations which have provided limited services to end users while acting as proofs-of-concept for the ESAP system. Notably, these include deployments of the ESAP core at ASTRON and SKAO as well as JupyterHub³³ and/or BinderHub³⁴ systems deployed at IAA-CSIC, JIVE, FAIR, and RuG. These systems have been instrumental both in integrating ESAP itself and in testing scientific workflows within the associated user communities.

Future goals

This section considers future work in the ESAP context. We consider first the natural technical evolution of the codebase – areas in which technical work beyond that scoped for ESCAPE could have substantial impact. Then we discuss the wider impacts that future support for ESAP could have on the EOSC and ESFRI community.

Technical development

The current ESAP system, as described in “Achievements” part, provides a robust and reasonably fully-featured environment. However, there are many opportunities for further enhancements that cannot be completed within the scope of ESCAPE. This include:

- 🔗 Improved support for provenance Tracking through the ESAP data management system, including robust minting of Persistent Identifiers (PIDs) for published results;
- 🔗 Support for sharing data between user soft help at form, enabling collaborative workflows;
- 🔗 Support for persistent development environments, in which users are able to store the state of their session and return to it in future;

31 <https://www.r3broot.gsi.de> and <https://redmine.cbm.gsi.de/projects/cbmroot>

32 <https://ui.adsabs.harvard.edu/abs/2019A%26A.632A.78J/abstract>

33 <https://jupyter.org/hub>

34 <https://binderhub.readthedocs.io/>

- Richer understanding of the links between data products, enabling features such as presenting science products together with the calibration data used to generate them;
- Federation between distributed ESAP instances.

In addition, ESAP remains endlessly adaptable to integrate with new external service offerings. Many opportunities for these arise: from the management of OpenStack³⁵-based virtual machines to integration with workflow- or function-as-a-service systems.

Wider prospects

This section addresses the potential wider impacts of ESAP: how can continued development of the system continue to further the promise of EOSC and ESCAPE?

Development of an ESCAPE/EOSC Virtual Research Environment

The ESAP vision, as described previously, provides the basis for a true Virtual Research Environment: an online collaborative system that provides all the tools and services that scientists required to conduct high-impact research. The flexible and scalable design of ESAP means that it can evolve into a role in which it provides seamless access to all of the various software and services which are developed by ESCAPE, by its successors, and across the whole EOSC ecosystem. However, assuming such a role requires two major ongoing activities.

Firstly, ESAP will require continued maintenance and integration effort. The platform is stable, solid and extensible, but – as with any network connected service – emergent issues will need to be resolved as they arrive. Further, while the plugin-driven ESAP system is extensible to address a wide range of use cases, development effort will be required to integrate new service plugins and to extend the ESAP core to support new service types when necessary.

Secondly, the ESCAPE project provides no long-term structural support for delivery of ESAP as an operational service. Within the context of the project, various ESCAPE partners have deployed infrastructure in support of ESAP development or in support of particular ESFRI use cases. However, these are inappropriate for providing high-availability services to a more general community: that would require not only dedicated infrastructure, but also support and maintenance services.

Sustaining ESAP in support of ESFRI development

Several of the ESCAPE-affiliated ESFRIs are currently at a stage in their development where they are making strategic choices about the long-term delivery of services to end users. For example, the Cherenkov Telescope Array Observatory (CTAO) is currently evaluating how to provide user-facing services in the context of its up-coming Science Data Challenge, while SKAO is beginning prototype developments for its network of Science Regional Centres, which will be observatory's primary means of delivering data to end users. Both of these infrastructures have been heavily involved in ESCAPE in general and in the development of ESAP in particular; it would be natural for ESAP technologies to play a major role in their future plans.

No individual ESFRI will find it advantageous to assume responsibility for a legacy software system, even if that system provides useful functionality. The value of ESAP to these projects is therefore dramatically increased if it is not simply a collection of source code, but an actively maintained and supported project with a clear governance model and a sustainable future.

Common standards for science platform development and interoperability

The last several years have seen an explosion of heterogeneous development in the field of “science platforms” (broadly defined). Tens or hundreds of projects generally providing some form of interactive analysis

35 <https://www.openstack.org>

environment with access to bulk storage and computing systems are being rolled out across a multitude of research domains³⁶. This ecosystem is disjointed and fragmented: although many of these projects are individually very technically advanced and provide compelling functionality, they are specialized and difficult to apply in other domains.

The ESAP experience suggests a way to move beyond this fragmentation of platforms into disconnected systems serving individual communities or infrastructures. Instead, ESAP proposes a model of common standards and interconnectedness among science platform efforts. By using ESAP, individual projects or communities which need particular capabilities can build on a common, interoperable technological basis, customizing and extending it to address just their particular needs.

Moving beyond that: as more platforms become centred around common technical standards, we can move to adopt common standards for access not only to data – as pioneered by the Virtual Observatory in the ESCAPE context – but to compute and other resources. This process has already started: members of the ESCAPE team have engaged with our colleagues in the IVOA to begin exploring concepts for the *Execution-Planner* system³⁷, which would standardize access to computing capabilities in much the same way as IVOA recommendations harmonize access to data.

Ultimately, by building upon the experience developed in ESAP and CEVO, one can imagine working towards a future federated network of science platforms, build on common standards, and providing the lowest possible barrier to entry to new service or data providers making their offerings available to a wide range of researchers.

36 For example, consider CANFAR Skaha, Rubin Science Platform, ESA Datalabs, NOIRLab Astro Data Lab, SciServer, and many others.

37 <https://github.com/ivoa/ExecutionPlannerNote>



ESCAPE ECO - Engagement, Communication and Citizen Science

The activities of Engagement and Communication (ECO), are divided into two parts. One is focussed on our outward-facing and inward-facing communication. The other one focuses on using crowdsourced data mining to engage external communities directly with ESCAPE scientific discovery, and is a two-way interaction, as shown in Figure 10 below. Here we discuss some of the lessons learned and best practices developed for our engagement and communication, particularly in the context of crowdsourced data mining through citizen science.

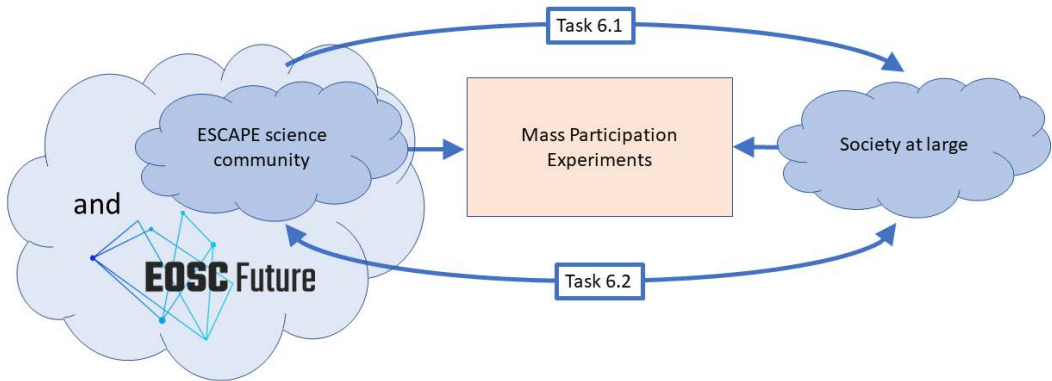


Figure 10: ECO work package tasks, adapted and updated from the figure in the Description of Work.

The ECO Vision

Our aim is intended to connect the ESCAPE science community to society at large, and vice versa. The first aspect is the conventional broadcasting mode, for execution of the public engagement and communication strategy, based around project results, serving all the work packages. This is the conventional route of blog articles, social media, web pages, news articles and so on. The other aspect however is more two-way. Our vision is to involve society at large with open science much more directly, which we achieve through mass participation experiments, i.e. citizen science. Our vision is to improve access to data and tools through citizen science crowdsourcing experiments for most of the facilities in the ESCAPE remit. Furthermore, now that ESCAPE is contributing to the EOSC-Future project to prototype a working EOSC, our vision is to extend the remit of these two way-activities to the wider EOSC-Future science areas and facilities.

Crucial to this vision is our definition of citizen science as scientifically-driven crowdsourced data mining and data collection, which happens to involve non-specialist volunteers. In particular, note what citizen science is not: outreach. Citizen science can of course help with outreach, but that is not what it is for. It is perhaps more appropriate to think of crowdsourced data mining as the application of a biological

computer. As with any other scientific tool or facility, such as a beamline or a spectrometer, there are science questions that are particularly well suited to the tool.

A central vision of the EOSC is to make scientific data FAIR. Implicit in this vision is that the FAIR data should also be useful, but this is far from being guaranteed, especially given the inter-disciplinary and multi-disciplinary remit of EOSC.

Even within the remit of ESCAPE's astrophysics / (astro)particle physics EOSC cell, there are many examples of conflicting data interpretations by users within and outside particular sub-disciplines. For example, [Daylan et al. 2016](#) (Physics of the Dark Universe, Volume 12, p. 1-23) reanalysed public sky survey data from the Fermi gamma-ray telescope, and interpreted a gamma-ray excess towards the Galactic centre as a signature of dark matter particle annihilation. This would be the natural location for such a signal, given the expected density peak in the dark matter distribution. However, the instrument team themselves responded to this claim ([Ackermann et al. 2017](#), Astrophysical Journal, Volume 840, Issue 1, article id. 43) by re-interpreting the claimed signal as an instrumental artefact. Without taking any view on the merits or otherwise of either side of this debate, it is clear that the usefulness of FAIR data will be limited by the supporting supplementary contextual information. This goes beyond the simplest expressions of metadata, and may require specialist training. The further that the data are from a user's subject specialism, the more curated the interaction has to be with that data. The most extreme example of this is citizen science.

Citizen science has the advantage of involving a much larger and more diverse scientific user community with the EOSC, and indeed the citizen community is one of the strategic priorities of the Strategic Research and Innovation Agenda³⁸ of the EOSC. As an example, our Galaxy Zoo Clump Scout project had a science team of just three academics but a community of nearly fourteen thousand volunteers, contributing nearly two million classifications. Citizen science is clearly the most readily accessible route to increasing the size of the EOSC user base by several orders of magnitude. In the Clump Scout project, as with all our citizen science projects, contextual educational and training resources are embedded into the volunteer workflows. This allows non-specialist volunteers to gain enough subject specialist knowledge for more comprehensive explorations of the data, and indeed on many such projects there are explicit links to external professional tools for this deeper engagement. Our aspiration is that ESCAPE Virtual Observatory tools will be common for deeper engagement with astrophysics citizen science (see below). There is also evidence for volunteers acquiring new scientific terminology that was not provided in training material, i.e. there is evidence that the activity has stimulated independent study (e.g. [Luczak-Roesch et al. 2014](#), International AAAI Conference on Web and Social Media; [Oesterlund et al. 2017](#), Academy of Management Annual Meeting Proceedings). For these reasons, we represent the non-specialist interaction with EOSC as shown schematically in Figure 11, with the science communities leading the way for non-specialists, and in the limiting case (but typically only in the limiting case), the non-specialists achieving an engagement with EOSC equivalent to the experts.

38 <https://eosc.eu/sria>

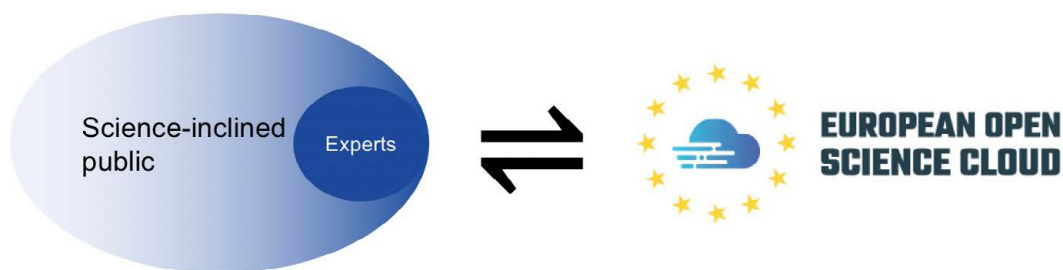


Figure 11: Schematic representation of experts leading the way for the non-specialist interactions with the EOSC.

For our more conventional purely outward-facing communication activities, we have a very wide range of activities and communication channels, including posters, flyers, roll-up banners, videos, social media accounts (Twitter, LinkedIn, YouTube), position papers, give-aways and press releases. We created a responsive and dynamic website, which evolved a great deal during the progress of the project. This site now contains a wide range of resources on ESCAPE, its Test Science Projects, its services, its facilities and more. We are in the process of updating this website to present the services in a more “work package agnostic” way, to suit the needs of likely website users beyond the funded period of ESCAPE. We created dedicated resources regarding artificial intelligence, partly because our citizen science crowdsourced data mining is closely linked to the classification and data mining opportunities from machine learning. These were promoted in a series of short animations commissioned by ESCAPE, the Sixty Second Adventures in Artificial Intelligence. The virtuous circle between machine learning and human classifications will be discussed in more detail below. We are in close communication with EOSC Future about communicating the results.

Achievements to date

Communication activities

ESCAPE has exceeded the engagement key performance indicators (KPIs) shown in Table 2 below, in many cases by a wide margin. The few KPIs still not yet met are nevertheless on course for being met by the end of the project.

KPI by end project (Month 48)	Progress by Month 40
6 200 Website visits/users	26 100
39 000 page views	91 600
125 registered users	160
700 tweets	926
420 twitter followers	657
85 LinkedIn followers	295
20 videos on YouTube	35
10 newsletters	11
37 Contributions to external events	76
3 Posters	4
3 Flyers	2

KPI by end project (Month 48)	Progress by Month 40
2 Rollup banners	2
20 Videos	35
3 Position Papers	2
4 Give-aways	3
4 Press releases	1

Table 2: Communications key performance indicators

Beyond the KPIs themselves, we have healthy statistics on community engagement, such as having 494 thousand impressions on Twitter, 1800 downloads of three joint EOSC Position Statement with Cluster Projects, and over 4.2 million citizen science classifications to date. We have also had active presence at events with other EOSC Clusters and ESFRIs, such as: the SSHOC events on EOSC collaborations and Research Data Alliance (October 2019) to discuss commonalities and collaborative solutions for community research data services; the SSHOC event on Realising the EOSC (November 2020) on teaching communities how to use the new tools and techniques developed to improve data FAIRness and prepare data and services to be aggregated into the EOSC Portal.

Galaxy Zoo: Clump Scout

Galaxy Zoo: Clump Scout is a new citizen science project that will help discover how galaxies form and evolve.

One of the main goals for modern observational cosmology is to discover and understand how galaxies and their constituent substructures have assembled and evolved throughout cosmic history. The diverse observed morphologies of individual galaxies are not only indicative of their current composition, but also encode a detailed record of their assembly histories, their past and ongoing star formation, and their interaction with local environments.

Galaxies grow by forming stars. Today, the Hubble Space Telescope can detect distinct star-forming structures inside the galaxies that populated the Universe when it was less than a quarter of its current age. These early galaxies look very different to their modern-day counterparts. Their disks are thick, turbulent and violent environments, where hundreds of new stars are born every year. Many also exhibit giant regions of enhanced star formation that appear as bright clumps in telescope images. In contrast, today's star forming galaxies are typically much more placid. Their disks are thin and well-ordered and clumpy star formation is much less common.

These profound differences raise obvious questions. Which physical mechanisms drove the observed evolution in star formation activity? Why are giant star forming clumps so much more common in the early Universe?

To understand why clumpy galaxies became so rare, we need to find and investigate as many examples as possible. One potential approach involves training modern deep learning algorithms that use deep learning to identify galaxies with clumps. However, appropriately labelled training data for clump detection is scarce and laborious to generate. Moreover, automatic algorithms struggle to operate effectively if their limited training datasets underrepresent the diversity of the data being analysed. In contrast, human beings working in collaboration can extrapolate successfully from a handful of examples.

To benefit from this impressive human capability, we used the Zooniverse platform to develop a new citizen science project called *Galaxy Zoo: Clump Scout*. The project invites the general public to examine

images of galaxies obtained by the Sloan Digital Sky Survey (SDSS) and annotate all the clumps they can see. By participating in *Galaxy Zoo: Clump Scout*, volunteer clicks will identify the locations of clumps within thousands of galaxies in the nearby Universe. The project uses a novel Bayesian aggregation algorithm that dynamically derives a consensus for the clump locations based on the annotations provided by multiple volunteers for the same image. The algorithm also estimates the reliability of the dynamic consensus, which helps to ensure completeness while avoiding spurious clump detections. *Galaxy Zoo: Clump Scout* represents one of the first large-scale studies of clumps in local galaxies.

In the future, new space telescopes like *Euclid* will image more than a billion galaxies. Using citizen science to manually check so many galaxies for clumps would take many years, even for the most dedicated *Clump Scout* volunteers. The speed of computer algorithms will be required to process such large volumes of data and we have adapted the *TensorFlow* Faster-RCNN implementation to detect clumps using 5-channel imaging SDSS data with promising results. However, there will always be galaxy images that confuse the computer algorithms and we'll need the help of the *Clump Scout* volunteers to step in when deep learning fails. Even more importantly, human beings inspecting images are much better at spotting any unusual or unexpected phenomena that single-minded algorithms would just ignore. Indeed, the history of citizen science is full of examples when keen-eyed volunteers make amazing, serendipitous discoveries. Projects like *Clump Scout* will help to maintain this tradition in the future.

The *Galaxy Zoo: Clump Scout* project has now concluded after collecting 1738822 classifications from 13762 volunteer citizen scientists who collectively annotated the locations of visible giant star-forming clumps in 85286 low-redshift galaxies. The results of the project have been analysed and will form the basis of journal publications that are currently in preparation. The science team published two blog articles (see <https://blog.galaxyzoo.org/2019/09/18/introducing-galaxy-zoo-clump-scout-a-new-citizen-science-project/> and <https://blog.galaxyzoo.org/2020/04/11/galaxy-zoo-clump-scout-a-first-look-at-results/>) which were shared on the ESCAPE website and promoted via the ESCAPE social media channels. Results have now been published in [Adams et al. 2022](#) (*Astrophysical Journal*, Volume 931, Issue 1, id.16).

Radio Galaxy Zoo: LOFAR

The *Low Frequency Array* (LOFAR) is a large interferometric array of radio telescopes located primarily in the Netherlands, but with outlying antennae dispersed across Europe. LOFAR is also a recognised science and technology pathfinder facility for the next-generation radio telescope, the SKAO.

Radio Galaxy Zoo: LOFAR (<https://www.zooniverse.org/projects/chrisrmp/radio-galaxy-zoo-lofar>) is a new citizen science project led by ASTRON in the Netherlands with substantial ESCAPE-funded support provided by the Zooniverse platform and the Open University (OU) The project invites volunteers to classify radio images extracted from the first data release of the *LOFAR Two-metre Sky Survey* (LoTSS) which covers 424 square degrees in the region of the HETDEX Spring Field. In this release, 325,694 individual radio sources were detected with a signal five times greater than a typical background noise fluctuation.

Classification entails attribution of distinct regions of radio emission to a single origin and (where possible) identifying an optical counterpart for the radio emission's source. By Zooniverse standards this is a very complicated analysis task, which requires consideration of multiple images, representing radio and optical data. Moreover, the degree of scientific comprehension that volunteers require to successfully provide the required classifications is more than typical Zooniverse projects, which often rely on somewhat mechanical "microtasks" that can be performed without complete understanding.

To render such complex classifications tractable for citizen scientists, the ESCAPE and Zooniverse teams have developed an advanced volunteer training and feedback system. The project uses a tutorial video paired with a dedicated training workflow that allows volunteers to mimic the classification process as demonstrated by one of the LOFAR project scientists. The training workflow presents subjects in the

same order as they appear in the video (unlike the normal random ordering employed by the Zooniverse platform) and volunteers receive real-time feedback in response to the annotations they provide. This is the most advanced training infrastructure that has been deployed using the Zooniverse project builder platform and the upgrades that have been developed by the OU and Zooniverse with ESCAPE support will be available for future citizen science projects to leverage. It has been shown that volunteers' confidence is a critical factor in citizen science projects, which improves classification accuracy and volunteer retention. The Radio Galaxy Zoo: LOFAR project is still running on the Zooniverse web platform. So far, 6839 citizen scientists have provided 171,961 classifications of 18,197 separate radio images to help associate relativistic radio jets with their host galaxies. The science team published two blog articles (<https://blog.galaxyzoo.org/2020/03/20/radio-galaxy-zoo-lofar-the-first-classification-results/> and <https://blog.galaxyzoo.org/2020/08/07/radio-galaxy-zoo-lofar-a-short-update/>) and one video (<https://daily.zooniverse.org/2020/05/20/around-the-zoo-radio-galaxy-zoo-lofar/>) which were shared on the ESCAPE website and promoted via the ESCAPE social media channels.

Links to other ESCAPE EOOSC Services

We have been working closely with Work Package 5 on the ESCAPE Science Analysis Platform (ESAP), specifically in integrating ESAP with the Zooniverse citizen science platform. ESAP users are currently able to query the Zooniverse back-end database, load the results into their ESAP Shopping Basket, and then send them to analysis services for further processing. Note that this activity is not included in the Description of Work, but was carried out with a view to facilitating a better integration of the ESCAPE services including citizen science. This ESAP Zooniverse access to crowdsourced data mining products includes the cross-cluster domains SSHOC, EOOSC-Life, ENVRI-FAIR, PANOSC, and ESCAPE, with over 100 active projects and around 2 million volunteers.

This is part of a wider model of cross-service integration across ESCAPE. Figure 12 shows this schematically. We start with the professional scientists working already on the ESCAPE science analysis platform, using data from the data lake. On that platform they are working with the data from their citizen science experiments and ultimately managing the citizen science projects themselves throughout the project lifecycle (discussed below). The volunteers themselves then have the ability to link out of the citizen science projects into the professional tools such as the Escape virtual observatory cells, armed with the new knowledge that they have acquired as part of the citizen science project that gives them the context of the scientific data. In summary, we have the science platform managing the data and the citizen scientists generating the data, and meanwhile the professional scientists would like of course to accelerate the classifications of their data with the help of machine learning. Therefore, the intention is that a professional scientist would work within a science platform, training machine learning algorithms based on the truth sets from the citizen science data. Having done that, one can use the machines to classify the most straightforward and unambiguous items to classify. Then the project will be able to refocus the human effort on the difficult edge cases that are the most sensible use of human effort. Therefore, one achieves a virtuous circle between human and machine classification.

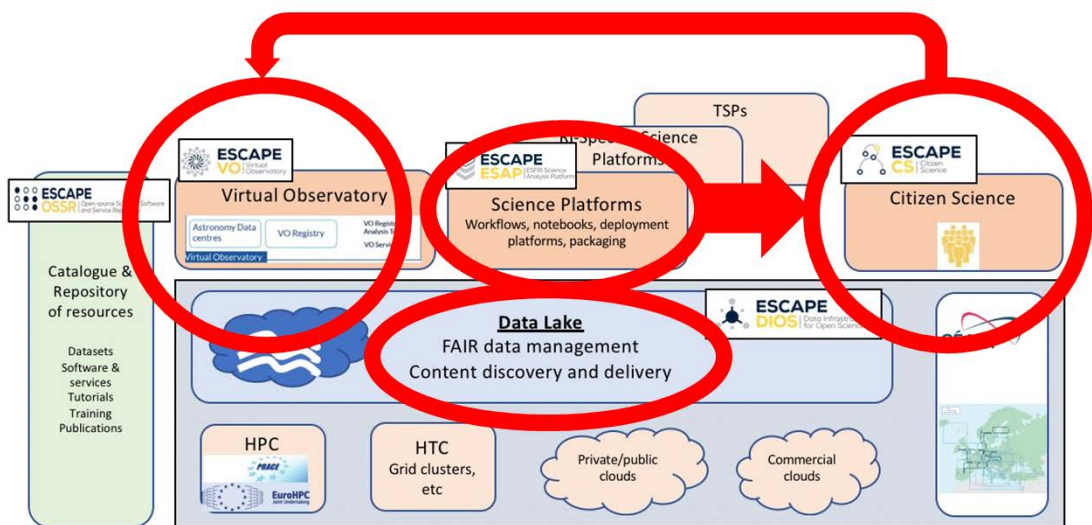


Figure 12: Schematic model of the connections between ESCAPE services, in their relations to citizen science. Adapted from a diagram by G. Lamanna and I. Bird.

Engagement with EOSC-Future

As part of our involvement in EOSC-Future, we extended the remit of our milestone Second ESCAPE Citizen Science Workshop to research domains beyond ESCAPE. As a direct result, we now have crowdsourced data mining projects at an advanced stage of development in the domains of the SSHOC and ENVRI-FAIR research infrastructure clusters. One is the project titled AIK, African Indigenous Knowledge, which is aimed at the preservation and visualisation of African indigenous knowledge for resilient food systems, tackling problems of food supply and food security in the developing world. The second project aims to characterise knitting patterns in domestic magazines from the early 20th century, which are a rich source of primary evidentiary material for social scientists. Both projects are at an early stage of development, and neither are yet ready for public dissemination, but both are on course for public release by the end of the funded period of the ESCAPE project.

Future Goals

Our broader objective is to better integrate the crowdsourced data analysis (citizen science) into the wider EOSC ecosystem, and this will be a significant focus of our future activities, including engagement with the EOSC Association Task Forces to facilitate the development of citizen science functionality and to support the needs of the wider science-inclined public in its engagement with EOSC.

The current ESAP data discovery portal provides functionality to identify Zooniverse classification data and mark it for subsequent download and processing by interactive or batch analyses launched by the ESAP. However, ESAP does not currently provide tools that enable straightforward generation and management of citizen science projects on the Zooniverse platform. Many of the experimental data generated by the ESCAPE partners and accessed via the ESCAPE data lake are complex, large in volume, and present significant analysis challenges that, despite being difficult to automate, could form the basis of exciting and engaging citizen science projects. Exposing data via the Zooniverse platform would not only enable valuable scientific analyses that are difficult to automate, but would also promote ESCAPE and ESCAPE science to the general public.

To minimise any perceived barriers to entry for researchers wishing to launch citizen science projects, we will develop a suite of easy-to-use tools that can be used to automatically create, initialise and manage new projects on the Zooniverse platform. Our tools will streamline interaction with the Zooniverse back-end (Panoptes) and advanced aggregation and retirement engine (Caesar) via their respective REST APIs. We will also provide generic template workflows that can be straightforwardly adapted to integrate Zooniverse projects with automated analysis pipelines and Deep Learning models. In particular, we will provide a template active learning workflow that can be used to optimize the generation of Zooniverse volunteer labelled-training data to rapidly train or retrain Deep Learning algorithms.

Conversion of experimental data into high-level “subjects” (images, videos etc.) that can be successfully interpreted and analysed by non-expert volunteers often requires substantial effort from researchers who manage citizen science projects. To streamline this process and reduce the workload of running a successful citizen science project, we will also provide tools and template workflows that generate attractive subject data, upload them to the Zooniverse servers and manage them effectively as the volunteer classifications accumulate.

Although we will focus on the data types and analyses typical of the ESCAPE partner experiments, we anticipate that the diversity of use cases we encounter and facilitate will be sufficient to make the tools and workflows we develop readily applicable to numerous other experimental datasets that will be generated as part of the EOSC-Future initiative. We intend to create:

- 🔗 Notebook and documentary materials demonstrating web-interface based and programmatic (scriptable) Zooniverse project management including project and workflow creation, subject creation and upload, adding training and feedback to subjects;
- 🔗 Notebook and documentary materials demonstrating integration with the Zooniverse's Caesar engine for advanced aggregation and efficient subject retirement;
- 🔗 Notebook demonstrating how to integrate Zooniverse projects with existing machine learning frameworks and combine volunteer classifications with machine learning predictions;
- 🔗 Notebook and documentary materials demonstrating how to set up an active learning framework to continuously train machine learning models using volunteer classifications of optimally selected subjects.

Ultimately, our vision is for the community to have access clear exemplars of planning, creating and managing crowdsourced data mining on EOSC, implementing machine learning in real time, across a wide range of scientific domains, which they can use as templates to deploy with ease. Through this, our vision is to increase the size of the community making real scientific engagement with EOSC by orders of magnitude, solving the difficult problem of usefulness of FAIR data by giving non-specialists a carefully curated and educationally supportive experience of EOSC.



ESCAPE to the FUTURE

Current networking within the Science Clusters should bring to future engagement in new common projects within the next Horizon Europe framework. Enhanced coordination with the EOSC Association is expected. It is not a project closure but the start-up of a second phase of ESCAPE.

Join the Community

 projectescape.eu

 projectescape.eu/contact-us

 twitter.com/ESCAPE_EU

 linkedin.com/company/projectescape/

 youtube.com/c/ESCAPEEU



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 Grant Agreement no. 824064.