POSITION PAPERS BY THE ESFRI CLUSTER PROJECTS

ON EXPECTATIONS OF PLANNED CONTRIBUTIONS TO EOSC

CONTENTS

Greetings 1

Executive Summary 1

Summary of Positions 2

Position Paper ESCAPE

Position Paper PaNOSC

Position Paper ENVRI – FAIR

Position Paper EOSC-Life

Greeting

Dear members of the Executive Board, dear reader,

the EOSC initiative already has a high impact on structured access to Open Data and with the current effluent of implementation projects the interdisciplinary exploitation of data can only improve. The future EOSC is shaped along action lines that are detailed in close collaboration between the EOSC governance boards and ESFRI.
Please find below four position papers on EOSC of the ESFRI cluster projects. A fifth paper by the SSHOC project is pending.

The following persons have been in contact with the EOSCsecretariat.eu regarding the papers:
Andy Gotz for the PaNOSC project
Andreas Petzold and **Ari Asmi** for the ENVRI-FAIR project
Niklas Blomberg for the EOSC-Life project
Giovanni Lamanna for the ESCAPE project
Ron Dekker for the SSHOC project

Executive Summary

As part of its engagement activities, EOSCsecretariat.eu is undertaking pooling of stakeholder views on the implementation of EOSC. This document presents a compilation of four position papers of ESFRI projects submitted to EOSCsecretariat.eu to that effect. Since the beginning of the year 2019, 5 ESFRI cluster projects have been launched to link to the European Open Science Cloud (EOSC). The 5 ESFRI Cluster projects aim together to implement interfaces, to integrate computer and data management solutions, to create cross-border and open cooperation spaces and to promote clouds via the EOSC portal for a larger user community.

As stakeholders of EOSC these projects were invited to contribute to the development and implementation process. These position papers are the tools for the projects to contribute directly to systems development of EOSC and identify ways for future synergies. While EOSC can leverage their domain expertise, their views can better drive innovations within EOSC towards the research community's needs.

The organization of the papers is aimed at highlighting:

- Expectations of the research community

- The added value of EOSC to the existing research data infrastructure

- Issues to be addressed by the EOSC executive board

- Commitments to EOSC

The overall expectation of these projects is that EOSC will enable the accessibility and re-use of research data, increase scientific value of research data, and deliver an interoperable environment of data infrastructures. The projects expect EOSC will bring the added values of the infrastructure for sustainable use of research data and a virtual research environment enabling real-time collaboration between researchers using FAIR data.

A summary of the common viewpoints of the projects (recommendations) is presented as follows.

- A clear and concise communication of the concept of EOSC is required.
- A common standard for FAIR data needs to be defined to create a common understanding across communities.
- A research community-centered bottom-up approach needs to be prioritized.
- EOSC needs to aspire to offer a continuous, trusted working environment and networking opportunities to the research community.
- EOSC needs to provide a long-term open data archive with high performance storage and computing services, to enable sustainable use of data beyond the life span of individual data infrastructures.
- The EOSC infrastructure needs to provide Cross-Europe AAI, high performance storage, computing, archiving, simulation and analysis services, in order to build A Minimum Viable Ecosystem. This will make the idea EOSC adoptable by (encourage participation) the research community.
- Beyond the provision of open data, EOSC needs to aspire to create a virtual research space for open science, where scientists can create content and collaborate.
- Collaborate with publishers to make open data a publication in its own right.
- EOSC is required to show a cost-benefit analysis to participating research communities.
- A documentation of guidelines on how to join and use EOSC services.
- The business model of the Federating core must be positioned at the national level, so that governments have the interest to encourage open science and define the national policies that can support it.

Summary of Positions

Further detailed requirements (issues) with our assumptions of target working groups, which can possibly need to address the issues are presented as follows.

Project / WG	ESCAPE	EOSC-Life	ENVRI-FAIR	PaNOSC	SSHOC*
Architecture	-EOSC needs to be organized as a federation existing resources. - Needs to include open source repository of analysis, computing and storage services. - Requires a federated repository.	-A user management and access services with e-infrastructure is required. - Open collaborative model to bring existing national cloud infrastructure together.	- Expose thematic data services and tools from the research infrastructure catalogues to the EOSC catalogue of services, COPERNICUS, GEO, and other end-users.	-EOSC should operate and sustain AAI, supporting AAI features implemented by PaNOSC and partners. -A distributed service for downloading and transferring data. -A high performance, long term, storage and computing resources for open data. -Cross-domain federated search capability. -Services for data simulation and analysis.	-AAI across Europe. -Personalized virtual workspace for researchers. -A platform service to integrate the SSHOC market place.
FAIR	-Implemented the IVOA standards to enable the use of data produced by compliant tools and services by the astronomy research community. -There is a need to scale up & plans to integrate the VO framework with EOSC.	-Consistent adoption of FAIRification and provenance services.	-Increase the potential for innovation of each research infrastructure by establishing a specific ENVRI-FAIR section in the EOSC service catalogue.	-Working on the full FAIR compliance of the PaNdata policy framework. - Working on extending metadata standards to enable wider usability of data. - Training scientists on FAIR data practices. -A common standard for FAIR data is required.	- Interoperability and FAIR guidelines.
Sustainability	-A cooperative framework for open data resulting in a dedicated infrastructure.	-EOSC Life working on a framework for open science policies.	-Increase the potential for innovation of each research infrastructure by establishing a specific	-A virtual open science environment, where content can be created can make EOSC more	-Helpdesk and training resources.

	- A repository for the sustainability of research data beyond the lifespan of projects.		ENVRI-FAIR section in the EOSC service catalogue.	attractive to the scientific community and sustain its use in the future. Long term financing plans for EOSC	
Landscaping	- A prototype federated data infrastructure covering astronomy and particle physics to be integrated with EOSC. - A platform service for data analytics will be implemented and offered as part of the EOSC catalogue.	-EOSC Life wants to creates new standards through the demonstrator programs e.g. the standardization of chemosensitivity screening.	-Establish cohesion with the global research infrastructure landscape, including research infrastructure clusters and regional/international initiatives in the environmental sector; maintain ENVRI community knowledge with particular consideration of developing integrated activities.		
Rules of participation	-An acknowledgement mechanism (certificate) for the EOSC collaborators.	-The thorough monitoring of all data and data services must be available in a manageable framework.	-Policies and standards with more extensive European policy (e.g. ISO 19115 INSPIRE) as well as with relevant international developments are in progress to align.		-Workflows to support participation.

*) Because the delivery of the SSHOC position paper is pending , the information on SSHOC was taken from the "EOSC Federating Core, Community Position Paper v1.0" which is currently open for public comments. See: https://www.eoscsecretariat.eu/eosc-liaison-platform/post/eosc-federating-core-updated-proposals-and-first-draft-community-position

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures

ESCAPE POSITION STATEMENT

TABLE OF CONTENTS

PREAMBLE **3**

WHAT IS THE RESEARCH COMMUNITY EXPECTING FROM EOSC? **4**

WHAT ADDED VALUE EOSC WILL BRING TO THE RESEARCH? **5**

ESCAPE Data Infrastructure for Open Science 6

ESCAPE Open-Source Software and Service Repository 7

ESCAPE connecting ESFRI projects through VO framework 8

ESCAPE ESFRI Science Analysis Platform 9

ESCAPE Engagement and Communication 10

WHICH MAIN ISSUES/KEY MESSAGES SHOULD BE CONSIDERED BY BOTH THE EOSC EXECUTIVE AND GOVERNING BOARDS? **11**

European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE) aims to address the Open Science challenges shared by ESFRI facilities (CTA, ELT, EST, FAIR, HL-LHC, KM3Net, SKA) as well as other pan-European research infrastructures (CERN, ESO, JIVE, EGO) in astronomy and particle physics.

PREAMBLE

This document summarizes the current views and expectations of astronomy and particle physics partners in ESCAPE about EOSC. It is written keeping in mind the various exchanges that are taking place with the EOSC secretariat, governance and working groups. It is also written to follow up the EC chaired EOSC stakeholders' concertation events aiming at gathering the scientific views on the implementation of EOSC.

The ESCAPE Executive Board addresses mainly three questions:

1. What is the research community expecting from EOSC?
2. What added value will EOSC bring to the research?
3. Which main issues/key messages should be considered by both the EOSC executive and governing boards?

WHAT IS THE RESEARCH COMMUNITY EXPECTING FROM EOSC?

1. From the ESCAPE perspective, EOSC will federate existing resources across national data centres, e-infrastructures, and research infrastructures, allowing researchers (and citizens) to access and re-use data produced by the ESFRI projects in Astronomy and Particle/Nuclear Physics for a multi-probe approach to understand the Universe; accelerating the discoveries and increasing scientific value by sharing data and by transferring knowledge within scientific communities.

2. The important volumes of data produced and managed by several of the ESFRI projects that are partners in ESCAPE imply investment in IC technology developments and related increase of costs for services, computing and storage resources to access their data. Therefore, ESCAPE partners aim to contribute to the provision of data services for the benefit of EOSC, while they expect that EOSC provides a sustainable environment in which data and data infrastructures are made interoperable and re-usable. Such a sustainability implies the establishment of a cooperative framework for open data, where research infrastructures, national and pan-European e-infrastructures co-develop and support together a dedicated infrastructure for data-research for the scientific-community-based needs in terms of an EOSC repository of open-source software for analysis, services, computing and storage resources.

3. The research-community-based foundation approach to build a global virtual research environment within EOSC for data interoperability and analytics is a way for ESCAPE partners to commit to building the EOSC and to contributing to the EOSC catalogues and portal. This would imply a combination of data produced from the facilities (namely the ESFRI research infrastructures encompassed within ESCAPE) and other data generated and shared by the concerned community. Therefore, a sharing of responsibilities for quality-certified scientific data is envisaged among the operators of the facilities and researchers in the longer term. ESCAPE partners would expect that EOSC defines and/or supports credit systems, acknowledgement, standard licences, and certification for the results of all researchers engaged in co-operative work that allow all to reap the benefits of open science. EOSC should be able to include and make sustainable the infrastructures and platforms that the community will deploy for such purposes.

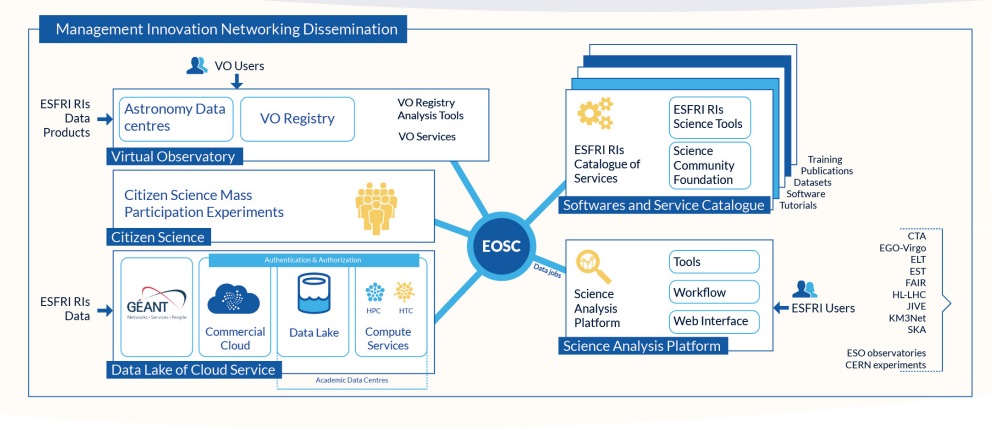
4. EOSC is expected to establish a reference federated digital repository for the sustainability of scientific data preservation independently of the lifetime of each individual research infrastructure. Such a repository would leave the management of the responsibility for the respective data and the financing of their share of the repository with the originators so that a commensurate open access plan can be supported.

WHAT ADDED VALUE
EOSC WILL BRING TO THE RESEARCH?

The following picture summarizes the ESCAPE Work Packages (WP) activities and expected delivered components of what one could call “a thematic cell” of the global EOSC. Such a “cell” in Astronomy and Particle/Nuclear Physics will enable EOSC to adopt transversally some services and e-infrastructures that will be useful also in support of other disciplinary “cells”.

The added value that EOSC will bring to our community, through the ESCAPE commitments, is the formal opportunity to build upon the “cell” a “virtual research environment”. There, researchers can upload their analyses in a notebook-like, reproducible and shareable style. Through this they will co-develop software and add mining to data, as well as running and improving workflows, in a real-time collaboration.

The following sections describe the main ESCAPE contributions to ESOC through its work programme



ESCAPE Data Infrastructure for Open Science

ESCAPE will design, implement, and operate a prototype data lake – a federated data infrastructure that will form the basis of an open access data service and science analysis environment for the ESFRI projects within the ESCAPE cluster covering astronomy and particle physics. It will propose such a solution as a key component of a future EOSC framework. The data lake concept enables the large, reliable, national research data centres to work together to build a robust cloud-like service to curate and serve data of CTA, FAIR, EGO-Virgo, HL-LHC, JIVE, KM3NeT and SKA at all scales up to the multi-Exabyte needs of such projects. These data centres (CERN, CNRS-CCIN2P3, DESY, GSI, IFAE-PIC, INFN, CNRS-LAPP-MUST, Nikhef, SURFSara, UG), partners in ESCAPE, have experience built up over a decade in particle physics within WLCG and in supporting major astronomy and astroparticle physics precursors pan-European research infrastructures (such as AMS, ANTARES, HESS, LOFAR, MAGIC, etc.).

Key outputs: *A mechanism required for large research infrastructures such as HL-LHC and SKA that in future will manage multi-Exabyte data sets, that will need to be served to global user communities in a scalable and performing way. A federated storage infrastructure implementing the FAIR data management principles at the base level, and form the basis for higher-level data preservation and access services delivered in the other work packages.*

Integration in EOSC. The Data Lake development leverages collaboration and integration of work and results from previous and ongoing frameworks (e.g. EOSC-hub). It will build on and integrate existing work from a variety of areas – the Research Infrastructures, previous EU projects, as well as using the current state of the art solutions in the appropriate areas and collaborating with ongoing work from GEANT, PRACE, and other proposed H2020 projects specifically addressing the European Open Science Cloud. The Data Lake will consist of an ecosystem of tools and services integrated into a reference implementation, while still providing, to the different science projects in ESCAPE, the flexibility to decide which ones to use.

ESCAPE Open-source scientific Software and Service Repository

ESCAPE deals with software services for open science data-analysis of the ESFRI facilities. The development of multi-messenger data analysis practices promotes activities for innovative methods, to maximise software re-use and co-development, to identify open standards for software release, to investigate data mining tools and new analysis technique. ESCAPE supports an open environment to guarantee cross-fertilisation and to develop community-specific data services that will be exposed under the EOSC catalogue of services under the FAIR principles.

Key outputs: *A sustainable open-access repository to share scientific software, digital libraries for data analysis, data-sets, data products including related user-support documentation, tutorials and training activities, which will be dynamically enhanced and maintained by ESCAPE ESFRI projects and included in the EOSC catalogue.*

Integration in EOSC. ESCAPE will make the software and service developed by the ESFRIs available to the scientific community via an open-source scientific software and service repository, which will be fully integrated into the EOSC environment. The services from the EOSC catalogue will be used where possible and thin-layer interfaces generated where needed. Central regulations compliant with the EOSC global ones will ensure the software and service quality via audits. Training activities and a help desk will provide support to the ESFRIs to devise regulations and adhere to them. At the same time, an EOSC-based approach allows for new innovative models of data exploitation, including data mining and deep learning techniques, making these accessible beyond the current expert communities. A competence group for innovations will be formed to steer these developments.

ESCAPE Connecting ESFRI projects to EOSC through VO framework

ESCAPE plans to: a) make the seamless connection of ESFRI and other astronomy and astroparticle research infrastructures to the EOSC through the Virtual Observatory (VO), b) refine and further pursue implementation of FAIR principles for astronomy data, and c) establish stewardship practices for adding value to the scientific content of ESFRI data archives. With the VO, Astronomy has built an operational interoperability infrastructure that has proven to be a great success for many aspects of astronomy data interoperability. The VO is an essential component of the astronomy data landscape, as has been strongly stressed in the ASTRONET Infrastructure Roadmap since its first publication in 2008. International astronomy data providers, in particular ground- and space-based telescopes, publish their data using the IVOA standards, and compliant scientific tools and services enable discovery, access and use of the data by the whole astronomy research community. Integrating the VO in EOSC implies the need to scale its framework to the biggest data sets that will be produced by the ESFRI and other projects.

Key outputs: *Assess and implement the connection of the ESFRI and other astronomy and astroparticle RIs to the EOSC through the Virtual Observatory framework, actively contributing to the setting up of the EOSC services and the provision of trusted data adhering to FAIR principles.*

Integration in EOSC. By leading the connection to the EOSC with the ESFRI facilities we will set the path for a new era of cross-disciplinary interoperability, and connections to the necessary computing resources. EOSC will facilitate the next step for the VO framework to realise its potential to scale to the biggest data sets that will be produced in particular by the ESFRI projects, and will enable use of VO data in scientific analysis platforms. ESCAPE aims to map the VO framework to the EOSC so that the VO enabled archive services from ESFRI will be interoperable. The integration of the VO registry and other standard maps will be key for discovery and reuse, access, deposition and sharing of data, as well as for data management curation and preservation.

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

Key outputs: *A platform-service for data analysis into EOSC and tailored to the requirements and the user needs of each of the ESFRI and other RI member of ESCAPE. It will be part of the EOSC catalogue.*

Integration in EOSC. Once data for analysis has been located and staged, and workflows have been defined, either by accessing the EOSC software repository or by the user directly, the next step is to deploy those workflows on the underlying processing infrastructure. For many of the involved ESFRIs and RIs, the data scales involved require significant computational resources (storage and compute) to support additional processing and analysis. The EOSC-ESFRI science platform therefore must interface to an underlying HPC or HTC infrastructure. Consequently, it is important to make efficient use of the full performance potential of the HPC centres, e.g. by optimizing the access to file systems from the data processing layer and by ensuring the portability of science applications with container solutions. As with the data, this infrastructure is likely to be large, widely distributed geographically, and definitely heterogeneous. Deploying user-initiated processing and analysis tasks on this HPC infrastructure while simultaneously maintaining interactivity and responsiveness in the analysis system will be a challenge and requires a mixture of dynamic resource allocation and optimization. ESCAPE will draw upon the "data-lake" design and implementation and will build upon existing EOSC-hub activities such as Federated High Throughput Computing, Scientific Workflow Management and Orchestration, and EOSC-hub AAI.

ESCAPE Engagement & Communication

ESCAPE develops and manages a programme of crowdsourced data mining via Citizen Science mass participation experiments, with additional goals of public engagement and education in parallel.

Key outputs: *Involve citizens directly in knowledge discovery with ESCAPE and the ESFRI facilities, improving transparency of the scientific process. A harmonised suite of Citizen Science mass participation experiments and online video material, also deploying machine learning to accelerate volunteer classifications.*

Integration in EOSC. Central to ESCAPE's implementation of the Open Science Cloud vision is innovation for the society at large, by improving access to all ESCAPE results and through ESCAPE to the EOSC more widely. The ESCAPE EOSC "cell" prototype will allow researchers to stage existing data collections for analysis including citizen science.

WHICH MAIN ISSUES OR KEY MESSAGES SHOULD BE CONSIDERED BY BOTH THE EOSC EXECUTIVE & GOVERNING BOARDS?

ESCAPE as a large thematic cluster has gathered the formal commitment of ESFRI, RIs and pan-European research organizations in building a prototype of this cell that should be seamlessly integrated with the EOSC core services. The consequence of this is an increasing attention and involvement of scientists from our community to be part of the endeavor. Our success and our results depend on the sustainability and the acknowledgement of the cluster actions. The EOSC executive and governing boards aim at exploring and establishing the most pertinent and effective format for the EOSC implementation. However, the bottom-up approach where the researchers are central in the implementation choices for the effectiveness of open-data-research through EOSC is fundamental and must be pursued. A cross-disciplinary approach is critical to guaranteeing that EOSC will be inclusive. The organization, orchestration and accessibility of underpinning heterogeneous e-infrastructures for data archive, access and analysis are important structural EOSC issues.

However, EOSC should be more than that and should not be limited to a marketplace provision, but be a continuous evolving working space where researchers should routinely find out their daily dashboard for trusted data research and continuous networking opportunities within a large scientific community. ESCAPE partners commit to that and naturally advocate the cluster as a sustainable ecosystem as well as a successful approach for effectiveness and inclusiveness. The combination of partners in the project naturally leverages economies in supporting a joint repository and favors longevity of the approach. In addition, a subsidiarity principle could be applied towards the scientific communities; in order to leverage the work of ESCAPE and other clusters, the EU is invited to facilitate Member States to guarantee core and sustained funding for the cluster role in EOSC and to set an example that may be implemented in other research contexts and/or new focus.

ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

projectescape.eu

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

PaNOSC position paper on the EOSC

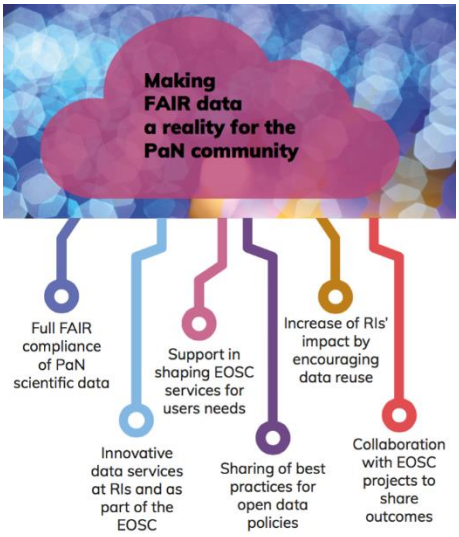
endorsed by the PaNOSC Executive and Management Board

Version 1.0 on the 22 November 2019

Introduction

PaNOSC is an INFRAEOSC-04 cluster project of 6 European Photon and Neutron sources on the ESFRI roadmap [1]. The science done at the research institutes represents a huge variety of scientific disciplines using photon and neutron sources to study almost any kind of material on a wide range of length scales – from angstroms to microns – and time scales. The ESFRI and national sources (an additional 15 RIs involved in the ExPaNDS project [2]) represent a large user community of roughly 30 000 scientists annually who use the Photon and Neutron sources in Europe for their research projects. The science carried out at these institutes addresses at least five out of the seven societal challenges defined by the EC [3] which range from climate change, renewable and efficient energy sources, to drug discovery.

PaNOSC Objectives



- *Participate in the construction of the EOSC by linking with the e-infrastructures and other ESFRI clusters.* ***Concretely*** PaNOSC is working with EGI and GÉANT, participating in the Architecture Working Group and EOSC meetings. PaNOSC is developing a shared vision with the other INFRAEOSC-04 cluster projects to identify common areas of collaboration.

- *Make scientific data produced at Europe's major Photon and Neutron sources fully compatible with the FAIR principles*. ***Concretely*** PaNOSC is updating the PaNdata data policy framework to be FAIR compliant and implementing Data Management Plans (DMPs).

- *Generalise the adoption of open data policies, standard metadata and data stewardship from 15 photon and neutron RIs and physics institutes across Europe.* ***Concretely*** PaNOSC partners are updating/adopting their data policies.

- *Provide innovative data and simulation services to the users of these facilities locally and the scientific community at large via the European Open Science Cloud (EOSC).* ***Concretely*** PaNOSC is making it possible to run bespoke services for the PaN community from Jupyter or through remote desktops and providing Jupyter and remote desktops as a service as part of their local infrastructure close to the data and on the e-infrastructures.

- *Increase the impact of RIs by ensuring data from user experiments can be used beyond the initial scope.* ***Concretely*** PaNOSC partners are working on providing richer metadata by extending the photon and neutron community metadata standard Nexus [4] and adopting electronic logbooks for capturing user experiments. The goal is to enable open data to be used

by the scientific community at large. PaNOSC will also provide training for scientists to understand FAIR and adopt FAIR practices for their data.

- *Provide training in the use of the involved facilities, services developed in PaNOSC, and data stewardship as an important tool for dissemination.* ***Concretely*** PaNOSC partners are extending the scope of the e-learning platform e-neutrons.org to cover the domains represented by all the PaNOSC partners (i.e. including the photon sources), and developing courses and training material for that platform. PaNOSC will hold workshops and a summer school.

- *Share the outcomes with the national RIs who are observers in the PaNOSC project, the community at large and the EOSC, to promote the adoption of FAIR data principles and data stewardship.* ***Concretely*** PaNOSC is collaborating closely with the ExPaNDS project which is implementing FAIR data for the national photon and neutron RIs. PaNOSC is providing training and engaging with users at User Meetings to explain and train.

EOSC Minimum Viable Ecosystem

PaNOSC sees EOSC as an opportunity to generalise the adoption of FAIR data practices at the 6 Photon and Neutron research facility partner institutes and eventually at all photon and neutron sources. Adopting FAIR data will enable data sharing across a wider community and the provisioning of services for remote data analysis. In order for these objectives to be realised the **EOSC must provide** the following services:

1. A common way of identifying, authenticating, and authorising users (**AAI**) across Europe. The EOSC should operate and sustain AAI as part of the EOSC infrastructure. The EOSC AAI should support the AAI features PaNOSC is implementing on the UmbrellaId AAI [6].

2. A service for **transferring and downloading data efficiently** (distributed and high bandwidth);

3. A solution for **long-term archiving of large quantities of open data** (petabytes) coupled to high-performance storage and compute resources for the (re)analysis of open data;

4. A **federated search capability** for searching and finding scientific data in a wide variety of domains;

5. A set of **services for data simulation and analysis** ranging from generic services like Jupyter notebooks to domain specific applications per scientific application in the PaN software catalogue [5]. These data services could be remote from the data source if the data are easy to move but should be available close to the data if the data are difficult to move.

The above services are considered as the **Minimum Viable Ecosystem** for the EOSC from the PaNOSC point of view. Only if the EOSC provided these services it would encourage and help many medium and small institutes, as well as individual scientific groups to make FAIR data available, to re-use FAIR data for new scientific findings, but as well to make appropriate use of the EOSC.

Once the implementation of FAIR data is standard practice, then it would be desirable for the EOSC to be extended to be more than a source of FAIR data. The EOSC could become the **GitHub of Open Science in Europe**. This means making it a platform for scientists to share their data analysis and workflows and link these back to open data and other workflows – either their own, or that of other scientists. To achieve this, it will be necessary to provide scientists with a personal space where they can create content (data analysis

recipes, workflows, publications), store analysed data and share their work with collaborators via a versioning system like git.

PaNOSC provides

The PaNOSC Research Institutes will be an essential part of the EOSC as sources of data and providers of data services.

The PaNOSC partners aim to provide:

1. **Petabytes** of raw and processed data in a wide variety of scientific domains
2. **Meta-data** that will create **FAIR** raw and processed scientific data
3. **Software** for generic and specific data simulation and data analysis
4. **Workflows and expertise** for reducing and analysing data
5. **Reference training material** and **training platform** for understanding photon and neutron science and associated handling of data
6. Liaison with large **user communities** of photon and neutron sources and their expectations for services

A summary of the PaNOSC work packages can be found in the publication [6] and on the website [4].

Feedback to the EOSC Executive Board

PaNOSC is a bottom-up approach to making data FAIR and making FAIR data sustainable for the Photon and Neutron community to help users do better science, be more efficient with the help of better data management and to make science more reproducible. The EOSC Executive Board can play an important role in this by bringing in a top-down approach in the following areas:

- provide a clear concise answer to the question “what is the EOSC?”. This should be incorporated in the architecture being developed by the EOSC working group dedicated to this
- define common standards for FAIR data so that the different scientific fields have a common approach and understanding, e.g., FAIRsFAIR could provide clear guidelines with examples on how to implement FAIR by different communities
- provide long-term sustainable plan for how the EOSC will be maintained and financed
- provide cloud resources for running data analysis workflows and simulations, ideally unlimited but at least enough to make a significant difference for users needing access to computing resources beyond what can be offered by the PaNOSC partners
- collaborate with publishers to generalise the requirement for citing data in publications and making open data a publication in its own right
- provide documentation and training material on how to join and use the EOSC
- do a cost-benefit analysis of what the EOSC provides to the Photon and Neutron communities and comparing the benefits with the cost thereof.

Example data

Some of the PaNOSC RIs have databases of data collected over the last decades that are currently under-exploited e.g. paleontology data in the http://paleo.esrf.fr is an example of processed data which are not widely known or exploited yet. These data are ideal for cross-disciplinary applications and linking up with data from museums and other scientific disciplines. PaNOSC will provide raw and processed data with metadata of a far higher quality. In turn, this will allow opportunities for useful data mining to take place, to the benefit of the community and its wider stakeholders. The EOSC will offer an excellent opportunity to make such data more widely known and used by different communities.

Picture 1: Synchrotron tomographic data (from the ESRF) used for the study of the Egyptian crocodile mummy 90001591 to establish the cause of death, its age at death, and its diet, demonstrating that it was a wild animal hunted to make a mummy. Data from original publication [7] can be downloaded from [8].

Other examples of open data are the ILL and ESRF data portals, respectively https://data.ill.fr and https://data.esrf.fr.

References

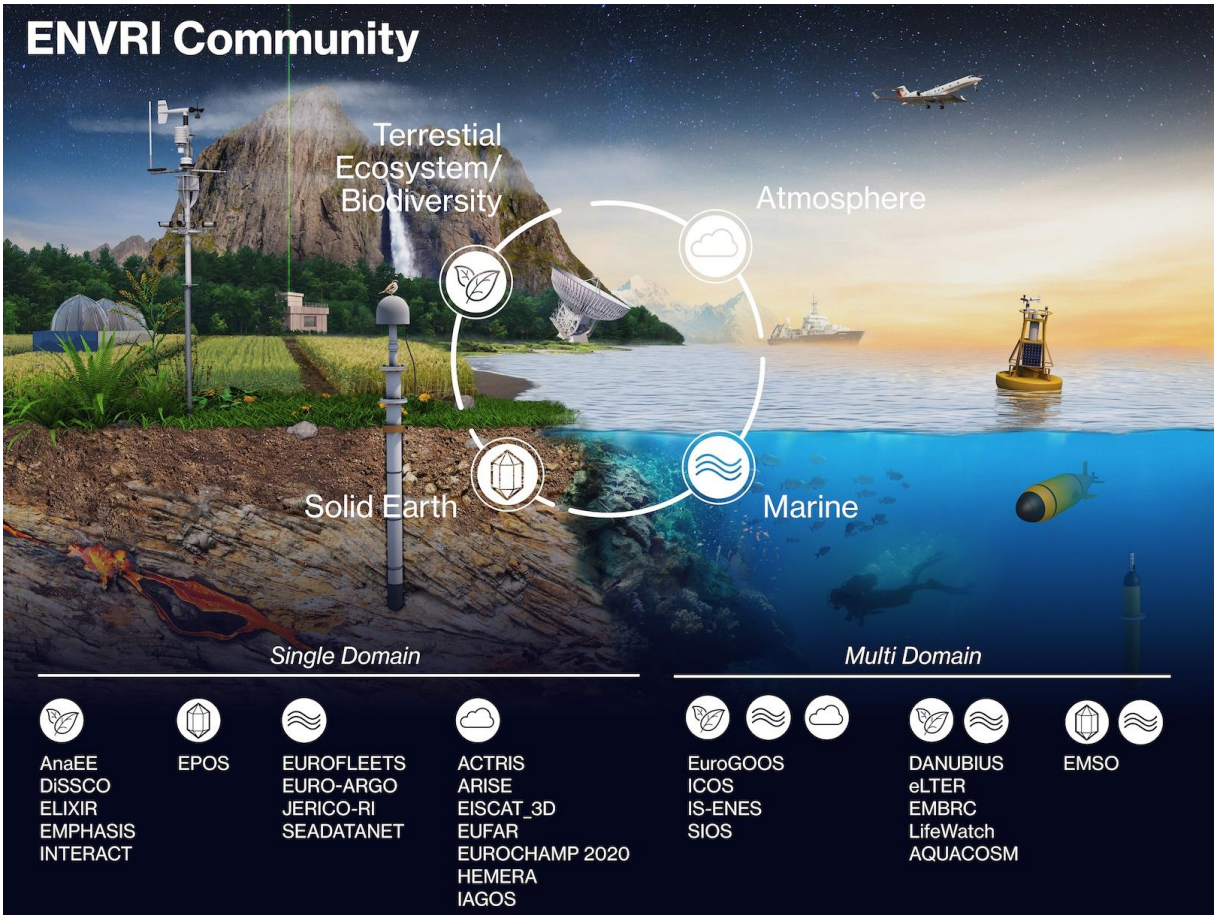
[1] ExPaNDS project – https://expands.eu
[2] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges
[3] Nexus metadata format - https://www.nexusformat.org/
[4] PaNOSC website – https://panosc.eu
[5] PaN software catalogue - https://software.pan-data.eu/
[6] "**Enabling Open Science for Photon and Neutron sources**" by A. Götz et. al., *ICALEPCS 2019 Pre-Proceedings*, http://icalepcs2019.vrws.de/papers/tubpl02.pdf
[7] Porcier S. M., Berruyer C., Pasqali S., Ikram S., Berthet D., Tafforeau P. « **Wild crocodiles hunted to make mummies in Roman Egypt: Evidence from synchrotron imaging** ». Journal of Archaeological Science, 1 October 2019. Vol. 110, p. 105009. DOI : https://doi.org/10.1016/j.jas.2019.105009
[8] http://paleo.esrf.fr/index.php?/category/2846

ENVRI-FAIR Position Paper on the EOSC

prepared by Andreas Petzold (a.petzold@fz-juelich.de) and Ari Asmi (ari.asmi@helsinki.fi)
endorsed by the ENVRI-FAIR Executive Board on 03 December 2019
Version 2.0 from 17 December 2019

ENVRI-FAIR Cluster Project Mission Statement and Objectives

European Environmental Research Infrastructures on the ESFRI level provide data, research products and services from key areas of the Earth system. These research infrastructures (RI) form the ENVRI Cluster and include the principal producers and providers of environmental research data and research services in Europe from the four segments of the Earth system - Atmosphere, Marine, Solid Earth, and Biodiversity/Terrestrial Ecosystems. The data, products and services provided by the ENVRI Cluster are crucial European contributions to the integrated global observation system monitoring the state of the Earth system and climate. They are vital for assessing past and defining future policies, as well as for the development of environment-friendly innovations and adaptation as well as mitigation strategies. The ENVRI Cluster represents the core component of the European environmental research infrastructure landscape with the ENVRI community as their common forum for collaboration and co-creation.



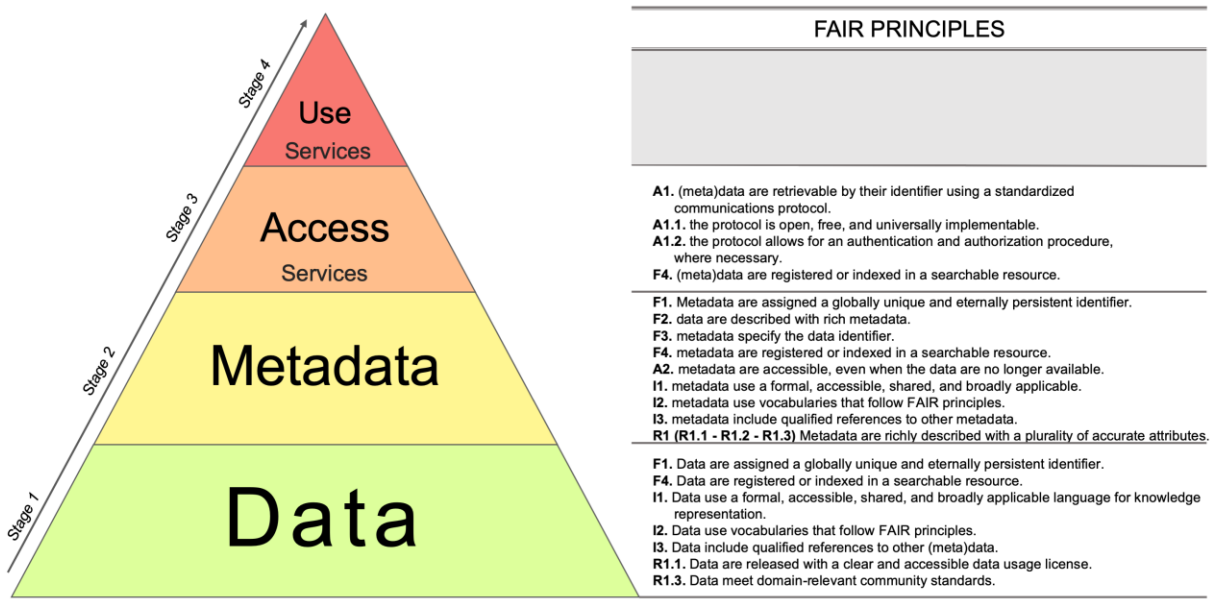
ENVRI-FAIR (www.envri-fair.eu) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824068

ENVRI-FAIR is the cluster project aiming towards the provision of services according to FAIR principles by the ENVRI community. It develops tools and resources for easy and seamless access to environmental data and services provided by ENVRI research infrastructures, with the high-impact ambition to prepare the foundations for the successful implementation of a federated machine-to-machine interface – the ENVRI-hub – to access environmental data and services provided by the contributing ENVRIs. The highest priority is on the provision of high-quality data using open licenses, standard mechanisms and protocols. The overall architecture is designed as strong integration at sub-domain level with a layer of integrated services on top which may serve as the nucleus for the ENVRI-hub. The hub forms the interface to the EOSC and will be realized as the services across ENVRIs and even between environmental subdomains become progressively more integrated [1].

ENVRI-FAIR Objectives are to

1. further develop common standards, protocols and policies for the data life cycle, including cataloguing, curation, provenance and service provision within the ENVRI Cluster, with specific consideration of the FAIR principles including interoperability, and of the tools and methods created during the preceding EU-projects ENVRI and ENVRIplus;
2. align these policies and standards with more extensive European policies (e.g. ISO 19115 INSPIRE) as well as with relevant international developments;
3. develop and implement the necessary tools for reaching Objective 1 in each research infrastructure, thereby adopting an open approach for sharing data and software;
4. improve the skills of research infrastructure personnel to develop and sustain knowledge on Research Data Management and FAIRness, including both cross-cutting and subdomain-specific knowledge, and on the FAIR infrastructures resulting from Objectives 1 and 2 through an extensive training;
5. increase the potential for innovation of each research infrastructure by establishing a specific ENVRI-FAIR section in the EOSC service catalogue, with the aim of stimulating common pre-commercial procurement processes and dissemination of outcomes and thus enhancing the uptake of research infrastructure services by private partners;
6. establish cohesion with the global research infrastructure landscape, including research infrastructure clusters and regional/international initiatives in the environmental sector; maintain ENVRI community knowledge with particular consideration of developing integrated activities;
7. expose thematic data services and tools from the research infrastructure catalogues to the EOSC catalogue of services, COPERNICUS, GEO, and other end-users.

The implementation strategy of ENVRI-FAIR follows the FAIRness maturity pyramid developed by EPOS [2]. The FAIR principles discussed by Wilkinson et al., 2016 [3] are analysed with respect to FAIRness of data, metadata and services, and are broken down into ascendant stages towards increasing maturity. These stages are reflected in the FAIRness implementation plans of the participating research infrastructures, taking into account the maturity of each research infrastructure. The maturity of the participating research infrastructures is assessed periodically in an approach inspired by the FAIRification process adopted by GO FAIR [4] and is applied in accordance with specific needs of the ENVRI cluster and those of the other ESFRI domains.



The FAIRness Maturity Pyramid [2].

ENVRI-FAIR Cluster Project and its relation to the EOSC

What is the ENVRI community expecting from the EOSC?

ENVRI-FAIR supports the view of EOSC and its services as a public good. We also consider that long-term and sustained funding is required to ensure the EOSC continues to exist and serve its users. This funding should also reflect the resources requested by the supported communities, with necessary periodic updates and related development initiatives.

To our understanding, the Federating Core with its components Compliance Framework, Hub Portfolio, and Shared Resources, will serve as the power house of the EOSC, whereas the majority of curation of research data remains the responsibility of the contributing research infrastructures.

We also consider that the Federating Core and Shared Resources can form an essential part of the ENVRI Research Infrastructures operation models, if the sustainability and access issues can be solved.

The ENVRIs play an important and multi-facetted role for EOSC as both providers of data and services of all kinds (e.g., data services, research products services) and as first users of services provided by EOSC. The relationship between RI funding and EOSC-derived funding for the services and resources provided for the RIs need to be detailed, together with the precise specification of services and resources provided for each facility.

What are the added values of EOSC to our community?

The primary goal of ENVRI-FAIR is the implementation and further development of data and research services at RI and subdomain (atmosphere, marine, etc.) levels while ensuring the highest possible level of standardisation at the whole environmental domain (cluster) level. ENVRI-FAIR implies motivation to share solutions and strategies, moving towards a shared approach to EOSC. Key added value to the RIs can be the wider findability and accessibility of RI data and research services for the greater public and thus beyond the ENVRI's "traditional" scientific communities. Additionally, the EOSC can offer the basis for the provision of platforms for scientist-developed virtual research environments, more extensive use of shared workflows (including their publication), and access of less resource-rich country researchers to these facilities.

Fundamental infrastructure components and metadata services (AAI, PID, provenance, workflow management, etc.) need to be integrated at sub-domain and RI levels, but also across the entire cluster. EOSC may provide generic solutions that can be tailored to specific RI needs and then adopted by the single RI.

The provision of resources such as repositories, high-performance computing (HPC) and high-throughput computing (HTC) resources and data management resources may foster the FAIRification process of the involved ENVRIs and particularly of those RIs at an early stage of their life cycle.

To fulfil these expectations, the provision of EOSC resources needs to be sustainable. Otherwise, the deactivation of services implemented at the RI level may pose a high risk on the RIs which have adopted EOSC services, because in this case their operational status will be threatened.

Key message for EOSC Boards - both executive and governing boards:

- ENVRIs can provide the EOSC with their amassed collective domain-specific knowledge and competencies that underlie all the data and other services provided.
- Making and sustaining data FAIR as well as sustaining maintenance of the respective infrastructures require resources and expertise which are available at the ENVRI scientific communities.
- Ensuring the coherence of methodologies and technologies for the FAIRification process across subdomains and ensuring the sustained functionality of provided services are essential; these processes can be strongly supported by EOSC resources.
- Sustainable and long-term funding of the observations and the production and provision of high-quality open and FAIR data and services are crucial.
- It is essentially needed that the financial model of the Federating Core must be positioned at a national level where governments have a political interest in encouraging open research and the means to define national policies that can support it.

ENVRI-FAIR Requirements Table [5]

What ENVRI-FAIR needs from the EOSC
Generic infrastructure services such as for AAI, PID, and provenance, for tailoring to specific Research Infrastructure needs and adoption by individual research infrastructures
Generic workflow management tools and services, for tailoring to specific Research Infrastructure needs and adoption by individual RIs
Access to shared resources such as repositories, HPC, HTC and data management tools
Standard APIs to support remote data discovery, access, and sharing
Provision of notebook-based environments which allow to access and integrate data services for the community
What ENVRI-FAIR can offer to EOSC
Collective domain-specific knowledge and competencies that underlie all the data and other services provided by the European ENVRIs
FAIR-based tools and resources for easy and seamless access to environmental data and services provided by the European ENVRIs
ENVRI-hub – a federated machine-to-machine interface to access environmental data and services provided by the contributing ENVRIs

Relevant Links and References

[1] Petzold, A., Asmi, A., Vermeulen, A., Pappalardo, G., Bailo, D., Schaap, D., Glaves, H. M., Bundke, U., and Zhao, Z.: ENVRI-FAIR - Interoperable environmental FAIR data and services for society, innovation and research (Version Camera ready), Proc. IEEE International Conference on eScience 2019, 1-4, DOI: http://doi.org/10.1109/eScience.2019.00038, 2019.

Accessible at https://zenodo.org/record/3462816#.XfjYJHsxlaQ

[2] Bailo, Daniele. (2019, July 10). Four-stages FAIR Roadmap - FAIR "Pyramid" DOI: http://doi.org/10.5281/zenodo.3299353

[3] Wilkinson, M. D., et al.: The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data, 3, 160018, DOI: 10.1038/sdata.2016.18, 2016.

[4] GO FAIR https://www.go-fair.org

[5] EOSC Federating Core Community Position Paper

https://www.eoscsecretariat.eu/eosc-liaison-platform/post/eosc-federating-core-updated-proposals-and-first-draft-community-position

http://tiny.cc/FedCorePPv1

EOSC-Life

EOSC-Life: Building an open, collaborative space for digital life science in Europe

DISCUSSION PAPER ON EXPECTATIONS AND PLANNED CONTRIBUTIONS TO EOSC

Horizon 2020
H2020-INFRAEOSC-2018-2
Grant agreement no. 824087
Type of action: RIA

1. The EOSC-Life project

EOSC-Life brings together the 13 Life Science ESFRI research infrastructures (LS RI) to a collaborative digital space for biological and medical research in Europe based on open science and FAIR principles.

The project will construct an EOSC space for life sciences researchers by publishing FAIR data resources ready for (re)use in the cloud, work with scientific software developers in our community to make the large ecosystem of tools and workflows available in EOSC and implement the policies and guidelines necessary for secure handling of sensitive data and preserving the trust of study participants. EOSC-Life will also consolidate existing user authorisation and authentication services into a unified LifeScienceLogin connected to underlying e-Infrastructure services.

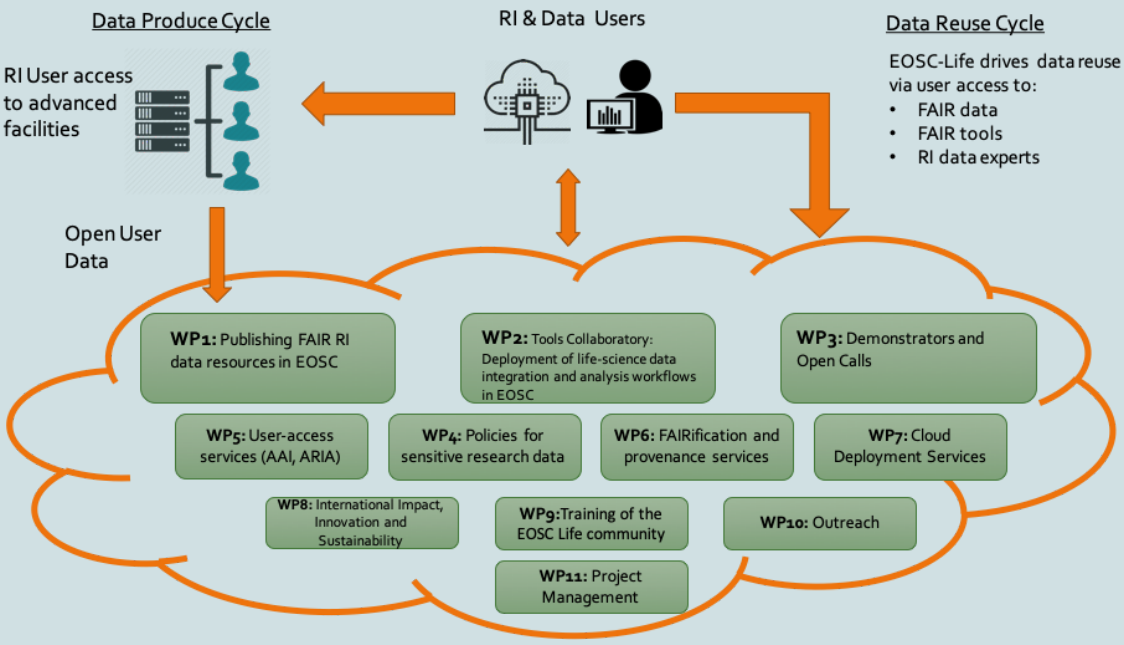
Figure 1. *EOSC-Life will give European scientists access to advanced data resources, analysis platforms, samples and support services throughout the European Research Area in full compliance with all ethical, regulatory and legal requirements. We will achieve this in a user-driven project where open calls for user research will allow our large user community to adopt advanced data management practice and access data integration and large-scale analysis tools in the cloud.*

Through open calls for user projects, EOSC-Life will seek partnerships and user projects that will help drive connection to our users' science. In these partnerships ('demonstrators') we will provide cloud resources and connect these projects with growing cadre of data experts in the 13 LS RI. The EOSC-Life project is designed to support a transformation of life science to better use digital resources and has an ambitious training programme to support the development of organisations, staff and users from the participating RIs. Thus, by the end of this project EOSC-Life will be established as the new norm for digital biology in Europe – accessible by Europe's 500,000 life scientists.

EOSC-Life's work plan drives FAIR data publishing and data reuse by research projects and scientific tool developers

EOSC-Life brings together the 13 Life Science RIs on the ESFRI Roadmap (LS RI). The project consortium comprises all the legal entities of the established LS RIs and include a set of national centres that provide access to leading international scientific service platforms - in total 69 partners and linked third parties in 14 countries. Collectively, the LS RI have national nodes in 23 European countries and EOSC-Life work through the established outreach and technical coordination functions within the RI.

The work plan is designed around 11 work packages where WP1-3 are outward-facing and drive participation of data resources, tool developers and applied user projects in the EOSC. These WPs are supported by policies (WP4), user management and access services (WP5), FAIR and data management services (WP6) and clouds (WP7). EOSC integration, stakeholder management, training and outreach is provided by WP7-10 and supported by project management (WP11).



WP1: Publishing FAIR RI data resources in the EOSC

WP1 drives the development of BMS RIs' data handling and integration capabilities and ensure that data from RI nodes and facilities is integrated in cloud compatible, FAIR-compliant data resources.

WP2: Tools Collaboratory: Deployment of life-science data integration and analysis workflows in EOSC

WP2 works with our software developer community to make life science tools ready for deployment in the EOSC following FAIR software principles. It builds an integrated EOSC environment by addressing three major points: software and tools packaging, workflow composition and execution, and registries.Through open hackathons WP2 (with WP1) will also

develop the skills needed in the tool developer community to make efficient use of FAIR data within EOSC.

WP3: Demonstrators and Open Calls for User Projects
WP3 connects life science users to EOSC. User projects will be selected (by peer-review) to guide and structure the work in order to allow implementation of data, tools, policies and support with concrete exemplary projects at hand.

WP4: Policies, specifications and tools for secure management of sensitive data for research purposes
WP4 address policies and guidelines for sensitive data storage, processing, sharing and reuse for research purposes.

WP5: User management and access services
WP5 develops our shared authentication and authorisation infrastructure (AAI) and user access management (ARIA). Service operations with e-Infrastructure providers will make the whole access and user management system fully interoperable with the EOSC.

WP6: FAIRification and provenance services
WP6 delivers the common services and standards (incl. ISO standards) needed to capture provenance and to make data FAIR. Consistent adoption across infrastructures is supported by FAIRmetrics and FAIRassist services.

WP7: Cloud Deployment
WP7 provides a set of integrated cloud resources (from national life science clouds and commercial clouds) that underpins the cloud-based FAIR RI data resources (WP1), workflows (WP2) and science demonstrators (WP3) that are being supported within the project. Via training of staff WP7 aims to establish a sustainable "ResOps" expert community for cloud-enabled workflows in LS RI.

WP8: International Impact, Innovation and Sustainability
WP8 integrate EOSC-Life with the developing EOSC governance and coordination structures, with the other EOSC cluster projects and link with relevant global initiatives to develop open science data clouds. WP8 will also extend and align the BMS RIs' joint quality management framework to the service management framework to be adopted by EOSC.

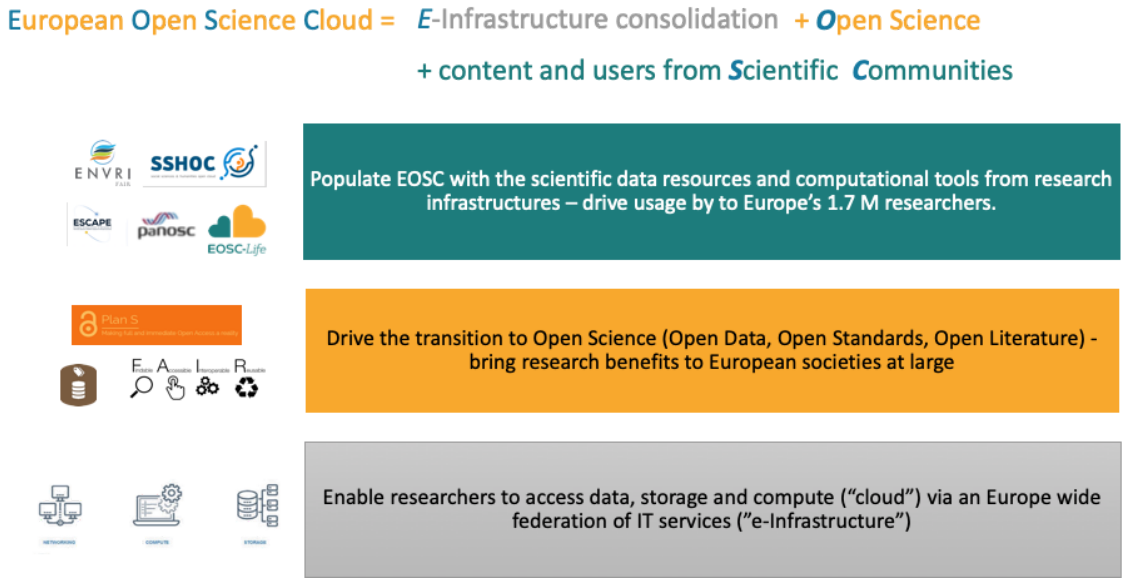
WP9: Training of the EOSC Life community
WP9 deliver the training for staff and users needed to enable effective data access and preservation for immediate and future sharing and re-use of data in the Biological and Medical Sciences. WP9 will provide hands-on training in using EOSC-Life tools, data resources and other services developed in the project, effective re-use of publicly available data, and best practices that users should adopt in managing their own data (WP6, FAIRassist).

WP10: Outreach provides the market research, web, social media and other outreach activities needed for dissemination and stakeholder engagement.

WP11: Project Management runs the project administration and management functions.

2. WHAT IS THE EOSC-LIFE COMMUNITY EXPECTING FROM EOSC

EOSC-Life represents an important step towards realising EOSC and aligns with the EOSC strategy (set out by the European Commission[1]) to federate existing research infrastructures and scientific clouds into a Europe-wide platform of cloud-based services. It brings together the LS RI around Open Science principles and will populate the emerging EOSC with content and users from the large European life science community. Thus, EOSC-Life closely aligns with the Implementation Roadmap for the European Open Science cloud[2]. The life science RIs, as distributed international organisations, are naturally federated with a rich ecosystem of national facilities that provide user-focussed services and tens of thousands of annual users, are an ideal foundation to create EOSC in the life sciences.



EOSC is constructed from three major components: consolidation of European e-Infrastructure Services into a federating core, the policies and services that drive transition to Open Science and the content and users from scientific communities - represented via European research infrastructures

Federation of distributed data resources from the full panoply of life science disciplines requires robust, yet lightweight and flexible data and metadata standards to produce "FAIR" data as the default BMS RIs output. These minimum metadata recommendations are captured in the EDMI (EOSC Dataset Minimum Information) guidelines. EOSC-Life will build on the EDMI standard[3], developed in EOSCpilot, for the publication of FAIR data resources.

[1] EC COM (2016) 178
[2] EC SWD(2018) 83
[3] using the bioschemas.org specification

Digital biology also needs the technical infrastructure to manage data from human patients, highly pathogenic organisms, as well as global genetic resources securely, with assurance and documentation that any processing and analysis is in line with the consent given and complies with all relevant legal and regulatory requirements. As the national implementations of the European General Data Protection Regulation (GDPR) progress and best practice in its usage in research is established, the life science community will need to adopt these policies and establish technical cloud infrastructure that meets their requirements. For example, capturing the data provenance from acquisition of biological material, through its processing and storage to the data generation and analysis is a key requirement to ensure secure data handling as well as to ensure the reproducibility of biomedical research for FAIR-Health. The regulatory requirements go beyond human health - compliance with the Nagoya protocol is imperative for natural resources.

ICT e-infrastructures are an important part of the life science RIs supply-chain for computational services and close partnerships are developed via several mechanisms. Importantly, our joint authorisation and authentication infrastructure (LS AAI) will incorporate core components developed by e-Infrastructures in the pan-European AARC2 project. EOSC-Life will access cloud infrastructure from coordinated national, international and commercial offerings which is closely aligned with developments of data distribution and cloud execution infrastructures developed in the EOSCHub and HNSciCloud projects. Thus EOSC-Life's work package on cloud access links national and commercial clouds, including EuroHPC and PRACE centers.

EOSC-Life will further develop the links between the life-science domain registries and the scholarly communication landscape with platforms like DataCite and OpenAIRE. e-Infrastructures are involved in EOSC-Life as third parties to deliver operational services according to specifications (developed within EOSC-Life or earlier cluster projects such as CORBEL). Budget for this ICT supply-chain is based on initial estimates developed in dialogue with e-Infrastructure service providers.

In summary, the EOSC-Life plans are built on the expectation that EOSC will deliver a pan-European set of federation services that are strongly linked into national centres. The EOSC Architecture should also provide core components (e.g. for user access/AAI) that can be readily incorporated into the many national, regional and locally funded services within the life science data ecosystem. A corollary of this vision of enablement and inclusiveness is that Rules of Participation should have low barriers to entry, allowing for diversity in access, and encourage open publication service quality and performance.

The EOSC-Life community would welcome a European-wide model for access to cloud and storage resources with multiple cost-recovery mechanisms to recognise the diversity in national and international funding schemes.

The EOSC-Life community considers that open access to publicly funded research data is a fundamental tenet of EOSC. Thus EOSC sustainability planning must consider the infrastructure costs associated with data curation and long-term storage and access. Open research data resources provide significant value to the scientific community, EOSC should go beyond quality-based certification schemes to develop mechanisms to capture this value and the impact of data resources to underpin future business-cases.

Architecture

EOSC-Life aims to provide a single gateway to EOSC for Europe's life scientists. The life science RIs have developed a long-term model for the operation of a common user authentication and resource authorisation infrastructure in the life-science community that incorporates e-Infrastructure services. The model is simple: e-infrastructures operate the transverse technical components of the AAI while the primary user interface, connections to services, management of groups, and resource authorisation is managed by the life sciences community. EOSC-Life will train service providers across the life science RIs and provide the user-facing policies needed for operations and EOSC compliance.

The BMS RIs have significant internal cloud infrastructure running at the national nodes (e.g. ELIXIR, BBMRI, INSTRUCT) and are partners in EOSC-Hub/EGI (FedCloud). For instance, the collective cloud capacity in ELIXIR's national clouds is over 60,000 cores and serves thousands of users (e.g. useGalaxy.eu have 2000 users every month). The life science RIs also utilise commercial cloud providers (AWS, Google, Azure) via the HN-SciCloud where a procurement model has been established.

While the aggregate compute capacity exposed in EOSC-Life is large, accessing this infrastructure is not always straightforward with divergent access procedures (including long-term sustainable models for transnational access to facilities and compute resources), resource allocation models and a lack of common standards for tools deployment and workflow execution. In EOSC-Life we will, through an open, collaborative model, bringing together existing national cloud infrastructures associated with BMS RI nodes and connect this to EOSC, adopting our shared AAI services and implementing the agreed standards for workflow and task execution.

EOSC-Life will drive the implementation of standards to make clouds compatible both within the life sciences globally (e.g. by using the GA4GH cloud standards) and with other science domains in EOSC. This aligns with and contributes to the overall EOSC strategy: a common platform (developed in EOSC Hub) that is accessible via machine-to-machine interfaces and offers access to all the EOSC shared resources.

EOSC-Life will align with the large and active European Galaxy community, support the development of novel Galaxy workflows for users and RI experts, and make sure that these can be reproducibly executed on the participating national clouds. This will allow us to establish an easy-to-use interdisciplinary data analysis platform in EOSC that leverages the rich set of published community tools and workflows. EOSC-Life will also work with the Common Workflow Language (CWL) open standard to allow interoperable tools and capture dependencies and workflow requirements for universal deployment.

Long-term use and uptake by the life-science community require that tools and workflows are findable and accessible, and further that they are citable so that author's credit is attributed fairly and accurately. Across the BMS RIs, and the international BMS community, tools and workflow registries are already available and in some instances, have been running for over a decade. In

EOSC-Life, we build on metadata schemas and know-how developed in previous EU projects such as wf4ever and will create a registry and discovery environment that overcomes fragmentation and supports sharing and reuse.

- Any EOSC AAI solution should take into account the Life Science AAI specifications developed in EOSC-Life WP5
- Data access solutions, personal work spaces, etc. must take into account the specific requirements for protecting personal and health data
- We need a better understand the business models underlying the different EOSC services (e.g. FAIRification vs long-term data storage)
- PIDs for datasets should allow for predictable, machine-actionable access to the underlying datasets (not just a landing page)

FAIR

The main objective for the EOSC-Life consortium is to bring the vast quantities of life science data into EOSC by equipping each RI to publish their data in a FAIR and sustainable manner. This will be achieved by clustering of data resources along scientific themes (across RIs) that capture research communities' interests.

Data publishing in EOSC-Life builds on the principles of EOSC developed in EOSCpilot to leverage existing repositories and minimum metadata specifications captured in the EDMI model. BMS RIs' data resources will be indexed in a domain data catalogue and made available for access and use in the cloud based on the needs of project demonstrators, user projects and an RI driven nomination procedure based on cloud-readiness and clustering into themes (to ascertain a diversity of data resources). We will also develop a digital assistant (FAIRassist, WP6) and deliver training to enable the on-going publication and improved access of the large number of data resources from national centres (e.g., the BBMRI-ERIC Directory links to 530 biobanks with over 1400 biological material and data collections in 19 countries). In EOSC-Life we have reserved budget to bring in RI data resources and clouds from national expert centres as third parties to the project based on the cloud readiness model and user project demand. This will create a nucleus of high value data resources in EOSC coupled to an enablement plan for continuously updating and expanding this set.

Beyond the generic architecture and meta-data standards (EDMI) for exposing data resources developed in EOSCpilot, we also need to expose training materials, workflows and other services for community reused via existing domain-specific registries and "aggregators". Examples in the life-sciences are the omicsDI data catalogue, TeSS for training materials, FAIRSharing for standards, and the BBMRI-ERIC Directory for samples.

EOSC-Life will provide the necessary training and development resources to adopt the EOSC data catalogue architecture within each BMS RI and provide an access route to the BMS RIs data catalogues within the EOSC (e.g., by further publishing their content into cross-disciplinary EOSC catalogues). The EOSC-Life is designed around a strategy where EDMI and the life-science component Bioschemas.org provide the minimum metadata guidelines, validation tools and

discovery services needed to make tools, workflows, training materials and other research assets FAIR and discoverable within EOSC.

Many different national and trans-national projects and initiatives work on standards and solutions for FAIR data (including EOSC-Life) and so clear guidance and expectations on requirements for FAIR data and services is required early on to avoid confusion within the life science user community.

Landscape

EOSC-Life brings together the 13 BMS RIs on the ESFRI Roadmap. The project consortium comprises all the legal entities of the established BMS RIs including a set of national centres that provide access to leading international scientific service platforms. The EOSC-Life consortium represents a truly continent-scale effort: collectively the BMS RI have national centres in 23 European countries.

There are often strong national collaborations between life science RIs. International alignment is therefore a key requirement from many national BMS RI Nodes and we expect that joint data solutions in the EOSC will drive further collaboration and harmonisation.

Landscaping across the regional, national and European levels is important, for instance the broad geographic involvement of national expert centres - research infrastructure nodes - in EOSC-Life. Access to advanced life science infrastructures is a key objective of the Smart Specialisation Strategies of many European regions. These efforts are set to directly benefit from the implementation of Europe-wide standards through EOSC-Life and through our established partnerships link these with ongoing global efforts (e.g. MIAPPE, NIH Data Commons, GA4GH, RDA). Also, new standards are emerging within EOSC-Life via the demonstrator programs. For example, MICHA aims at the standardization of chemosensitivity screening of cancer samples involves several RIs including EATRIS, OPENSCREEN and INSTRUCT.

Rules of Participation

The rules of participation in EOSC – defining the rights, obligations and accountability of the EOSC actors (notably data producers, service providers, data/service users) - are of particular relevance to the life science community. The life science RIs, in daily operations and in this project, need to closely monitor the development of applicable legal frameworks (e.g. GDPR, clinical trials directive, copyright rules, and data security). We also need to consider our different service models (e.g. from open access databases to carefully regulated access to governmental BSL4 pathogen laboratories) and the differences in established rules and regulations (e.g. animal welfare, Nagoya protocol on access benefits sharing). Thus, EOSC guidelines covering a variety of service providers (e.g. public vs private), agreed tools, specifications, catalogues and standards ('EOSC shared resources'), principles for regulating transactions in the EOSC (e.g. cost recovery for laboratories, data storage and high-capacity compute) are areas where life science RIs jointly are an important stakeholder.

- The RoP have to work across very different scientific communities with very heterogeneous requirements
- RoP should not micromanage service and data provision

- RIs have been providing data/computational services for decades with large, established user communities (in many cases >100k users/month) and complex operations (e.g. EGA archive distribute >2PB of access controlled data every year), the experience and requirements of these services must be taken into account.
- Diverse funder requirements on service delivery must be considered.

Sustainability

The European Open Science cloud brings together the Open Science policies in H2020 with the ICT services needed to create a digital single market in Europe. These are two major initiatives - that operate with very different drivers, business cases and community engagement mechanisms. The draft EOSC sustainability document does not separate out these components and as a result the document doesn't provide clarity on the necessary operating models.

EOSC-Life considers Open Data / Open Literature (i.e. Open Science) components of EOSC are a public good - following on the OECD and EC recommendations the EOSC creates an opportunity to create a layer of open access resources as a foundation for the Digital Single Market for research. These data and literature resources will thus need to be funded to deliver on the Open Science policies. This is a major challenge not addressed in the document - the data resources are largely operated by organisations outside the e-Infrastructures that to date have made up the core of EOSC investments.

Other components of EOSC - a large part of the Federating Core and cloud/storage resources - are not likely to be open access / public goods but rather operate on a cost-recovery basis (or possibly as a "club good" where a number of funders pools resources for access). This could e.g. be cloud resources offered on cost-recovery basis on top of open data (e.g. Copernicus Sentinel provides an exemplar) or pooled resources that are offered as "free-at-point-of-use" via an excellence-based resource allocation system. Cloud and Storage cannot be a public good - there needs to be mechanisms for cost-control.

Thus, there is a need for a clear separation of drivers, business-models and concerns early in the document - this would avoid the confusing argumentation of open data resources amidst argumentation for cost-recovery of depletable (rivalrous) goods such as cloud and storage. We would also like to see a clear statement on Open Science that differentiates between open data access and the (possibly cost-recovered) analytics of open data.

2. ADDED VALUE OF EOSC TO THE LIFE SCIENCE COMMUNITY

The lasting impact of the EOSC-Life ambition to connect data, tools, clouds and people across geographies into an open collaborative digital space for digital biology will be a step change in data driven life science for Europe. Through the European Open Science Cloud scientists – in industry and academia - will have access to data from advanced technology platforms with the tools and expertise that allow integration across disciplines and geographical barriers. Good ideas and excellent science will not be limited by the individual scientist's local access to leading facilities. EOSC-Life has a dedicated work package on International Impact, Innovation and Sustainability (WP8) to engage all relevant stakeholders for collaborative development. In the

long run, the life science community should benefit from enhanced reproducibility via the improved data management techniques provided by EOSC-Life.

EOSC provides an opportunity to further converge the “classical” e-infrastructure services (storage, compute, connectivity) towards the specific needs of the life science community - characterised by highly dynamic, heterogeneous, short-lived data, research driven by both small labs and large-scale initiatives.

The data, tools, workflows and other resources linked to the EOSC via the BMS RI will naturally be multidisciplinary and cross borders. The cloud environment to be created will allow open development of tools and workflows and support cloud-based innovation as a result of dedicated cloud resources made available through WP7. A shared login system (AAI) will open up facilities for transnational access. The data made available to the EOSC via EOSC-Life (WP1) will be accessible (with controls where appropriate) to all European researchers.

The user base of the BSM RIs is large, consisting of Europe’s 500 000 life scientists and their global collaborators. Documentation, training and communication activities are instrumental to enable the dissemination of EOSC-Life’s opportunities to participation and access to the data, workflows and working practices resulting from the project. The project contains a dedicated Training Work Package (WP9). Training material will be produced by the technical work packages and delivery will be supported by a coherent training programme that link with on-going RI efforts:

- Training of life scientists in proper data stewardship tailored to their respective discipline (training in data management topics such metadata, data quality management, data acquisition, data transfer, data curation, FAIR data, data publication for open access, ...)
- Training of a new generation of data support experts in the life sciences who speak the same “language” across disciplines (computational scientists, data analysts, IT infrastructure experts, software developers, data curators, ...)

Life science RI experts are involved in global standardisation and interoperability initiatives and, in some cases, also partners on US-based projects within the NIH BD2K programme. Work in this project will align with, and contribute towards efforts in e.g. GA4GH, BD2K/NIH FAIR Data Commons, FORCE11, GO FAIR, Global BioImaging, Galaxy, Bioconda and RDA to further develop community standards. In addition, many of the large data resources in WP1 (see ELIXIR Core Data Resources46) operate long-standing global collaborations for data exchange and metadata standards. The development of pan-European standards, services and training programmes must be coordinated with emerging global standards in an active process where the European community contribute to and, where appropriate, lead the development.

3. KEY MESSAGE FOR EOSC EXECUTIVE AND GOVERNING BOARDS

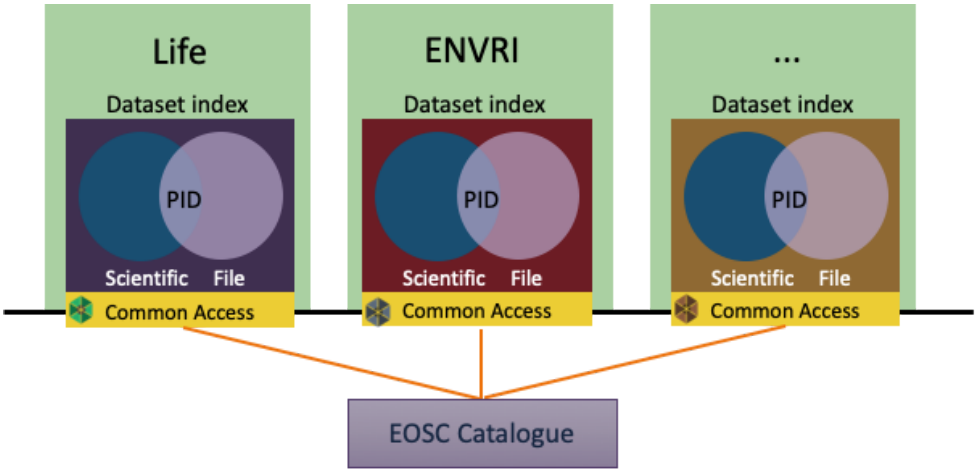
Although EOSC-Life mainly represents the LS research community, many of the driving principles for the interaction with EOSC can be generalised to other research communities:

- Focus on user requirements (e.g. in the area of personal data) and providing added value to them (not technology first, not policy first). The construction of EOSC must be science-driven and center around user needs. Open calls for partnerships, participation in European (e.g. Horizon Europe) projects and directed outreach (e.g. to ERC grantees) are all useful mechanisms to engage and enlist the science community in the construction.

- ESFRI research infrastructures (across all domains) have established user communities accessing a large number of national facilities, thus the ESFRI RIs naturally provide federating capabilities and user access mechanisms for EOSC.

The construction of EOSC requires a careful balance between community-specific and general EOSC services that recognises the significant and long-standing investments into community services while gradually driving a convergence towards a pan-European federating core.

EOSC FAIR services should be based on high-level guidelines for FAIR data that can be translated into community-driven specifications for the broad and complex life-science data landscape. The EOSC Architecture needs to support aggregation models that enable disciplines to elaborate rich discipline-specific meta data descriptions. Thus, EOSC-Life would like to see a hierarchical model for data discovery based on common minimal information (EDMI) that is abstracted from richly annotated, discipline specific registries into high level EOSC-wide catalogues.



In the context of discussions around business models for the EOSC and in discussions on the long-term sustainability of ESFRI Research Infrastructures, suggestions are increasingly being made through multiple channels for database operators to charge users for access to data. This model conflicts with Open Data principles[4] and policies and sits in contrast to the current business models of life science data resources. If major European funders abandon Open Data principles it would bring dire consequences for the interlinked ecosystem of life science data and the ability of users to access resources. We note that the data residing in open access research data archives and other resources are not owned by those operating the database - they are merely the

[4] e.g. *"The vision underlying the Commission's strategy on open data and knowledge circulation is that information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full. This means making publicly-funded scientific information available online, at no extra cost, to European researchers and citizens via sustainable e-infrastructures, also ensuring long-term access to avoid losing scientific information of unique value."* (*https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0108:FIN:EN:PDF*)

custodians of those data. Seeking to charge users to access data would break trust with the community, dis-incentivise future support of and deposition into the data resource, and would have the net effect of reducing the publishing of data and tools in the public domain in future.

In conclusion, the landscape of service providers that make up EOSC is broad and complex, including e-infrastructures, domain-specific Research Infrastructures (such as ESFRI RIs) and also industry. For the European Open Science Cloud (EOSC) to be a long-term success, it must bring value to both users and service providers within a sustainable framework based on Open Data, Open Source and Open Science. This means that European scientists will have access to advanced data services, technology platforms, samples and support services throughout the European Research Area and that the resulting data will be openly accessible for reuse through the European Open Science Cloud in full compliance with all ethical, regulatory and legal requirements.