D1.1 - CCC (Community-based Competence Centre) First Survey

Project reference no.	101129751
Project acronym	OSCARS
Project full name	Open Science Clusters' Action for Research and Society
Horizon Europe Call	HORIZON-INFRA-2023-EOSC-01-01
Project coordinator	Giovanni Lamanna
Coordinating organisation	CNRS LAPP
Project website	https://oscars-project.eu/
Deliverable no.	D1.1
Deliverable type	Report
Dissemination Level:	Public
Version:	1
Status:	Completed and submitted
Contractual delivery date	30/09/2024
Actual delivery date	30/09/2024
EC Project Officer	Enrico Pellizzari

	Name	*Partner*	*Activity*
Responsible partner	Jordi Bodera Sempere	ESRF	OSCARS WP1 leader
Contributor 1	Gary Saunders	EATRIS	OSCARS WP1 co-leader
Contributor 2	Romain David	ERINHA	OSCARS WP3 leader
Contributor 3	Giovanni Guerrieri	ESCAPE	OSCARS WP1 ESCAPE representative
Contributor 4	Jonathan Tedds	ELIXIR	OSCARS LS RI representative
Contributor 5	Anca Hienola	FMI	OSCARS WP3 co-leader

Funded by the European Union

Contributor 6	Joaquín López	LifeWatch	OSCARS WP3 ENVRI representative
Contributor 7	Vasso Kalaitzi	CESSDA/KNAW DANS	OSCARS WP1 participant (SSHOC)
Contributor 8	Panagiota Starida	CESSDA	OSCARS WP1 participant (SSHOC)

DELIVERABLE ABSTRACT
This deliverable is a report covering a survey made within the clusters and newly formed CLOCCs to obtain a first analysis of what competences are available and what the community needs are.

TERMINOLOGY

Terminology/Acronym	*Definition*
Cluster Open Science Competence Centres (CLOCC)	A CLOCC is a virtual hub dedicated to fostering research excellence through training and knowledge transfer. The CLOCCs are community-based initiatives supported by a collaborative network of people in the context of the Science Clusters providing expertise, best practices and services in relation to Open Science, and the promotion of cross-disciplinary collaboration.
Competence	“Competence is a demonstrated ability to apply knowledge, skills and attributes for achieving desirable results”, as per the European e-Competence Framework definition.
Community-based Competence Centre (CCC)	CCCs or Competence Centre is an alternative name for the CLOCC.
ENVRI	The Environmental Science Cluster.
ESCAPE	The Astronomy and Particle Physics Science Cluster.
Life Science Research Infrastructures (LS-RI)	The Life Science Cluster.
OSCARS	Open Science Clusters’ Action for Research and Society, the name of this project.
PaNOSC	The Photon and Neutron Open Science Cluster.

European Research Infrastructure (RI)	European Research Infrastructure
Science Cluster	The Science Clusters are a group of Research Infrastructures (RIs) related through the domain of science on which they operate. There are five Science Clusters
SSHOC	The Social Sciences and Humanities Open Cluster.
FAIR	FAIR stands for Findable, Accessible, Interoperable and Reusable. FAIR usually relates to the FAIR principles and FAIR data, key components to enable Open Science.
Open Science	Is the movement to make scientific research and its dissemination accessible to all levels of society.

Contents

Executive summary

This report presents a compilation and analysis of answers from participants in each of the five European Science Clusters (SCs) regarding their shared competencies and the needs of their user base in relation to Open Science. Each of the five Science Clusters is distinct, possessing unique competencies, but they all share certain core aspects, particularly those central to previous EOSC-related projects, such as training, data management, FAIR principles, and Open Science policies.

When it comes to user needs, these are more varied than the clusters' available competencies, with fewer commonalities between them. A notable point is that a Science Cluster may identify as a need a competence it has already listed as available. This suggests that while SCs have strong expertise in certain areas (e.g., data management), they recognise the potential for further development and improvement.

OSCARS WP1 - Cluster Open Science Competence Centres will use the insights from this report to shape a roadmap of proposed activities, ensuring that each of the five competence centres focuses on supporting the most critical areas.

In parallel, OSCARS WP2 - Composable Research Infrastructure Services in EOSC will consider the needs of the clusters and their communities to ensure that the delivered composable services provide maximum added value.

1. Introduction

The European Research Infrastructure Science Clusters have grown out of five collaborative projects funded by the European Union in 2019 to link European Strategy Forum on Research Infrastructures *(*ESFRI) and other world-class Research Infrastructures (RIs) to the European Open Science Cloud (EOSC). The services developed by the clusters and other outcomes of the projects are cornerstones of the emerging EOSC fabric and support both disciplinary communities and multidisciplinary initiatives with harmonised models for access to data, tools, workflows and training.[1] Each cluster unites multiple RIs in their specific scientific domain:

- Astronomy and particle physics: ESCAPE[2]
- Environmental science: ENVRI[3]
- Life science: LIFE SCIENCE[4]
- Photon and neutron science: PaNOSC[5]
- Social sciences and humanities: SSHOC[6]

Representatives of the five European Research Infrastructure Science Clusters taking part in the OSCARS project and its Work Package 1 (WP1) are working to create and operate Cluster Open Science Competence Centres (CLOCCs), with the aim to share better the competences (useful skill-set or know-how) that enable open science.

The definition of a CLOCC used within the OSCARS project is "*a virtual hub dedicated to fostering research excellence through training and knowledge transfer*", adding that our CLOCCs "*are community-based initiatives supported by a collaborative network of people in the context of the Science Clusters providing expertise, best practices and services in relation to Open Science, and the promotion of cross-disciplinary collaboration*".

One of the key aspects to create CLOCCs that are fit for purpose is to survey the landscape to know what competences are available and what the community needs are, which is the focus of this report.

[1] See also https://science-clusters.eu/
[2] https://projectescape.eu/
[3] https://envri.eu/
[4] https://lifescience-ri.eu/home.html
[5] https://www.panosc.eu/
[6] https://sshopencloud.eu/

This report will also help the OSCARS project, and readers, to understand the current status at the different Science Clusters with regards to training, resources and knowledge sharing, with a particular emphasis on the topics that enable Open Science.

2. Competences available

The OSCARS WP1 contributors have contacted partners and colleagues in their science cluster in many different research institutes to identify what competences are available within each science cluster and community. The competences that have been listed repeatedly by the science clusters are listed in the section “existing competences in the five European Research Infrastructure Science Clusters”.

a. Existing Competencies in the five European Research Infrastructure Science Clusters

Whilst all five Science Clusters and their communities are different, they have all developed over time and thanks to the participation in many EOSC-related initiatives the following competences:

- Training development and workshop organisation
- Software development
- Data Management

b. Astronomy and particle physics specific competences available

The 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, ET, FAIR-GSI, HL-LHC, KM3NeT, SKA), two pan-European organisations (CERN, ESO), an ERIC (JIV-ERIC) and a pan-European research infrastructure (EGO-Virgo) in astronomy and particle physics. ESCAPE has cultivated extensive expertise in data management, analysis, and preservation, leveraging advanced platforms and tools to support the complex needs of astronomy and particle physics research, while ensuring compliance with FAIR principles and catalysing collaborative scientific innovation across Europe.

- **Distributed Data Management**:
    - Rucio serves the data needs of modern scientific experiments. Large amounts of data, countless numbers of files, heterogeneous storage systems, globally distributed data centres, monitoring and analytics are addressed in a single, modular solution.
    - FTS3 is the service responsible for globally distributing the majority of the LHC data across the WLCG infrastructure. It is a low level data movement

service, responsible for reliable bulk transfer of files from one site to another while allowing participating sites to control the network resource usage.
    - HiPS is an IVOA standard for the description, storage and access of large sky survey data across multiple international nodes.
    - Expertise on the publication of data associated with scientific journal articles.

- **Analysis Frameworks for Scientific Computing**:
    - The ESFRI Science Analysis Platform is 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 members of ESCAPE.
    - The Virtual Research Environment is an analysis platform developed at CERN serving the needs of scientific communities involved in European Projects. Its scope is to facilitate the development of end-to-end physics workflows, providing researchers with access to an infrastructure and to the digital content necessary to produce and preserve a scientific result in compliance with FAIR principles.

- **Interoperability Frameworks for Data and Services:**
    - The Virtual Observatory is an international astronomical community-based initiative. It aims to allow global electronic access to the available astronomical data archives of space and ground-based observatories and other sky survey databases.
    - Expertise on the definition of disciplinary interoperability standards
    - Expertise on training for implementation of standards.
    - Expertise on the development of tools and services for the distribution of astronomical data to wide audiences of researchers, educators and public.

- **Preservation and re-interpretation of analysis workflows and results**:
    - Zenodo, a CERN service, is an open dependable home for the long-tail of science, enabling researchers to share and preserve any research outputs in any size, any format and from any science.
    - The Open-source Scientific Software and Service Repository (OSSR) is an open-access repository for software collaborative development, uptake and reuse. The OSSR is intended for scientific researchers to share and find software and services.
    - REANA is a reusable and reproducible research data analysis platform. It helps researchers to structure their input data, analysis code, containerised environments and computational workflows so that the analysis can be instantiated and run on remote compute clouds.

- **Widening scientific participation**: The ESCAPE Citizen Science (CS) is an astronomy and astroparticle physics programme of crowdsourced data mining, creating truth sets for machine learning. It trains and educates both the scientific community and the wider science-inclined public in the usage and implementation of the ESCAPE services and ESFRI facilities, in line with the FAIR principles.

c. Environment science existing competences

The ENVRI (Environmental Research Infrastructures) community is a European collaboration of environmental research infrastructures (RIs) that collectively provide data, services, and tools for studying the Earth's environment across four domains - atmosphere, marine, geosphere, and biodiversity. Throughout its past projects, such as ENVRI, ENVRIplus, and ENVRI-FAIR, the ENVRI community has cultivated a suite of advanced competencies that enhance data management, interoperability, service provision, and collaborative efforts among European environmental research infrastructures. These competencies, developed and refined over time, can be categorised into several key areas:

- **Data Management and Interoperability:**
  - **Establishment of Common Data Frameworks**: ENVRI has pioneered the development of standardised frameworks for data management across diverse environmental RIs, ensuring consistency in data storage, description, and accessibility practices.
  - **Implementation of Interoperability Solutions**: The cluster has adopted and implemented robust interoperability standards and protocols (such as OGC and ISO) to facilitate seamless data exchange and integration across RIs. This includes the development of standardised metadata schemas and ontologies to support cross-domain data sharing and utilisation.
  - **Techniques for Data Harmonization:** ENVRI has advanced methodologies for harmonising heterogeneous data formats and standards, enabling effective integration of diverse data sources and fostering interdisciplinary research.

- **Development of Tools and Services:**
  - **Creation of Data Access and Discovery Platforms**: The cluster has developed sophisticated portals and platforms, such as the ENVRI Data Discovery Portal, which provide unified access points and advanced search functionalities for data retrieval across multiple RIs.

- **Advancement of Visualization and Analytical Tools**: ENVRI has focused on the development of state-of-the-art tools for data visualisation, processing, and analysis, empowering researchers to interpret complex datasets and derive meaningful scientific insights.
- **Provision of Virtual Research Environments (VREs)**: The establishment of customizable VREs has enabled collaborative research, providing integrated access to data, analytical tools, and computational resources tailored to specific scientific communities.

- **ENVRI Reference Model and Architecture:**
  - **Development of the ENVRI Reference Model (ERM)**[7]: ENVRI has designed a conceptual model that outlines the core components and functionalities required across environmental RIs. The ERM provides a shared architectural framework that guides the development and integration of tools and services within the cluster.
  - **Creation of Common Software and Middleware:** The cluster has produced shared software components and middleware solutions that support the ERM, ensuring seamless interoperability, efficient data management, and the deployment of integrated services.

- **Training**[8] **and Capacity Building:**
  - **Organization of Training Programs and Workshops:** ENVRI has conducted numerous training sessions, workshops, and tutorials to enhance the competencies of researchers, data managers, and RI operators in areas such as data management best practices, FAIR data principles, and the use of ENVRI tools and services.
  - **Development of E-Learning Resources:** The cluster has created comprehensive online training materials and e-learning modules, covering a wide range of topics including data curation, interoperability standards, and specific ENVRI services and tools.
  - **Facilitation of Community Engagement and Knowledge Exchange**: ENVRI has fostered platforms for community engagement, knowledge sharing, and collaboration among researchers, data scientists, and RI staff, including the organisation of conferences and networking events.

[7] https://doi.org/10.1007/978-3-030-52829-4_4
[8] https://envrihub.vm.fedcloud.eu/training

- **Support for FAIR Data Principles**:
    - **Promotion of FAIR Data Compliance**: The cluster has formulated guidelines and best practices to ensure that data managed within the ENVRI framework adheres to the FAIR principles, enhancing the usability and accessibility of data.
    - **Development of FAIRness Assessment Tools**[9]: ENVRI has created a tool for assessing and improving the FAIRness of data, providing actionable feedback and recommendations to optimise data management practices.
    - **Alignment with FAIR Principles in the ENVRI Hub:** The ENVRI Hub[10], the central platform for data and service access, has been developed in strict alignment with FAIR principles, offering a user-centric interface for efficient data discovery and utilisation.

- **Development of the ENVRI Hub:**
    - **Creation of a Centralised Access Platform**: ENVRI has established the ENVRI Hub as a comprehensive, centralised access point for data, services, and tools across the cluster. This Hub integrates various Research Infrastructures (RIs) and provides a unified interface for researchers, facilitating streamlined access to a vast array of environmental data and resources.
    - **Integration with the European Open Science Cloud** (EOSC): The ENVRI Hub is designed to align with and contribute to the EOSC initiative. By integrating its data and services with EOSC, the ENVRI Hub ensures broader accessibility and interoperability of environmental research data across Europe, fostering collaboration and data sharing on a continental scale. This integration supports the EOSC vision of creating a trusted, open environment for research data and tools, enhancing the global reach and impact of ENVRI's offerings.
    - **Incorporation of Diverse Services:** The ENVRI Hub facilitates the seamless integration of multiple data access services, analytical tools, and Virtual Research Environments (VREs), creating a cohesive and versatile research environment. It includes: ENVRI Catalogue of Services[11], ENVRI Knowledge Base[12], ENVRI Training Gateway, ENVRI Science Demonstrators[13], Hands-On Area[14] and ENVRI Dashboards[15].

[9] https://trainingcatalogue.envri.eu/course/23
[10] https://envrihub.vm.fedcloud.eu
[11] https://envrihub.vm.fedcloud.eu/cservicesmain
[12] https://search.envri.eu
[13] https://envrihub.vm.fedcloud.eu/sciencedemonstrators
[14] https://envrihub.vm.fedcloud.eu/hands_on
[15] https://envrihub.vm.fedcloud.eu/dashboards

- **Influence on Policy Development:** ENVRI has expertise in shaping policies related to data management, interoperability, and open science within the European research landscape. In addition, ENVRI engages with policymakers and regulatory bodies, offering expertise and data-driven insights to inform environmental policy and regulation.

- **Integration Across Environmental Domains:**
    - **Facilitation of Cross-Domain Data Integration:** ENVRI has developed methodologies and tools that support the integration of data across various environmental domains (atmospheric, marine, terrestrial, and biodiversity), promoting interdisciplinary research and a holistic understanding of environmental systems.
    - **Development of Domain-Specific Services:** The cluster has tailored services and tools to address the specific needs of different environmental research domains, while maintaining the capability for interoperability and cross-domain data sharing.

- **Alignment with EOSC and other International initiatives**: by ensuring compatibility with EOSC standards and participating in international initiatives, ENVRI supports the development of a cohesive, federated infrastructure for open science, enabling researchers worldwide to benefit from shared resources and knowledge in environmental research.

d. Life science specific competences available

The Life Sciences Research Infrastructures (LSRI) cluster provides a comprehensive range of competencies aimed at integrating life sciences data and resources across Europe into the EOSC. These competencies have been crucial for advancing open science and fostering collaboration in the life sciences community:

- Integration of Life Sciences Data: Competency in integrating diverse life sciences data from various European Research Infrastructures into the EOSC. This integration has enhanced data accessibility and interoperability, supporting cross-disciplinary research and collaboration.
- FAIR Data Management: Expertise in implementing and promoting FAIR (Findable, Accessible, Interoperable, Reusable) data principles. LSRI provides tools, guidelines, training and support for ensuring that life sciences data are managed according to these principles, improving data reuse and research efficiency.

- Cloud-Based Tools and Services: Proficiency in developing and deploying cloud-based tools and services tailored for life sciences research. These resources support a wide range of data analysis, processing, and management tasks, enabling researchers to perform complex computational work collaboratively and efficiently.
- Support for Pandemic Research: Capacity to mobilise and provide critical resources for urgent research needs, exemplified by its support for COVID-19 research. LSRI offers access to data, analytical tools, and computational resources essential for understanding the virus and contributing to global pandemic responses.
- Training and Capacity Building: Competency in delivering training and educational programmes designed to enhance the skills of life sciences researchers. LSRI cluster organises workshops, webinars, and other training initiatives to help researchers effectively utilise the tools and data available.
- Open Science Advocacy: Strong advocacy and support for open science practices within the life sciences community. LSRI provides the infrastructure and resources necessary to promote transparency, collaboration, and reproducibility in scientific research.
- Collaboration Across Research Infrastructures: Expertise in fostering collaboration among European Research Infrastructures. LSRI plays a key role in harmonising standards, sharing best practices, and creating a more integrated and cooperative research environment across Europe.
- Community Building: Capability in establishing and maintaining a vibrant community of researchers, data providers, and infrastructure operators. The LSRI continues to collaborate, share resources, and contribute to the ongoing development of the European Open Science Cloud.

These competencies have made LSRI a cornerstone of the European research ecosystem, equipping the life sciences community with the tools, knowledge, and collaborative frameworks needed to advance scientific discovery in a more open and integrated manner.

e. Photon and neutron science specific competences available

PaNOSC is the Science Cluster aggregating the Photon and Neutron (PaN) RIs, which are extremely powerful sources of X-rays and neutrons used for a wide range of research domains, spanning from cultural heritage to life science and material science. PaNOSC RIs produce petabytes of high-value datasets of unique samples that visiting scientists bring to the facilities to be analysed. PaN RIs have a privileged position that enables them to manage the data produced for the scientists and have developed over time many competences:

- Data management: general expertise in managing scientific data (from the facilities researchers, visiting scientists or industrial private research)
- Data Management Plans (DMP): their implementation and usage to better understand the volume of data and treatment to be done during and after the onsite research
- Data policy: expertise in getting an Open Science data policy adopted
- Developing metadata catalogues, ontologies and standards: to better classify and manage scientific data. PaN RIs produce data from heterogeneous scientific domains and thus have built great expertise in using ontologies and metadata to better data classification.
- Establishing open access metrics: from publications to better understand the impact of the research done at the facilities
- Application of Persistent Identifiers (PIs): like DOIs to uniquely and permanently identify datasets
- Open source software: to enable collaboration and open science. Competences are available in choosing the right policy and engaging with the community to contribute to the software.
- Compressing data and handling different data formats: using HDF5 and other formats to analyse and store data
- Data visualisation: via desktop and web applications
- Data processing (coherent imaging, scattering, spectroscopy,tomography, etc.): including using High Performance Computing (HPC) on CPU and GPUs
- Provisioning software: using CERN Virtual Machine File System (CVMFS) and containerisation to easily deploy and execute diverse software
- Transferring large volumes of data: to facilitate moving terabytes of experimental data that is located in our data centres
- Virtual Research Environments (VRE): VISA (Virtual Infrastructure for Scientific Analysis) is the VRE used in PaNOSC to enable analysing large experimental data sets with the relevant specific software in our data centres. VISA allows working collaboratively and can be complemented by Jupyter Hub.
- Pipeline automation services: e.g. for creation of raw files, data ingestion and automatic data processing.

f. Social sciences and humanities-specific competences

The domain of Social Sciences and Humanities (SSH) is extensively supported by the SSH Open Cluster (SSHOC) community. Since 2022, SSHOC consists of several thematic research

infrastructures and various other actors in the SSH domain. Over time and particularly during and after the SSHOC project (2019-2022), this community developed, organised and supported both general and SSH-specific competences.

An overview of existing Competence Center (CC) Building Blocks has been created to provide insight into the range of competences covered for the SSH domain. These competences, while applicable to a broader understanding of Open Science, are largely geared towards the needs of the SSH community. The areas of competence include:

1. **Connecting the SSH Research Communities to Digital Infrastructures for Research**

   A large suite of tools and services have been developed to support research activities in SSH and help researchers improve data literacy competences, including:

   - Discovery services for research data, tools, publications, workflows and training materials, e.g. SSH Open Marketplace, CLARIN Virtual Language Observatory, DANS Data Station SSH
   - Tools for data collection, processing and analysis, e.g. CLARIN Virtual Collection Registry, Language Resource Switchboard
   - Tools and services for data preparation, depositing, sharing and archiving, e.g. CLARIN Standards Information System
   - Services for the management of SSH vocabularies, e.g. SSH Vocabs Service

2. **Training and Capacity Building**

   Expertise and knowledge sharing are realised through training to support the SSH research communities and data stewards/curators, including:

   - Train-the-Trainer workshops for SSH trainers cover general and disciplinary topics related to open science, the FAIR principles, and research data management (RDM). This approach creates multipliers within the SSH research community.
   - Evaluation and implementation of FAIR practices for the development and sharing of training and learning resources, e.g. upcoming paper in “Towards FAIRification of learning resources and catalogues”, FAIR SSH Working Group
   - Discovery services and registries for the hosting, developing and sharing of training and learning resources, e.g. DARIAH-Campus
   - Development, curation and maintenance of worldwide courses, programmes and other activities in Digital Humanities, e.g. DH Course Registry
   - Guidelines and support for the development and delivery of training activities
   - Registries of SSH trainers’ networks, e.g. SSHOC Trainers’ Network, CLARIN Trainers’ Network

3. SSH Guidelines & Best Practices

Within the SSH research community, guidelines and best practices are being produced and shared on competences related to:

- Data policy, data management and ensuring the FAIRness of SSH data and software, e.g. CESSDA Data Management Expert Guide, The Open Handbook of Linguistic Data Management
- Use of Persistent Identifiers and citation practices for data and software
- European legislation and ethical protocols related to copyright, data protection and handling of sensitive data, e.g. CLARIN Legal Information Platform
- Data archiving practices and services specific to the SSH domain, e.g. CESSDA Data Archiving Guide
- Open Science practices for the Arts and Humanities, e.g. digitalisation of cultural heritage artefacts (archives, books, newspapers, museum collections)
- Adoption and use of SSH-specific data and metadata standards, e.g. in relation to data interoperability between domains, data deposition formats for language resources, etc.

4. SSH Community Engagement and Technical Support

Ongoing engagement and technical assistance for the SSH community is achieved through:

- Impact stories and case studies, blog posts, interviews, podcasts, e.g. CLARIN Impact Stories, DARIAH Impact Case Studies …
- Helpdesks, forums, FAQ sections, live interactive sessions and contact pages
- Collaborative platforms with citizen science, e.g. VERA by OPERAS.
- Zenodo communities, e.g, DARIAH, CLARIN, Research Data Management, etc.

3. Community-specific needs

The OSCARS WP1 contributors have contacted their partners and colleagues in many different research institutes participating in their science cluster to identify what are the needs within their science cluster and community. The items that have been listed repeatedly can be found in the subsection “Recurrent needs across the five clusters”

a. Recurrent needs across the five clusters

The OSCARS project contributors have identified the capacity to build interoperable data and services as a key recurrent need in the five clusters. The OSCARS and previous projects involving the Science Clusters are pushing forward interoperability as a key requisite to enhance Open Science, however there has been very limited success delivering sustainable interoperable or composable services.

The Science Clusters have also repeatedly highlighted the following needs:

- Standardisation and interoperability
- Ethics guidance
- Data management
- Sustainable funding

It is not surprising to find that while competences exist in data management, there is the perceived need to go further and expand these competences more widely among the research centres and the scientific communities they serve.

It is worth noting that while sustainable funding is not a competence itself, the science clusters have repeatedly identified funding as a key need for their Open Science strategies.

These are important leads to be followed for the CLOCCs, so competences in these topics (and many more specific topics listed below) are taken into account for future actions of the CLOCCs.

b. Astronomy and particle physics specific needs

The ESCAPE community is dedicated to addressing the specific needs of the astronomy and particle physics research communities by advancing key areas essential for managing and utilising the vast amounts of data generated in these fields. ESCAPE focuses on enhancing data management, access, and security to support scientific discoveries. More specifically:

1. **Distributed Data Management**: There is a pressing need for a scalable and unified Large Scale Distributed Data Management System capable of handling data at the Exabyte scale. This system would:
   a. Manage the entire data life-cycle, from raw data injection to complex multi-tiered asynchronous data transfers.
   b. Address global replication rules and access policies while supporting seamless data delivery to facilities providing computing resources to research infrastructures (RIs) and other diverse providers.
   c. Integrate with metadata systems or offering metadata services to enhance data management.

2. **Data Access and Analysis Framework for Scientific Computing**: To improve the efficiency of scientific research, there is a need to develop a robust platform that enhances data processing capabilities for end-users.
   Such a framework would:
   a. Facilitate streamlined access to data and software, utilising interactive notebooks and advanced data management services. The goal is to accelerate the research process, reduce the time required to achieve initial results, and ensure the reproducibility of scientific experiments.
   b. Build a shared community of developers and operators.

3. **Authentication, Authorization, Identity Management and Cyber-Security**: Establishing standardised practices for authentication, authorization, and identity management is crucial for fostering collaboration within the European research community. By implementing a common layer of trust, ESCAPE seeks to achieve:

   a. Secure and seamless access to shared resources, ensuring the integrity of identities across diverse platforms. This initiative would continue to align with and contribute to the work being done within the European Open Science Cloud (EOSC) and related AAI federations, in collaboration with service providers and European Science Clusters.
   b. Compliance with relevant data protection regulations, such as GDPR and HIPAA, on the storage and transmission of user information is required.

c. Environment science specific needs

The complexity and diversity within the ENVRI community require specific strategies to advance open science and data interoperability. The achievements/competencies listed above in Chapter 2.3 provide a foundation, while the needs indicate the next steps required to expand and refine that foundation. Below are the specific needs of the ENVRI community, focusing on enhancing data integration, standardisation, the adoption of advanced technological frameworks and sustainability.

Data Integration and Interoperability

- **Advanced Metadata Systems:** Development of comprehensive metadata schemas tailored to integrate data from the wide range of environmental science disciplines. These systems should enhance the discoverability and usability of data across different platforms, ensuring that data can be efficiently shared and reused within and beyond the ENVRI cluster.
- **Standardised Data Formats:** Adoption and promotion of standardised, domain-specific data formats to facilitate seamless data exchange and interoperability between different research infrastructures within the ENVRI cluster.
- **APIs for Infrastructure Federation:** Development and deployment of robust APIs that enable the federation of disparate environmental research infrastructures. These APIs should allow for the seamless integration of data services across multiple platforms, enhancing the capability of ENVRI to operate as a unified ecosystem within EOSC.
- **Machine Actionable Tools:** Ensuring that all tools and services within the ENVRI cluster are as machine actionable as possible. This would involve developing interfaces and protocols that allow machines to autonomously perform tasks such as data discovery, retrieval, and processing, thereby enhancing efficiency and scalability.

Scalable and Secure Data Infrastructure

- **High-Performance Computing (HPC) Resources:** Access to scalable HPC resources is crucial for processing the large-scale environmental datasets typical in this field, such as high-resolution climate models and satellite imagery. These resources must support the computational demands of intensive environmental analyses.
- **Robust Data Storage Solutions:** Implementation of secure and resilient data storage systems that can handle the large volumes of sensitive and complex environmental data, ensuring long-term data preservation and accessibility.

- **Standardised Data Management Plans (DMPs):** Development and enforcement of standardised DMPs across all ENVRI infrastructures. Standardising DMPs ensures that data is managed consistently and in alignment with best practices, supporting the goals of interoperability and FAIR data principles.

FAIR Data Implementation

- **FAIR Data Services:** Enhancing support for the full implementation of FAIR principles across all data generated within the ENVRI cluster. This includes providing specialised tools and training for data stewardship and management to ensure that environmental data is fully compliant with FAIR principles.
- **Data Stewardship Expertise:** Developing a network of data stewards with specific expertise in environmental sciences. These stewards will support researchers in managing the data lifecycle, from collection to archiving, ensuring adherence to FAIR principles and promoting best practices in data management.
- **Incorporation of Scientific Knowledge Graphs (SKGs):** Gradually incorporating SKGs into ENVRI infrastructures to enhance the organisation and linking of data across diverse sources. SKGs provide a structured way to represent knowledge and data relationships, improving the ability to search, integrate, and analyse complex datasets.

Community Building and Support

- **Collaborative Platforms:** Creation and enhancement of virtual research environments and collaborative platforms that enable scientists from various environmental disciplines to share data, tools, and methodologies seamlessly. These platforms should facilitate interdisciplinary collaboration and the integration of diverse data sources.
- **Targeted training and Capacity Building:** Offering targeted training programs and workshops designed to build expertise in open science practices, data management, and the use of advanced technologies in environmental research. Special focus should be given to training on the use of federated APIs, data integration tools, and the implementation of SKGs.

Open Science and Innovation

- **Open Access to Environmental Data:** Promoting and implementing open access policies that ensure the broad public availability of environmental data, e.g., for supporting the computing different essential climate variables. This approach

supports greater societal engagement and enables interdisciplinary research efforts that address global environmental challenges.

- **Innovative Tool Development:**To enhance effectiveness, the ENVRI community needs to collaboratively continue to develop cutting-edge analytical tools and services that address specific challenges in environmental sciences, including real-time climate change modelling, large-scale biodiversity assessments, and comprehensive ecosystem monitoring.

Policy and Governance

- **Regulatory Compliance Support:** Providing comprehensive guidance and resources to help researchers navigate complex legal frameworks related to environmental data, including regulations on data protection, biodiversity rights, and environmental impact assessments.
- **Ethical Standards for Environmental Research:** Establishing and promoting clear ethical guidelines for conducting environmental research, particularly in studies involving protected areas and vulnerable ecosystems. This includes ensuring that all data and methodologies adhere to high ethical standards.

Sustainable Funding Models

- In order to continue their collaboration, ENVRI community requires sustainable funding models to ensure the ongoing operation, maintenance, and development of its research infrastructures, supporting their long-term growth and adaptability to evolving scientific needs. This collective effort is needed to strengthen ENVRI's overall capabilities in processing complex environmental data and facilitate innovative solutions for advancing scientific research and informed policy-making.
- **ENVRI EOSC node:** ENVRI requires funding to support the potential construction of an ENVRI EOSC node. In order to pursue this achievement, it is crucial, firstly, to have a clear definition of what constitutes a node within the EOSC framework, and secondly, EOSC needs to become a stable and well-defined entity rather than a constantly shifting objective. A clear and consistent framework will ensure reliability and enable long-term planning, implementation, and integration efforts, making it easier for stakeholders like ENVRI to align their goals and resources effectively.

By addressing these specific needs the ENVRI cluster will be well-positioned to maximise its impact on environmental science and policy-making. These advancements will support ENVRI's integration into the broader European Open Science Cloud, ensuring that it contributes effectively to both scientific innovation and societal needs.

d. Life science specific needs

The European Life Sciences Research Infrastructures (LS RIs) support a diverse and multidisciplinary community, each with specific needs that are essential for advancing research, collaboration, and innovation. These needs vary depending on the focus of the research infrastructure, the type of data and resources they manage, and the objectives of the scientific community they serve. Below are some LSRI cluster needs identified (this work is being advanced in a manuscript in preparation to categorise and combine in some instances):

- More advanced and well recognized Standardised Data Formats: Different life sciences communities, such as genomics, proteomics, and imaging, require data to be stored in standardised formats to facilitate sharing and reuse across disciplines.
- Interoperable Databases: Researchers need life sciences oriented databases that can easily communicate with one another, allowing for seamless integration and cross-referencing of data from different sources (with a focus on interoperable terminologies).
- Scalable Computing Resources: Life sciences research often involves large datasets, particularly in genomics, imaging, and systems biology. The community requires access to high-performance computing (HPC) facilities that can scale according to the computational demands of these projects.
- Secure Data Storage Solutions: Sensitive data, such as patient information and proprietary research findings as well as Dual Use Research Concerns (DURC) and resources, must be securely stored with robust access controls to meet ethical and legal requirements.
- Customised Bioinformatics Tools: Different research fields within life sciences need specialised bioinformatics tools for data analysis, including tools for sequence alignment, protein structure prediction, and metabolomic analysis.
- User-Friendly Interfaces: To ensure that researchers from various backgrounds can efficiently use these tools, there is a need for intuitive interfaces that do not require advanced computational skills.
- Targeted Training Programmes: Researchers need training tailored to their specific areas of interest, such as genomics, imaging, or clinical research, to effectively use the tools and resources provided by LS RIs.
- Capacity Building in Emerging Technologies: As new technologies, like CRISPR and AI-driven analytics, emerge, there is a need for ongoing education to ensure that researchers are proficient in these cutting-edge methodologies.

- Collaborative Research Platforms and Virtual Research Environments (VRE): There is a strong need for platforms that facilitate collaboration between researchers from different disciplines, institutions, and countries. These platforms should support project management, data sharing, and communication and VREs should support interoperability of data, workflows and terminologies.
- Networking Opportunities: The life sciences community benefits from conferences, workshops, and online forums that enable researchers to share their findings, discuss challenges, and form new collaborations.
- Support for Regulatory Compliance: Researchers need guidance and resources to ensure their work complies with European regulations, particularly concerning data protection (e.g., GDPR) and the ethical use of biological samples.
- Ethical Review and Oversight: There is a need for infrastructure that supports the ethical review of research proposals, especially in fields like clinical trials and genetic research.
- Incentives for Data Sharing: To promote open science, there is a need for systems that reward researchers for sharing their data, such as citation systems for datasets or recognition in research assessments.
- Tools for Ensuring Data FAIRness of life sciences data: Researchers need tools and support to ensure their data meet the FAIR (Findable, Accessible, Interoperable, Reusable) principles, facilitating broader use and impact.
- Bridging Basic and Clinical Research: There is a need for platforms that connect basic research findings with clinical applications, supporting translational medicine and the development of new therapies.
- Patient and Public Engagement: Infrastructures must support the inclusion of patients and the public in research processes, ensuring that research aligns with public needs and ethical standards.
- Long-Term Funding Models: The community requires sustainable funding models to ensure the continued operation and evolution of research infrastructures.
- Maintenance of Data Repositories: There is a need for reliable, long-term data repositories that ensure ongoing access to research data beyond the lifespan of individual projects.
- Tools for Environmental Monitoring: Specific infrastructures supporting environmental and agricultural sciences need tools for ecosystem monitoring, plant phenotyping, and climate impact studies.
- Interdisciplinary Data Integration: Researchers in these fields need to integrate data from diverse sources, such as remote sensing, climate data, and genomic studies, to address complex environmental challenges.

These community-specific needs highlight the diverse and evolving requirements of the LSRI community. Addressing these needs is crucial for maximising the impact of the European Life Sciences Research Infrastructures and fostering a collaborative, innovative, and sustainable research environment.

e. Photon and neutron science specific needs

The PaNOSC community consists of the support staff at the PaN RIs (scientists, IT, user office, etc.), the users (visiting scientists to the facilities) and data scientists interested in the open data acquired and available in the PaN RIs data portals.

The needs are thus different between these three categories within the community and so must be the strategies to address them and engagement.

Key support staff at PaN RIs are

1. the scientists that guide the users onsite to help them use the facilities' instrumentation to perform their experiments
2. the user office that engages with them when applying for beamtime (to perform their experiments)
3. the members of the IT teams that manage the research data through its lifecycle and offer the IT services that enable the data analysis.

It is worth noting that not all PaN RIs have the same level of maturity when it comes to competences that support open science, and among them, the key needs are:

- Data policy and data management
    - To provide the right environment and tools to enable FAIR data and Open Science
- FAIR principles awareness and implementation
    - To ensure that users learn about FAIR and how to make their data FAIRer.
- Persistent Identifiers and DOIs
    - To ensure that users cite the DOIs in their papers.
- Data processing
    - To guide users on the tools and services available to analyse the experimental data
- Metadata & ontologies
    - To support the inclusion of relevant metadata on the datasets

    - Data Stewards to support the implementation of best practices by scientists
- Data Management Plans
    - To support the users' creation and update of their DMPs
- Virtual Research Environment
    - To provide a single stop shop where to analyse the data with all the relevant software. VISA is the VRE used by PaNOSC.
- Data Transfer
    - To enable large datasets to be moved (e.g. to the university of the user)

Users and data scientists on the other hand have similar needs, in particular Dr. Bridget Murphy, Vice President of the European Synchrotron and FEL User Organisation (ESUO) highlighted that "user friendly tools for (meta)data capture, storage and analysis are key to achieving open data. Some important aspects are: sample PIDs, metadata, data formats, ELNs, experimental logs, repositories for raw and analysed data, reference data bases, a platform to share data reduction, analysis and visualisation software, along with standards for software production, verification and sharing. For users it is important that there is consistency across the sources." She also explained that "Education and training is vital not only for awareness and acceptance but also in order to encourage users to adapt the tools available".

f. Social sciences and humanities specific needs

The existing Competence Centre Building Blocks of the SSH CLOCC have been developed and curated using a bottom-up process, including regular SSH community consultation. They reflect the community-specific needs, which can be grouped into the following categories:

- expertise and knowledge exchange, including continuous professional development and training opportunities, activities, materials and tools in support of training;
- guidance and training regarding the use of AI in SSH education and research;
- guidance and best practices with regards to research data management, legal and ethical issues when handling sensitive data, data and metadata standards, citation practices for data and software;
- networking, cross-domain collaboration, awareness raising and advocacy;
- guidance on how to demonstrate that SSH research has a tangible impact;
- tools for engaging with non-academic stakeholders and co-creating new values for society to respond to global challenges.
- harmonisation of vocabularies used in the SSH communities

- support for multilinguality
- publication support and open access
- adequate information and outreach regarding the availability of SSH support, services and tools;
- appropriate support and sustainable funding for developing and maintaining domain-specific services and tools.

4. Competence gaps

The five European Research Infrastructure Science Clusters (SCs) have identified several user needs that are not yet adequately addressed by existing competences, highlighting key gaps to be filled. One prominent need is for more advanced training in Open Science practices, such as data sharing and open access publishing, beyond current offerings. While all SCs demonstrate strong capabilities in data management, users are seeking enhanced support for integrating FAIR principles across diverse disciplines.

Another gap lies in the provision of interoperable tools for data analysis and reuse, which are not yet fully developed to meet the growing complexity of scientific research. Additionally, many users require more tailored support for complying with evolving Open Science policies and legal frameworks.

Addressing these unmet needs will require the development of more sophisticated infrastructure, improved cross-cluster collaboration, and expanded training programmes to ensure that all clusters and their communities can fully participate in the European Open Science Cloud (EOSC).

5. Conclusion and next steps

Work has commenced on establishing the five Cluster Competence Centres (CLOCCs), each designed to meet the specific needs of their scientific communities while addressing key transversal topics such as Open Science and FAIR data principles. These CLOCCs will serve as vital hubs of expertise, facilitating knowledge-sharing across clusters and enhancing support for Open Science practices.

During the OSCARS project, we expect the CLOCCs to maximise the sharing of available competences and actively address the evolving needs of their respective communities. In particular, the CLOCCs will focus on developing advanced training programmes, improving data management capabilities, and fostering collaboration between clusters to drive innovation.

Next Steps

1. **Tailored Development**: Continue building CLOCCs with a dual focus—addressing the specific needs of individual scientific communities while ensuring alignment with broader Open Science objectives and FAIR data standards.
2. **Roadmap**: The Science Clusters will collaborate to create a draft roadmap of activities and links to be established between the CLOCCs, their scientific communities and other stakeholders.
3. **Competence Sharing**: Promote greater collaboration across the five clusters to ensure that best practices, tools, and expertise are shared effectively, minimising duplication and maximising impact.
4. **Position Papers**: Each Science Cluster will prepare detailed position papers outlining their long-term operational and funding requirements. These papers will play a critical role in securing sustainable funding and ensuring the CLOCCs' long-term viability.
5. **Integration into EOSC**: As the CLOCCs evolve, their contributions will become increasingly important within the European Open Science Cloud (EOSC) ecosystem, helping to build a more robust, interconnected infrastructure for European research.

By following these steps, the CLOCCs will become important pillars of the EOSC, addressing both current and future challenges in Open Science and data management.