Funded by
the European Union

TOWARD FULL IMPLEMENTATION OF THE KM3NeT RESEARCH INFRASTRUCTURE

KM3NeT-INFRADEV2 – HORIZON – 101079679

Report on ICT development, data management and open science

KM3NeT-INFRADEV2 GRANT AGREEMENT DELIVERABLE: D4.3	
Document identifier:	KM3NeT-INFRADEV2-WP4-D4.3-vfinal
Date:	31/12/2025
Work package:	WP4 Data Management and Open Science
Lead partner:	FAU
Document status:	FINAL
Dissemination level:	PUBLIC
Document link:	https://www.km3net.org/km3net-eu-projects/km3net-infradev2/infradev2-outputs/

ABSTRACT

This deliverable presents the outcomes of Work Package 4 (WP4), focusing on Data Management and Open Science for the KM3NeT neutrino telescope. A distributed computing and storage infrastructure was deployed, integrating Rucio, DIRAC, CVMFS, and containerized workflows to ensure reproducible large-scale data processing. The KM3NeT Open Science System was established to provide FAIR access to data, software, and workflows, including a dedicated Open Data Center (ODC), reproducible analysis repositories, and educational resources. In parallel, real-time alert pipelines were developed for multi-messenger astronomy, enabling rapid dissemination of astrophysical neutrino events via the General Coordinates Network.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the REA can be held responsible for them.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner and WP	Date
From		FAU, WP4	01/12/2025
Author(s)	Vincent Cecchini Francisco Salesa Greus Jutta Schnabel Francisco Vazquez de Sola	IFIC, WP4 IFIC, WP4 FAU, WP4 NWO-I, WP4	
Reviewed by	Victoria Ciarlet Thaon Laurent Aphecetche Riccardo Bruno	CNRS, WP1	24/11/2025
Approved by	Paschal Coyle KM3NeT IB	CNRS, WP1	31/12/2025

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	15/10/2025	WP4 draft	T. Gal, K. Graf, J. Schnabel, V. Cecchini, F. Salesa Greus, F. Vazquez de Sola
2	21/11/2025	Version ready for review	T. Gal, K. Graf, J. Schnabel, V. Cecchini, F. Salesa Greus, F. Vazquez de Sola
3	18/12/2025	Reviewed version	
4	31/12/2025	Final version	

IV. APPLICATION AREA

This document is a formal deliverable of the Grant Agreement of the project, applicable to all members of the KM3NeT-INFRADEV2 project, beneficiaries and third parties, as well as its collaborating projects.

V. TERMINOLOGY

AAI	Authentication and Authorization Infrastructure
ACME	Astronomy Cloud-enabled Multi-messenger Environment (EU project)
ANTARES	Astronomy with a Neutrino Telescope and Abyss environmental RESearch
ARCA	Astroparticle Research with Cosmics in the Abyss
CERN	Conseil Européen pour la Recherche Nucléaire (European Organization for Nuclear Research)
CCLyon	Centre de Calcul de l'IN2P3 (Lyon Computing Center)
CSA	Coordination and Support Action
CTA	Cherenkov Telescope Array
CVMFS	CERN Virtual Machine File System
DL	Data Level
DL3	Data Level 3 (science-ready data format)
DIRAC	Distributed Infrastructure with Remote Agent Control
DMP	Data Management Plan
EGI	European Grid Infrastructure
EOSC	European Open Science Cloud
ESCAPE	European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures
GCN	General Coordinates Network
GNN	Global Neutrino Network
HEIG	High Energy Interest Group in the IVOA
HPC	High-Performance Computing
HTC	High-Throughput Computing
ICT	Information and Communications Technology
IVOA	International Virtual Observatory Alliance
KM3NeT	Cubic Kilometre Neutrino Telescope
LHC	Large Hadron Collider
MoU	Memorandum of Understanding
ODC	Open Data Center
ORCA	Oscillation Research with Cosmics in the Abyss (KM3NeT low-energy detector)
PB	Petabyte
Rucio	Distributed Data Management System
RI	Research Infrastructure
VO	Virtual Observatory
VODF	Very high-energy Open Data Format

WLCG	Worldwide LHC Computing Grid
WP	Work Package
Zenodo	Research data repository operated by CERN

VI. LIST OF FIGURES

VII. PROJECT SUMMARY

The Kilometre Cube Neutrino Telescope (KM3NeT) is a large Research Infrastructure (RI) comprising a network of deep-sea neutrino telescopes in the Mediterranean Sea with user ports for Earth and sea science instrumentation. During the EU-funded Design Study (2006-2010) and Preparatory Phase (2008-2012), a cost-effective technology was developed, deep-sea sites were selected and the Collaboration was formed in 2013. This proposal constitutes a second INFRADEV project dedicated to KM3NeT in order to implement an efficient framework for mass production of KM3NeT components, accelerate completion of its construction and provide a sustainable solution for the operation of the RI during ten or more years. Following the appearance of KM3NeT on the 2016 ESFRI Roadmap and in line with the recommendations of the Assessment Expert Group, this project addresses the Coordination and Support Actions (CSA) to prepare a legal entity for KM3NeT, accelerate its implementation, establish open access to the RI and its data, and ensure its sustainability by implementing an environment-friendly operation mode and evaluating the Collaboration socio-economic impact.

VIII. EXECUTIVE SUMMARY

The INFRADEV2 project has delivered the essential technological foundations for a robust, scalable, and interoperable data-management and open-science ecosystem for KM3NeT. Significant progress was achieved in building a distributed infrastructure capable of supporting the rapidly growing data volumes from the ORCA and ARCA detectors. Key components—including Rucio for distributed data management, DIRAC for workload orchestration, CVMFS for software distribution, and a Snakemake-based workflow system—were deployed and validated across European e-infrastructures. Together, these systems will enable KM3NeT to store, process, and replicate its data reliably while supporting diverse user communities.

A dedicated Open Data Center (ODC), implemented via a Dataverse instance, now serves as the central platform for FAIR, long-term access to KM3NeT's public research products. The Collaboration also established a structured open-science review process to ensure high-quality, reproducible, and well-documented public releases. In parallel, major advancements were made in multi-messenger readiness: KM3NeT's online alert infrastructure was successfully integrated with the General Coordinates Network, enabling rapid dissemination of neutrino alerts using community-standard formats.

The work carried out has also revealed the requirements for future scalability. Long-term projections indicate that KM3NeT will reach the 100-PB storage scale and require increasing compute capacity, reinforcing the need for expanding Tier-1 resources and automating data workflows. Authentication and authorization emerged as a critical area for further consolidation, with token-based access through IndigoIAM identified as a key step for ensuring future interoperability. Additional development of the Open Science System, including adoption of VODF/IVOA standards and Rucio-based access for large public releases, is planned to align KM3NeT with European Open Science Cloud practices.

Overall, the project has successfully met its main objectives and positioned KM3NeT on a strong technological footing for long-term data taking, open-science dissemination, and integration into the global multi-messenger astronomy community.

IX. TABLE OF CONTENTS

1. Introduction

The computing and data processing environment of the KM3NeT Collaboration centers around the processing and analysis of the scientific data produced by the ARCA and ORCA detectors. To fully exploit this data for the full data-taking time of the detectors and beyond, not only a solid data-processing scheme for large volumes of data is needed, but also an infrastructure to suitably access the data for both KM3NeT Collaboration members and outside researchers, and to facilitate the development of diverse analysis pipelines.

In Work Package 4 (WP4) on *Data Management and Open Science*, the project has developed and implemented the foundations for a sustainable, scalable, and interoperable ICT ecosystem supporting KM3NeT's scientific mission. The work package addressed three major objectives: establishing robust workflows for data management, processing, and storage across Grid and Cloud resources; developing data models and software frameworks enabling asynchronous, multi-instrument open-data analyses; and ensuring KM3NeT's integration into global online multi-messenger networks. Activities were structured into three tasks, each targeting a different component of the data life cycle—from raw data handling and user access, to open-science dissemination, to real-time alert generation.

Workflow and Data Management

Task 4.1 concentrated on building a coherent workflow and data-distribution system to support large-scale processing in coordination with European e-Infrastructures. Key achievements include the deployment of Rucio for distributed file management and the piloting of DIRAC for distributed computing, enabling reliable data placement and flexible workload execution across heterogeneous Grid resources. Central data processing was consolidated within a Snakemake-based workflow, improving reproducibility and operational consistency. Furthermore, prototype user access to the system was established through the integration of Indigo IAM for authentication and authorization.

Open Science System

Task 4.2 focused on the development and validation of KM3NeT's open-data ecosystem, covering data generation, quality assurance, archiving, documentation, and access procedures. A major result of this task is the establishment of an Open Data Center based on Dataverse, which provides robust facilities for multifaceted data storage and metadata management. The task also delivered exemplary data sets accompanied by user documentation to facilitate external scientific use, and contributed to the development of data-format standards for high-energy astroparticle physics, enhancing interoperability with other experiments.

Global Multi-Messenger Alert System

Task 4.3 developed the online data-management components required for the Collaboration's participation in global multi-messenger networks. Within this task, two online event-selection pipelines were implemented: one for the identification of single high-energy events and another for multiplet detection, enabling the generation of astrophysical alerts. A further major achievement was the integration of KM3NeT alerts into the General Coordinates Network (GCN), ensuring rapid dissemination of information and establishing KM3NeT as a fully contributing partner in real-time multi-messenger astronomy.

In all three tasks, the goals of the INFRADEV2 project could be reached, enabling an infrastructure that shows the basic capabilities to handle the challenge of decades of data taking and analysis as envisioned by the KM3NeT Collaboration. However, further developments and efforts to roll out the services to their full capacity are needed and will be outlined below.

2. Data management, processing and storage

One of the first deliverables of the INFRADEV2 project was to produce a detailed Data Management Plan, based on the original plan created as an output of the first INFRADEV project. The document gathered information on the data to be collected and processed, and the relevant standards and methodologies to be applied. The resulting document was reviewed in Q3 2023 by a panel of experts in the field, with an updated version provided by Q1 2024. It served as the guideline for the developments described in this document.

2.1 Requirements for multisite computing

Raw data production

The detailed description of our data types can be found in the Data Management Plan. The largest by data rate is the raw data coming from the detector itself (including optical, acoustic and calibration data), at about 300 Mb/s per DU - eventually amounting to about 10 Gb/s per BB. The shore stations perform online filtering which reduces this data stream by a factor of about 10^4, reducing the data rate to around 0.5TB/day, or O(100TB)/year. This raw data is passed from the shore stations to the Tier 1 computing sites, see Figure 1.

Tiered processing

As reported in the DMP, data processing is structured in a tiered approach following the CERN categorization, with Tier 1 sites handling the central data processing and data storage, and multiple Tier 2 sites being available for additional high-level data analysis, machine learning, simulation studies and other computing-intensive activities. Therefore, distributed computing is currently being implemented for Tier 1 infrastructures which consist of CC-IN2P3 (FR), CNAF (IT) and SURF-SARA (NL). Tier 2 computing sites are usually offered by member institutes to a limited number of users, and require access to data and software. Tier-2 will rely on in-kind contributions from the sites hosting the Collaboration members interested in performing advanced analyses. This is expected to be a higher cost in computing than storage resources.

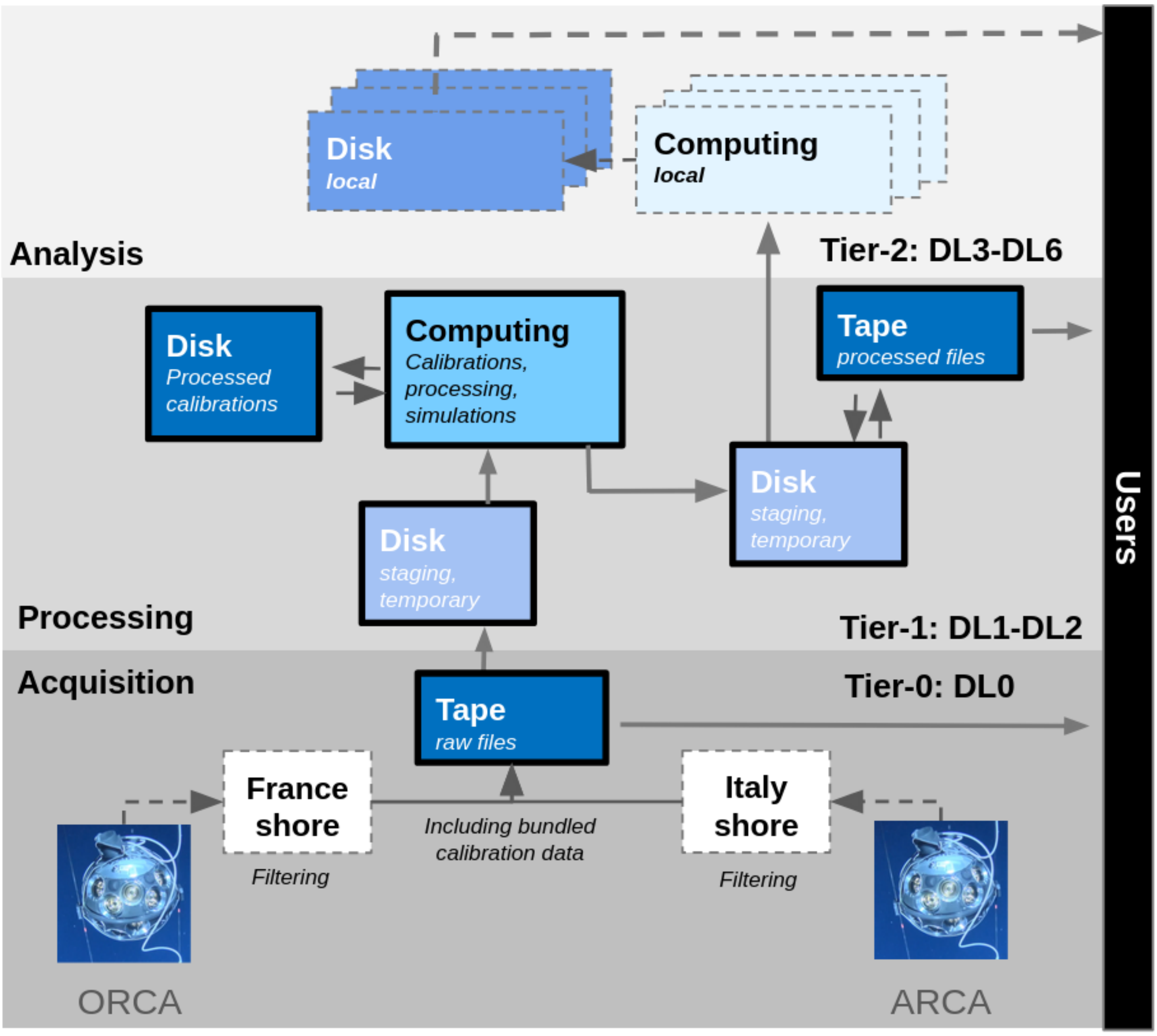

Figure 1: KM3NeT's data flow using a distributed infrastructure. Solid lines and squares represent usage of distributed resources through Rucio or Dirac, dashed ones represent local resources. Projections for storage

The projected storage needs for KM3NeT increase substantially over the coming years, driven by the annual data rate from the ARCA and ORCA detectors—of order petabytes when combining raw data and simulation. To ensure long-term data preservation and reliability, replication is required according to data criticality: non-reproducible data must be stored with two additional copies, while important but reproducible data require at least one extra copy. Altogether, this leads to processed and simulated data sets requiring approximately four times as much storage as the raw data itself, considering also multiple versions of some key data. Additional storage demand arises from archived copies of earlier processing stages, as required by the experiment's data-retention rules, as well as from instrument-related data and future simulation campaigns. Taken together, these factors imply a rapid growth of storage needs, reaching the order of 100 PB of total storage once the detectors are in full operational mode, which requires large resources especially for tape storage.

Projections for computing

KM3NeT's computing needs are expected to grow significantly as the detectors reach full operational capacity. While extrapolation from current processing needs is difficult, processing the annual data volume from the complete detector configuration will require on the order of 10 MHS06h. This baseline demand is further amplified by the necessity to reprocess the data multiple times—typically two to three iterations—to incorporate updated calibrations and improved reconstruction algorithms. Additional computational load arises from increasingly complex analysis workflows, including those employing machine-learning techniques with high resource requirements. Altogether, these factors imply that in steady-state operation, the computing needs of KM3NeT will easily exceed several 100 MHS06 hours per year for all Tier 1 and Tier 2 computing.

2.2 Implementation of the distributed infrastructure

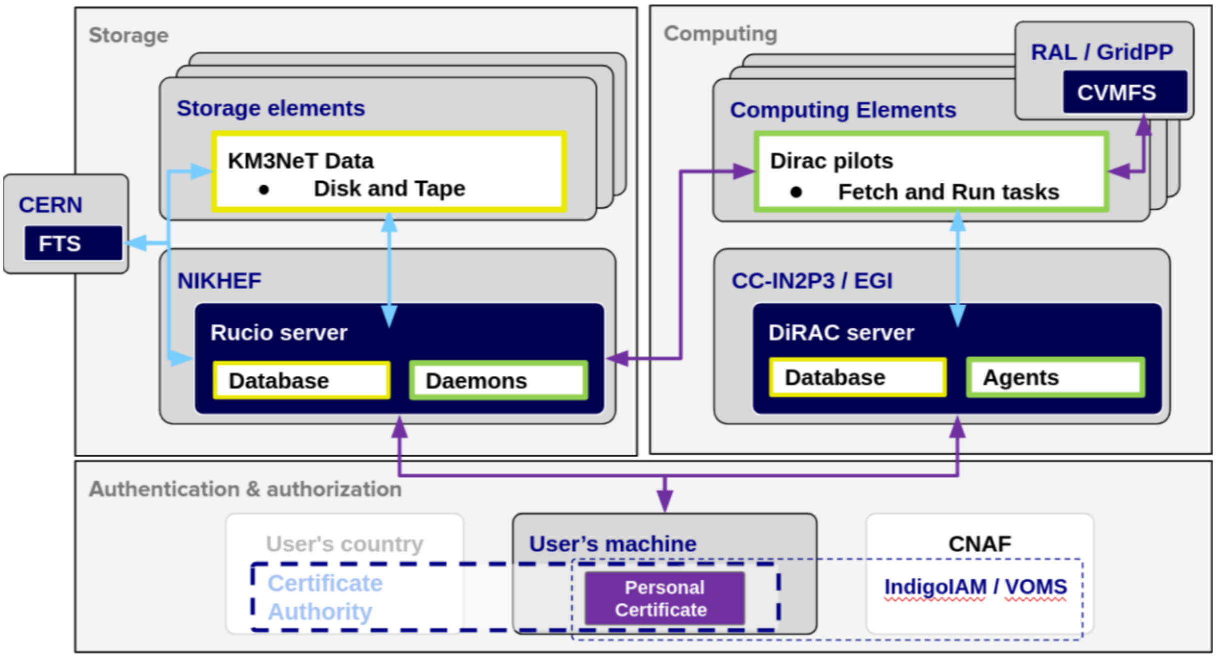

Figure 2: KM3NeT's distributed computing infrastructure. Purple arrows indicate connections made explicitly by the user, while cyan arrows indicate connections managed automatically by Dirac and Rucio.

Distributed Storage

The project established a functional distributed storage system centered on a dedicated Rucio instance deployed on a Kubernetes-managed server at Nikhef. This instance, provided as an in-kind contribution through a Dutch eScience project, now serves as the central interface for all KM3NeT data. Rucio's data identifiers, dataset and container hierarchy, and metadata system allow users to locate and handle files without knowledge of underlying storage paths or transfer protocols. A significant achievement was the ingestion of existing tape and disk data from CC-IN2P3 directly into Rucio, together with the migration of calibration data previously stored in GitLab LFS to grid-accessible storage. As a result, the full set of inputs required for data processing—raw files, calibration information, and a large fraction of already-generated Tier-1 products—is now discoverable and accessible through the Rucio catalogue. Writeable storage endpoints have also been configured for receiving future Tier-1 production output.

Automated interaction with CERN's FTS service was established in connection with Rucio to handle background data transfers, especially replication between disk and tape storage and enforcement of the desired number of replicas. Cron-jobs were configured for automatic ingestion of new Tier-0 and Tier-1 data and for cleaning up single-file rules to prevent unnecessary load on the server, ensuring that files remain registered only if grouped into datasets.

Software Deployment

CVMFS remained the backbone of software deployment across KM3NeT's distributed computing environment. The Collaboration relied on the CVMFS server hosted at RAL by GridPP, which ensured uniform availability of software environments needed for reading raw data and executing processing steps at all partner sites. Containers used in these environments were generated through GitLab CI/CD pipelines, which also carried out automated testing and documentation generation. The resulting images were stored in the KM3NeT Docker registry and are now provided via the KM3NeT Gitlab Docker registry, with public access granted only to selected projects while internal containers remained protected behind Collaboration authentication.

As workflows evolve, further improvements may be required to streamline the release process, strengthen integration between containers and workflow management tools, and automate quality-control mechanisms for future analysis environments.

Distributed Computing

The project integrated the EGI DIRAC service as the primary framework for managing KM3NeT's distributed computing resources. Users can submit jobs specifying execution scripts and resource needs, and DIRAC matches these with available Computing Elements through pilot-based scheduling. Because DIRAC does not provide an in-depth data and software management environment, the integration of CVMFS for software distribution and Rucio for handling large input and output files was essential. Only small files can be handled through the DIRAC sandbox, making coordination with Rucio a core part of the computing model. EGI's commitment to maintaining the DIRAC server and providing user support represents a long-term in-kind contribution to KM3NeT.

A major technical development was the extension of the Snakemake-based data processing workflow to operate within this distributed environment. Rules were added to automatically upload and download files via Rucio, enabling full processing chains—including event triggering, simulation, calibration, and reconstruction—to run on worker nodes. A fallback mechanism for output uploads ensures that files are not lost in the event of Rucio transfer failures. This proof-of-concept system was successfully demonstrated in early 2025 during a collaboration-wide training workshop.

Authentication

Work toward a unified Authentication and Authorization Infrastructure (AAI) focused on deploying an IndigoIAM instance at CNAF to act as an OIDC token provider for the Collaboration. The intended setup would combine OIDC tokens with existing KM3NeT SAML authentication and draw membership and authorization information directly from the KM3NeT user database, creating a single source of truth for collaboration identity management. While the service is technically prepared, it has yet not been fully adopted operationally, as a formal MoU between KM3NeT and CNAF and implementation of eduGAIN for KM3NeT user identification vis-a-vis the CNAF user management is still pending.

During the project, temporary reliance on X.509 certificates remained necessary. Members obtained certificates from national Certificate Authorities where possible, or from CERN if they held a full CERN account. Authorization was handled through IndigoIAM acting as a VOMS service, which required users to hold an INFN account. Rucio and DIRAC currently maintain separate user databases, with new users added manually by maintainers.

2.3 Current challenges and future developments

Actions following the DMP review

With the above presented setup, some of the core goals of the actions pointed out in the review of the DMP were reached. These include

- To improve the data processing efficiency starting from the bottlenecks of the data processing chain, which was addressed in the Collaboration by introducing a run-based data processing scheme based on snakemake which is now enabled in the distributed computing setup;
- To create a legal agreement with the main computing centres and consider becoming an observer to the LHCONE and WLCG collaboration, which was pursued by joining WLCG currently by participating to open meetings, and establishing MoUs with EGI and the computing centres individually;
- To implement the outcomes of the ESCAPE project if possible, and check the use of other CERN-related software for KM3NeT, which is followed up by KM3NeT as a member of the ESCAPE collaboration through actively engaging in the ESCAPE Data Lake developments and common workshops on Rucio and IndigoIAM;
- To set up monitoring of services and resources, which will be included in the services offered by EGI;
- To review the current status of AAI in KM3NeT, investigate possibly unmet needs and corrective actions with the goal of smooth operation in a fully Distributed Computing environment, which is addressed in the implementation of IndigoIAM.

Current challenges

Implementation challenges for distributed storage remain with the scaling of the storage and computing pledges in years to come to ensure sufficient capacities at the Tier 1 sites. On a shorter timescale, the full data management schema which is facilitated by the Rucio file catalogue needs to be clarified to ensure that ingestion workflows, legacy data, calibration databases, and newly produced files are consistently referenced within the system.

For distributed computing, several challenges limited broader adoption within the project timeline. Additionally, tight coupling between DIRAC, Snakemake, and Rucio will be required to work around job queue length limits, which will constrain us as our data volumes increase with the detector. This point requires continued development to ensure stability and maintainability.

Last but not least, implementation challenges were substantial for authentication. The unexpected termination of the Sectigo contract for European CAs at the end of 2024 left many Collaboration members without access to certificates for several months, until a new agreement with HARICA was established. This interruption underscored the urgency of transitioning to a future robust token-based system. Additional complications stem from the uneven pace of adoption of OIDC tokens across grid sites, requiring KM3NeT to maintain X.509 support for the foreseeable future, which also poses a problem for the use of Rucio by all KM3NeT members who then require a certificate for data access.

Future developments

While KM3NeT was able to take a huge step towards a distributed computing setup, full adoption of the implemented services and future developments will be essential to ensure the smooth operation of scientific analyses in KM3NeT in the future. These developments will be pursued in close exchange with other research infrastructures of similar scale and computing needs, inter alia in the context of the ESCAPE research cluster, and by adapting those technologies to the specific needs in KM3NeT.

The framework for the expansion of the computing infrastructure is given by legal agreements which are currently drafted after the establishment of the AISBL legal entity for KM3NeT. These include on the one hand slim but functional MoUs for Tier 2 computing providers to ensure data security for institutes willing to contribute their resources to KM3NeT computing, and on the other hand individual and detailed MoUs with the Tier 1 computing sites to ensure adequate services are provided for a stable operation of the KM3NeT data processing. An already signed MoU with EGI addresses the seamless integration of the infrastructure shared between the contributing Tier 1 sites.

Future developments for distributed storage will focus on completing replication policies once sufficient storage is available, expanding automation, and ensuring long-term maintainability of the Rucio instance as data volumes increase toward full detector operation. Stress tests like those coordinated in ESCAPE data challenges will serve to ensure stable running in the future, while a long-term pledge to the hosting of the Rucio instance and human resources to maintain it will further be a key point for the sustainability of the computing setup in KM3NeT.

In distributed computing, future developments include the integration of a new [Rucio–Snakemake plugin] that will eliminate the need for custom transfer rules and make workflows natively aware of file locations in distributed storage. This will allow more efficient job splitting and improved tracking of workflow state across grid sites.

For authentication, the priority lies with enabling IndigoIAM to function as a full authentication provider for the Collaboration by fully implementing the SAML interface to the KM3NeT database. Once established, the system will allow automatic synchronization with Rucio and DIRAC databases, eliminate dependency on national CAs, and support the broader transition of the grid community toward token-based authentication. Especially the interplay between the AAI service and Rucio will need developments to enable a fine-grained user access and role management, authentication without grid certificates, and providing a user-friendly computing environment reaching as far as enabling open science as described in the next section. As similar developments are pursued e.g. in CTAO and by other ESCAPE partners, KM3NeT can benefit hugely from broader support that is expected to be given to these shared technologies.

3. Data, software and services for Open Science

The KM3NeT Open Science System is intended to provide an integrated framework for FAIR (Findable, Accessible, Interoperable, Reusable) access to scientific data, software, and workflows. It connects data products, interfaces, analysis environments, and aggregators within a coherent open-science infrastructure.

During the first INFRADEV project, a prototype setup for Open Science had been established and was described in the DMP. This served as the basis for the efforts to validate and develop the system further, covering data generation along specific use cases, establishing quality assurance, documentation, and dedicated access paths.

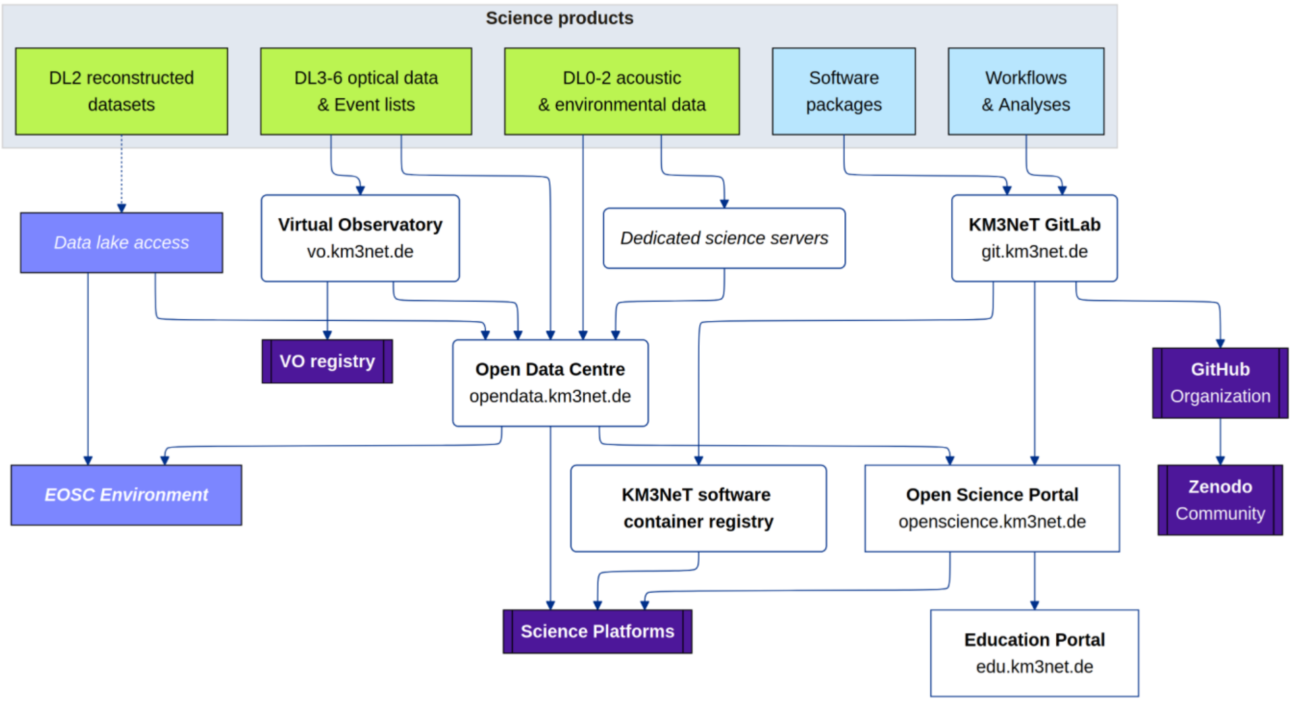

Figure 3: Overview of the KM3NeT Open Science System.

3.1 The Open Science System

Science Products

The KM3NeT Open Science System encompasses a broad range of science products that extend beyond reconstructed data. In the context of open science, these products include datasets, software, and complete scientific workflows. KM3NeT considers multiple classes of data for open access, reflecting different stages of the analysis process and various scientific objectives. Fully reconstructed datasets form the core of these products, representing standard particle-detection results derived from optical hit patterns and containing reconstructed particle properties. Environmental and acoustic data, collected for calibration and multidisciplinary science use cases, provide additional information relevant to the scientific exploitation of the detector. High-level, event-based analysis products, derived from simulations or advanced workflows, supplement KM3NeT's published scientific results and offer users deeper insight into the analysis chain.

In parallel, open-source software is released to support reading, processing, and analysing scientific data. This includes scientific software packages for particle simulation, specialised analysis tools, and full analysis frameworks. The Collaboration also publishes complete analysis repositories, which provide installation-ready environments, data-access tools, and the processing and analysis scripts required to reproduce KM3NeT scientific results. These repositories integrate data, software, workflows, and documentation designed for reproducibility of scientific results.

Interfaces

A set of interconnected interfaces forms the operational backbone of the KM3NeT Open Science System. The KM3NeT Open Data Center (ODC), which was at the focus of the infrastructure development in this work package, serves as the central hub, hosting the Collaboration's open datasets and associated metadata. Through its search and API-based access mechanisms, the ODC ensures that KM3NeT outputs conform to FAIR principles. Complementing this, the Virtual Observatory (VO) server provides access to astrophysical and neutrino datasets via IVOA-compliant protocols.

Software development, documentation and maintenance of analysis tools are carried out within the KM3NeT GitLab, which also hosts the analysis repositories provided to the community. Continuous integration pipelines automatically generate and validate containerised software environments, which are published both in the KM3NeT Docker registry and partially on CVMFS. These services enable users to execute reproducible workflows independent of local computing environments.

Aggregators

To maximise global visibility and integration, KM3NeT relies on several external aggregators that facilitate the integration with established research platforms. Public software repositories in GitLab are mirrored to the KM3NeT organisation at GitHub, ensuring exposure to the wider open-source community and enabling collaborative development. Zenodo is used to archive datasets and software releases, assign persistent DOIs, and register KM3NeT outputs within long-term archiving infrastructures like the ESCAPE Open Software and Service Repository through the KM3NeT Zenodo community. The KM3NeT VO node is registered in the International Virtual Observatory Alliance Registry of Registries, ensuring that datasets and services hosted through the VO server become discoverable and interoperable within the global Virtual Observatory ecosystem.

User Access and Analysis Level

User-facing portals and platforms serve as the main entry points for researchers, educators, and citizen scientists engaging with the KM3NeT Open Science System. The Open Science Portal, implemented via GitLab Pages, consolidates documentation, policy information, and links to all open-data services, functioning as a central access hub for the ecosystem. Alongside it, the Education Center offers structured learning materials, from introductory scientific demonstrations to advanced research tutorials, enabling broad uptake of KM3NeT data in both educational and professional contexts.

Support for external analysis platforms further enhances accessibility. BinderHub and Jupyter-based services, such as MyBinder, allow users to run KM3NeT analysis repositories directly in a browser without installing local software, as repositories are configured to automatically launch the required analysis environment. Within the Virtual Observatory domain, tools such as TOPCAT and Aladin provide interactive exploration of tables, images, and sky maps, enabling users to integrate KM3NeT data with external astronomical catalogues. These services collectively ensure that the KM3NeT Open Science System is usable by diverse communities and can be integrated into scientific workflows.

3.2 Open Science Publications and Quality Assurance

Developing along use cases

The development of the KM3NeT system has been strongly guided by concrete use cases. In this context, data from the predecessor experiment ANTARES plays an important role, as it provides example data for KM3NeT by offering full data sets and high-level analysis outcomes from mid-2005 up until the last year of its operation in 2022, which the KM3NeT Collaboration has agreed to preserve alongside its own data.

The transition of ANTARES data into the KM3NeT ecosystem has been streamlined by the substantial overlap between the two collaborations, including shared researchers and infrastructure, as both projects utilize the CCLyon HTC and collaborative platforms such as GitLab. As a result, ANTARES data is treated as example datasets within the KM3NeT Data Management Plan (DMP) and can be used to develop open access solutions to large data sets from Rucio. Additionally, ANTARES software and data products have been preserved and will be adapted to ensure continuity in Open Science practices, including conversion to KM3NeT data formats for broader dissemination alongside KM3NeT data. This approach addresses also the challenge of long-term preservation by calling for strategies to also ensure future accessibility to the ANTARES datasets alongside KM3NeT's own observational data.

Apart from the data and workflows from ANTARES, KM3NeT has published various analysis repositories and software which are detailed below and served to improve publication and quality procedures.

Exemplary Open Science Products

- The **KM3NeT-CTA analysis** of sources in the Milky Way represents the first fully reproducible analysis publication within the KM3NeT Collaboration, demonstrating parallel publication of a scientific paper and its associated code repository in the multimessenger domain based on simulated data. It leverages Jupyter notebooks to reproduce plots and analytical results, and scripts for regeneration of pseudodata sets using batch computing. The analysis is publicly available on both GitHub and Zenodo.
- The **KM3-230213A repository** supplements the publication "Observation of the Ultra-high-energy Cosmic Neutrino with KM3NeT" in Nature, providing open data and methods to reproduce selected outcomes from the paper. The repository uses Jupyter notebooks to guide users through reading the provided data and exploring an interactive event display, offering hands-on access to the analysis of the PeV-range neutrino event. With clear instructions for environment setup and the option to run directly via MyBinder, KM3-230213A exemplifies reproducible research for selected experimental data from KM3NeT even before its full operational deployment.
- **gSeaGen** is a C++ application developed by the KM3NeT Collaboration for simulating high-energy neutrino-induced events detectable in neutrino telescopes. It generates events for all neutrino flavors, accounting for track-like and shower-like topologies, and simulates neutrinos from diffuse atmospheric sources as well as point-like or extended astrophysical sources. The code is maintained in a public GitLab repository, providing an essential tool for neutrino event simulation and detector studies.

The Open Data Center as main data hub

The KM3NeT Open Data Center (ODC) has been established as the main hub for all public science products, including data, software, and publications. During this project, the pilot ODC has been replaced with a sustainable solution building on an open source data repository software. During the evaluation phase, several alternatives for hosting and managing open science content were tested, including Zenodo, CKAN, and Dataverse. Based on assessments leveraging flexibility in metadata handling, long-term software support and test implementations with existing open science products, a dedicated Dataverse instance was deployed at opendata.km3net.de, capable of handling both open and restricted data sharing. The ODC is designed to serve as a central registry for KM3NeT scientific outcomes including publications, plots, and software alongside the data directly hosted or registered there. Key development tasks are ongoing, including authentication setup, defining standard entry types, and implementing wrappers for the API interface. Users can already access ODC entries via the openkm3 Python client, which also provides redirection to Virtual Observatory (VO) entries, ensuring interoperability with the full KM3NeT Open Science system.

Procedures and quality assurance

In order to ensure quality of the open science products, the Open Science Committee, established during the first INFRADEV project, set up a review process within KM3NeT that follows a structured checklist to ensure high-quality, reproducible, and accessible research outputs and is filled during testing of the product by the reviewers. Each repository is required to be publicly available on KM3NeT GitLab or another recognized platform, and the relevant scientific working group reviews and approves the content prior to publication. The repository structure and documentation are assessed, including the presence of a README with execution instructions, clearly organized folders (e.g., /src, /data), code comments, and an appropriate open license and citation file. Execution practices are reviewed to ensure analyses can be run independently, preferably using software packaging, notebooks with detailed descriptions, environment specifications (e.g. Conda), or containerization. Data handling is checked to guarantee that publicly available data are properly referenced, dummy data are provided when necessary, standard data formats are used, and relevant metadata are included. Finally, consistency between repository content and corresponding publications is verified, a DOI is assigned, and the repository is published in Zenodo with proper reference to related publications. Any issues identified during the review are tracked on GitLab and resolved before publication.

3.3 Future developments

As open science happens in a fluid environment that is strongly driven by large-scale projects and dedicated initiatives across multiple disciplines and scientific infrastructures, the future development of the Open Science System is largely guided by these collaborative efforts. The currently most significant initiatives are summarized below to give an outlook to anticipated future developments.

Data Lake and ESCAPE environment

Following similar developments in the ESCAPE community, and with the internal setup for a large-volume file catalogue using Rucio and by providing access via IndigoIAM already in place, the basic infrastructure is enabled to provide large-volume open data access in the future. In close connection with similar CERN-driven developments, KM3NeT will pursue its goal to provide access to full releases of KM3NeT particle event data as envisioned in its open science policy by establishing light-weight authentication through the AAI system and following the development of open-access Rucio endpoints for large data sets. However, with these developments aiming for a functioning setup by the end of the decade, the implementation falls outside the scope of this project, but can be expected to be in time for the full deployment of the KM3NeT detectors.

Virtual Observatory extension and VODF

As protocols in the IVOA are focused on observations of single sky objects which do not take into account the low event count rate and full sky observation approach of e.g. neutrino astronomy, the Virtual Observatory introduced a High Energy Interest Group (HEIG) to which KM3NeT contributes and which will define requirements for the representation of high energy astrophysics data in the VO. It focuses on the development of use cases for data discovery, access and visualization, identification of metadata concepts needed by high energy astrophysics data that are not currently supported by the VO, and contributes to updates and additions to the relevant parts of the IVOA standards framework.

In a related initiative, KM3NeT contributes to the development of the Very high energy Open Data Format (VODF) which is based on the Gamma Astronomy Data Format (GADF) and implements common standards for high-level scientific products from gamma ray and neutrino detectors such as instrument response functions and particle event lists. KM3NeT aims to provide its data in accordance with both IVOA and VODF standards once available.

Software developments and common frameworks

Common analyses call for integrated and co-developed software environments. Therefore, KM3NeT aims to foster common developments and both use and provide contributions to widely-used open scientific software, e.g. with the IceCube collaboration in frameworks like GraphNet for Neural Network development for neutrino telescopes or gammapy for astronomical analysis with gamma-ray astronomy. With the use of GitLab and Github, community-based development is easily facilitated, and software quality can be addressed by using tools based on Github actions or CICD pipelines for testing and evaluation.

Science Platforms

The VO services and data provided through the Open Data Center serve as a basis to integrate KM3NeT products in multi-messenger pipelines like the MMODA platform developed in the ACME project in the future. As the KM3NeT Open Science System is constructed such that it adheres as far as possible to standard data and infrastructure approaches widely used in the community, future developments for cross-experiment pipelines like this can be facilitated easily.

4. Integration into the global online multi-messenger network

KM3NeT is pursuing a real-time program to participate in multi-messenger astrophysics, by following the alerts issued by partner experiments and to send them early triggers of potential neutrinos of astrophysical origin. The follow-up part of the program, based on a triggered analysis of the KM3NeT data, has been addressed in section 3.6 in the report D4.1 on the Data Management Plan.

4.1 The KM3NeT online alert system

The KM3NeT online alert system refers to the standalone identification by KM3NeT of interesting neutrino candidates, and their broadcast to the multi-messenger community, on a timescale of about one minute. It gathers all the modules allowing for close to real-time reconstruction and classification of the data (common with the follow-up program), the various analysis pipelines to identify events with a possible astrophysical origin, a preliminary search for interesting counterparts, and the dissemination of the alert. Next to the neutrino candidate search, a supernova detection system (SN processor & module) is implemented, which falls outside the scope of this project.

The real-time reconstruction and classification of the data is done at the shore stations of ARCA and ORCA. The online events of both detectors are then written in a database hosted at the ORCA shore station, MM_EVENT_DB. For the record, in our computing system, MM_EVENT_DB stores the events reconstructed in close to real-time, while MM_DB stores the information related to the alerts (both for follow-up and alert sending).

The alert identification starts with a first step of pre-selection, where a set of generic cuts are applied (they differ only according to the detector, ARCA vs ORCA). All the events passing this pre-selection are then forwarded to the analysis modules. There are about 500 events/month/detector.

At the second stage of the system, the various analysis modules run in parallel. As input, they take the events selected by the pre-selection module and, based on the computation of a False Alarm Rate (FAR) for the given event, classify the event as a candidate alert or discard it. Each analysis module is tuned to select about 1 candidate alert/day.

On top of this analysis, a final selection module is in charge of

1. performing a fast simulation to assess the error region of the alert (and prepare its HEALPix skymap);
2. searching for astrophysical counterparts to this alert;
3. merging the results of the various analysis pipelines;

4. reducing the number of alert candidates to a few per month;
5. ensuring the safe saving of the alert in our computing facility.

The first two steps are run as individual submodules:

- The computation of the error region is done through a fast simulation of a few thousand events coming from the same direction as our alert, with a similar energy range.
- The search for astrophysical counterparts is done by cross-matching the sources found in the error region of our alert from various catalogs, with a focus on Blazars, GRBs, TDEs and µQuasars. An association probability is attributed to each source, depending on its angular distance to the alert, its flux/luminosity and, if available, timing/variability information.

Finally comes the broker part: the insertion of a new alert in the database MM_DB (by the final selection module) triggers the alert sending program. At first, the alert is retrieved from the DB, along with the associated event. This information is then formatted as JSON and VOEvent notices and sent to GCN. Additionally, an internal reporting of the alert is done by mail to a subscriber list, and through the KM3NeT Rocket.Chat service in the channel "Online physics alerts".

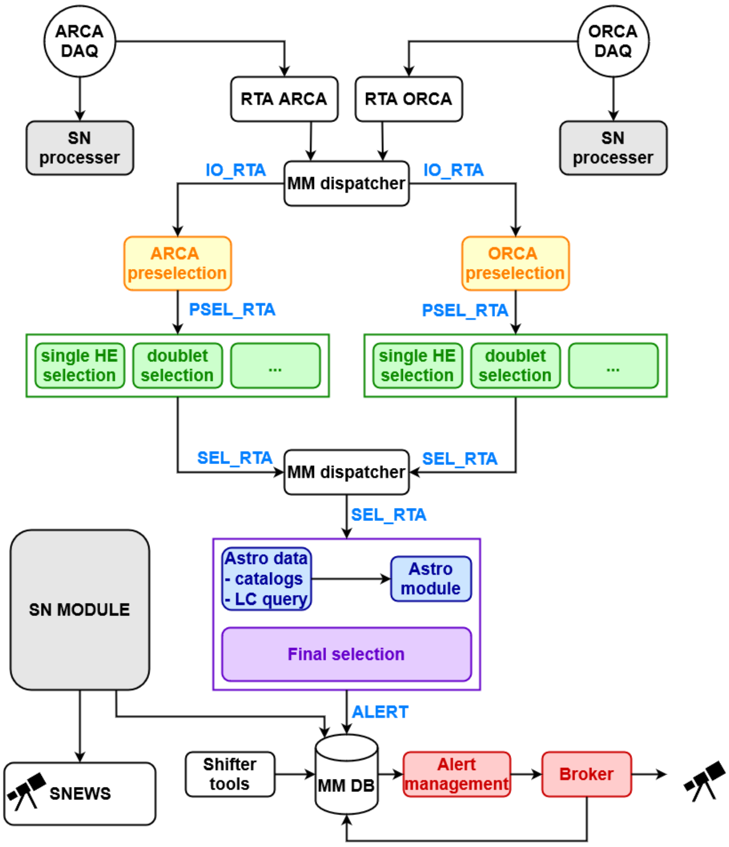

Figure 4: Overview of the KM3NeT Online Alert System.

4.2 Analyses workflows

The two pipelines under development are the "single high-energy" event selection and the "multiplet" selection. The first one aims at identifying cosmic neutrino candidate events by prioritizing the candidates according to their level of significance. The second one searches for events coming from compatible sky locations within a given time window. Both are data-driven, selecting the events having a low occurrence probability.

The decision whether to send an alert is based on the FAR — a number asserting the probability to have a more significant event over a given period. The alert threshold is set at 1 per month per pipeline: all events having an occurrence probability below 1 per month will be reported, i.e. 3.9e-7 Hz. Because 2 pipelines for 2 detectors are running, 4 alerts per month are expected on average. The FAR value of the alert is stated in the notice, allowing the final user to decide on a more stringent threshold to follow up our alerts, depending on their requirements.

If more than one pipeline is triggered by the same neutrino candidate, the one with the lower FAR is used to fill the notice.

4.3 Multi-messenger integration into global networks

The alerts will be publicly broadcast through GCN in both JSON and VOEvent formats, via the Kafka protocol. The JSON messages are sent with the Kafka topic *gcn.notices.km3net* and the VOEvent are sent to the topic *gcn.notices.km3net.voevent*.

The physical alerts are shared through the official server “gcn.nasa.gov” and test alerts are sent through the server “test.gcn.nasa.gov”.

For its core content, the JSON notice uses the "neutrino" schema defined in agreement with the IceCube Collaboration, to provide a unified template with variables relevant to neutrino telescopes. This schema is extended with information specific to KM3NeT, like the identified counterparts and additional variables related to the neutrino event. The schema and two examples for a singleHE event and a multiplet event can be found at the GCN notice definition for KM3NeT.

The same content will be provided in VOEvent notices, differing only in the embedding format.

4.4 Future developments

Several updates should be envisioned for the future. The first and most urgent one is the implementation of a quasi-online calibration, that is mandatory to minimize the angular error of our alerts. That would take into account the time and position calibration in the online reconstruction chain, to follow as closely as possible the evolution of the detector.

Additionally, the event-resimulation (from the error-region definition) could be used to set flux limits in our alerts. On top of that, it could be used to optimize the selections at the analysis level and/or to set up a refined analysis, based on likelihood techniques, to probe the astrophysical origin assumption.

In parallel, one could think of expanding the time-scale to search for multiple neutrinos of common origin, in a so-called “long-term analysis” pipeline. In the case where a neutrino is detected just above the alert threshold, being able to search for a significant excess over a long period of time would help to increase the significance of the current event, therefore enhancing the interest for additional follow-up. Then, one could think of the inclusion of cascade events in the analysis, despite their low angular resolution, in the various analyses searching for multiple neutrinos.

Still with the aim of cross-checking information and increasing the sample size, working with other neutrino telescopes, like the IceCube Collaboration, to develop a subthreshold stream is scientifically motivated, as long as a way is found to share the relevant information. Indeed, detecting more than one neutrino with two telescopes increases the significance of the detection, as it is when detecting more than one neutrino with multiplet/long-term pipelines.

Finally, it would be interesting to have a pipeline running our analysis on the data reconstructed with the regular calibration (sometimes called “offline” data) and the full run-by-run simulation, to refine the alerts and update the information in a published alert catalog.

5. Conclusion

Across Work Package 4, the INFRADEV2 project has successfully established the core foundations of a sustainable, scalable, and interoperable data-management and open-science ecosystem for KM3NeT. The project delivered major advancements in distributed data handling, open-science services, and real-time multi-messenger integration, enabling the Collaboration to manage the increasing data volumes of the ORCA and ARCA detectors and to participate fully in global astrophysical networks. A functional distributed computing and storage infrastructure was deployed through the integration of Rucio for data management, DIRAC for workload orchestration, and CVMFS for software distribution, together with a Snakemake-based workflow supporting reproducible processing across European e-

Infrastructures. The Open Data Center, implemented using a dedicated Dataverse instance, now provides a robust platform for FAIR access to KM3NeT's public datasets, software, and documentation, complemented by a structured quality-assurance and review process for all open-science products. In parallel, the project achieved a major milestone in multi-messenger astronomy by implementing KM3NeT's online alert system and integrating it into the General Coordinates Network, enabling the timely publication of neutrino alerts using community-standard formats.

The project's work has also clarified key requirements for the long-term evolution of KM3NeT's ICT environment. The expected rise in storage and computing demands—reaching the 100-PB and 100-MHS06/year scales, respectively—will require scaling of Tier-1 pledges, continued development of automated data-management workflows, and sustained support for the distributed processing framework. Authentication emerged as a critical area where further consolidation is needed, with the transition to token-based access through IndigoIAM identified as essential for future usability, security, and interoperability. Similar forward-looking developments were outlined for the Open Science System, including the adoption of community standards such as VODF and IVOA protocols, the extension of Rucio-based data access to large public releases, and deeper integration with cross-infrastructure science platforms developed in ESCAPE and ACME.

Overall, the project has met its main objectives and delivered the technological basis required to support decades of data taking and scientific exploitation by KM3NeT. The implemented systems demonstrate the feasibility, robustness, and integrability of the chosen approaches and will be brought into full production mode in the near future. Continued development—particularly in authentication, large-scale storage, workflow automation, and open-data services—will be essential to bring the infrastructure to full operational maturity, but the work completed within INFRADEV2 provides a solid and future-proof foundation for KM3NeT's scientific mission and its contribution to the European Open Science Cloud and the global multi-messenger community.