Funded by
the European Union

TOWARD FULL IMPLEMENTATION OF THE KM3NeT RESEARCH INFRASTRUCTURE

KM3NeT-INFRADEV2 – HORIZON – 101079679

Report on the definition of data models and interfaces

KM3NeT-INFRADEV2 GRANT AGREEMENT DELIVERABLE: D4.2	
Document identifier:	KM3NeT-INFRADEV2-WP4-D4.2-vfinal
Date:	31/12/2024
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

In view of the cross-disciplinary environment of KM3NeT research, it will be necessary to adapt software frameworks for data analyses that are well suited for a broad community of users. Also, interfaces to transfer KM3NeT open data to a common data lake infrastructure are set up. In this report, the development of standardised open data models, their documentation and integration in common scientific workflows are addressed and an intermediate status of their implementation is given in order to promote codevelopment of integrated multimessenger workflows and promote the FAIR sharing of KM3NeT data.

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	09/12/2024
Author(s)	Vincent Cecchini Tamas Gal Kay Graf Francisco Salesa Greus Jutta Schnabel Francisco Vazquez de Sola	CSIC, WP4 FAU, WP4 FAU, WP4 CSIC, WP4 FAU, WP4 NWO-I, WP4	
Reviewed by	Artur Ukleja Maurizio Spurio Riccardo Bruno	AGH University of Bologna INFN	18/12/2024 23/12/2024 25/12/2024
Approved by	Paschal Coyle KM3NeT IB	CNRS, WP1	31/12/2024

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	09/12/2024	WP4 draft	T. Gal, K. Graf, J. Schnabel (FAU), V. Cecchini, F. Salesa Greus (CSIC), F. Vazquez de Sola (NWO-I)
2	16/12/2024	Version ready for review	V. Ciarlet (CNRS)
3	27/12/2024	Reviewed version	A. Ukleja, M. Spurio, R. Bruno
4	06/01/2025	Final version	V. Ciarlet (CNRS)

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
ANTARES	Astronomy with a Neutrino Telescope and Abyss environmental RESearch
CERN	*Conseil Européen pour la Recherche Nucléaire* (European Organization for Nuclear Research)
CSA	Coordination and Support Action
CTA	Cherenkov Telescope Array
DL	Data Level
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
IVOA	International Virtual Observatory Alliance
KM3NeT	Cubic Kilometre (km3) Neutrino Telescope
LCG	LHC Computing Grid
LHC	Large Hadron Collider
MoU	Memorandum of Understanding
RI	Research Infrastructure
VO	Virtual Observatory
VODF	Very high energy Open Data Format
WLCG	Worldwide LHC Computing Grid
WP	Work Package

VI. 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.

VII. EXECUTIVE SUMMARY

This report summarises the efforts related to documentation for high-level KM3NeT data classes and their use in the KM3NeT data interfaces. The procedures established serve to ensure transparency and stability in the use of the KM3NeT data and quality in future development of common data interfaces in the world-wide scientific community.

VIII. TABLE OF CONTENTS

1. Introduction

To adhere to the FAIR principles[1] and make data and scientific information easily available, user-friendly and efficient data formats and interfaces for the sharing of data have to be defined and made available for a world-wide community. During the construction phase of KM3NeT, it is therefore necessary to well document and define the relevant structure and configuration of the infrastructure to build for interoperability. Following the development of the Open Science System of KM3NeT, the current status of definitions and documentation of the relevant system parts is outlined here and the strategic approach to further develop the data formats and document them is defined.

2. Relevant data and interfaces

KM3NeT Data Levels

In the KM3NeT Data Management Plan (DMP)[2], KM3NeT data is categorized into specific data levels (DL) based on the processing stage of the data. Publication of data and the related software are governed by the data release policy, currently under discussion in the collaboration, which is partially following the CERN Open Data policy[3]. Here, relevant data is identified as those related to publication, and full particle event lists and the associated auxiliary data e.g. from simulation. Translating this to the KM3NeT DLs, relevant levels are:

- *DL2: Reconstructed data*: fully reconstructed events, both from the real experiment and simulations.
- *DL3: Science-ready data*: selected event lists and related instrument response functions
- *DL4-6: Advanced high-level data*: neutrino source analysis products, maps and catalogues.

Regarding the online real-time data processing pipeline, reconstructed alert data and related notices are also assigned to these levels. Beyond the neutrino event data, the KM3NeT detectors also provide additional data related to the deep-sea setting and their environment, which can be of scientific use in other domains like maritime research. Although these interfaces might be developed in the future, there are currently no instances provided and therefore, this document focuses on neutrino data only. However, the documentation strategy and data format approaches outlined here can similarly be applied to future handling of environmental data and other approaches to processing of big data samples.

[1] https://www.go-fair.org/fair-principles/

[2] see D4.1 – Report on the review of the KM3NeT Data Management Plan by external experts

[3] https://cds.cern.ch/record/2745133

Interfaces

Data servers

The currently available interfaces for open data and software are described in the DMP, and include a dedicated interface to the Virtual Observatory to cater to the astronomy community, a data server for general providing provision, and also external platforms that are used for long-term archiving:

- The Virtual Observatory server[4] is registered to the Virtual Observatory and provides data in the formats defined by the International Virtual Observatory Alliance[5].
- The Open Data Center[6] is a server based on the DataVerse software[7] which is currently customised to host the broad variety of KM3NeT data. With a flexible configuration for metadata and grouping options and a full API it serves as backbone for future data access for high-level analysis data.
- Data and software are also added or linked to public repositories like Zenodo[8] to enable easier findability of the data and provide high-level data archiving.

Computing environment

The processing of DL1 and DL2 data is provided by KM3NeT Tier-0 and Tier-1 resources, see the KM3NeT DMP. Those are HPC, HTC or WLCG resources, to be managed through EGI servers with the Dirac grid middleware for data processing. In the overlap with other communities, KM3NeT was active in defining the framework of the ESCAPE Data Lake[9]. This should allow both for internal data processings and also for providing them to other communities and scientists. Distributed grid storage for KM3NeT data at Tier-0 and Tier-1 levels will be accessible using the Rucio middleware as data interface to the distributed file management, which is in full compliance with internal and external needs as e.g. the data lake. This also includes the authentication and authorization instance (AAI), where a KM3NeT instance will be set up which allows for AAI between Rucio, WLCG and data lake resources.

Online Alert system

The Online Alert system is a framework dedicated to the real-time analyses of the KM3NeT data. It aims to identify interesting neutrino candidates, extract high-level information from their raw data (such as time and coordinates), and disseminate them publicly. For the detailed description of the data acquisition reconstruction and analysis, see the DMP.

To ensure a wide adoption by the community, as well as by all potentially interested users and observatories, the alerts will be broadcasted through the GCN Notice services, see below. The

[4] http://vo.km3net.de/

[5] IVOA standards: https://ivoa.net/deployers/

[6] https://opendata.km3net.de/

[7] Dataverse: https://dataverse.org/

[8] https://zenodo.org/communities/km3net

[9] R. Dona and R. DiMaria, *The ESCAPE Data Lake*, in EPJ Web of Conferences 251, 02060 (2021) https://doi.org/10.1051/epjconf/202125102060

alert content will be provided in two formats, *i.e.* JSON (a well-established internet standard) and VOEvent (a well-established standard in the astronomy community). Along with the real-time broadcasting, GCN offers an interactive platform to search the previously sent notices. At this time, and because GCN is considered a robust and reliable service, no other interface is foreseen to publicly offer our real-time alerts.

3. Documentation and development strategies

Data format documentation

The main I/O data classes and their structures in KM3NeT are provided in a centrally organized code repository in the KM3NeT GitLab[10] as a C++ libraries, which are used in software that read and generate input and output files. The documentation is auto-generated, and standard parameter and value definitions are added as CSV tables. These CSV are used to create include files for C++, Julia and Python. In addition to the C++ library and the definition include files, the data format specification document is semi-automatically generated, which is used as first reference for the currently valid definitions. This document's aim is to also provide human readable specifications for the future. Through the use of the versioning system in GitLab and by manually assigning version tags to the main data format documents, the change history of the data format definitions can easily be tracked. At the time of writing, the content of the data class repository is consolidated.

During the construction phase of KM3NeT, the interfaces are developed using common software quality standards adhering to the *KM3NeT Software Quality Plan*[11]. Once established, documentation, including maintenance and updates of it, is subject to the *Quality Plan for KM3NeT*[12]. The latter document defines the roles of actors, the tools, and the organisation of quality means for the whole project.

Defects in the definitions will be traced by *Non-Conformity Reports*, changes in the design via *Design Change Requests* and be handled by the quality team as well as the responsible working group coordinators (*Computing and Software* in this case).

The quality process is supported by templated web forms for all quality document types, as well as a document management system. The process is traced via a *Configuration Item Documentation List* which defines the applicable and reference documents.

[10] See https://git.km3net.de/common/km3net-dataformat/

[11] KM3NeT_SOFT_2016_001-SoftwareQualityPlan_kgraf_v7.pdf

[12] KM3NeT_QA_2023_004_Quality_Plan.docx (internal)

Initiatives for data format development and integration

The following initiatives and collaborations serve to show that data formats and interfaces are embedded in a dynamic field in which constant development and exchange is crucial. Therefore, KM3NeT is actively pursuing this exchange.

Alerts: General Coordinates Network

The dissemination of the alerts benefits from the GCN and VOEvent standardisation of notices content. However, neutrino astronomy has special needs concerning the alert information, due to its specific detection methods. To meet them, an informal discussion channel has been set up between the IceCube and KM3NeT collaborations in charge of alert sending. Its purpose is to develop a common core content for our alerts, in order to provide the same set of minimal information to the final user (event name, time, position, probability, *etc.*). Therefore, this neutrino alert core will be an extension of the GCN JSON schema. The definition of this core schema is based on the experience we are acquiring by defining our own alert, together with intensive exchanges with the IceCube members on what are the most high-level variables that we share and we can truthfully offer publicly.
The final objective will be to have a neutrino schema in the GCN git repository for core JSON schema. That way, the neutrino telescope community would get the opportunity to use it as a basis for their own alerts (if they aim to provide their alerts).

Astronomy data: Virtual Observatory

In the IVOA, a common initiative of high-energy particle experiments has been launched in November 2024 to enhance the inclusion of event-like data like gamma rays and neutrinos through a High-Energy Interest Group (HEIG)[13] to which KM3NeT contributes. Here, the extension of standards to include simulation-based information, creating metadata for low-count rate experiments and improving discovery within an astronomy context are targeted.

Common data for neutrino detectors: Global Neutrino Network

In the context of the Global Neutrino Network (GNN)[14], a working group has been formed to explore the standardisation of output formats in order to facilitate the easier sharing of data between the high-energy neutrino experiments. This initiative focuses on all common types of data and analyses. One of the first outcomes is e.g. the integration of a reader for KM3NeT data in the GraphNet framework[15] for machine learning that is used by the IceCube collaboration.

General data integration: ANTARES legacy data

The ANTARES collaboration, having operated a similar infrastructure close to the ORCA site until 2022, has decided to hand over their legacy data to the KM3NeT collaboration. Offering

[13] Virtual Observatory and HIgh Energy Astrophysics: https://ivoa.net/documents/Notes/VOHE/20241112/index.html
[14] https://www.globalneutrinonetwork.org/
[15] https://github.com/graphnet-team

this data through the KM3NeT infrastructure, converting it to KM3NeT formats to ensure easy and future access, and including workflows to handle the data will greatly facilitate the development of open data standards. Parts of these developments are tackled in the ACME project[16] and will also connect e.g. to the HEIG developments.

4. Conclusion

The documentation of the high-level data formats in KM3NeT is tackled alongside the internal documentation of data classes. Through this common documentation, standard definitions are made available. However, during the development process of the KM3NeT data interfaces, the coordination of data standards within larger community efforts is crucial for the development of combined scientific studies and thus pursued in cooperation e.g. with the IVOA and in the GCN. Change procedures to KM3NeT standard definitions will follow the internal quality assurance rules. While currently still in the setup phase, a well-documented and transparent use of the KM3NeT data formats can thus be ensured. This approach facilitates reproducibility of the KM3NeT scientific results and enhances collaboration within the research community. Additionally, it ensures that data can be easily accessed and understood by new researchers, promoting continuous innovation and discovery in high-energy physics and astronomy disciplines.

[16] Astrophysics Center for Multimessenger studies in Europe: https://cordis.europa.eu/project/id/101131928

IX. REFERENCES

Websites:

European Union. (2024). "Astrophysics Center for Multimessenger studies in Europe", on CORDIS. Accessed on the 16/12/2024 at: https://cordis.europa.eu/project/id/101131928

GitHub. (n.d.). "GraphNet. Graph neural networks for neutrino telescope event reconstruction". Accessed on the 16/12/2024 at: https://github.com/graphnet-team

GoFair. (n.d.). "Faire pinciples". Accessed on the 16/12/2024 at: https://www.go-fair.org/fair-principles/

Global Neutrino Network. (n.d.). "GNN. The Global Neutrino Network". Accessed on the 16/12/2024 at: https://www.globalneutrinonetwork.org/

International Virtual Observatory Alliance. (n.d.). "Information fo Deployers". Accessed on the 16/12/2024 at: https://ivoa.net/deployers/

KM3NeT GitLab Server. Available at: https://git.km3net.de/common/km3net-dataformat/

KM3NeT Open Data Center (n.d.). ANTARES 2007-2017 data. Accessed on the 16/12/2024 at: http://vo.km3net.de/

The Dataverse project. (2024). "The Dataverse project. Open source research data repository. Accessed on the 16/12/2024 at: https://dataverse.org/

Zenodo. (n.d.). "KM3NeT Collaboration". Accessed on the 16/12/2024 at: https://zenodo.org/communities/km3net/records?q=&l=list&p=1&s=10&sort=newest

Reports:

CERN. (2020). *CERN Open Data Policy for the LHC Experiments*. Available at: https://cds.cern.ch/record/2745133

Dona, R. and Di Maria, R. (2021). *The ESCAPE Data Lake*, in EPJ Web of Conferences 251, 02060 (2021). Available at: https://doi.org/10.1051/epjconf/202125102060

KM3NeT Collaboration (2024). *Report on the results of the review of the KM3NeT DMP by external experts*. Available at: https://www.km3net.org/wp-content/uploads/2024/04/KM3NeT-INFRADEV2-WP4-D4.1-v.approved.pdf

IVOA. (2024). Virtual Observatory and High Energy Astrophysics Version 1.0. Available at: https://ivoa.net/documents/Notes/VOHE/20241112/index.html