Ref. Ares(2020)6226876 - 02/11/2020

KM3NeT INFRADEV – H2020 – 739560

KM3NeT report on rules and conditions for data access

KM3NeT-INFRADEV GA DELIVERABLE: D4.7	
Document identifier:	**KM3NeT-INFRADEV-WP4-D4.7_v1.0**
Date:	**11 April 2020**
Work package:	**WP4**
Lead partner:	**FAU**
Document status:	**Final draft**
Dissemination level:	**Internal**
Document link:	

Abstract

The KM3NeT Research Infrastructure will, over a period of at least a decade, produce a large amount of unique scientific data that are to be made available to the scientific communities concerned and to the broader general public. This requires the set-up of tools, procedures, documentation and rules to provide this service. Although access to open science data produced by the KM3NeT collaboration is per definition open, the various components of the open data system require regulations regarding access modes, usage of underlying software, and training according to the specific user group. These rules and conditions for data access are presented in this document.

I. Copyright notice

Copyright © KM3NeT Collaboration

II. Delivery slip

	Names	Partner/WP	Date
Author(s)	Jutta Schnabel	FAU	
Approved by			

III. Document log

Issue	Date	Comment	Author/Partner
0.1	28/01/2020	First draft	Jutta Schnabel/FAU
1.0 1.1 1.2	04/04/2020	Comments and edits by J. Hofestädt and U. Katz	J. Hofestädt/FAU U. Katz/FAU
1.3	11/04/2020	Comments by C. Markou, edits by U. Katz	C. Markou/NCSR Demokritos U. Katz/FAU

IV. Application area

This document is a deliverable for the grant agreement of the project, applicable to all members of the KM3NeT-INFRADEV project, beneficiaries and third parties, as well as its collaborating projects.

V. Terminology

ADQL = Astronomical Data Query Language
ARCA = Astroparticle Research with Cosmics in the Abyss
(KM3NeT neutrino astroparticle physics telescope)
CSA = Coordination and Support Actions
DaCHS = GAVO Data Center Helper Suite
EduGAIN = EDUcation Global Authentication INfrastructure
EOSC = European Open Science Cloud
ESFRI = European Strategy Forum on Research Infrastructures

FAIR = Findable Accessible Interoperable Reproducible
FITS = Flexible Image Transport System (Data format)
GAVO = German Astrophysical Virtual Observatory
GDPR = General Data Protection Regulation
HDF = Hierarchical Data Format
IVOA = International Virtual Observatory Alliance
MoU = Memorandum of Understanding
ORCA = Oscillation Research with Cosmics in the Abyss (KM3NeT neutrino particle physics detector)
SAMP = Simple Application Messaging Protocol
SCS = Simple Cone Search
TAP = Table Access Protocol
VO = Virtual Observatory

VI. List of figures

VII. List of tables

VIII. Project summary

KM3NeT is a large Research Infrastructure that will consist of a network of deep-sea neutrino detectors in the Mediterranean Sea with user ports for Earth and Sea sciences. Following the appearance of KM3NeT 2.0 on the ESFRI roadmap 2016 and in line with the recommendations of the Assessment Expert Group in 2013, the KM3NeT-INFRADEV project addresses the Coordination and Support Actions (CSA) to prepare a legal entity and appropriate services for KM3NeT, thereby providing a sustainable solution for the operation of the research infrastructure during ten (or more) years. The KM3NeT-INFRADEV is funded by the European Commission's Horizon 2020 framework and its objectives comprise, amongst others, the preparation of Open Data Access (work package 4).

IX. Executive summary

The KM3NeT Research Infrastructure will, over a period of at least a decade, produce a large amount of unique scientific data that are to be made available to the scientific communities concerned and to the broader general public. This requires the set-up of tools, procedures, documentation and rules to provide this service. This document lists the rules and conditions for access to and usage of KM3NeT open data and the according computing platforms.

X. Table of Contents

1. Introduction

KM3NeT is a large Research Infrastructure (RI) that will consist of a network of deep-sea neutrino detectors in the Mediterranean Sea with user ports for Earth and Sea sciences. The main science objectives, a description of the technology and a summary of the costs are presented in the KM3NeT 2.0 Letter of Intent [1].

KM3NeT will open a new window in the study of the Universe, and at the same time pursue important research into the properties of neutrinos. With the ARCA telescope, KM3NeT scientists will search for neutrinos from distant astrophysical sources such as supernovae, gamma ray bursts or active galactic nuclei. Using the exact same technology, the ORCA detector will provide data of unprecedented quality on neutrino oscillations, exploiting neutrinos generated in the Earth's atmosphere. Arrays of thousands of optical sensors will detect the faint light generated in the deep sea by charged particles originating from collisions of the neutrinos with atomic nuclei. The facility will also house instrumentation for Earth and Sea sciences for long-term and on-line monitoring of the deep-sea environment and the sea bottom at a depth of several kilometres [2].

The KM3NeT Collaboration has developed a data policy plan [3] reflecting the research, educational and outreach goals of the facility. During an embargo time of two years, access to the data will be restricted to the KM3NeT Collaboration for processing and calibrating the raw data, and securing their quality and correctness. During this period, access to the data is exclusively granted to the Collaboration members as a return for constructing, maintaining and operating the facility. The Collaboration commits itself to generating high-quality reconstructed event data suited for a wider user community during the embargo period. These data will subsequently be made publicly available under an open-access policy on a web-based service and will not only allow the public to validate the scientific results presented by the Collaboration but also to perform individual analyses.

The contribution of KM3NeT to the body of scientific knowledge will depend to a large extent on the quality of the analysed data. A data management plan that ensures a correct handling of the KM3NeT data along all the production and processing chains has been presented in [3]. From an open-access perspective, the data management plan should also be compliant with the FAIR (Findable - Accessible - Interoperable - Reproducible) data principles, which were specifically designed to enable and enhance the reuse of scholarly data by third parties [4], [5].

In this document, the conditions for access to KM3NeT data and supplementary material are introduced. These conditions follow closely the requirements envisioned for the scientific user communities interested in data access and are governed by the standard software environments developed within these communities. In particular, access to astrophysical data will closely follow procedures laid out for the Virtual Observatory (VO) environment, whereas software access and usage is granted within the developing standards of the astrophysics and particle physics communities relying on collaborative software development platforms and interfederated services.

2. Data products, platforms and access levels

2.1. KM3NeT Open Data products

2.1.1. Neutrino event data

Information on particle detection within KM3NeT is recorded as so-called “events”. In the simplest case, an event contains the reconstructed direction and time of particle transition through the detector. In a more complex case, information on all detected photons within the event duration are stored together with the additional information about detector conditions.

For most science use cases, providing information about reconstructed particle direction, timing and further particle properties like reconstructed energy or particle type suffices for the external user. This data can be represented as a simple vector of parameter values per event, which leads to full datasets being provided as tables of neutrino events. These tables can be stored and provided in a multitude of mostly text-based formats, alongside annotations for the parameter description and dataset content. If formatting follows the established standards of the Virtual Observatory (VO) [6], easy use in VO-compatible software is guaranteed.

For full event data containing information on detected photons, multiple tables have to be provided containing information on the individual photon detections. As this format does not fall under the regime of the VO, custom solutions regarding the file format have to be developed, keeping in mind the target of interoperability and useability for data analyses from experiments in the closer science community. Therefore, widely used formats like HDF5 and FITS are used to store this data. In addition to that, software for reading, displaying and analysing these events are made available together with the data.

2.1.2. Supplemental distributions for physics analysis

For the full understanding of neutrino datasets, additional information beyond the detected neutrino events is necessary, like the probability distributions of neutrino detection for a given scientific target object or the background expectation. As these probability distributions are generally generated from extensive Monte Carlo simulations, providing access to the full simulation is not feasible and not desirable for the external scientists. Therefore, this information will be provided in a simplified form through web services published together with the datasets, which allows creating probability distributions of neutrino detection for a given scientific target within the scope of the dataset. For a more in-depth investigation, software for simulation of neutrino events will also be made available for those external users who have sufficient computing power at their disposal to generate simulated data sets of appropriate statistics.

2.1.3. Software and services

Dedicated KM3NeT software has been and will be developed throughout the commissioning and operation of the experiment covering the full range of detector operation, data taking, data processing, and simulation. Software that is used for event reconstruction and simulation is aimed to

be made available to outside users, as it provides tools for the processing and analysis of KM3NeT open data. In addition to that, software solutions that are not restricted to use in the experiment, e.g. for workflow or data management, or specific scientific programs are part of the open regime as open software.

This software can be downloaded and installed by the user. However, the software will also be provided as software containers, which allow platform-independent and easy application by the user by running unmodified KM3NeT software that passed the internal KM3NeT quality control.

2.1.4. Training material

To facilitate easy understanding of open data handling, tutorials and open courses, providing an introduction to KM3NeT data taking and software applications, are an integral part of the open data regime. Apart from the documentation provided for software, data and services, a virtual education portal offers open courses with video tutorials, schematics and descriptions.

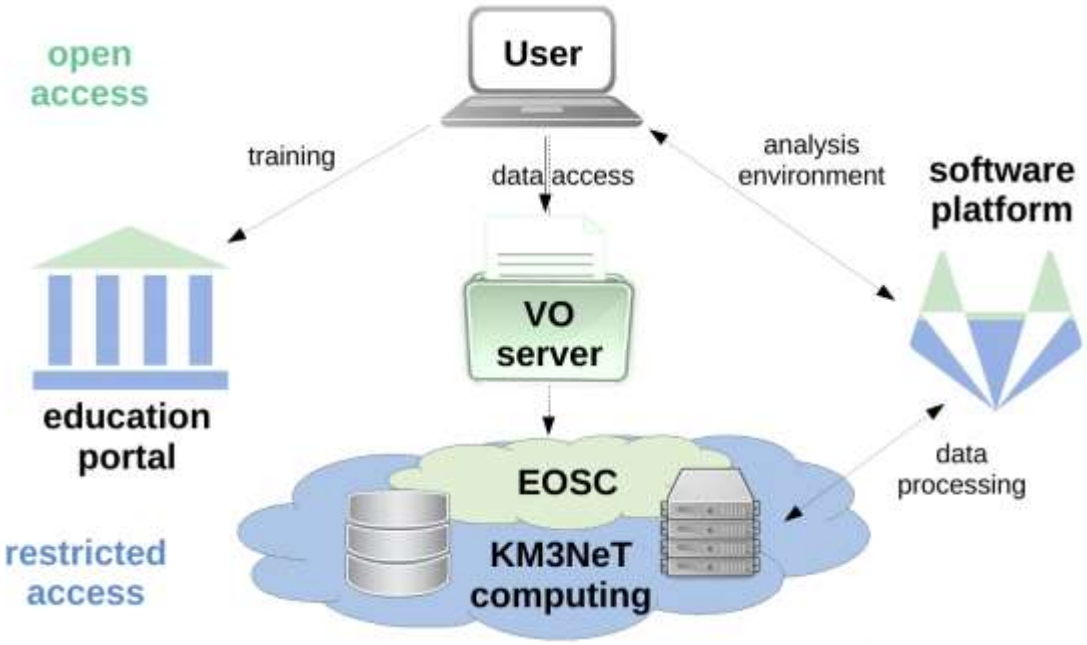

Figure 1: Access modes and levels to KM3NeT Open Data platforms.

2.2. Open Data Users

Access rights to the KM3NeT data products can generally be divided into restricted access data products and software, to which only KM3NeT members have access, and open data, to which, by definition, no restriction applies. However, for the different user groups listed below, different conditions might apply to access resources not initially contained in the open data regime. This list

does not include users of the alert system for mulit-messenger astronomy, as this system is covered in the dedicated work package WP7 of this project.

- **General public**: Meaning everybody not included in the user groups below. Here, easily understandable tutorials need to be accessible, based mostly on web interfaces.
- **Software developers**: Typically not interested in the data, but in co-development of software and/or own contributions to that software.
- **External scientists**: Interested in access to full simulation sets and full event datasets beyond the datasets published through the virtual observatory.
- **Associated scientists**: Cooperating on specific scientific or technical aspects of KM3NeT beyond the scope of open data under regulation of bilateral agreements, e.g. MoUs[1].
- **Internal scientists**: KM3NeT members have access to all components of KM3NeT operation, data taking and analysis.

2.3. Open Data servers and platforms

2.3.1. Virtual Observatory Data Server

Publication to the VO is handled by a dedicated server running the DaCHS software which is designed to provide all standard protocols for data exchange within the VO. For tabled events in a source catalogue, the Simple Cone Search (SCS)[2] regulates the access to the catalogue content through querying the tables using right ascension, declination and a radius around a given celestial position. The information can either be accessed using a web form, through a client application using the Table Access Protocol (TAP) provided by the International Virtual Observatory Alliance (IVOA), or treating the server as database and querying the resources through a specialised, SQL-based query language, the Astronomical Data Query Language (ADQL). Findability of the datasets is guaranteed through registration of the datasets to the central registry of the IVOA. Therefore, access to the datasets is guaranteed using VO tools.

The services for supplemental information to the datasets will also be made available through the VO server and be accessible without restrictions. For communication with standard VO applications, these services can use the Simple Application Messaging Protocol (SAMP), which ensures easy communication between VO-compatible applications.

2.3.2. Software repository and containers

KM3NeT software is hosted on a dedicated server running the GitLab [7] software. This software repository allows software versioning, download, deployment, documentation and collaborative development. While the user registration to the KM3NeT GitLab instance is restricted to KM3NeT members by authentication through OAuth connecting to a centrally managed KM3NeT user identity, GitLab allows projects to be published without restriction using such open projects. Software that is

[1] As the content of such MoUs does not fall into the scope of Open Access Data, this case will not be further discussed here.

[2] For these and the following standards, see [6]

considered as open software by KM3NeT after an initial release procedure is made available as open projects. These projects are also mirrored to GitHub[3] for easy findability. The software can therefore be downloaded and installed without restriction, and documentation is openly available.

GitLab allows the creation and storing of Docker [8] images from the continuous integration process, which can be provided through a dedicated container registry for each project. These software containers can be accessed through local instances of the Docker software, which is available as freemium software as a service. Therefore, all users can employ software containers of the KM3NeT open software without restrictions or additional authentication processes through Docker clients.

Contribution to KM3NeT open software is possible through the collaborative software tools provided on GitHub, where each registered user can suggest code modifications and use a ticketing system for feedback.

2.3.3. Virtual Education Centre

Training material is provided on the Virtual Education Center, which consists of a server running the education management system LifterLMS [9] on Wordpress [10]. Material is provided in courses which may either be protected (i.e. users must enrol to these courses), or freely available. While KM3NeT members are required to authenticate with their KM3NeT user identity to access restricted courses for internal training, all information relevant to the open data regime is provided in free courses without the necessity of enrolment.

2.3.4. Computing centre access and data sharing

External access to copies of large data samples and simulation releases stored at KM3NeT host computing centres is envisioned after KM3NeT has entered regular data taking. This access is not covered in the VO regime due to restrictions on data format and volume. However, the KM3NeT Collaboration is working towards common solutions for data access with the astronomy and particle physics communities within the framework of the European Open Science Cloud (EOSC) [11]. Access to the data would probably be regulated by registration to an interfederated identification service, e.g. eduGAIN. The software environment provided by KM3NeT within the EOSC would include the possibility to request full sets of measured or simulated data from distributed KM3NeT resources, to pull containerised software from a centralised software registry and to apply KM3NeT or third party software. In all cases, the aim is to provide high-quality data samples in an open manner to the full science community, such that effective scientific cooperation can be ensured.

[3] see https://github.com/KM3NeT

3. Rules and conditions

3.1. Publication procedures

3.1.1. Releases of neutrino event data and simulation

Once KM3NeT reaches standard operation mode, neutrino event data will be made publicly available after an embargo period of two years after the end of each full year of data taking. Dataset processing will follow a standardised procedure ensuring full data provenance, testing and monitoring of the datasets, creation of metadata, as well as complete documentation of the published dataset. Before publication, an internal review process will decide on the fulfilment of quality standards and approve the final publication of the data catalogues and associated services.

3.1.2. Software releases as open software

The KM3NeT Collaboration supports and has decided to implement open source access to its software. Software packages are released as official KM3NeT software if they meet the internally defined software standards regarding coding style, documentation, testing, storage, installation and versioning. An internal review process will also decide on the publication of software as official KM3NeT software, as well as for the following major releases, which ensure compatibility and quality control within the KM3NeT software and data environment.

3.1.3. Supplementary material

Supplementary material will be published upon production during workshops or as dedicated online training courses. As the courses build on open KM3NeT resources, no additional publication procedure is foreseen.

3.2. Access rights by user group

For the open access platforms introduced above, user role assignment allows group-specific access modalities to the different data products offered on these platforms. In Table 1, an overview over the access rights is given. The short notation for the software repository indicates that extended access to collaborative software tools is granted through the mirrored software repositories on GitHub, which is open for registration to any user. In the EOSC environment, additional access to scientific datasets is granted through an interfederated authentication, like eg. EduGAIN [12].

User group	**VO server**	**Software repository**	**Virtual Education Center**	**EOSC access**
all	full access	open projects / container images	open courses	-

software developers	full access	community tools (GitHub)	open courses	interfederated authentication
external scientist	full access	open projects / container images	open courses	interfederated authentication
KM3NeT user	full access	full (KM3NeT authentication)	full (KM3NeT authentication)	full (KM3NeT authentication)

Table 1: Access rights to KM3NeT Open Data platforms by user group.

3.3. Privacy

For all KM3NeT open data services, established privacy standards including data protection and minimal data collection on usage are met and are in compliance with the GDPR (EU) 2016/679 [13]. Statistics on service access are anonymised, and no additional information will be required from the users to gain access to open data products.

4. References

1: Adrián-Martínez, S. et al. (2016). Letter of Intent for KM3NeT 2.0. Journal of Physics G: Nuclear and Particle Physics, 43 (8), 084001.

2: The KM3NeT Project: https://www.km3net.org/

3: KM3NeT-InfraDev (2017). The KM3NeT Data Management Plan (D4.1 KM3NeT2.0_WP4_D4.1).

4: The FAIR principles: Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; et al. https://www.nature.com/articles/sdata201618. *Scientific Data*. p. 160018.

5: The FAIR principles: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf

6: IVOA standards: http://ivoa.net/deployers/intro_to_vo_concepts.html

7: GitLab: https://about.gitlab.com

8: Docker: https://www.docker.com

9: LifterLMS: https://lifterlms.com/

10: Wordpress: https://wordpress.org/

11: European Open Science Cloud: https://www.eosc-portal.eu/

12: eduGAIN: https://edugain.org/

13: General Data Protection Regulation: https://eur-lex.europa.eu/eli/reg/2016/679/oj