KM3NeT INFRADEV – H2020 – 739560

KM3NeT report on documentation strategy, environment and software

KM3NeT-INFRADEV GA DELIVERABLE: D4.6

Document identifier:	**KM3NeT-INFRADEV-WP4-D4.6_v1.0**
Date:	**19/11/2019**
Work package:	**WP4**
Lead partner:	**FAU**
Document status:	**Final**
Dissemination level:	**Public**
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. For all aspects of the open data access system, including data processing methods, data structure, access and usage examples, sufficient documentation for the effective use of the open data must be provided. In this document, the documentation strategy for the different components is described.

I. Copyright notice

II. Delivery slip

	Names	Partner/WP	Date
Author(s)	T. Gal, J. Hofestädt, U. Katz, J. Schnabel	FAU	05/11/2019
Reviewed by	S. Navas	IB	19/11/2019
Approved by	PMB		03/12/2019

III. Document log

Issue	Date	Comment	Author/Partner
0.1	04/11/2019	First draft	T. Gal, J. Hofestädt, U. Katz, J. Schnabel / FA
0.2	05/11/2019	Includes comments/amendments by T. Gal, J. Hofestädt, U. Katz, J. Schnabel	T. Gal, J. Hofestädt, U. Katz, J. Schnabel / FAU
1.0	19/11/2019	Includes comments from internal IB review by Sergio Navas, from H. Yepes Ramirez and from R. van der Meer	U. Katz

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

ARCA = Astroparticle Research with Cosmics in the Abyss (KM3NeT neutrino astroparticle physics telescope)
CSA = Coordination and Support Actions
CTA = Cherenkov Telescope Array
DaCHS = Data Center Helper Suite
DOI = Digital Object Identifier
EOSC = European Open Science Cloud

ESCAPE	=	European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures
ESFRI	=	European Strategy Forum on Research Infrastructures
FAIR	=	Findable Accessible Interoperable Reproducible
FAQ	=	Frequently Asked Question
FITS	=	Flexible Image Transport System
HTTP	=	Hypertext Transfer Protocol
IVOA	=	International Virtual Observatory Alliance
GAVO	=	German Astrophysical Virtual Observatory
ORCA	=	Oscillation Research with Cosmics in the Abyss (KM3NeT neutrino particle physics detector)
PMT	=	Photomultiplier Tube
RI	=	Research Infrastructure
SCS	=	Simple Cone Search
SOAP	=	Simple Object Access Protocol
TAP	=	Table Access Protocol
UCD	=	Universal Content Descriptor
VO	=	Virtual Observatory
WDSL	=	Web Service Description Language
W3C	=	World Wide Web Consortium
XML	=	Extensible Markup Language

VI. List of figures

VII. List of tables

None

VIII. Project summary

KM3NeT is a large Research Infrastructure that will consist of a network of deep-sea neutrino telescopes 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. For all aspects of the open data access system, including data processing methods, data structure, access and usage examples, sufficient documentation for the effective use of the open data must be provided. In this document, the documentation strategy for the different components is described.

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 Letter of Intent for KM3NeT 2.0 [1].

KM3NeT will open a new window to our Universe, but also will forward the 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, extracted from measurements of 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 interactions of 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 management plan [3] reflecting the research, educational and outreach goals of the facility. From an open-access perspective, the data management plan also is set up such as to 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].

Efficient use of the KM3NeT open data by users who are not members of the KM3NeT Collaboration will require complete, self-contained and continuously maintained documentation of the data with all related aspects, such as access methods, software, example programs and scientific background. In this document, an overview over different approaches to documentation of open access data is given as starting point for a documentation strategy. Documentation relies on the one hand on metadata definition as introduced by the IVOA [5], and incorporates on the other hand software application documentation standardized for common development as driven by the ESCAPE project [6]. In KM3NeT, the different levels of open access data call for a multi-layer approach to documentation, including reference documents and e-learning tools. The path to generation of these various sources of documentation is described in the last part of this document.

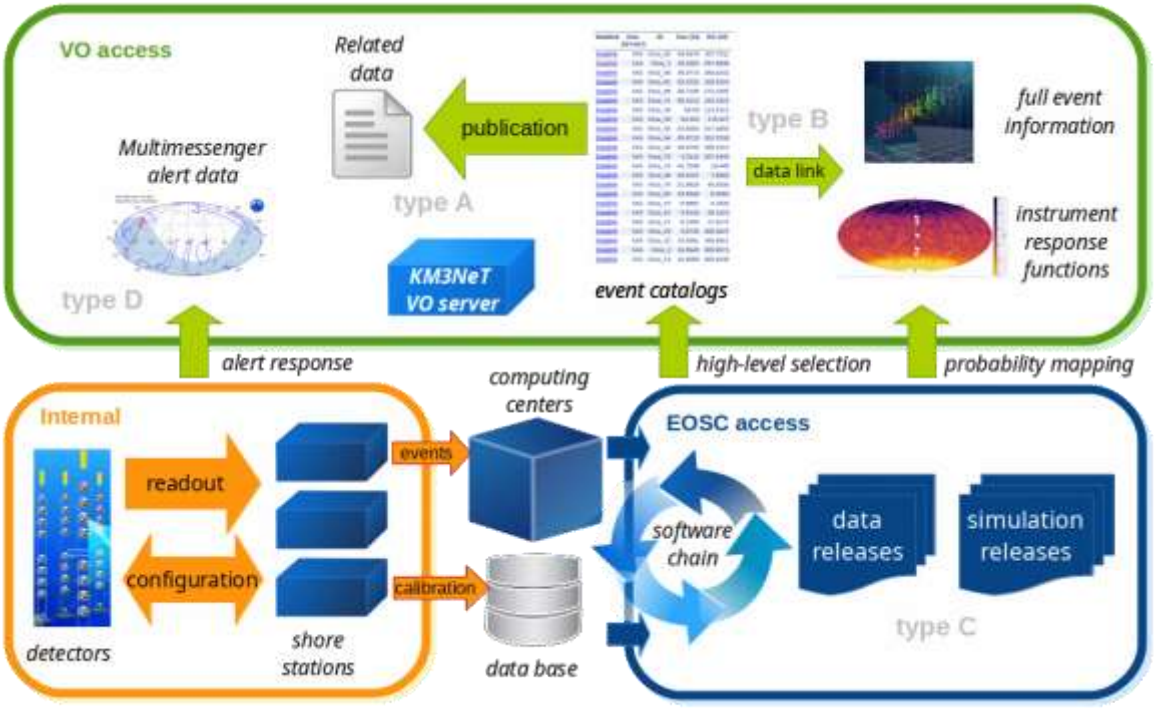

Figure 1: Open access data levels in KM3NeT [7]

2. Objects of documentation

The KM3NeT Collaboration will produce scientific data products through the generation of high-level data consisting of samples of events obtained through processing the filtered data by software applying calibration procedures, identifying events (such as neutrino reactions) and reconstructing their parameters (such as the neutrino directions and energies). As introduced in Deliverable 4.5 [7], KM3NeT open data access can be divided into different access levels, see **Figure 1**.

2.1. Open access data products

High-level data will be arranged into versioned datasets containing reconstructed events, where each event includes all relevant parameters related to event classification, reconstruction and detector running conditions. In order to evaluate the expected event yield for an assumed neutrino flux for a given scientific scenario, these data sets are accompanied by sets of simulated events that reproduce the detector response for the given data taking periods and particle interaction types. Both measured and simulated event samples are expected to be made public within the European Open Science Cloud (EOSC) currently under development in the ESCAPE project [6]. Publication follows closely the

standards set up by the International Virtual Observatory Alliance (IVOA) [8], helping open access data providers and users from the science community to find, share, access, and use data from different observatories through the Virtual Observatory, see **Figure 1**.

Every scientific analysis performed with these data will require a dedicated optimisation of the event selection based on the reconstruction and classification parameters, and therefore produce a specific subset of the full high-level event sample as derived data product related to a specific scientific scenario or publication.

From a documentational point of view, an implementation of the FAIR principles requires the provision of full information for usage of the open data. However, depending on the data product, two levels of data publication can be distinguished.

2.1.1. Reduced astrophysics event data

As a minimum requirement for data publication, KM3NeT particle detections can be described by providing galactic coordinates per particle detection, which can be augmented by adding energy estimates, errors on reconstruction parameters and further parameters relevant to detector state and particle detection. Descriptions of these parameters can be provided e.g. within the Virtual Observatory protocols. Here, no deeper documentation of event generation, processing and detector principles is required, and documentation can be limited to describe data components and detection principles and conditions without the need to provide the user with the full information about data processing on a user level.

2.1.2. Full event data

If event information is published not only on the level of reconstructed particle parameters, but is to include full information of photon detection within the optical array, a full understanding of photon generation, detection and event reconstruction algorithms becomes necessary. This level of data access is no longer included in the Virtual Observatory regime, but requires additional standards for software documentation, metadata structures, and data models and data sharing. Efforts to harmonize open data standards between the different ESFRI projects is currently underway within the ESCAPE project.

2.2. Data usage

Implementation of the FAIR principles must ensure the reproducibility of the high-level data processing. Documentation must thus provide both information on the data components and the processing software and workflow. Also, from a user's perspective, tools to access, use and display the information contained in the data must be provided and be well accessible. Therefore, an extensive documentation including usage examples have to be provided both for KM3NeT specific software used in high-level data processing as well as for applications handling the standardized data products of KM3NeT in a wider physics context. While the task of providing documentation of the KM3NeT-specific software falls into the full responsibility of the KM3NeT Collaboration, sufficient linkage and referencing to external software capable of KM3NeT data reading and handling must also be provided.

3. Documentation approaches

3.1. Documentation in the Virtual Observatory

A thorough documentation of open data is part of the FAIR data principles, which are fully endorsed in the Virtual Observatory standards that govern data exchange and access in the astrophysics community. Standardization within the Virtual Observatory builds on internet standards, using HTTP and XML within the Simple Object Access Protocol (SOAP) or Web Service Description Language (WDSL) for the description of web services [8]. Standards are discussed and defined within the Virtual Observatory Alliance, based on the standardization efforts of the World Wide Web Consortium (W3C) and are published as a combination of text documents, XML schemes and reference implementations. Resources and services within the VO, i.e. data and interfacing applications, have to comply with the VO standards. Data resources are accompanied by resource metadata according to the standard data model, and services follow standard protocols in publishing service metadata to ensure correct interfaces of all services. All these protocols are accessible and documented within the IVOA, and resources and services published under the VO regime implement said standards. As compliance with the VO standards also lies at the core of KM3NeT data publication, they are used as a minimal benchmark for the KM3NeT data model development and documentation.

3.1.1. Data Model and resource description

Within the Virtual Observatory, standardized identifiers and metadata are allocated both to the elementary components of a data set and to the complete data set. In case of KM3NeT data, which uses an event-based data format describing mostly single particle detections within the detector, the corresponding VO standards are implemented in the Simple Cone Search (SCS), which requires right ascension and declination as minimum operational parameters. Here, the standard protocols ensure consistency in the usage of coordinate system and by default refer to well-defined standard coordinates. Unique identifiers, so-called “Universal Content Descriptors” (UCDs) for standardization of additional information per data element are provided according to astrophysical data models for the different astrophysical regimes. In case of KM3NeT, the description of transient observational data as event-type VO data could be used. However, this standard is primarily developed to describe measurements of short-term astrophysical phenomena like supernovae and not for single particle detections. Therefore, an extension of the VO data model to enable a full description of single high-energy particle detection data as produced by KM3NeT or CTA is currently under development in collaboration with IVOA members and members of the corresponding experiments within the ESCAPE project.

Sets of this data are called “catalogs” within the Virtual Observatory and are treated as resources. Resources as data catalogs are allocated their own set of descriptors, which include information on the instrument and generation of the data, authorship and additional references. Further standardization is applied in the registration of the catalog to the VO registry, which acts as a universal pointer to all available data. On the one hand, the resource description contains information of the available interfaces to the data, e.g. by providing the possibility for the user to retrieve the data through the Table Access Protocol (TAP). Furthermore, the data provider itself is registered to act as host for VO

data to the central VO registry. In order to add a registry of resources to the Virtual Observatory, the VO protocols must be implemented by the data provider. The KM3NeT Collaboration will act as a data provider to the VO and will therefore closely follow the VO standards using VO-provided server protocols and software by running a DaCHS server [9].

3.1.2. Data Access and sharing

For end user access to VO published data, a variety of desktop applications and interfaces have been developed to retrieve, manipulate and pass on this data. Applications like Topcat [10] for access to tabled data or Aladin [11], which provides an interactive sky atlas and visualization methods, follow the Virtual Observatory standards. These software packages are well documented and adopted in their basic features to the VO regime. Therefore, gaining a fundamental understanding of software structure and interfaces is possible by following VO documentation. However, for the end user, data access is the main goal, which is guaranteed through the availability of the tools as open source software and abundant documentation of the usage of the applications themselves. The tools also allow for accessing the accompanying metadata to the data catalogues and can display all information shared in VO-compatible formats like VO Tables or FITS [12].

3.2. Software documentation

KM3NeT software is hosted on an internal GitLab [13] instance and software related to open access data is planned to be made available as open software. The KM3Pipe framework [15] can serve as an example reference for the software documentation strategy adopted.

KM3NeT software is required to use technical documentation embedded in the source code like doxygen [16] or sphinx [17], which auto-generates detailed references and documentation of the full source code. In addition to that, nomenclature of classes, functions and variables must follow a standardized format to ensure a consistent code style. This enables easy understanding of specific software functionalities and adapting of example scripts for a wider use.

Each software package is required to be accompanied by documentation in a wiki or as webpages hosted on git, which includes a quick start and a basic user manual. Beyond this minimum standard, the extended software documentation includes example scripts, an extended user guide, description of the overall software architecture and FAQs as already available in KM3Pipe. A feedback and development mechanism is already available via GIT, which not only allows for a formalized procedure to tackle problems, but also serves as a reference for users facing problems.

A ChangeLog to track all relevant software changes for different software versions is required. Transparency is ensured as software releases are defined using annotated Git tags consisting of a version string in a standard format to mark notable stages of the development. The version number follows the SemVer 2.0 [18] specification. Version numbers and sub-numbers are increased according to the impact of the changeset. Here, increasing the main version number is used only if backwards compatibility of the software version is broken, while the sub-numbers indicate introduction of new features or only minor bug fixes.

3.3. Repositories & Virtual Education Centre

Beyond description of the individual components of open access data like data elements, catalogues and user applications, contextualization of the various parts is necessary to create an easily accessible overview and quick access to the relevant reference material.

3.3.1. Guided tours at the Virtual Education Centre

A general overview over the KM3NeT experiment with introductory material to detector construction, science goals, data taking procedures and information relevant to the outside scientist is provided in the Virtual Education Centre [19]. The platform provides e-learning courses both for KM3NeT members as well as open access courses for both the science community and the broader public. While not intended as a repository for documentation per se, the platform aims at providing a low-threshold access to the various scientific resources by providing overview schematics, hands-on exercises and quizzes for various educational levels. KM3NeT data will thus not only be made available to the expert scientist but also for high-school education and the interested public.

3.3.2. Standard terminology and glossary

Throughout the open data regime, a multitude of abbreviations and domain-specific terminology is used. Where applicable, KM3NeT aims to follow community standards in the nomenclature of open access data components. In case of experiment-specific terms, the KM3NeT Collaboration aims to standardize the terminology. A glossary of all relevant terms serves as an easily accessible pointer towards the relevant information on data and software components and is made centrally available for users of open access data, published software, and the Virtual Education Centre.

3.3.3. Reference documents

Beyond the linked and component-specific documentation of KM3NeT data, also document-based documentation will be provided to accompany the open-access data components. These include user manuals made available as online documents, reference publications including descriptions of data taking procedures, scientific results, event catalogue descriptions, and software reference documents. These documents will be made available via the KM3NeT webpage, through online scientific archives like e.g. arxiv.org, or by publication in scientific journals, depending on the nature of the document and awarded object identifiers (DOIs) [20] where applicable. These documents will be linked from the various data products, software and e-education resources and made centrally available.

4. Documentation strategy

The strategy to generate open access data documentation consists of creating relevant metadata for the open data elements, documentation of relevant KM3NeT-specific software, and the provision of reference documents and education materials to access documentation available outside KM3NeT and to facilitate the use of the open access data.

4.1. Data description through rich metadata

4.1.1. Metadata generation and access

Metadata related to data generation, detector status and environmental conditions are stored in large abundance during data taking. These will be reduced to a minimal number of easy to handle and descriptive parameters which are added to the event information. In addition to that, data processing, applied software and the scientific scope of full event datasets, documented as metadata during the generation of event samples, is made available alongside the full data sets. These are handled according to VO standards where applicable, otherwise they are either made available via referencing in the FITS format header or alongside the data sets as external data set description using references to an extended KM3NeT data model yet to be defined.

4.1.2. Metadata definition

Metadata on the content and use of the KM3NeT data products fall into two categories. On the one hand, parts of astrophysical data fully fall into the regime of the VO and follow the VO standards, in which rich metadata usage is key to a thorough documentation. On the other hand, open data which exceeds the scope of the VO by either its scientific application, being e.g. related to neutrino oscillation research, or since it contains e.g. detection-specific parameters not describable within the VO, will have to be allocated regime-specific metadata. This regime-specific metadata will, as minimum, clearly be defined within the KM3NeT context. However, where applicable, efforts will be made to standardize this metadata in exchange with the wider science community. Efforts are made to develop these standards within the ESCAPE project, where both an extension of the VO data model for high-energy particle experiments is discussed, and a harmonization of data model standards (especially regarding common features with the CTA data model) is aimed for.

4.2. Usability through user guidelines and references

4.2.1. User manuals and example code publication

As described above, extended software documentation is already provided for a large part of the KM3NeT software. Here, access to KM3NeT data generally happens beyond the scope of the VO protocols. While not all software is relevant for users of open access data, the software that is made available for the end user will be accompanied by example scripts, which come e.g. as Jupyter notebooks [21]. Within the ESCAPE efforts, it is planned to make this software centrally available and provide a computing environment where it can be easily executed. Due to the community feedback functionality provided by the bug tracker mechanism Git Issues [13], the request for specific solutions for data usage can be easily picked up and documented. Requests via Git Issues can lead to the implementation of additional user scripts and helps to keep documentation up to date and relevant for the individual user.

4.2.2. Reference document as publications

Publications on the general design, detection mechanisms and data management within KM3NeT have already been made available, e.g. in the Letter of Intent [1] or the data management plan [3]. Various publications on further aspects of the work of the KM3NeT Collaboration are continuously published and made available on various document servers according to the publication channel. Linkage to these reference documents will be assured via the KM3NeT website, through metadata descriptions to published data sets where applicable, in the online courses at the virtual education centre, and in the references of software functionalities and packages. These linkages are considered as standardized procedure within the KM3NeT Collaboration.

4.2.3. Easy accessibility of documentation through the Virtual Education Centre

The generation of content for the Virtual Education Centre is driven by workshops both for internal use and for the wider public. Materials, video and audio recordings and exercises from these events are transformed to online courses and provide links to the relevant background material and example scripts. In addition to that, further material is published on the Virtual Education Centre, introducing VO tools through referral to their user documentation and demonstrating the application of these common tools to KM3NeT open data. The creation of online teaching material is an integral part of internal and external training events and therefore creation of the content is kept dynamic to reflect the current status of development and data usage options.

5. 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.

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 International Virtual Observatory Alliance. https://ivoa.net

6: European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures. https://projectescape.eu

7: KM3NeT-InfraDev. 2019. Report on monitoring and quality control setup

8: IVOA. Introduction to VO Concepts. http://ivoa.net/deployers/intro_to_vo_concepts.html

9: GAVO Data Center Helper Suite (DaCHS). http://vo.ari.uni-heidelberg.de/docs/DaCHS/

10: Tool for Operations on Catalogues And Tables (Topcat). http://www.star.bris.ac.uk/~mbt/topcat/

11: Aladin Sky Atlas. https://aladin.u-strasbg.fr/

12: The FITS Support Office at NADA/GSFC. https://fits.gsfc.nasa.gov/

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

15: KM3Pipe. https://pypi.org/project/km3pipe/

16: Doxygen. http://doxygen.nl/

17: Sphinx Python Documentation Generator. https://www.sphinx-doc.org

18: Semantic Versioning 2.0.0. https://semver.org:

19: KM3NeT Education Portal (Virtual Education Center). http://edu.km3net.de/

20: The Digital Object Identifier System. https://www.doi.org/

21: Project Jupyter. https://jupyter.org/