KM3NeT INFRADEV – H2020 – 739560

KM3NeT report on monitoring and quality control setup

KM3NeT-INFRADEV GA DELIVERABLE: D4.5

Document identifier:	**KM3NeT-INFRADEV-WP4-D4.5_v2.0**
Date:	**25/09/2019**
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. A process of central importance is the generation and quality control of the public data. This document describes the quality control protocols that will be followed in the generation of open access data, and the plan to make these data compliant with the FAIR data principles.

I. Copyright notice

II. Delivery slip

	Names	Partner/WP	Date
Author(s)	Rodrigo G. Ruiz Jutta Schnabel	FAU	
Approved by	IB + PMB		25/09/2019

III. Document log

Issue	Date	Comment	Author/Partner
0.1	29/07/2019	First draft	Rodrigo G. Ruiz/FAU Jutta Schnabel/FAU
1.0	31/07/2019	Comments and edits by T. Gal, K. Graf, D. Dornic, U. Katz and J. Hofestädt	Rodrigo G. Ruiz/FAU Jutta Schnabel/FAU Uli Katz/FAU Kay Graf/FAU Tamàs Gal/FAU Jannik Hofestädt/FAU Damien Dornic/CPPM
1.1	26/08/2019	Comments by Rob van der Meer, edits by R. G. Ruiz, U. Katz	Rodrigo G. Ruiz/FAU Uli Katz/FAU
2.0	25/09/2019	Comments by Cristiano Bozza (internal reviewer), IB	Uli 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

ARCA	=	Astroparticle Research with Cosmics in the Abyss (KM3NeT neutrino astroparticle physics telescope)
CI/CD	=	Continuous Integration / Continuous Delivery-Deployment
CSA	=	Coordination and Support Actions
EOSC	=	European Open Science Cloud
ESFRI	=	European Strategy Forum on Research Infrastructures
FAIR	=	Findable Accessible Interoperable Reproducible
HDF	=	Hierarchical Data Format
IRF	=	Instrument Response Function
IVOA	=	International Virtual Observatory Alliance
GAVO	=	German Astrophysical Virtual Observatory
MIT	=	Massachusetts Institute of Technology
ORCA	=	Oscillation Research with Cosmics in the Abyss (KM3NeT neutrino particle physics detector)
PMT	=	Photomultiplier Tube
RI	=	Research Infrastructure
ROOT	=	Rapid Object Oriented Technology
VO	=	Virtual Observatory

VI. List of figures

VII. 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).

VIII. 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 describes the quality control protocols that will be followed in the generation of open access data, and the plan to make these data compliant with the FAIR data principles.

IX. 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 on our Universe, but also 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, 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 a certain embargo time (e.g. two years, to be ratified by the KM3NeT Collaboration) 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, the exploitation of 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 chain 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 data production and processing chain that will lead to the generation of KM3NeT open-access data is briefly reviewed. A plan to monitor quality control in the generation of open-access data is presented, and a plan for ensuring the compliance of the KM3NeT open-access data with the FAIR data principles is described.

A report on the implementation of the monitoring and quality control setup will be part of Deliverable D4.8, *Report on implementation and test of the open data system, including data generation, monitoring, archiving, example programs and access*.

2. The KM3NeT data processing chain: from PMT hits to open data

Every time the analogue signal of a PMT passes a threshold that typically corresponds to 0.3 photo-electrons, the signal is digitised into a data structure named hit that contains the timestamp corresponding to the leading edge of the analogue pulse, the time-over-threshold, and the PMT address. The hit rate per PMT is dominated by optical background and is about 8 kHz, which corresponds to a data production rate at the order of about 10 Gb/s for each detector. This data stream is sent from the detector to a computer farm on shore where it is filtered to obtain events of interest for physics analyses. The data filtering is done by running trigger algorithms, which select clusters of causally connected hits that are compatible with the different neutrino interaction topologies. The filtered data are saved on persistent storage for further analysis. As a result of the filtering, the data volume is reduced by three orders of magnitude w.r.t. the incoming data stream from the detector. Real-time processes, running in parallel to the data writing, provide information to an online physics analysis framework. Alerts of potential neutrino detections are sent to other observatories, and space and time correlations with external observations are investigated.

The processing of filtered data includes calibration and reconstruction of physics events that are finally stored in ROOT files [13]. The KM3NeT collaboration will produce scientific data products through the generation of high-level data consisting of samples of events including e.g. neutrino directions and energies. The quality of the scientific result depends on the accuracy of the detector calibration and on the reliability of the reconstruction algorithms; these aspects represent the scientific dimension of data quality control.

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 planned 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) [7], 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 the perspective of open-access data and data quality control, three main categories of high-level datasets should be considered. In addition to this standard processing of neutrino events, an immediate response relying on data from the running detector is needed for multi-messenger alerts.

A) **Datasets for peer-reviewed publications**: A dataset will be generated for each analysis published by the KM3NeT collaboration. Each of these datasets will contain the list of events

selected for corresponding publication, and each event will contain all the parameters related to the selection process. In addition to the event lists, further information necessary for the scientific interpretation of the published results will be made available. Full event data sets will be contained in data types B and C, see below.

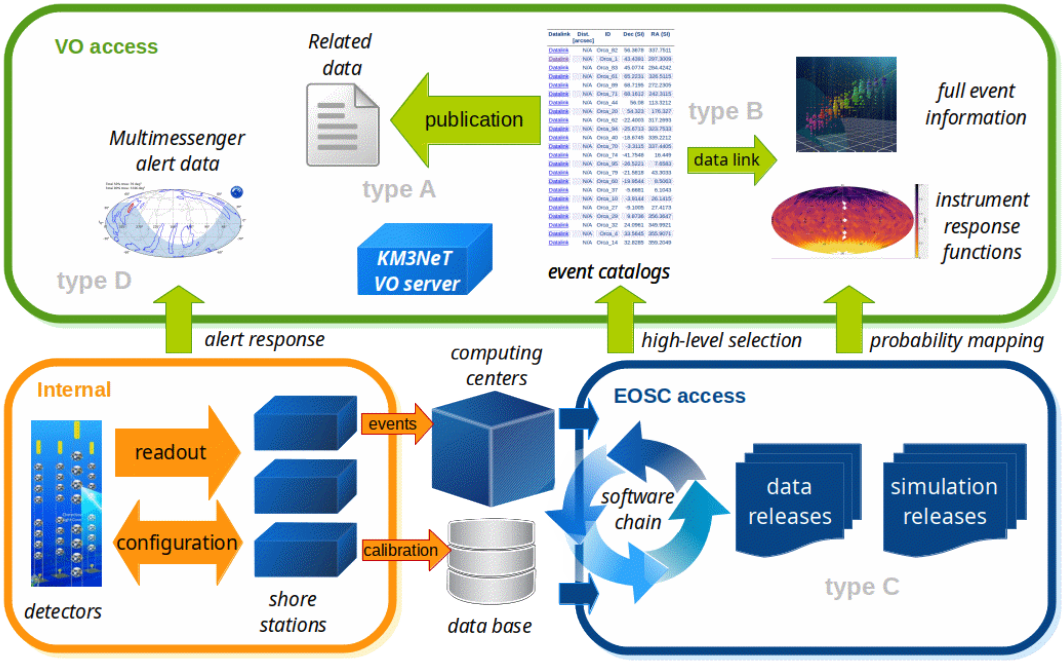

Figure 1: Open access data levels in KM3NeT

B) **Reduced high-level datasets:** These datasets are extended versions of the previous ones where all events defined as relevant for science studies for a given data taking period are included, and where each event contains at least all the parameters relevant for analyses corresponding to datasets of type A. Instrument Response Functions (IRFs) derived from simulation will be provided as a service to ensure usability for a wider science community without adding the burden of having to deal with the full simulation samples. Parametrisations of the systematic uncertainties, which are particularly relevant for neutrino oscillation analyses, will be also provided. While the access to astrophysically relevant data will be fully managed within the protocols of the Virtual Observatory, the access to datasets beyond the scope of astrophysics will be tailored considering the needs of the communities involved. Licensing issues will be addressed in deliverable 4.7, *Report on rules and conditions for data access*.

C) **The full set of processed data samples** and related simulation sets are considered for full publication. However, as both data volume and possible scientific application go beyond the scope of the Virtual Observatory protocols alone, access and software requirements can only be met within the scope of the EOSC, within which access rights, software availability and portability and data handling will be addressed.

D) **Datasets related to alert generation and follow-ups:** The main goal of multi-messenger astronomy is the simultaneous detection of different cosmic messengers produced in transient astrophysical events, which can be as short as a few seconds. To meet this objective, different observatories cooperate in generating and disseminating real-time alerts and following up incoming alerts. A key factor in this process is to generate the alerts as fast as possible. For this reason, and since the cooperation with other observatories may require to share the information in a specific format, these datasets will likely be produced by different processing chains, and shared in different platforms from the ones above. However, publication under the Virtual Observatory umbrella is aimed for by the KM3NeT collaboration.

3. Monitoring and quality control setup in the generation and publication of open access data

From the open-access data perspective, the KM3NeT data quality plan must ensure (i) that the published data are reliable and (ii) that the data are understandable and usable by external users. The high-level datasets described above will be made publicly available after being transformed into a format suited for public use.

3.1. Ensuring reliability in the production of data for open access

The generation of the open data will be done in an automated way by a chain of computer programs specifically designed for this purpose, which will run in batch processing. In order to guarantee the reliability of the output, a quality control protocol is being developed that will monitor the data processing.

3.1.1. Quality of the data processing chain

To ensure quality in the software development, the KM3NeT collaboration has set up a software quality plan that includes a version control system based on git [8]. This allows to keep track of all the changes implemented in the different programs and to generate metadata containing the versions of the different software packages used to produce each dataset, as well as changelog files containing information about the changes between different software versions. As an extension of the version control system, a Continuous Integration/Continuous Delivery (CI/CD) system has also been set up that allows to perform automatic tests on every modification introduced to the software. In addition to that, test productions will be used to validate the correctness of the software functionality. To ensure that the software documentation is up to date, the documentation will be embedded within the source code and dedicated tests will be designed within the CI/CD context to verify its consistency with the code. Software is made available through docker images [9] in a central software registry, which ensures full reproducibility of the data processing chain and its results, independent of the specific computing environment. Also, the implementation of a workflow governing regime using a workflow description language [15], [16] for the standard processing of KM3NeT events is currently investigated and aimed for. This should also be linked to rich metadata description for all data products and relevant processing inputs for the chain, ensuring both full documentation and traceability of the data

processing steps. Performance of the full processing chain is monitored through logging and benchmarking.

3.1.2. Production of open data sets

Performance of the program chain

Monitoring the performance of the program chain refers to the process of verifying that no software execution errors occur during the data transformation, and also that the output contains the expected results and is complete. The exchange formats are tailored to guarantee that the values of the parameters are exactly conserved during transcoding. The presence of execution errors will be automatically monitored by scripts that will check the log files produced during the software execution.

Event selection for open access event releases

For all events contained in the data releases and derived data products, a quality flagging according to the detector and data taking conditions during the measurement will be applied. The pass criterion for events to be contained in the published data will be internally established and constantly evolving, for which also a versioning and reviewing process will be established. For each generation of the open access data, a quality flag will be set, its definition and the rationale behind it documented, and the results benchmarked.

Data type specific quality criteria

A dedicated verification of the contents of the output files will also be envisaged. Since the level of detail and format of the open data is expected to differ for each of the dataset types described in section 2, the implementation of this verification depends on the type of the data set. Datasets of type A will be obtained after discarding a significant fraction of the reconstructed events by applying cuts on different reconstruction parameters in order to optimise the physics results. As a side effect the data volume will be significantly reduced, allowing to publish the full sets of data. For such datasets, the verification of the VO published event list contents will consist of a comparison with the files for internal use, where a one-to-one cross check for each event will be performed. This cross check will ensure that both files contain the same events, and that the values of the parameters are exactly the same in both files for each event. A unique identifier for each triggered event is assigned to the event, through which the inclusion of the events in different data sets can be traced. As these data sets correspond to KM3NeT published analyses, these analyses will additionally have undergone stringent peer reviewing procedures set up by the KM3NeT collaboration to ensure their compliance to scientific standards.

The level of detail for the published datasets of type B is currently under discussion. An unbiased Monte Carlo dataset containing the full list of simulated events with all the reconstruction parameters would be excessively large, and probably not necessary for the vast majority of the potential users. A possible solution under discussion is the publication of a full list of events where each event contains only the most relevant parameters, and probability distributions for signal neutrino events of the same parameters obtained from Monte Carlo simulations. In such a case, an additional cross check will verify the one-to-one correspondence between the provided distributions and those obtained from the full simulation samples. Publication of data sets of type A and B will fully concur with Virtual Observatory standards.

Access to the full releases of measurement and simulation data as regarded for type C data will no longer fall under the VO regime due to the currently limited VO capability to handle vast amounts of event-based data. However, the access will be standardized in line with the EOSC efforts. Here, access to the necessary software and documentation will be granted to external users for producing simulated data themselves, and for fully exploiting the potential of the processed data sets. The publication of new data releases will include an internal quality process, where consistency both with the physics expectations and with the latest detector modelling are checked, and versioning and documentation in relation to previous releases will be added. An additional test must be carried out before publishing new datasets, where the consistency with overlapping datasets will be verified. As an example, the corresponding tool will verify that all the datasets corresponding to peer reviewed publications can be retrieved from the unbiased dataset for the same data taking period.

The decision path with respect to the extent of publication of simulation data is detailed in Section 4.

3.1.3. Continuous improvement of data quality

Generally, quality management for software development and for addressing issues related to data quality are handled by applying best-practice coding conventions and a ticketing pipeline in git, through which alerts on inconsistencies, necessities for documentation or requests for further development are directed to the corresponding working group.

3.2. Compliance with the FAIR principles

Although the FAIR data principles define a series of criteria that data and metadata should meet in order to enhance their public usage [4], the KM3NeT collaboration is working to ensure that the data for internal usage are also FAIR compliant. In the following sections, the internal compliance with the FAIR principles is detailed, and the strategy for open access data compliance with FAIR is described.

3.2.1. Metadata for findability and accessibility

A database has been implemented which houses data and metadata related to different aspects of the KM3NeT research infrastructure. Amongst others, this database hosts metadata related to the data taking runs, and calibrations, as well as detector identifiers needed to find the existing data. Data storage at high-performance computing clusters is tracked and files are identifiable through a unique numbering system, where filenames contain the detector identifier and the run number. The metadata contain all the necessary information to track the data in each file down to the original raw data from which they were produced. Additionally, the information about the trigger parameters used in each run is also contained in the metadata. Metadata for software contain complete information about the software versions used to produce each data file as well as information about the computing environment.

Metadata are currently stored within the processed file, although future options for external storage of metadata are investigated to comply with high-performance data management systems like Dirac [10] within the EOSC. External metadata storage will also secure the future provenance if outdated data sets are deleted.

3.2.2. Standardization for access, interoperability and reusability

Currently two different frameworks are maintained, documented and developed for official use within the KM3NeT Collaboration that allow to use the data: the KM3pipe framework, which is developed in python language [11]; and the Jpp framework which is a C++ based software design. Complementing file storage in a ROOT-based format, an HDF5 format definition for both low and high-level data is envisaged, so that the data can be accessed by open source libraries without additional dependencies. All KM3NeT processing software is available in portable environments for use in docker [9] or singularity [14] to ensure portability, and partly available under MIT license.

Introduction of semantic metadata according to established conventions by the World Wide Web Consortium and extensions by the IVOA will further enhance the interoperability of the data processing chain and products.

3.2.3. Compliance of open access data with FAIR principles

The properties which make the KM3NeT data compliant with the FAIR principles will be propagated in their transformation to the open-access datasets. During the last few months, contacts have been made with the German Astrophysical Virtual Observatory (GAVO) [12], which is the German contribution to the IVOA. The first conversations between KM3NeT and GAVO members have been focused on the required standards for the publication of datasets corresponding to searches of cosmic neutrinos in the Virtual Observatory. By respecting and developing these standards it is ensured that the provided data will comply with the FAIR principles.

4. Decision points and mechanisms, next steps

The decision points and mechanisms as defined in deliverable 4.3 [17] are still valid and apply to the decisions in the context of this document.

A report on the implementation of the monitoring and quality control setup will be part of Deliverable D4.8, *Report on implementation and test of the open data system, including data generation, monitoring, archiving, example programs and access*.

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, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; 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: European Science Cluster of Astronomy & Particle Physics https://projectescape.eu/

7: International Virtual Observatory Alliance: http://www.ivoa.net/

8: GITLab software: https://about.gitlab.com/

9: Docker on OpenStack: N. Agarwal, B. Moreira (2014). doi:10.5281/zenodo.11783.

10: Distributed Research utilizing Advanced Computing: https://dirac.ac.uk/

11: KM3Pipe 8.13: Tamas Gal, Moritz Lotze (2019). doi:10.5281/zenodo.2662139.

12: German Astrophysical Virtual Observatory: https://www.g-vo.org/

13: ROOT: https://root.cern.ch/

14: Singularity: https://sylabs.io/docs/

15: Common Workflow Language: https://www.commonwl.org/

16: Workflows for e-Science: Taylor, I.J., Deelman, E., Gannon, D.B., Shields, M. doi:10.1007/978-1-84628-757-2

17: KM3NeT-INFRADEV: KM3NeT report on conceptional design of open data generation, archiving, test programs and access. KM3NeT2.0_WP4_D4.3