# Publishing Multi-Purpose Data Sets from KM3NeT

Jutta Schnabel[0000−0003−1233−7738] ⋆

Friedrich-Alexander Universität Erlangen-Nürnberg
jutta.schnabel@fau.de https://ecap.nat.fau.de/

**Abstract.** The KM3NeT neutrino detector, a water Cherenkov experiment to detect relativistic charged particles, is currently under construction at two deep-sea locations in the Mediterranean Sea. As cross-domain experiment between neutrino, astro-particle and astrophysics, data processing and data publication from KM3NeT draws on computing paradigms and standardization from all fields. In this contribution, key considerations for the provision of multi-purpose open data sets, interface options and interoperability requirements are presented.

**Keywords:** KM3NeT · neutrino astronomy · neutrino physics · open data

## 1 Introduction

The ARCA and ORCA detectors designed, constructed and operated by the KM3NeT Collaboration for the detection of high-energy neutrinos, follow identical technical design, data acquisition and processing principles. However, the detectors aim for two main scientific goals from different fields of physics. While ARCA is intended for the detection of high-energy neutrinos from the cosmos, spanning a wide range of astrophysical research topics, ORCA targets neutrino oscillation research utilizing atmospheric neutrinos and is therefore rooted in particle physics. The KM3NeT science cases interlink to a wide range of fundamental physics, contributing e.g. to dark matter searches or multi-messenger follow-ups of gravitational wave alerts, see [2].

Implementing open science standards in KM3NeT therefore is not only a necessity to ensure full scientific exploitation of the generated data both inside KM3NeT and in the general physics community, it also requires a leveled and diversified approach to meet the respective needs of the various communities. This includes the generation of accessible and interoperable open science products, harmonizing data and software standards and an ongoing search for

⋆ for the KM3NeT collaboration

best-practice solutions to explore the technical, organizational and scientific limitations of data sharing while aiming for maximum transparency of scientific procedures and outcomes.

## 2 Data generation and processing

The method to detect high-energy neutrinos in KM3NeT is based on the measurement of Cherenkov photons from charged particles generated in neutrino interactions in or around the instrumented volume of the detector in the deep sea. Pressure-resistant glass spheres, so-called digital optical modules (DOMs), each containing 31 photomultiplier tubes, register individual photons with timing precision on the nanosecond scale. In its final configuration, each building block will consist of 115 detection lines with 18 DOMs each, transferring all instrument readout, especially PMT signals above a pre-set threshold, to the shore station. As a single DOM sends digitized PMT readouts and control data at a rate of up to 100 $\mathrm{Mb\,s^{-1}}$, this continuous data stream needs to be heavily reduced before further processing and storage. In the final KM3NeT configuration, two building blocks for ARCA and one building block for ORCA with denser spacing to target lower-energy GeV-scale neutrinos are planned.

Data processing therefore follows a tier-based approach [8], where initial filtering for particle interaction-related photon patterns (triggering of photon "hits") serves to create data at a first event-based data level. In a second step, processing of the events, applying calibration, particle reconstruction and data analysis methods leads to enhanced data sets, requiring a high-performance computing infrastructure for flexible application of modern data processing and data mining techniques.

For physics analyses, derivatives of these enriched data sets are generated and their information is reduced to low-volume high-level data which can be analysed and integrated locally into the analysis workflow of the scientist, see Figure 1. For interpretability of the data, a full Monte Carlo simulation of the data generation and processing chain, starting at the primary data level, is run to generate reference simulated data for cross-checks at all processing stages and for statistic interpretation of the particle measurements.

### 2.1 Event-based data generation

Data processing at the DAQ level follows paradigms of particle physics and utilizes computing and software methodological approaches of this community. At the shore stations, event triggering in the Data Acquisition (DAQ) system leads to a significant reduction of the data stream. The data stream also includes relevant instrumentation readouts for a comprehensive understanding of data taking conditions. Photon-related information is written to ROOT-based [3] tree-like data structures and accumulated during a predefined data taking time range of usually several hours (so-called data runs) before being transferred to

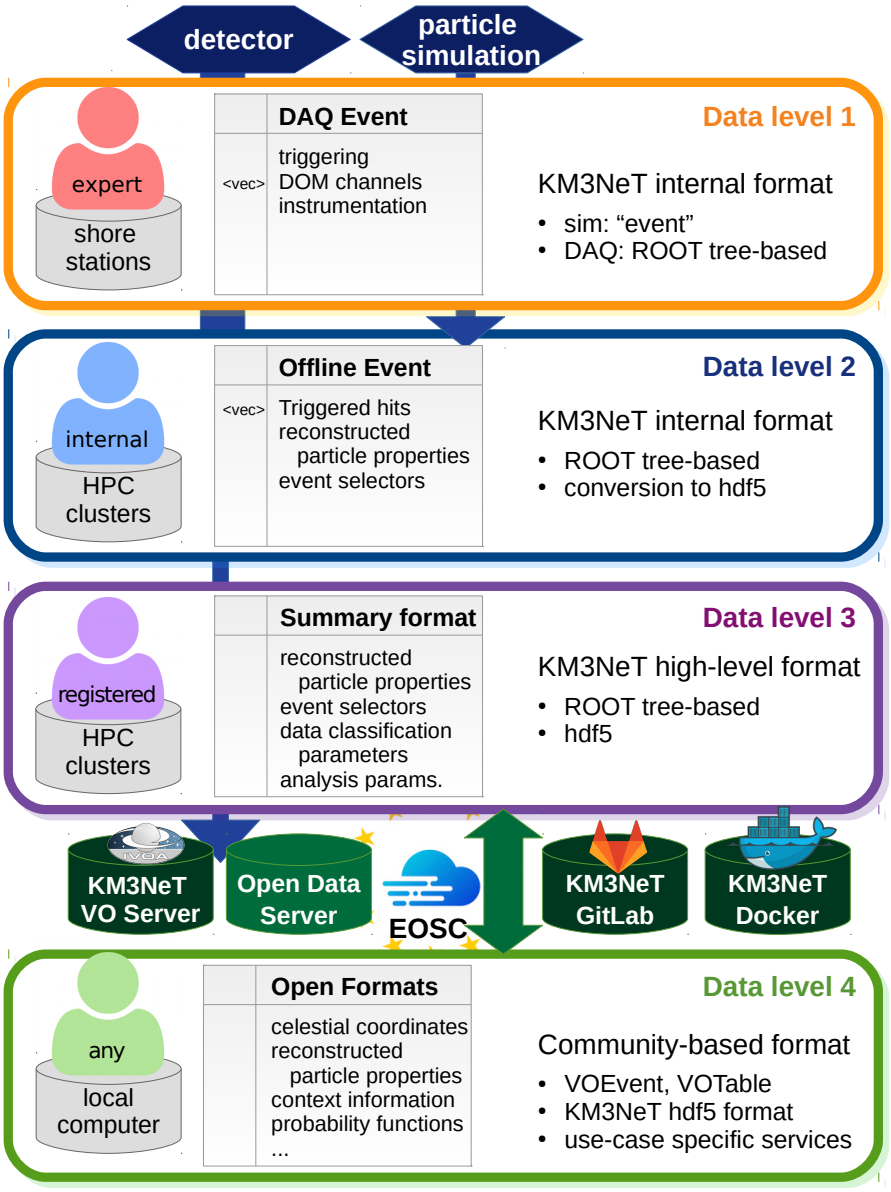


**Fig. 1.** KM3NeT data levels related to open data publication, including data format description, user access rights and open data publication layer.

high-performance computing (HPC) clusters. Instrumentation and environmental data collected at the detector site are stored separately in a central database. Acoustic and other environmental data serve as basis for Earth and Sea-science initiatives. Access to this information following an Open Science approach is under development, however, it will not be covered in the scope of this report.

Both the complex process of neutrino detection in a natural environment and the low expected count rate of the cosmic neutrino signal in comparison to atmospheric background events necessitates the full modelling of particle generation, detector response and data processing. To this end, a dedicated simulation chain, starting form cosmic air-shower particle generation or astrophysical neutrino flux assumptions, replicates the complete data-processing pipeline. At the event generation level, photon distributions induced by these particles within the detection volume are generated, masked by a simulation of the detector response and treated to the same processing as measurements starting from the second data level of the offline event format.

### 2.2 Event data processing

Processed event data sets at the second level represent input to physics analyses, e.g. regarding neutrino oscillation and particle properties, and studies of atmospheric and cosmic neutrino generation. Enriching the data to this end involves probabilistic interpretation of temporal and spatial photon distributions for the reconstruction of event properties in both measured and simulated data, and requires high-performance computing capabilities. Due to the distributed infrastructure of the KM3NeT building blocks and the contribution of computing resources from various partners, data processing will, in the final detector configuration, necessitate a federated computing approach, the implementation of which is prepared through containerization of the required software and testing of distributed resource management approaches. In this context, the use of a middleware like e.g. DIRAC[1] is explored, again linking closely to the particle physics community.

Access to data at this level is restricted to collaboration members due to the intense use of computing resources, the large volume and complexity of the data and the members' primary exploitation right of KM3NeT data. However, data at this stage is already converted to HDF5[2] format as a less customized hierarchical format. This format choice increases interoperability and facilitates the application of data analysis software packages used e.g. in machine learning and helps to pave the way to wider collaborations within the scientific community utilizing KM3NeT data.

[1] Distributed Infrastructure with Remote Agent Control Interware, http://diracgrid.org/

[2] The HDF5 file format, https://www.hdfgroup.org/

### 2.3 High level data and data derivatives

**Summary formats and high-level data** As mostly information on particle type, properties and direction is relevant for the majority of physics analyses, a high-level summary format has been designed to reduce the complex event information to simplified arrays which allow for easy representation of an event data set as a table-like data structure. Although this already leads to a reduced data volume, these neutrino data sets are still dominated by atmospheric muon events at a ratio of about $10^6 : 1$. Since for many analyses, atmospheric muons are considered background events to both astrophysics and oscillation studies, publication of low-volume general-purpose neutrino data sets requires further event filtering. Here, the choice of optimal filter criteria is usually dependent on the properties of the expected flux of the signal neutrinos and performed using the simulated event sets.

**Event simulation derivatives as service** To correctly judge the statistical significance of a measured neutrino event rate, the full high-level simulation data sets are used in KM3NeT internal studies to ensure a high accuracy of the count rate estimate. As handling these large data sets is impractical for inter-experimental studies, but the information is crucial for the interpretability of the data, parameterized distributions of relevant observables need to be derived from the simulation data sets and offered as services. Even in absence of significant neutrino measurements in the construction phase of KM3NeT, offering sensitivity estimates as in [4] for given models is beneficial for the development of common research goals and the development of a corresponding open service is currently under investigation.

## 3 Meeting the Open Science challenge

### 3.1 Publishing FAIR data

The widely-accepted paradigm for open science data publication requires the implementation of the FAIR principles [7] for research data. This involves the definition of descriptive and standardized metadata and application of persistent identifiers to create a transparent and self-descriptive data regime. Interlinking this data to common science platforms and registries to increase findability and the possibility to harvest from the data through commonly implemented interfaces is as mandatory as is the definition of a policy standard including licensing and access rights management. In all these fields, the standards of KM3NeT are currently developing, including the implementation of a data management plan, the installation of a data provenance model including the application of workflow management, and the setting of data quality standards. In this development process, the application of existing standards especially from the astrophysics community, the development of dedicated KM3NeT software solutions and the integration of the KM3NeT efforts developed during the KM3NeT-INFRADEV

project[3] are integrated into the ESCAPE project[4], which forms the main development environments for open data publication in KM3NeT.

### 3.2 Approach to the Virtual Observatory standard

The Virtual Observatory (VO) standards[5] serve to create an interface between astronomy-related data resources from astrophysics experiments which act as data providers. The focus is on the scientific end user to easily interface from their personal computer with the provided data sets. The KM3NeT collaboration is a data provider to the VO and operates a data server[5] running the DaCHS software[1]. The well-developed data sharing regime of the VO serves well as a guideline for the implementation of astrophysical data sharing in the KM3NeT collaboration. However, considering the role of neutrino physics just at the edge of use for astronomical studies, KM3NeT data integration also meets some limitations considering the scientific usability of the provided data sets. In addition to that, publication of data through VO standards is clearly limited to astronomy-related data in a celestial reference frame.

**Neutrino sets in the VO** Tabulated high-level neutrino event data can be provided through the VO registry, utilizing access protocols like the Table Access Protocol (TAP) and query languages like the Astronomical Data Query Language (ADQL). To query these data sets related to astronomical sources, the Simple Cone Search (SCS) protocol allows to pick specific events according to particle source direction, using Unified Content Descriptors (UCDs) to identify the relevant table columns. The underlying data format is the VOTable which allows for metadata annotation of data columns. As the DaCHS software provides input capabilities for various formats like FITS[6] or text-based tables on the server side, a common KM3NeT open event table format can be chosen quite independently and the interface adapted such that high-level neutrino data sets can be both offered through the VO and alternative access protocols, as long as the required metadata description is handled adequately.

VO standards are at the current stage not fully adapted to the inclusion of neutrino data and require development of metadata standards for easy interpretability of the data, a matter which is targeted within the ESCAPE project. Open questions in this regard are the linkage of observation probabilities to a given event selection, the inclusion of "non-observation" in a given field of view and within a given time as relevant scientific information to retrieve from the services, and the introduction of a dedicated vocabulary for the description of neutrino data. This vocabulary will need to be developed within KM3NeT as a matter of internal standardization, however, the process will draw guidance from the VO expertise and framework.

[3] see https://www.km3net.org/km3net-infradev/

[4] European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures, https://projectescape.eu.

[5] at http://vo.km3net.de/

[6] Flexible Image Transport System, https://fits.gsfc.nasa.gov/.

**Multimessenger alerts** Single or stacked neutrino events of high astrophysical signal probability will be selected in KM3NeT to trigger an alert to other observatories indicating a possible target for multimessenger observations [6]. The VOEvent format, together with the VOEvent Transport Protocol as implemented in the Comet software[7] will be used to distribute these events as outgoing alerts. As the format is specifically tailored to the use in multimessenger alerts, indicating a quite restricted scientific target, the providing of context information for the events can be specifically adapted to this use case. However, harmonization of metadata standards like parameter descriptors and event identifiers in reference to the full neutrino event sets will also have to be implemented.

**Providing simulation-driven services** Providing context information on a broader scale in the form of e.g. sensitivity services and instrument response functions alongside the VO-published data sets is still under investigation. On the one hand, VO access protocols like TAP facilitate the use of standardized queries on services. On the other hand, integrating those services with the data sets in a meaningful and user-transparent way, e.g. through VO Datalinks, still requires further deliberation. Therefore, the development of these services will be use-case driven and also include the application of similar services for studies in other fields of KM3NeT research beyond astrophysics.

### 3.3 Developing interfaces

**Data modelling and access** Publishing interfaces to data sets requires the generation of an accessible data model according to the W3C[8] standards to well define both interface and data format, and will be provided alongside the KM3NeT data. For a basic KM3NeT open data format, the requirements on the format and access methods are building on current software solutions already in use within the KM3NeT collaboration. Interfacing to full tabular data sets beyond the VO requires the choice of a standard data format and the provision of a dedicated but easily usable software interface to this data. Here, a python-based approach is favoured due to the wide use of the related tools in the science community, with access to data sets ensured e.g. through the astropy package[9] for FITS tables or python data analysis packages with HDF5 reader capabilities like pandas[10]. Usage examples and exemplary workflows for analyses will be provided as Jupyter[11]-notebooks.

**Software requirements** As KM3NeT software development also follows an Open Software approach, dedicated high-level analysis software is made available to facilitate proper handling of the research data. The software is provided

[7] J. Swinbank, Comet, https://comet.transientskp.org.
[8] https://www.w3.org/
[9] The Astropy Project https://www.astropy.org/
[10] see https://pandas.pydata.org/
[11] see the Project Jupyter https://jupyter.org/

as Docker[12] or Singularity[13] virtual containers on the respective KM3NeT registry server, with usage examples and source code made available through the KM3NeT Gitlab[14] instance. The easy deployment of the software should enable users to locally analyse smaller data sets on their personal computer either in a containerized environment or by installing, if preferred, custom-made python-packages for KM3NeT data handling. Access to services and data sets will probably be implemented through the creation of a dedicated REST-API[15] to the KM3NeT open data server and facilitated by providing dedicated function wrappers. This software development will also be integrated into the development of the Open Software and Service Repository in the ESCAPE project.

**Data access** However, this approach is only feasible for small-scale data sets. Beyond this, access to external computing resources is required and computing needs to be moved to an integrated cloud environment allowing user authentication, easy management of the software environment and running of custom scripts on data stored by federated data providers. This transition from a locally restricted user level to integrated open science computing is currently under development for the European Open Science Cloud (EOSC), and the software choices for the KM3NeT Open Science environment as outlined above aim to integrate into this development. Here, containerized software will be made available, while larger data sets can be queried from the EOSC data lake, and the use of Jupyter notebooks for the use in an Open Science Portal is currently also envisioned.

## 4 Conclusion

KM3NeT data generation and publication draws from two different worlds of experimental physics by being rooted in particle physics especially in the generation of the data, while targeting both particle and astrophysics on the research level. Providing meaningful interfaces to research products like data sets and flux sensitivity estimates and instrument response functions can therefore not only rely on ready-made data sharing paradigms like the VO standards but holds the potential for new developments in order to make high-energy neutrino data FAIR in both worlds. Here, the ESCAPE and INFRADEV projects provide the ground for these common endeavors from both communities and the developing KM3NeT Open Science standards are integrated into this process. As the EOSC environment is also aimed at fostering interaction between scientists and co-development of projects and software, the scientist as user is here expected to become a potential partner in research and development, shifting the focus from

[12] see https://www.docker.com/
[13] see https://sylabs.io/singularity/
[14] see Gitlab (https://about.gitlab.com/) at https://git.km3net.de/
[15] Fielding, R.T., Architectural Styles and the Design of Network-based Software Architectures, UCI, 2000

the sole providing of the data to the interaction amongst researchers. Therefore, both at the current development stage and in the future, KM3NeT Open Science first and foremost builds on the knowledge-sharing and common ideas in the wider science community.

## References


1. Demleitner, M. et al., Virtual Observatory publishing with DaCHS. Astronomy and Computing, Volume 7, p. 27-36., published 2014. https://doi.org/10.1016/j.ascom.2014.08.003
2. Adrián-Martínez, S et al., Letter of intent for KM3NeT 2.0. Journal of Physics G: Nuclear and Particle Physics, Volume 43, Number 8, published 2016. https://doi.org/10.1088/0954-3899/43/8/084001
3. Brun, R., Rademakers, F., ROOT - An Object Oriented Data Analysis Framework. Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389, p. 81-86, published 1997. http://root.cern.ch/
4. Aiello, S. et al., Sensitivity of the KM3NeT/ARCA neutrino telescope to point-like neutrino sources. In: Astroparticle Physics, Volume 111, p. 100-110, published 2019. https://doi.org/10.1016/j.astropartphys.2019.04.002
5. Arviset, C. et al., IVOA Architecture. IVOA Note 2010-11-23. http://www.ivoa.net/documents/Notes/
6. F. Huang, F. et al., Realtime Multi-Messenger Program of KM3NeT. XXIX International Conference on Neutrino Physics and Astrophysics, July 2020.
7. Wilkinson MD et al., The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific Data, Volume 3, 160018. Published 2016 Mar 15. https://doi.org/10.1038/sdata.2016.18
8. Hofestädt, J., Computing in the KM3NeT Research Infrastructure. In: VLVnT-2018, EPJ Web of Conferences Volume 207, 08001. Published 2019 May. https://doi.org/10.1051/epjconf/201920708001