KM3NeT INFRADEV – H2020 – 739560

KM3NeT report on conceptional design of open data generation, archiving, test programs and access

KM3NeT-INFRADEV GA DELIVERABLE: D4.3

Document identifier:	**KM3NeT-INFRADEV-WP4-D4.3**
Date:	**06/10/2018**
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. In addition, the data archive needs to be set up and the user access interface and authorization established. In this document, we describe the conceptional design of the open data generation including test programs, the concept of the archiving models and open data access.

I. Copyright notice

II. Delivery slip

	Names	Partner/WP	Date
Author(s)	D. Stransky, K. Graf and U. Katz	FAU / WP4	06/10/2018
Approved by	PMB and KM3NeT IB		04/10/2018

III. Document log

Issue	Date	Comment	Author/Partner
1.0	25/05/2018	Initial draft derived from KM3NeT-INFRADEV-WP4-D4.2_v1.4	D. Stransky/FAU
1.1	12/07/2018	Include remarks as received by WP4	D. Stransky/FAU
1.2	08/08/2018	Finishing touches	K. Graf/FAU
1.3	27/08/2018	More finishing touches	U. Katz/FAU
1.4	31/08/2018	Version for internal review	U. Katz/FAU
1.5	23/09/2018	Version including Rob van der Meer's comments	U. Katz/FAU
2.0	06/10/2018	Version after including comments from IB review (C. Markou, S. Navas)	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

ANTARES	=	Astronomy with a Neutrino Telescope and Abyss environmental RESearch (first deep-sea neutrino telescope)
ARCA	=	Astroparticle Research with Cosmics in the Abyss (KM3NeT neutrino astroparticle physics telescope)
ASTERICS	=	Astronomy ESFRI & Research Infrastructure Cluster
e-IRG	=	e-Infrastructure Reflection Group
EOSC	=	European Open Science Cloud
ESFRI	=	European Strategy Forum on Research Infrastructures
ESS	=	Earth and Sea Sciences
IB	=	Institutional Board (KM3NeT governing body)
IVOA	=	International Virtual Observatory Alliance
GAVO	=	German Astrophysical Virtual Observatory
GNN	=	Global Neutrino Network
LTER	=	Long-Term Ecosystem Research
OM	=	Optical Module
ORCA	=	Oscillation Research with Cosmics in the Abyss (KM3NeT neutrino particle physics detector)
PMB	=	Project Management Board
PMT	=	Photomultiplier Tube
RDA	=	Research Data Alliance
RI	=	Research Infrastructure
VO	=	Virtual Observatory
W3C	=	World Wide Web Consortium

VI. List of figures

none

VII. List of tables

none

VIII. Project summary

KM3NeT-INFRADEV

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. In addition, the data archive needs to be set up and the user access interface and authorization established. In this document, we describe the conceptional design of the open data generation including test programs, the concept of the archiving models and open data access.

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 (Adrián-Martínez, 2016).

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 from 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 depth of several kilometres (KM3NeT Collaboration, 2017).

The KM3NeT Collaboration has developed a data policy plan (KM3NeT-InfraDev, 2017) reflecting the research, educational and outreach goals of the facility. For a certain embargo time (e.g. two years, to be ratified by the KM3NeT Collaboration) after data taking, the processing, quality control and exploitation of the data is granted to the collaboration members as a return for constructing, maintaining and operating the facility. During this phase, each collaboration member has full access rights to all data, software and know-how. The collaboration commits itself to process the data during the embargo phase so as to generate high-quality calibrated and reconstructed event data suited for a wider user community. These data will be made publicly available after the embargo time 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 prompt dissemination of scientific or methodological results achieved during the embargo time is in the responsibility of the KM3NeT Collaboration. The scientific responsibility and the publication rights for results derived from public data is with the scientists performing the corresponding analyses. The KM3NeT Collaboration offers analysis support to external analysers on their request, and after scrutinising the validity of the respective analyses. In this case, both the external scientists and the KM3NeT Collaboration will author the resulting publications.

This document details the conceptual design of the open data generation, archiving, test programs and access. The concepts presented have been validated versus quantitative estimates of key parameters such as data rates, storage and CPU requirements. A quantitative account of these performance parameters will be presented together with the technical implementation of the concepts in deliverable D4.8, Report *on implementation and test of the open data system, including data generation, monitoring, archiving, example programs and access*. The concepts are based on e-infrastructure commons developed by KM3NeT-INFRADEV in cooperation with the KM3NeT Collaboration, and are guided by expert group recommendations (ASTERICS, e-IRG, RDA, W3C, etc.) and a best-practice approach. Ultimately, the open access to KM3NeT data should be integrated in the EOSC. The implementation of that goal is, however, outside the scope of the KM3NeT-INFRADEV project. All concepts described in the following are consistent with the goal of EOSC integration.

2 Open data generation and monitoring

2.1 Data generation

One of the guiding principles of the KM3NeT data and software handling under the FAIR data principle (European Commission - Directorate-General for Research & Innovation, v3,2016) is to enable the reproducibility and re-usability of techniques and results. For data generation and processing, this means the full data provenance (see Chapter 2.4) is maintained, in this case via meta-data stored directly in the data files or in the central database of the project.

The first step of the data generation is the online filtering of events out of the bit stream arriving from the deep-sea infrastructures of ARCA and ORCA in an on-shore computer farm. The raw data emerging from this filtering is further processed together with the meta-data. The meta-data contains data-taking information, detector conditions, calibration data, software identifiers, etc. With automated scripts, single triggered events are processed by several reconstruction algorithms that allow the estimation of the direction, energy deposition and time of a triggered physics event. The event data is grouped in – typically hours-long – detector runs and is written to disk in a ROOT data format (CERN, n.d.). A data volume of roughly 50 000 reconstructed neutrino events and 1.6 billion reconstructed atmospheric muon events is expected per year.

In addition to detector data, Monte Carlo simulations of the detector are generated with the same settings and conditions as present during the real data taking. The simulation data go through the same processing steps as the detector data in order to provide meaningful interpretation of the real data. The reconstructed Monte Carlo events are written to disk together with the Monte Carlo truth information, using the same data format as for the real data.

Different resource level scenarios are envisioned:

a. A high-level scenario for fast and easy access, simple example and analysis programs;
b. A low-level scenario for comprehensive and detailed low-level analyses;
c. Special scenario for individual analyses within the scientific community.

The former will only contain the most important information while the latter includes more detailed event information. All scenarios include a mild selection cut for neutrinos so that the bulk of background events – mostly atmospheric muon events – is rejected. As a first implementation, only the high-level scenario will be made publicly available for the full data set. In addition, data leading to publications will be made available, where the level of detail and thus the required resources will be set depending on the analysis.

In addition to an event identifier, the high-level data includes the following reconstruction parameters and their error estimates where appropriate: declination, right ascension, zenith angle, azimuth angle, time, energy, number of hits, inelasticity, event, and other (important) reconstruction parameters characterising the event. Additional event information may be provided based upon user feedback, such as the integrated charge of all PMTs in the event, etc. As several reconstruction algorithms are available, for each parameter only the result from the best-suited algorithm is provided together with

uncertainties. Documentation of the definition of the parameters and of their corresponding error estimates will be provided along with the data, see Chapter 0.

For simulated events, it is envisioned to provide data for both scenarios, as well. However, the Monte Carlo scenario to be made public is not decided yet. The following scenarios exist:

1. Full scenario with data for all runs, including detector conditions (so-called run-by-run simulation);
2. Reduced scenario including only data of characteristic runs (e.g. once per week or month);
3. Provision of simulation software and necessary input parameters, as well as test tools for quality control.

For all data the detector efficiencies and sensitivities with respect to direction, energy and time will be provided.

The KM3NeT detector provides data important for earth and sea sciences (ESS). The interest of this community in the data is assessed in WP8 of the KM3NeT-INFRADEV project. The relevant neutrino telescope data include optical activity and acoustic data. Regarding the optical activity, minimum bias data can be provided in the low-level scenario provided via occasional runs (e.g. 10 min/day for full detector). In addition mean detector rates (per PMT or OM) for all time slices and/or runs are provided.

2.2 Monitoring and quality control

The data generation steps are performed automatically and are permanently monitored with quality-control routines. This includes test programs run on the output of the whole data-generation chain. Human interaction is only needed if warnings or errors occur.

The monitoring includes reconstruction of selected events with a dedicated, fast reconstruction algorithm (which also serves for disseminating alerts in case of neutrino observations with a high probability of cosmic origin) as well as a visualisation of specific monitoring data.

2.3 Data publication

The ROOT format is used internally by the KM3NeT Collaboration. For the open data generation, it will be converted with automated scripts to HDF5 (HDF Group, n.d.) and FITS (Wells, Greisen, & Harten, 1981), standards defined by the ASTERICS project (ASTERICS, n.d.).

The data is made publicly available after an embargo time of expectedly two years under an open-access policy on a web-based service, see Chapter 0. The embargo time is given the KM3NeT Collaboration as a return for hosting and operating the facility and allows the collaboration to process and exploit the generated data first. For the low-level scenario, special web-based services will be needed, which will be developed together with EOSC efforts.

It is planned to publish the data on a yearly basis including updated versions of old data in case of substantially improved reconstruction algorithms. Exceptional updates are performed in case of special events. The required amount of disk space per year for an assumed low-level scenario including hit data is estimated to be several hundreds of GB per year for the full detector. This data estimate includes atmospheric muons and run-by-run MC simulation events as well as the optical background rates (but not the acoustic data). This corresponds to a maximum disk space usage of 32 TB after ten years including yearly versioning. The assumed data format is HDF5.

Together with the data, the software to read, analyse and eventually simulate the data will be provided – thus creating an open data access platform. This will be done in a software repository, following (or preferably integrating in) the example of (8). The final goal is an open access data and software platform within the EOSC.

2.4 Data provenance

In order to be able to retrace the data processing history, the versions of the software release used for data taking, calibration, processing, etc. of the data contained in a file will be stored as meta-data in the same file. For example, the previous data levels or the calibration data used to generate a file will be identifiable in this way. Thus, it will be possible to fully identify, find and access the data, which a file depends on, by using the metadata contained in the file itself, and to rerun the programs with which the file has been produced. In addition, the provenance information can be used in queries during data access to filter for specific things.

To generate the data provenance, the developments of the IVAO provenance data model (IVOA, n.d.) are followed.

3 Archiving and access

For the first implementation of the high-level scenario, the data is to be provided in a virtual observatory using IVOA services (International Virtual Observatory Alliance, n.d.). The data will be stored as VOTables (IVOA, n.d.). As an initial step, the ANTARES data has been made available on GAVO (German Astrophysical Virtual Observatory, n.d.).

To ease the access to the data and software, they will be accessed via a web-based user environment. In order to get access permission to the data and the service, users have to confirm an online user agreement. This agreement states the terms and conditions of access and includes rules for authorship, scientific responsibility and property rights. These issues are addressed and prepared in the current project and are to be finalised subsequently in consultation with funding agencies and stakeholders. For establishing access to the environmental data, also cooperation with the LTER project is envisaged.

The data user environment will include examples for how to access the archive as well as a documentation for how to use the interface and the available data. In addition, example analysis programs will be provided that can be applied on the data via the user environment. Users may also apply their own analysis programs on the data via this environment. Furthermore, the service contains a feedback mechanism for user questions.

As a design example, the IceCube Public Data Access portal (IceCube Collaboration, n.d.) is considered.

Additional remarks:

- User environment, web interface, example programs and documentation are continuously checked for completeness, consistency, functionality, etc. These checks will be automated and will require human interference only if problems will be detected.
- Best-practice cooperation models are worked out, especially within the GNN (Global Neutrino Network, n.d.), ASTERICS (ASTERICS, n.d.) and EOSC projects[1].

[1] KM3NeT participates in the ESCAPE proposal to the INFRAEOSC-04-2018 call under the H2020-INFRAEOSC-2018-2020 call.

4 Decision points and mechanisms

The development of the software solutions for open data access is a continuous process pursued inside work package 4 (WP4) of KM3NeT-INFRADEV and regularly presented to the PMB and the KM3NeT Collaboration. Some decisions on key parameters and scope of this work, however, need to be be taken during the remainder of the KM3NeT-INFRADEV project that need to be endorsed by the current KM3NeT governance or, alternatively, may become part of the ERIC statutes. These are:

1. The **embargo time** during which the KM3NeT data remain proprietary (see Section 1). As a working hypothesis, an embargo time of two years is currently assumed. A well-justified concrete proposal, however, will require an assessment of the latency caused by processing, calibrating, and verifying the raw data, of producing Monte Carlo simulations, and of performing the key science analyses, as well as the definition of these analyses. In cooperation with the KM3NeT Collaboration, WP4 will submit a corresponding proposal to the PMB by mid-2019. The resulting recommendation is to be endorsed by the KM3NeT IB and fed into the ERIC negotiations.

2. The **level of data to be made public** (see Section 2.1). This decision is tightly bound to the availability of human and computing resources for implementation of open data access. Even though this implementation will not be part of KM3NeT-INFRADEV, a guideline on the objectives to be targeted by KM3NeT-INFRADEV will be important. It is expected that a corresponding decision will be prepared by the PMB and taken by the KM3NeT IB by early 2019.

3. The **Monte Carlo scenario** (see Section 2.1). This decision requires a detailed scrutiny of existing simulation software and strategies; in particular, it will be necessary to evaluate which simulation strategies and results will be required for an optimal use of open-access data. A proposal will be prepared by WP4 and the KM3NeT Collaboration in summer 2019 and submitted to PMB and the KM3NeT IB.

5 References

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

ASTERICS. (n.d.). *ASTERICS homepage*. Retrieved 06 2017, from https://www.asterics2020.eu

CERN. (n.d.). *ROOT Data Analysis Framework homepage*. Retrieved 06 2017, from https://root.cern.ch

European Commission - Directorate-General for Research & Innovation. (v3,2016). *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020.*

European Open Science Cloud. (n.d.). *EOSC for Research Pilot Homepage*. Retrieved 06 2017, from https://eoscpilot.eu

German Astrophysical Virtual Observatory. (n.d.). *GAVO Webpage*. Retrieved 08 08, 2018, from http://www.g-vo.org

Global Neutrino Network. (n.d.). *GNN Webpage*. Retrieved 08 08, 2018, from https://www.globalneutrinonetwork.org

HDF Group. (n.d.). *HDF5 support page*. Retrieved 06 2017, from https://support.hdfgroup.org/HDF5/

IceCube Collaboration. (n.d.). *Public Data Acesses*. Retrieved 08 08, 2018, from https://icecube.wisc.edu/science/data/access

International Virtual Observatory Alliance. (n.d.). *IVOA Webpage*. Retrieved 08 08, 2018, from http://www.ivoa.net

IVOA. (n.d.). *IVOA Provenance Data Model*. Retrieved 08 08, 2018, from http://www.ivoa.net/documents/ProvenanceDM/index.html

IVOA. (n.d.). *VOTable Format Definition*. Retrieved 08 08, 2018, from http://www.ivoa.net/documents/VOTable/

KM3NeT Collaboration. (2017). *KM3NeT Homepage*. Retrieved May 29th, 2017, from https://www.km3net.org

KM3NeT-InfraDev. (2017). *The KM3NeT Data Management Plan.* KM3NeT-InfraDev.

Wells, D. C., Greisen, E. W., & Harten, R. H. (1981, 06). FITS: A Flexible Image Transport System. *Astronomy and Astrophysics Supplement Series, 44*, 363-370.