KM3NeT INFRADEV – H2020 – 739560

Report on implementation and test of the open data system, including data generation, monitoring, archiving, example programs and access

KM3NeT-INFRADEV GA DELIVERABLE: D4.8

Document identifier:	**KM3NeT-INFRADEV-WP4-D4.8_v0.4**
Date:	**11/12/2020**
Work package:	**WP4**
Lead partner:	**FAU**
Document status:	**For endorsement by IB**
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. This report provides an overview for the full system of open science services from KM3NeT, as well as the measures to ensure the sustainable operation of the Open Science System, including its testing.

I. Copyright notice

II. Delivery slip

	Names	Partner/WP	Date
Author(s)	T. Gal, S.R. Gozzini, K. Graf, S. Hallmann, J. Hofestädt, S. Raab, J. Schnabel Z. Aly, F. Huang R. G. Ruiz D. Stavropoulos	FAU CNRS FOM NCSR-D	20/10/2020
Approved by			

III. Document log

Issue	Date	Comment	Author/Partner
0.1	04/11/2020	First draft	T. Gal, S.R. Gozzini, K. Graf, S. Hallmann, J. Hofestädt, S. Raab, J. Schnabel / FAU Z. Aly, F. Huang / CNRS R. G. Ruiz / FOM D. Stavropoulos / NCSR-D
0.2	17/11/2020	Internal working group review	R. G. Ruiz / FOM
0.3	01/12/2020	Additional review	U. Katz / FAU
0.4	11/12/2020	Corrections for IB endorsement	S. Navas / UGR

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 the collaborating projects.

V. Terminology

ADF	Acoustic Data Filter
ADQL	Astronomical Data Query Language
ANTARES	Astronomy with a Neutrino Telescope and Abyss environmental RESearch project
API	Application Programming Interface
ARCA	Astroparticle Research with Cosmics in the Abyss
CCSN	Core-Collapse SuperNova
CI/CD	Continuous Integration / Continuous Delivery-Deployment
CSV	Comma Separated Value (data format)
CWL	Common Workflow Language
DAQ	Data AcQuisition
DIRAC	Distributed Infrastructure with Remote Agent Control
DOI	Digital Object Identifier
DOM	Digital Optical Module
DU	Detection Unit
ELOG	Electronic Logbook
EMSO	European Multidisciplinary Seafloor and water column Observatory
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, Reusable
FITS	Flexible Image Transport System
GAVO	German Astrophysical Virtual Observatory
GCN	Gamma-ray Coordinates Network
GUI	Graphical User Interface
GW	Gravitational Waves
HDF	Hierarchical Data Format
HE	High energy
HPC	High Performance Computing
HRV	High-Rate Veto

IVOA	International Virtual Observatory Alliance
JSON	JavaScript Object Notation
LoI	Letter of Intent
MC	Monte Carlo
OAI	Open Archives Initiative (OAI-PMH, OAI Protocol for Metadata Harvesting)
ODC	Open Data Centre
ORCA	Oscillation Research with Cosmics in the Abyss
OSC	Open Science Committee
PMT	PhotoMultiplier Tube
RA	Right Ascension
REST-API	Representational State Transfer – Application Programming Interface
RMS	Root Mean Squared
RofR	Registry of Registries
RoI	Region of Interest
ROOT	Rapid Object Oriented Technology
SCS	Simple Cone Search
SNEWS	SuperNova Early Warning System
TAP	Table Access Protocol
TNS	Transient Name Server, reporting mechanism for astronomical transients
UHE	Ultra-High Energy
UID	Unique IDentifier
URL	Uniform Resource Locator
UTC	Coordinated Universal Time
UUID	Universally Unique IDentifier
VO	Virtual Observatory
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
ZTF/LSST	Zwicky Transient Facility / Large Synoptic Survey Telescope

VI. List of figures

Figure 1: Acoustic data acquisition system, from Deliverable 8.2.....13

Figure 2: Overview over data levels.....17

Figure 3: Example quality control parameter (mean of of the PMT rate in Hz) for the ORCA4 test data set.....22

Figure 4: Average reduced χ^2 for quality reference observables for agreement between data and simulation at data level 1 (triggered data).....24

Figure 5: Average reduced χ^2 for quality reference observables for agreement between data and simulation at data level 2 (reconstructed data).....24

Figure 6: Overview for the Open Science System.....30

Figure 7: CI Pipelines of the km3pipe projects showing a failing test in a Python 3.6 environment.....39

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 road map 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, among others, the preparation of Open Data Access (work package 4).

IX. Executive summary

The KM3NeT Research Infrastructure will, over a period of at least a decade, produce a large amount of unique scientific data that are to be made available to the scientific communities concerned and to the broader general public. This requires the set-up of tools, procedures, documentation and rules to provide this service. This report provides an overview for the full system of open science services from KM3NeT, as well as the measures to ensure the sustainable operation of the Open Science System, including its testing.

X. Table of Contents

I. Copyright notice....2
II. Delivery slip....2
III. Document log....2
IV. Application area....2
V. Terminology....3
VI. List of figures....4
VII. List of tables....5
VIII. Project summary....5
IX. Executive summary....5
X. Table of Contents....6
1. Science in KM3NeT....9
1.1. Scientific targets....9
1.1.1. Astrophysics....9
1.1.2. Neutrino physics....10
1.2. Detector....11
1.2.1. Detector design....11
1.2.2. Data acquisition....11
1.3. Sea science....12
1.3.1. Environmental data....12
1.3.2. Acoustic data....14
1.4. Multi-messenger astrophysics....14
1.4.1. Importance of neutrinos for multi-messenger studies....14
1.4.2. KM3NeT multi-messenger neutrino alerts....15
1.4.3. KM3NeT alert types....15
1.5. Detector and event simulations....16
1.6. Data processing....16
1.6.1. Event data processing....16
1.6.2. High level data and data derivatives....18
2. Meeting criteria for FAIR data....18
2.1. Making data FAIR....18
2.1.1. Requirements for FAIR data....18
2.1.2. Compliance with the FAIR principles....18
2.2. Workflow management....19

2.2.1. Workflow description 19
2.2.2. Computing resource management 19
2.2.3. FAIR workflows 20
2.3. Data models 20
2.3.1. Resource description 20
2.3.2. Identifiers and content description 20
2.4. Data quality 21
2.4.1. Data quality control procedures 22
2.4.2. Calibration 23
2.4.3. Simulations and event reconstruction 24
2.5. Publication procedures 25
2.5.1. Establishing the Open Science Committee 25
2.5.2. Software publication 25
2.5.3. Data publication 27
3. The KM3NeT Open Science System 28
3.1. Platforms for open science 28
3.1.1. Open science products 28
3.1.2. KM3NeT servers and platforms 29
3.1.3. Interfaces 31
3.2. Open data sets and formats 31
3.2.1. Particle event tables 32
3.2.2. Multi-messenger alerts 33
3.2.3. Supplementary services and data derivatives 34
3.2.4. Acoustic data 35
3.3. Platforms and servers 36
3.3.1. The Virtual Observatory server 36
3.3.2. The KM3NeT Open Data Centre 37
3.3.3. Gitlab & Docker server 38
3.4. Registries and archiving 40
3.4.1. Virtual Observatory registry of registries 40
3.4.2. DataCite and Zenodo 41
3.4.3. Software integration 41
3.5. Documentation and tutorials 42
3.5.1. Documentation 42
3.5.2. Virtual Education Centre 42
3.6. Accessing data 43
3.6.1. The Open Science Portal 43
3.6.2. Python environment 43
3.7. Further development of the Open Science System 44
3.7.1. Development of Virtual Observatory standards 44
3.7.2. Data citation and access 44
3.7.3. Software licensing 45
3.7.4. Workflow management 45

3.7.5. EOSC Integration....45

4. Example programs and use cases....45

4.1. ANTARES 2007-2017 point source analysis....45

4.1.1. Use case description....45

4.1.2. Data set....46

4.2. KM3NeT use case....47

4.2.1. Analysis of an atmospheric muon-dominated dataset....47

4.2.2. Run selection and online sample....47

4.2.3. Processing to metadata enriched HDF5 format....47

4.2.4. Analysis examples....47

5. Tests of the Open Science System....48

5.1. Automated software testing....48

5.1.1. Unit testing....48

5.1.2. Integration testing....48

5.1.3. System testing....49

5.2. Manual testing of the system....49

5.3. Test of the Virtual Education Centre....49

6. Conclusion....49

1. Science in KM3NeT

We start this report with a short summary of the science goals of KM3NeT, the detector functionality and the genesis of the data, so as to provide the reader with the basic information underlying the Open Data system.

1.1. Scientific targets

The KM3NeT detectors will continuously record data produced by the propagation and interaction of particles arriving from every direction. The neutrinos of astrophysical interest, i.e. those from extra-terrestrial origin, need to be identified in the background of atmospheric neutrinos, i.e. those created in the Earth's atmosphere by interactions of cosmic-ray particles. Access to cosmic neutrino data is of high importance for a wide astrophysics community to relate cosmic neutrino fluxes to observations by other neutrino observatories or using other messengers, and to compare them with theoretical predictions. The atmospheric neutrinos carry information on the particle physics processes in which they are created and on the neutrinos themselves. These data are also relevant for a wide astroparticle and particle physics community. Finally, KM3NeT will monitor marine parameters, such as bioluminescence, currents, water properties and transient acoustic signals and provides user ports for Earth and Sea sciences.

1.1.1. Astrophysics

The main science objective of KM3NeT/ARCA, see Section 1.2.1, is the detection of high-energy neutrinos of cosmic origin. Neutrinos represent an alternative to photons and cosmic rays to explore the high-energy Universe. Neutrinos can emerge from dense objects and travel large distances, without being deflected by magnetic fields or interacting with radiation and matter. Thus, even modest numbers of detected neutrinos can be of utmost scientific relevance, by indicating the astrophysical objects in which cosmic rays are accelerated.

The design of the KM3NeT/ARCA detector has been optimised to target astrophysical neutrinos at TeV energies and above in order to maximise the sensitivity to detect neutrinos from the cosmic ray accelerators in our Galaxy. In a neutrino telescope like ARCA, two main event topologies can be identified by, firstly, the "track" topology indicates the presence of muons produced in charged current muon neutrino interactions and tau neutrino interactions with muonic tau decays. Muons are the only class of particles that can be confidently identified, because they are the only particles that are able to traverse long distances (up to several km) in matter and therefore appear as tracks in the detector. Secondly, the "shower" topology refers to particle showers from neutral current interactions of all three neutrino flavours, or from charged current interactions of electron and tau neutrinos with non-muonic tau decays. For tau neutrinos at sufficiently high energies (> 100 TeV), the produced tau lepton can travel several metres before decaying, resulting in two distinguishable individual showers. This allows for the identification of tau neutrinos. All neutrino flavours can be used for neutrino astronomy.

The preferred search strategy is to identify upward-going tracks, which unambiguously indicates neutrino reactions since only neutrinos can traverse the Earth without being absorbed. This allows an effective filtering against charged particles from the atmosphere. A

neutrino telescope in the Mediterranean Sea on the Northern hemisphere of the Earth is well suited for this purpose, since most of the potential Galactic sources are in the Southern sky. Atmospheric neutrinos from all directions form a background to astrophysical measurements, which requires searching for an excess of events at high particle energies or from a particular celestial direction, using space-time clustering, or investigating temporal coincidences with transient messengers.

Besides all-flavour neutrino astronomy, i.e. investigating high-energy cosmic neutrinos and identifying their astrophysical sources, additional physics topics of ARCA include:

- multi-messenger studies,
- particle physics with atmospheric muons and neutrinos,
- indirect searches for dark matter.

The ARCA detector allows to reconstruct the arrival direction of TeV-PeV neutrinos to sub-degree resolution for track-like events and ~2 degrees for shower-like events. The energy resolution is ~0.27 in $\log_{10}(E)$ for muons above 10 TeV, while for showers a ~5% resolution on the visible energy is achieved. In order to reach these resolutions, typically a set of quality selection criteria are applied based on the output of the event reconstructions. Further details on the detector performance can be found in the the Letter of Intent (LoI, J.Phys.G 43 (2016) 8, 084001).

1.1.2. Neutrino physics

Neutrinos have the peculiar feature that they can change from one flavour to another when propagating over macroscopic distances. This phenomenon of neutrino flavour change is known as 'neutrino oscillation'. The Nobel Prize in Physics of the year 2015 was awarded to T. Kajita and A. B. McDonald for the discovery of neutrino oscillations, which shows that neutrinos have mass. One open question is the so-called 'neutrino mass ordering'. It refers to the sign of one of the two independent neutrino mass differences, the absolute value of which has already been known for more than two decades.

The main science objective of KM3NeT/ORCA, see Section 1.2.1, is the determination of the ordering of the three neutrino mass eigenstates by measuring the oscillation pattern of atmospheric neutrinos. Atmospheric neutrinos are produced in cosmic-ray air-showers in the Earth atmosphere. When produced on the other side of the Earth and traversing the Earth towards the detector, atmospheric neutrinos oscillate, i.e. change their flavour between production and detection. The oscillation pattern in the few-GeV energy range is sensitive to the neutrino mass ordering and other oscillation parameters.

Besides determining the neutrino mass ordering, additional science topics of ORCA include:

- testing the unitary of the neutrino mixing matrix by studying tau-neutrino appearance,
- indirect searches for sterile neutrinos, non-standard interactions and other exotic physics,
- indirect searches for dark matter; testing the chemical composition of the Earth's core (Earth tomography),
- low-energy neutrino astrophysics.

The detector design of the KM3NeT/ORCA has been optimised for atmospheric neutrinos in the 1-100 GeV energy range in order to maximise the sensitivity for the neutrino mass ordering. For neutrino oscillation measurements with KM3NeT/ORCA, the capability to differentiate the two event topologies, i.e. the track-shower separation power, is crucial. The projected detector performance of the ORCA detector is summarised in detail in the LoI.

1.2. Detector

The KM3NeT Research Infrastructure will consist of a network of deep-sea neutrino detectors in the Mediterranean Sea with user ports for Earth and Sea sciences.

1.2.1. Detector design

The KM3NeT neutrino detectors employ the same technology and neutrino detection principle, namely a three-dimensional array of photo sensors that is used to detect Cherenkov light produced by relativistic particles emerging from neutrino interactions. From the arrival time of the Cherenkov photons, which is measured with nanosecond precision, and the position of the sensors, which can be measured with an accuracy in the order of 10 cm, the energy and direction of the incoming neutrino, as well as other parameters of the neutrino interaction, can be reconstructed. The main difference between different detector designs is the density of photo sensors, which is optimised for the study of neutrinos in the few-GeV (ORCA) and TeV-PeV energy range (ARCA), respectively.

A key technology of the KM3NeT detectors is the Digital Optical Module (DOM), a pressure-resistant glass sphere housing 31 small 3-inch photo-multiplier tubes (PMTs), their associated electronics and calibration devices. The segmented photo-cathode of the multi-PMT design allows for uniform angular coverage, single-photon counting capabilities and directional information on the photon arrival direction. The DOMs are distributed in space along flexible strings, one end of which is fixed to the sea floor and the other end is held close to vertical by a submerged buoy. Each string comprises 18 DOMs. The strings are connected to junction boxes that provide connections for power and data transmission.

A collection of 115 strings forms a single KM3NeT “building block”. The modular design allows building blocks with different spacings between strings/DOMs, in order to target different neutrino energies. Three building blocks are foreseen in KM3NeT: two KM3NeT/ARCA blocks, with a large spacing to target astrophysical neutrinos at TeV energies and above; and one KM3NeT/ORCA block, to target atmospheric neutrinos in the few-GeV range.

The ARCA (Astroparticle Research with Cosmics in the Abyss) detector is being installed at the KM3NeT-It site, 80 km offshore Capo Passero at the Sicilian East coast (Italy) at a sea bottom depth of about 3450 m. About 1 km^3 of seawater will be instrumented with ~130000 PMTs. The ORCA (Oscillation Research with Cosmics in the Abyss) detector is being installed at the KM3NeT-Fr site, 40 km offshore Toulon (France) at a sea bottom depth of about 2450 m. A volume of about 8 Mton is instrumented with ~65000 PMTs. Technical details on the detector design are given in the LoI.

1.2.2. Data acquisition

The readout of the KM3NeT detector is based on the ‘all-data-to-shore’ concept, in which all analogue signals from the PMTs that pass a reference threshold are digitised. This data contain the time at which the analogue pulse crosses the threshold level, the time that the pulse remains above the threshold level (known as time-over-threshold, or ToT), and the PMT address. This is typically called a *hit*. All digital data (about 25 Gb/s per building block) are sent to a computing farm onshore where they are processed in real time. The recorded data is dominated by optical background noise from Cherenkov light from ^{40}K decays in the seawater as well as bioluminescence from luminescent organisms in the deep sea. Events of scientific interest are filtered from the background using designated software, which exploit

causality requirements applied to time-position correlations of hits. To maintain all available information for the offline analyses, each event contains a snapshot of all the data in the detector during a suitable time window around the event.

For calibration purposes summary data is written out, containing the count rates of all PMTs in the detector (with a sampling frequency of 10 Hz). This information is used in the simulations as well as in the reconstruction to take into account the actual status and optical background conditions of the detector.

In parallel to the optical data, acoustic data and instrument data are recorded. The main purpose is position calibration of the DOMs, which is necessary as the detector elements move under the influence of sea currents. The acoustic data includes the processed output from the piezo sensors in the DOMs and from the hydrophones in the base modules of the strings. The instrument data includes the processed output from the compasses, temperature sensors and humidity sensors inside the DOMs.

During operation the continuous data stream sent by the detector is split into time intervals, called *runs*, with typical durations of a few hours. This is done for practical reasons of the data acquisition. In addition, this procedure allows to select a set of run periods with high-quality data based on the monitored detector status, environmental conditions and derived data quality parameters. The calibration for timing, positioning and photon detection efficiency is done offline using the calibration data.

1.3. Sea science

1.3.1. Environmental data

The KM3NeT research infrastructure will also provide connections for instrumentation for Earth and Sea sciences for long-term and on-line monitoring of the deep-sea environment. Until now, measurements in the deep sea are typically performed by deploying and recovering autonomous devices that record data over periods of months to years. This method is severely constrained by bandwidth limitations, by the absence of real-time interaction with the measurement devices and by the delayed access to the data. A cabled deep-sea marine observatory, like KM3NeT, remedies these disadvantages by essentially providing a power socket and high bandwidth Ethernet connection at the bottom of the sea. This is an important and unique opportunity for performing deep-sea research in the fields of marine biology, oceanography, environmental sciences and geosciences. To this end, both the French and Italian KM3NeT sites are nodes of the European Multidisciplinary Seafloor and water column Observatory EMSO (emso.eu).

The sea science instrumentation modules will host sensors that provide real-time monitoring of a plethora of environmental parameters including temperature, pressure, conductivity, oxygen concentration, turbidity and sea current. Additional instrumentation including a benthic crawler, a seismograph, a deep-sea Germanium gamma detector and a high-speed, single-photon video camera for bioluminescence studies are also envisioned to be installed. Furthermore, the KM3NeT optical modules themselves provide invaluable data on deep-sea bioluminescence and bioacoustics monitoring of the local cetacean populations. Another example is the possibility to use the optical fibres in the main electro-optical cables, that run for many tens of kilometres along the seafloor, for seismological studies.

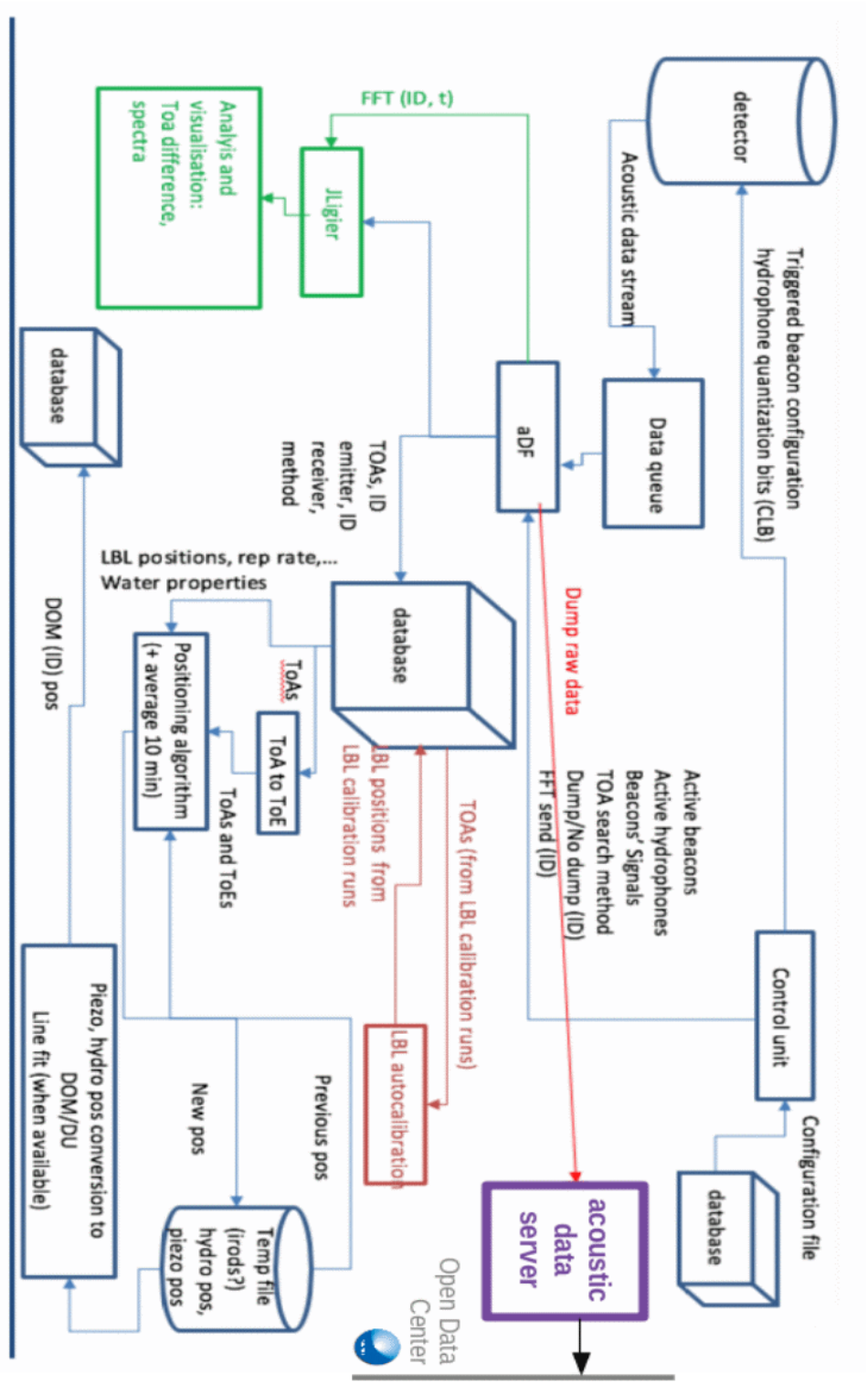

Figure 1: Acoustic data acquisition system, from Deliverable 8.2.

1.3.2. Acoustic data

The telescope is also equipped with acoustic sensors: 1 piezoelectric ceramic sensor is installed in each DOM, 1 hydrophone on each DU base and on the junction boxes of the seafloor network. The main purpose of these sensors is to provide real time positioning of each DOM with cm accuracy. The hydrophone sensitivity (omnidirectional) is about -173 dB re 1V/µPa over the frequency band between few tens Hz and 70 kHz, which makes this sensor also suitable for interdisciplinary studies. For example, acoustic signals from whales and dolphins can be detected with the acoustic sensors.

The default KM3NeT DAQ has been designed to identify only the (known) acoustic signals emitted along a long baseline of acoustic sensors from the full data stream (Figure 2). An extension of the DAQ is foreseen in the future, permitting online analysis, display and recording of the underwater noise spectrum through a dedicated channel. A large number of use cases has been identified, as described in Deliverable 8.1.

At this stage, the unfiltered acoustic data can be made available directly from the ADF node, which is provided in several formats through a REST-API on a data server integrated in the acoustic data processing system of KM3NeT.

1.4. Multi-messenger astrophysics

The multi-messenger approach in astrophysics means looking for two or more cosmic messenger signals to study the transient phenomena in our Universe, such as gamma-ray bursts, the outburst of active galactic nuclei, fast radio bursts, supernova explosions, etc. Using multiple messengers greatly extends our understanding of the Universe compared to using one single channel. The cosmic messengers include electromagnetic waves, cosmic rays, gravitational waves, and neutrinos.

Some of the most important open questions in astrophysics are the origin of astrophysical neutrinos, the origin of cosmic rays, acceleration mechanisms of high energy cosmic rays, etc. Multi-messenger studies can help answering these questions.

Up to now, there are three successful multi-messenger detections involving high-energy neutrinos:

1) In 1987, the observation of the supernova 1987A, where neutrinos are observed in neutrino experiments about 2 or 3 hours before the visual observations.
2) 30 years later in 2017, the signals of the gravitational wave and the electromagnetic observations of a gamma-ray burst observed by the gamma-ray satellites Fermi and Integral.
3) TXS 0506, where for the first time a blazar was identified as a neutrino source.

1.4.1. Importance of neutrinos for multi-messenger studies

Among those multiple messengers, neutrinos are an important type of messenger. Neutrinos are neutral and only interact via gravity and weak interactions. Neutrinos point back to their sources where they were created.

For example, cosmic rays are charged particles, thus they are deflected by the galactic magnetic fields. Cosmic ray observatories can detect them but their observed arrival directions do not point back to their sources. During the propagation of cosmic rays,

neutrinos are produced during the interaction of cosmic rays and the extragalactic background light. Since neutrinos are not bent by magnetic fields, they can act as good tracers for studying the propagation of cosmic rays.

Looking for coincidences of neutrinos and electromagnetic or GW counterparts may also reveal sub-threshold events that otherwise do not generate interest within each single observatory, or even reveal new sources.

Because neutrinos travel with nearly the speed of the light, a real-time or near real-time alert system based on neutrinos (with a good angular resolution) is possible. This is vital for the follow-ups of some high energy transient sources that are time-dependent with the flux quickly varying. For example, a real-time neutrino alert will be able to point to a direction for space electromagnetic observatories for which only a small sky coverage per observation is possible (e.g. Fermi-LAT) to conduct their search in a timely fashion.

1.4.2. KM3NeT multi-messenger neutrino alerts

With the Southern sky including the Galactic Centre in view and a good angular resolution, KM3NeT will contribute greatly to the multi-messenger investigations.

- For the search of astrophysical neutrinos, KM3NeT will be able:
    1. to receive external alerts, i.e. alerts generated by external partner experiments (e.g. gravitational waves alerts from LIGO/Virgo, neutrino alerts from other neutrino experiments) via GCN and search for correlated neutrinos in KM3NeT;
    2. to send alerts (to GCN) on neutrinos observed in KM3NeT, including multiplet alerts, possible astrophysical neutrinos, any neutrinos found in coincidence with external alerts. The alerts will be used by external partner experiments to conduct their correlation search/follow-up.

- For the search of MeV core-collapse supernova (CCSN) neutrinos, each KM3NeT DOM acts as a detector. CCSN neutrino interactions lead to higher counting rates of individual PMTs and an increase of the number of coincident hits in the same optical module. Identifying this signal requires a method different from the usual neutrino event reconstruction route, and there is a separate alert system called the SuperNova Early Warning System (SNEWS), so the KM3NeT supernova alerts are not discussed in here. KM3NeT is already connected to SNEWS.

1.4.3. KM3NeT alert types

The alert types include:

- MeV Core-Collapse Supernova alerts (SNEWS), already online currently.
- Multiplet alerts: Multiple neutrinos coinciding in direction within some time window (this suggests a potential neutrino source).
- High-energy neutrino alerts: Potential neutrinos from astrophysical sources (the higher the energy, the higher the probability of being of astrophysical origin).
- Any neutrinos correlated with external alerts.

Other alerts to be defined, or more subcategories divided from the High-Energy Neutrino alerts if necessary (e.g. track HE, cascade HE alerts).

For alert data formatting and sending, see the data format definition (Section 3.2.2).

1.5. Detector and event simulations

In order to evaluate and test physical models of neutrino interactions and productions and explore the quality of detector performance and its understanding, a variety of simulations need to be performed and compared to the data taken with the KM3NeT detectors. The full functionality of the detector is modelled, including photomultiplier tube characteristics, the front-end electronics, the high throughput data distribution over heterogeneous networks and also physical properties of the environment (like seawater, atmosphere…) and the materials used. All these are carefully taken into account when simulating the overall detector response to particle interactions.

The first level of event simulation starts with the incoming primary particle: neutrinos produced in cosmic events or atmospheric neutrinos and muons produced by cosmic radiation in the Earth's atmosphere. The second level of the simulation chain takes care of the propagation of these particles and additional particles produced along their way through the atmosphere, Earth and seawater - depending on their travel path - until they reach the detector volume. In the final step, the light produced by the particles is simulated and propagated to the optical modules where hits are generated, digitised and passed through the above mentioned hardware response simulation.

The simulation is implemented in a run-by-run simulation strategy, where the detector response is simulated individually for the conditions of each run. Since large statistics are required for precise analyses, the simulation data will significantly exceed the real data in volume.

As handling these large data sets is impractical for inter-experimental studies, but the information is crucial for the interpretation of the data, parametrised distributions of relevant features 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 for given models is beneficial for the development of common research approaches. The development of a corresponding open service is currently under investigation.

1.6. Data processing

Data processing follows a tier-based approach, where causally correlated hits are used to select data with potential scientific interest, which are arranged into events. In a second step, processing of the events, applying calibration, particle reconstruction and data analysis methods lead 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 integrated locally into the analysis workflow of the scientist. 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 statistical interpretation of the measurements.

1.6.1. Event data processing

Hit-related information is written to ROOT-based tree-like data structures and accumulated during data runs before being transferred to high-performance computing (HPC) clusters.

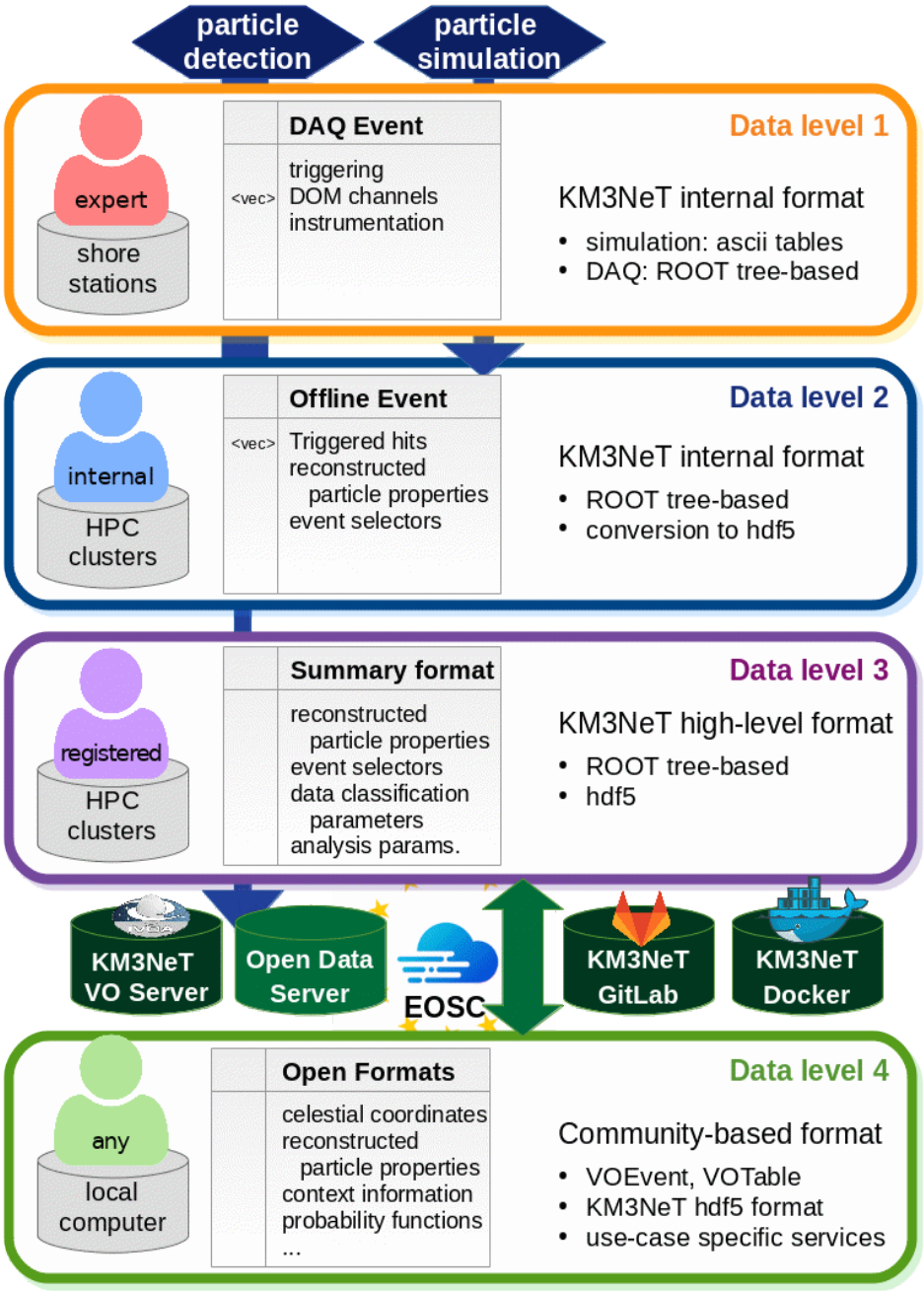

Figure 2: Overview over data levels.

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 hit distributions for the modelling of event properties in both measured and simulated data, requiring high-performance computing capabilities.

Access to data at this level is restricted to collaboration members due to the intense use of computing resources, and to the large volume and complexity of the data. However, data at this stage is already converted to HDF5 format as a less customized hierarchical format. This format choice increases interoperability, 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 if this might become feasible in the future.

1.6.2. High level data and data derivatives

Summary formats and high-level data

As mostly information on particle type, reconstructed 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, 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 made using the simulated event sets. A full description of the data processing and management has been provided in Deliverable 4.1, with an overview of the data processing being presented in Figure 2.

2. Meeting criteria for FAIR data

2.1. Making data FAIR

2.1.1. Requirements for FAIR data

The widely accepted paradigm for open science data publication requires the implementation of the FAIR principles for research data. It involves the following:

- definition of descriptive and standardised **metadata** and application of persistent identifiers to create a transparent and self-descriptive data regime,
- linking this data to common science platforms and **registries** to increase findability,
- possibility to harvest from the data through commonly implemented **interfaces** and
- definition of a **policy standard** including licensing and access rights management.

In all these fields, the standards of KM3NeT have been defined or initiated during the KM3NeT-INFRADEV project. In this development process, the application of existing standards especially from the astrophysics community and the development of dedicated KM3NeT software solutions are pursued in close cooperation with the ESCAPE project. The European Open Science Cloud (EOSC) is foreseen to be a main environment for open data publication in KM3NeT.

2.1.2. Compliance with the FAIR principles

The FAIR principles provide a solid set of requirements for the development of an open data regime. Following the FAIR requirements, the following solutions have been established in KM3NeT to enable FAIR data sharing and open science.

Findable data

- **Unique identifiers** have been defined for digital objects within KM3NeT, including data files, software, collections and workflow steps, as well as identifiers for relevant data sets like particle detection ("event") in the detector.
- At the publication level, extended **metadata sets** are assigned to each published data product.
- The datasets can both be accessed via UID directly on the data servers as well as through external community-relevant **repositories**.

Accessible data

- The data can, at this point, be directly accessed via a webpage and through a REST-API where data cannot be offered through VO protocols.
- **No authentication** is implemented as yet, although in the future an authentication scheme is aimed for to allow access to unpublished data sets for scientists involved in project-based cooperations.
- High-level data sets will be mirrored and ultimately transferred to long-term data repositories for archiving.

Interoperable data

- Vocabularies and content descriptors are introduced that draw on external standards like those of the VO or W3C where possible.
- **Documentation on the metadata** and vocabularies is provided.
- Metadata classes are designed to allow for cross-referencing between different digital objects and extended metadata.

Reusable data

- Licensing standards for data, software and supplementary material have been introduced.
- Basic **provenance** information is provided with the data, which serves as a development starting point to propagate provenance management through the complex data processing workflow in the future.

2.2. Workflow management

2.2.1. Workflow description

The simulation and raw data processing chains consist of workflows that are executed on heterogeneous clusters of computers. The harmonisation of workflow interfaces is one of the bigger challenges due to the large variety of software being used in individual processing steps. The benefits of a standardised interface lie in clarity and simplified maintenance. The Common Workflow Language (CWL) is currently being evaluated as an abstraction layer to describe analysis workflows and tools. It provides scalability and portability across many different software and hardware environments.

2.2.2. Computing resource management

The second layer of workflow management targets the distributed processing of data on heterogeneous computing infrastructures. For this purpose, the DIRAC interware has been

chosen which provides a common interface to a multitude of resource like grids, cloud systems and clusters of computers, offering a variety of interoperability options. DIRAC is tested to be used to set up a data driven processing infrastructure for all stages in the pipelines including the combination of raw data and Monte Carlo simulations.

2.2.3. FAIR workflows

In the combination of CWL and DIRAC software, both the detailed description and modularisation of the individual workflow steps and the data driven optimisation of installation and distribution of the workflows with full control of the processing environment is planned to be achieved. This would build the backbone of FAIR data processing and would facilitate the development of a solid data provenance scheme in KM3NeT. While the implementation of a full-scale data processing environment goes beyond the scope of this project, this software implementation is currently explored in the ESCAPE project in cooperation with other ESFRIs.

2.3. Data models

Metadata definition lies at the core of FAIR data, as it governs both the understanding of the data as well as the interoperability through access protocols. While some software can be used almost as it is, especially regarding the well-developed interfaces in the Virtual Observatory, the different data types and science fields of KM3NeT require a flexible approach and application of different types of software. In order to meet these requirements, metadata and class definitions are developed within KM3NeT, drawing on well established standards, e.g. of the W3 Consortium, scientific repositories or of the IVOA.

Data published via the KM3NeT Open Data Centre (ODC) is annotated as KM3OpenResource, which includes basic metadata for resource content, accessibility and identification. As a resource, it can be provided either as part of a collection, e.g. data set, or as related to an analysis, or as part of a stream of similar objects, e.g. of alert data, they are grouped in the server as KM3ResourceCollection or KM3ResourceStream to facilitate findability. Further details on these first data classes are documented in a developing Git project. In the future, further classes will be introduced and adapted governing e.g. the scientific workflow as discussed in Section 1.6.1.

2.3.1. Resource description

The **KM3OpenResource** class serves as base class to describe any KM3NeT open resource, be it a plot, dataset or publication. The information gathered here should be easily transformable to publish the resource in repositories like the Virtual Observatory or Zenodo based on DataCite. Therefore, resource description metadata as provided in the KM3Resource class is based on standardised formats like the Dublin Core standard, adding relevant entries from other formats including the VO Observation Data Model Core Components regarding the metadata specific to the scientific target, and the VOResource description and Zenodo resource description for general resource metadata.

2.3.2. Identifiers and content description

Identifiers serve to uniquely label and address digital objects. While Digital Object Identifiers (DOIs) are of long-standing use in the scientific community, these public identifiers have to

link to an KM3NeT-internal identification scheme which allows the user to back-track the data generation and link between various data products related to a scientific target or publication. In addition to this, an ordering scheme for class definitions and content descriptors helps in the interpretation of a specific digital object. To this end, the **kid** and **ktype** have been introduced.

kid

The kid is a unique identifier which follows the UUID schema. The UUID is ideally assigned at the generation of the digital object and stored in the metadata set or header of the digital object. It is the goal to use kid assignment at all steps of data processing, including those generating data products before publication level, and it has been implemented for all open science products.

ktype

The **ktype** serves as a content descriptor and is defined as a string with a controlled vocabulary of words separated by ".", starting with "km3." as in reverse domain name notation. The selected vocabulary comprises domain names, class and sub-class names and, in some cases, identifiers for class instances, like

```
km3.{domain}.{subdomains}.{class}.{subclasses}.{instance}
```

e.g. "km3.data.d3.optic.events.simulation" for a data set of processed optic event data (data level d3) from Monte Carlo simulation (indicating a file class) or "km3.params.physics.event.reco.reconame.E" indicating the parameter definition of the reconstructed energy of particle events from a reconstruction algorithm named "reconame".

Particle event identifiers

Identifiers are used to uniquely label e.g. different settings of software and hardware or annotate data streams. At the data aggregation level, an identifier therefore has to be introduced to uniquely identify a particle detection in one of the KM3NeT detectors.

Due to the design of the data acquisition process, these events can be uniquely identified by:

- the detector in which they were measured, assigned a **detector id**,
- the run, i.e. data taking period during which it was detected, assigned a **run id**,
- the **frame index**, indicating the numbering of the data processing package in the DAQ system on which the triggering algorithms are performed and
- the **trigger counter**, i.e. the number of successes of the application of the set trigger algorithms.

The internal KM3NeT event identifier is therefore defined as

```
km3.{detector_id}.{run_id}.{frame_index}.{trigger_counter}
```

2.4. Data quality

The processes involved in the KM3NeT data processing chain can be grouped into a few main categories. Although the ordering of these categories is not strictly hierarchical from the point of view of data generation and processing, in the context of a general discussion one could safely assume that a hierarchical relation exists between them. From bottom to

top, these categories would be data acquisition, detector calibration, event reconstruction, simulations and finally scientific analyses based on the data processed in the previous categories. The quality of the scientific results produced by KM3NeT will be affected by the performance of the different processes involved in the lower levels of the data processing chain.

In order to implement a complete and consistent set of data quality control procedures that span the whole data processing chain, it is required to have a complete, unambiguous and documented strategy for data processing at each of the aforementioned process categories. This includes the setting of data quality criteria which are initiated at the highest data level of the processing chain, and propagated towards the lowest levels, as described in Deliverable 4.5. For each of the aforementioned categories a KM3NeT working group has been established with the charge to develop a suitable quality strategy. It is therefore not possible to provide a full setup for data quality control at this point. Nevertheless, there have been copious software developments devoted to quality control along the different stages of the data processing chain. In the following, a description of some of the existing quality control tools and procedures is given. This description could be conceived as an incomplete prototype for a data quality plan. The implementation of these procedures into an automated workflow requires the design and implementation of a standardised data processing workflow which meets software quality standards. This does not yet exist either.

2.4.1. Data quality control procedures

Online Monitor

During the data acquisition process, the online monitoring software presents real time plots that allow the shifters to promptly identify problems with the data taking. It includes an alert system that sends notifications to the shifters if problems occur during data taking that require human intervention. The online monitor uses the same data that are stored for offline analyses. This implies that any anomaly observed during the detector operation can be reproduced offline.

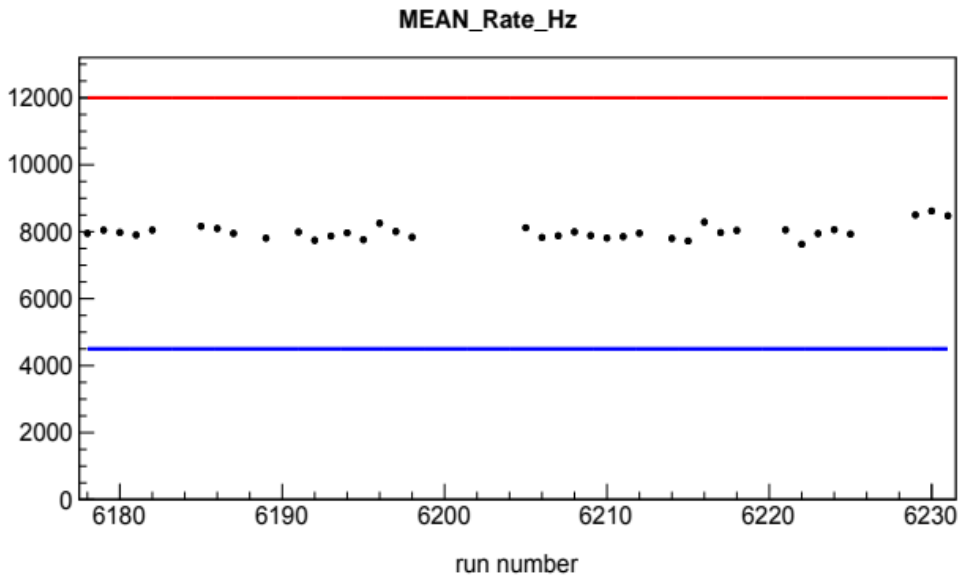

Figure 3: Example quality control parameter (mean of the PMT rate in Hz) for the test data set with ORCA with 4 detection units.

Detector Operation

As explained in the detector section, photon-detection related data from the detector operation are stored in ROOT files and moved to a high performance storage environment. The offline data quality control procedures start with a first analysis of new files which is performed daily. It mainly focuses on, but is not restricted to, the summary data stored in the ROOT files. The summary data contain information related to the performance of the data acquisition procedures for each optical module in the detector. As a result of this first analysis, a set of key-value pairs is produced where each key corresponds to a specific data quality parameter and the value represents the evaluation of this parameter for the live-time of the analysed data. The results are tagged with a unique identifier and uploaded to the central database. In the present implementation the analysis is performed for each available file where each file corresponds to a data taking run, although this may change in the future as the data volume generated per run will increase with the detector size.

A further analysis of the results stored in the database includes the comparison of the values of the different parameters to reference value ranges, allowing for a classification of data periods according to their quality. In addition, the evolution of the different quality parameters are monitored and made available to the full collaboration as reports. Currently this is done every week by the shifters, and the reports are posted on an electronic log book (ELOG). Figure 3 shows an example of the time evolution for a quality parameter during the period corresponding to the data sample that is provided together with this report. The selected runs correspond to a period of stable hit rates and have quality parameters that were within the allowed tolerance.

2.4.2. Calibration

The first step in the data processing chain is to determine the detector calibration parameters using the data obtained from the detector operation. These parameters include the time offsets, gains and efficiencies of the PMTs as well as the positions and orientations of the optical modules. The PMT time offsets and the positions and orientations of the optical modules are used in later stages of the data processing chain for event reconstruction, as well as by the real time data filter during the detector operation. While the event reconstruction requires an accurate knowledge of these parameters, the algorithms used by the real time data filter depend rather loosely on them, and the performance is not dependent on variations occurring within a timescale of the order of months.

Nevertheless, it is still necessary to monitor them and correct the values used by the trigger system if necessary. The performance of the detector operation also depends on the response of the PMTs, which is partly determined by their gains. These evolve over time, and can be set to their nominal values through a tuning of the high-voltage applied to each PMT. Monitoring the PMT gains is therefore also necessary to maximise the detector performance. Additionally, the PMT gains and efficiencies are also used offline by the detector simulation. Within the context of data quality assessment, software tools have been developed by KM3NeT that allow to monitor the parameters described above and to compare them to reference values, issuing warning when necessary. The reference values should be determined by the impact of mis-calibrations on the scientific results of KM3NeT, a task which at this point has been started to be addressed. The arrangement of these tools into a workflow is embedded in the improvement of an underlying calibration strategy. This has not been done, and the work is therefore on hold.

2.4.3. Simulations and event reconstruction

Once the calibration constants have been determined, the data processing chain continues with the event reconstruction, and with the simulation and reconstruction of an equivalent set of events where the information in the summary data is used to simulate the data taking conditions. The simulation of particle interactions and propagation is done by dedicated software, while the detector simulation and event reconstruction is done by the Jpp software package. As a result of the simulation chain, a ROOT file is obtained which has the same format as the ROOT file produced by the data acquisition system and contains events obtained after the simulation of the detector trigger. Both data sets are identically processed by the reconstruction software, which produces ROOT formatted files. The comparison between data and simulations is an important parameter to measure the quality of the data and can be done at trigger or at reconstruction level.

In both cases, the comparison follows the same strategy: the ROOT files are used to produce histograms of different observables and these histograms are saved into new ROOT files. A set of libraries and applications devoted to histogram comparisons has been developed in Jpp. These implement multiple statistical tests that can be used to determine if two histograms are compatible, as well as the degree of incompatibility between them. Additionally, tools have been developed that summarise the results into a number per file, which represents the average result after comparing all the observables. For the example provided here, the discrepancy between data and Monte Carlo is measured through the calculation of the reduced χ^2 for each observable, and the summary is given as the average reduced χ^2 of all the compared observables for each file. Figures 4 and 5 show the value of this parameter for the data and simulations comparisons at trigger and reconstruction levels.

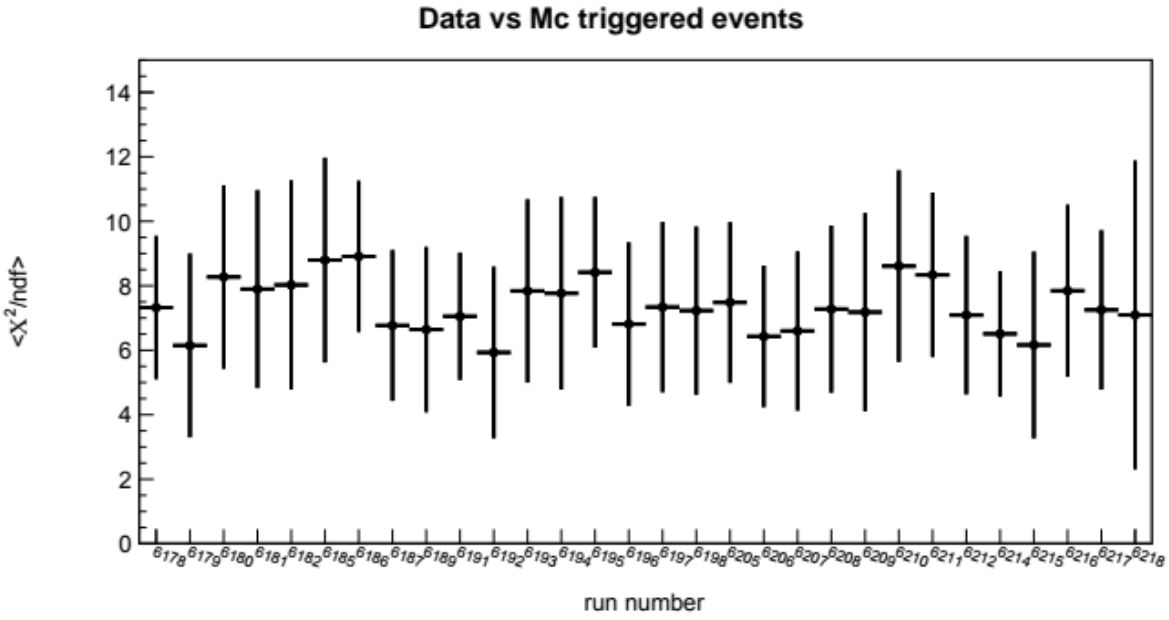

Figure 4: Average reduced χ^2 for quality reference observables for agreement between data and simulation at data level 1 (triggered data).

These tools also allow for the calculation of other data quality metrics by comparing the data with figures of merit for different observables.

The contents of the files produced by the reconstruction routines are the main ingredients for physics analyses, though these analyses are typically done with dedicated software frameworks which require a special formatting of the data. The data format used for physics analyses is called 'aanet' format, which is also based on ROOT. Control mechanisms are thus needed to ensure consistency between the files produced by Jpp, and the aanet formatted files. Consistency between Jpp and aanet files can be verified by producing distributions of the different parameters and by verifying that these distributions are identical.

2.5. Publication procedures

Software and data publication procedures assure the quality and interoperability of data and software, the reproducibility of scientific results, and the cross-fertilization between communities based on trust in the publicised science products. The following procedures have been proposed to and partly initiated by the KM3NeT Collaboration and will serve as a basis for the implementation of publication standards.

2.5.1. Establishing the Open Science Committee

Currently, the KM3NeT Collaboration reviews publications and contributions to conferences via the Publication Committee and Conference Committee. These committees are focusing primarily on the analysis logic, the used scientific methods and textual/graphical representation of papers, presentations and posters, while the underlying data and software are not fully scrutinised. To overcome this gap, the KM3NeT Institute Board has decided to install an Open Science Committee. It will focus on the explicit review of the data and software included in or used for publications, and also on general purpose releases of software and data in the open science regime.

2.5.2. Software publication

Installation of publication procedures involves the definition of quality standards, establishing of the review procedure and implementation during a transition phase.

Software quality standards

KM3NeT members are recommended to follow software development standards for their software projects and provided guidelines (see Deliverable 4.9) already during development of the software project.

Core requirements

Out of these recommendations, the following conditions are core requirements that have to be met for publication:

- Storage of the source code on the KM3NeT central Gitlab instance with documentation, testing and containerisation enabled via continuous integration.
- The software is made available containerised in a **Docker container**.
- The software project is **documented**:
  - The software is listed in the internal wiki entry pointing to the Gitlab hosted documentation.
  - Inline code documentation with automatic documentation generation is installed and a reference guide for the API is provided.

  - Additional documentation includes a “Getting started” with installation instructions and a description of the core concepts.
- A **change procedure** is established including versioning following Semantic Versioning 2.0.0 and changelog following the Gitlab workflow.
- **Coding standards** depending on the programming language are followed, using
  - the C++ Style Guide, also for ROOT-based software,
  - the Python Style Guide PEP-8,
  - the Java Style Guide
  - and other established coding standards for further programming languages.
- the software is licensed according to the KM3NeT licensing rules.

Recommendations

Beyond the core requirements, the following standards are recommended, but not required by the publication procedures:

- A full tutorial is added, including Getting started / Guides / Concepts / API.
- Additional examples for applications including benchmarks and test data are provided.

Publication procedures

1. Declare software as software candidate and pre-review

- The authors and the relevant working group coordinator define the software as software candidate.
- Metadata is added to the software for reviewing and documentation of the publication process.
- Referees are assigned to the software release candidate by the Open Science Committee.

2. Internal development to meet standards (after pre-review)

Compliance of the software with the quality standards is evaluated by the referees in a pre-review and the software adapted where necessary.

3. Reviewing process

- The assigned referees report on their final review of the software after adaptations and their recommendation at the working group meetings.
- The report is forwarded to the Open Science Committee for cross-checks.
- The Open Science Committee distributes the report to the collaboration for feedback.

4. Publication

- The KM3NeT-recommended license is applied and the copyright set to “the KM3NeT Collaboration”.
- The software is released through GitHub and Zenodo.

5. Maintenance

A maintainer is assigned to the software to reply to issues on the software and initiate necessary changes.

For major further releases of the software, a shortened review process is followed.

Implementation

Although already software has been produced at various levels in the KM3NeT Collaboration, the currently used and published software does, to a large degree, already follow the software quality standards but has not yet been officially released as KM3NeT products. Therefore, a transition procedure is implemented, where current software release candidates can follow a shortened review and publication procedure if the software has already been used for achieving published scientific results and software standards are followed.

2.5.3. Data publication

The data sets used for scientific publications are to be approved by the collaboration as well as the software involved in the data processing. In order to enable the publication of the relevant data alongside a publication, the data have to meet the following standards and go through the review process described below.

Standards

Data access and FAIRness

Data needs to be accessible and well documented.

- At time of publication, the full processing flow for the data must be accessible and reproducible for all members of the Collaboration:
  - All software versions and scripts must be included in a Gitlab repository.
  - The relevant environment and data locations have to be referenced in the Gitlab repository, which must be accessible for cross-checks.
- The data processing must be documented with the corresponding metadata, referencing identifiers of the processing data entities and activities.
- High-level data must be provided in a format that is suitable for data sharing, i.e. a KM3NeT public data format or custom format with sufficient metadata.

Data quality

- Standard tests must be performed on the completeness of the data processing and sets (sanity checks).
- Automatic testing is to be used where available.
- Tests must be performed on the validity of the data and consistency of the content. These tests depend on the nature of the data and can e.g. be done through data/MC comparisons for event data.

Product type specific quality requirements

As workflows and data derivatives are also a part of the public data and production process, additional criteria will apply to specific data products:

- **Event data** requires a full description of the processing workflow and annotation of all published event parameters.
- **Derived plots and functions** require a description of the basic data set from which data set they have been generated, including time range, parameter-based selections or the theoretical model.
- **Workflows** have to include software versions, processing scripts and annotations for all major steps.

Procedure

1. Declare data as publication candidate

- The data set and the intended publication is discussed in the respective working group and announced to the Open Science Committee (OSC).
- Access to the relevant workflow and data is granted as required by the quality standards. The intended publication date is declared.
- The data project is annotated with publication-relevant metadata and tags for the review process.
- Referees are assigned.

2. Testing and modifications

- Referees review the data processing and testing, and require modifications where necessary.
- A short report on the review is produced, including the reference to the data.

3. Review

- The referee report is forwarded to the OSC for review.
- The report is forwarded to the Publication Committee for review in the analysis context (paper, proceedings, etc.) and endorsed in the analysis context where applicable.
- The report is forwarded to the Collaboration for notice.

4. Publication

- Where applicable, an embargo is respected for the data publication if relevant within the analysis context or required by general policies.
- Upon approval by the Publication Committee, the data set is made available via the KM3NeT Open Science System, a unique identifier and version are assigned and the data is registered with public repositories.

3. The KM3NeT Open Science System

3.1. Platforms for open science

The KM3NeT Open Science System seamlessly integrates with the computing environment of the KM3NeT Collaboration, providing data, software and documentation to all relevant parts of the scientific analysis workflow.

3.1.1. Open science products

The products offered from the various servers, introduced below, do not only include data, but also software and supplementary material.

Data sets

Only high-level data sets and derivatives are offered in the open science platforms. These include:

- Particle events from **optical neutrino detection**, which is the primary data production channel for KM3NeT.
- **Multi-messenger alert data** which is broadcast in real time for events with high scientific relevance in multi-messenger searches.
- **Environmental data:** from both calibration-relevant data from the deep-sea detector and acoustic data which is relevant to Sea Science.

High-level derivatives

For the evaluation of the significance of the data for a given analysis target, additional information crucial to the interpretation of the data is published, including:

- Binned and parametrised information drawn from **simulations**, e.g. instrument response functions or sensitivities.
- High-level data **summary information** from dedicated analyses, e.g. public plots.

Software and workflows

Software and workflow examples facilitate the processing of the open data and form the basis for community-oriented developments. Offered products are:

- **KM3NeT-related software** for data processing, simulation and analysis, see Section 3.4.3.
- **Interface packages** for data access, currently based on Python, see Section 3.6.2.
- **Workflow descriptions** as Jupyter notebooks or annotated workflows for data handling, see Section 4.

Tutorials and Documentation

Documentation is set up to provide a step-by-step approach to the use of KM3NeT data. Also, specific details for user-oriented specific information is provided, namely:

- an **overview description** of the KM3NeT open data system,
- **documentation** of all components of the Open Science System,
- training and **online courses** for guided approaches to the use of the data.

3.1.2. KM3NeT servers and platforms

The products are published through various platforms tailored to the product-specific requirements.

Data servers

- Astrophysics-related data are offered through the **KM3NeT Virtual Observatory server** which runs the VO-integrated DaCHS software.
- All services and data not integrable into the VO due to their scientific domain or technical nature are offered through the KM3NeT Open Data Centre offering webpages and a REST-API to query the data products.

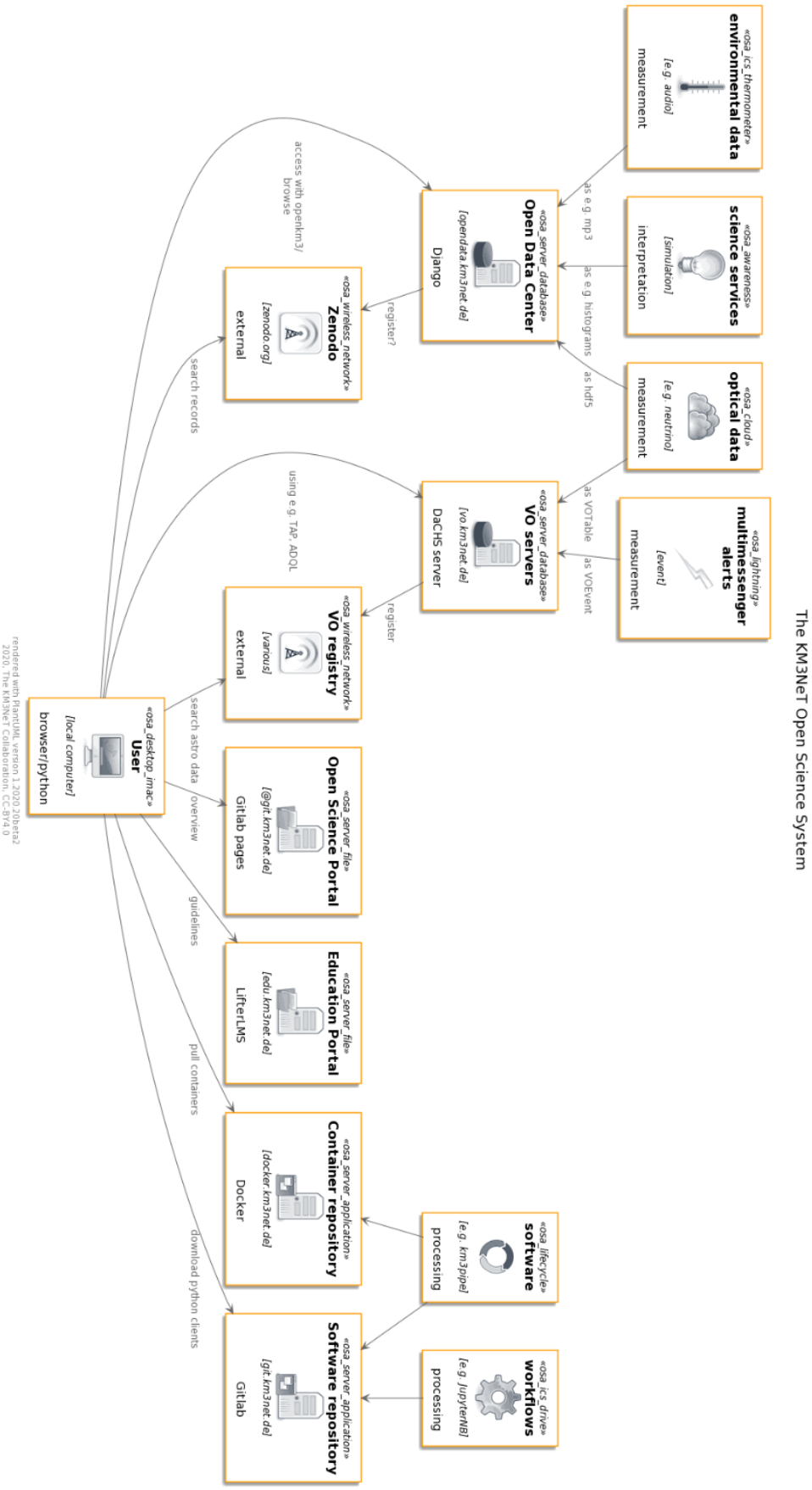

Figure 6: Overview for the KM3NeT Open Science System.

Software servers

- All software is offered through a KM3NeT-operated **Gitlab server** which hosts all software and documentation projects.
- Containers with key software are packaged and made available through a **Docker server**.

Documentation & Tutorials

Documentation is available for all parts of the system and generally offered in close connection with the documented product. Dedicated platforms offering centralised access to knowledge are:

- the **Education portal** for tutorials and webinars;
- Gitlab projects and pages, especially the **Open Science Portal** repository which documents the KM3NeT Open Science System.

3.1.3. Interfaces

For the end user, various access options to data, software and documentation are offered.

Repositories

Data is registered to large centralised data registries which make the KM3NeT data findable by the wider community. These registries include

- the **VO registry** to which all data sets and services registered through the VO server are pushed, and
- the Zenodo repository or alternative solutions to provide DOIs and integrate with **DataCite.**

Web interfaces

The platforms themselves offer various options to access the data, namely

- webpages, query forms and REST-APIs using both Graphical User Interfaces (GUIs) and machine-accessible endpoints, and
- domain-specific access protocols like TAP or ADQL for the VO-related data, see Section 3.3.1.

Software clients

Clients for user-friendly handling of the data are offered by third parties as well as by KM3NeT, including

- VO-specific access tools like TOPCAT or the Aladin Sky Atlas, and
- Python packages from third parties like pyvo or dedicated KM3NeT software like the openkm3 package, see Section 3.6.2.

3.2. Open data sets and formats

As all of the following data is published, inter alia, via the Open Data Centre, the data sets are all enriched with metadata following the KM3OpenResource description, see Section 2.3.

3.2.1. Particle event tables

Data generation

For particle event publication, the full information from the event reconstruction is reduced to a "one row per event" format by selecting the relevant parameters from the data level 2 files. The event and parameters selection, metadata annotation and conversion of parameters to the intended output format is performed using the *km3pipe* software. A prototype provenance recording has also been included in this software, so that the output of the pipeline includes already the relevant metadata as well as provenance information. The software allows writing of the data to several formats, including text-based formats and HDF5, which are the two relevant formats used in this demonstrator.

Data description

Scientific use

Particle event samples can be used in both astrophysics analyses as well as neutrino oscillation studies, see the KM3NeT science targets. Therefore, the data must be made available in a format suitable for the Virtual Observatory as well as for particle physics studies.

Metadata

The events, for which relevant parameters like particle direction, time, energy and classification parameters are included in the event table, are enriched with the following metadata.

Metadata type	content
Provenance information	list of processing steps (referenced by identifier)
Parameter description	parameter name, unit, type, description, identifier
Data processing metadata	start/stop time, detector, event selection info
Publication metadata	publisher, license, creation date, version, description

Technical specification

Data structure

The general data structure is an event list which can be displayed as a flat table with parameters for one event filling one row. Each event row contains an event identifier.

File format

For the tabulated event data, various output formats are used depending on the platform used for publication and the requirements for interoperability. The formats defined here are not exclusive and might be extended according to specific requests from the research community in the future.

For HDF5 files as output, various options exist to store metadata, as several tables can be written to the same file and each table and the file itself can hold additional information as attributes to the file or table. Therefore, metadata that should be easy to find and read for the user have been stored in a separate "header" table, while metadata that is more relevant for the machine-based interpretation of the data has been stored as attributes.

In the case of a text-based table, CSV files are generated that are accompanied by a metadata file.

output format	provenance	parameters	data taking	publication
HDF5	file header	table header	table header	“header” table
CSV table	metadata file	metadata file	metadata file	metadata file

Interfaces

VO server: If the neutrino set is relevant for astrophysics analyses, a text file is generated and the metadata mapped to the resource description format required by the DaCHS software, with the simple cone search (SCS) protocol applied to it. In the ODC, the event sample is recorded as KM3OpenResource pointing to the service endpoints of the VO server. Thus, the data set is findable both through the VO registry and the ODC, and accessible through VO-offered access protocols.

KM3NeT Open Data Server: In the current test setup, event files that are not considered relevant in an astrophysics context, like the test sample from the ORCA detector containing mostly atmospheric muons, are stored on the server, and registered as KM3OpenResource. While this practice is acceptable for the relatively small datasets available now, the design of the server will in the future also provide the functionality to point to external data sources and to interface with storage locations of extended data samples.

3.2.2. Multi-messenger alerts

Data generation

Data generation and scientific use have been described in Section 1.4. The output of the online reconstruction chain is an array of parameters for the identified event as JSON *key: value* dictionary, which then is annotated with the relevant metadata to match the VOEvent specifications.

Data description

The event information can, depending on its specific use, be divided into the following data or metadata categories.

(Meta)data type	content
Event identification	event identifier, detector
Event description	type of triggers, IsRealAlert (or test)
Event coordinates	time, right ascension, declination, longitude, latitude
Event properties	flavour, multiplicity, energy, neutrino type, angular error box 50%, 90% (TOC), reconstruction quality, probability to be neutrino, probability for astrophysical origin, ranking
Publication metadata	publisher, contact

Technical specification

Data structure & format

The VOEvent is stored as XML file which contains central sections of *WhereWhen*, *Who*, *What*, *How* and *Why*.

VO Event specifications

Section	Description	(Meta)data
<Who>	Publication metadata	including VOEvent stream identifier
<WhereWhen>	Space-time coordinates	event coordinates offered in UTC (time) and FK5 (equatorial coordinates) and detector location
<What>	Additional parameters	event properties, event identifier
<How>	Additional information	description of the alert type
<Why>	Scientific context	details on the alert procedure

Interfaces

The alert data is distributed to the GCN. The alert data will be the neutrino candidates in VOEvent format, which is the standard data format for experiments to report and communicate their observed transient celestial events, facilitating follow-ups. The alert distribution is done via Comet which is an implementation of the VOEvent transportation protocol.

Beyond this, there are also other receivers that can be implemented but are less convenient, e.g. the TNS for the optical alerts, the ZTF/LSST broker for the optical transients, the Fermi flare's advocate for the Fermi blazar outbursts.

For public alerts, KM3NeT will also submit notice and circular for dissemination after a short in-person check.

3.2.3. Supplementary services and data derivatives

Data generation

Providing context information on a broader scale in the form of e.g. instrument response functions alongside the VO-published data sets is still under investigation and highly dependent on the specific information. Therefore, additional metadata is required for the interpretation of the format.

Data description

Scientific use

Models and theoretical background information used in the analysis are provided, e.g. accompanying data sets (as for the ANTARES example dataset), to interpret the data sets. Alternatively, probability functions for theoretical predictions and drawn from simulations are considered for publication, including e.g. instrument response functions.

Metadata

Metadata here must be case specific:

- Description of the *structure of the data* (e.g. binned data, formula), which will be indicated by a content descriptor ktype and accompanied by type-specific additional metadata.
- Description of the *basic data set* from which the information is derived, its scope in time and domain-specific constraints, e.g. definition of the simulation parameters.

- Description of all relevant *parameters*.

Technical specification

Data structure & format

The data is provided as CSV table or JSON with the relevant metadata provided alongside the data in a separate text file or in a header section.

Interfaces

Interpretation of the plot or service data is provided using the *openkm3* package, which loads the data as KM3OpenResource from the ODC and interprets it according to the ktype. The relevant data can the be accessed either as array or, where applicable, directly be rendered to a plot using matplotlib, which can then be edited further.

3.2.4. Acoustic data

Data generation

Acoustic data acquisition, as described in Section 1.3.2, offers a continuous data stream of digitised acoustic data that will undergo a filtering process according to the scientific target of the data. Currently, the raw acoustic data can be offered as example data and to researchers interested in sea science. Snippets of acoustic data with a duration of a few minutes are produced at fixed intervals and directly offered, after format conversion, from a data server integrated in the acoustic data acquisition system and made accessible through a REST-API. Integrating this data stream in the Open Science System is a good example of demonstrating the use of a data stream offered externally through the ODC and with a growing number of individual data sets.

Data description

Scientific use

The data can be used, after triggering and filtering, for acoustic neutrino detection, detector positioning calibration and identification of marine acoustic signals, e.g. originating from whales. In the unfiltered form, the acoustic data might primarily be of interest for sea science for the search of maritime signals.

Metadata

- *Publication metadata* is added during record creation at the ODC.
- *Instrumentation & data taking settings* are offered for each data package through a separate endpoint of the REST-API.

Technical specification

Data structure & format

Each data package consists of the same data, recorded in custom binary format (raw), which is formatted to wave and mp3 files. Additionally, statistical properties of the data snippets are offered in a separate stream.

format	endpoint	description	return format
raw	/raw	custom binary format	application/km3net-acoustic

mp3	/mp3	MPEG encoded data	audio/mpeg
wave	/wav	wave format data	application/octet-stream
psd	/psd	array with waveform mean, median, 75% and 95% quantile	application/json

Interfaces

For each file, a KM3OpenResource is registered in the ODC. All resources belonging to the same data type are grouped using the KM3ResourceStream as metadata class, pointing to all resources of the data stream through the kid unique identifier, see Section 2.3.2. All streams belonging to the acoustic data service are grouped as KM3ResourceCollection. Thus, each single resource can be addressed as well as the logical connection between the resources preserved.

The data is directly accessible through the ODC webpage views or using openkm3 as client from a Python interface.

3.3. Platforms and servers

3.3.1. The Virtual Observatory server

The Virtual Observatory environment facilitates data publication and sharing in the astrophysics community according to the standards set by the IVOA. The VO protocols implement the FAIR data principles and ensure an environment for scientific use and public access to data from astrophysics observatories.

Dedicated software and exchange protocols are set up by participating data centres and programs tailored to the users' needs grant easy user access to the diverse data through standardised labelling and description of the data. The KM3NeT Collaboration is a data provider to the VO and operates a data server at http://vo.km3net.de/, running the DaCHS software.

Implementation

Software package

The GAVO Data Centre Helper Suite (DaCHS) is: "a publishing infrastructure for the Virtual Observatory, including a flexible component for ingesting and mapping data, integrated metadata handling with a publishing registry, and support for many VO protocols and standards". It can be applied for KM3NeT purposes as it is and allows to interface various types of data and formats to the VO.

Standards for metadata

Entries to the VO registry are annotated using the Resource Metadata standards of the IVOA. This metadata is required for publication. In the KM3NeT Open Science System, the KM3OpenResource class used to label all resources also contains entries which can be matched to the VO metadata standards, facilitating an automatic adaptation of the standard KM3NeT internal data format to the VO metadata.

Data conversion and upload

Format conversion

In order to transform event-based data to a VO-compatible format, standard scripts have been set up to convert neutrino event tables and the according metadata into VO-compatible format and to add the required metadata to the data set. The procedure to publish a data set includes:

- Labelling the data set according to VO standards with information about origin, authorship, parameter formats and standardised event identifiers. In the DaCHS software, this is handled through a resource description file, to which information from the KM3OpenResource description is cast.
- Selection of the service interface, i.e. the protocols which are offered through the various endpoints of the server to the Virtual Observatory. This interface is also defined in the resource description.

Publication procedure

Publication involves uploading and listing the data set as available in the registry of the server, which is handled through simple administration commands in DaCHS.

Interlink to registries

In order to declare the entries of the local server to the IVOA Registry of Registries, the KM3NeT Collaboration registered the server, obtaining a IVOA identifier (ivo://km3net.org) used in the VO context to uniquely identify the data provider.

Example data

As KM3NeT does not produce astrophysics data yet, an alternative data sample from the ANTARES experiment was used to set up the system. The ANTARES Collaboration has already published two data sets to the VO by using the services of the German Astrophysical Virtual Observatory (GAVO) where the DaCHS software is run and maintained. The most recent public data sample was only made available through the ANTARES website and thus did not match the FAIR criteria. Using the ANTARES 2007-2017 point source neutrino sample, the KM3NeT VO server could be registered to the VO and hosts this data set as example data.

Data access

Tabulated high-level neutrino event data can be accessed utilising access protocols like the Table Access Protocol (TAP) and query languages like the Astronomical Data Query Language (ADQL). To query data sets related to astronomical sources, the Simple Cone Search (SCS) protocol allows to pick specific events according to a given particle source direction.

3.3.2. The KM3NeT Open Data Centre

For all data not publishable through the IVOA, the KM3NeT Open Data Centre is serving as interface and/or server to the data. For the setup of this server, the Django framework has been used.

Implementation

Software package

Django is a Python-based free and open-source web framework that follows the model-template-views (MTV) architectural pattern. Models implement e.g. the KM3OpenResource description, which serves as basic data description class. Templates allow the display of the information through views, which are accessible via web browser. In addition to this, REST-API endpoints can be defined to allow data management and querying the data through web requests. It also offers an admin interface which enables manual adaption of the data via the GUI, and media storage which can serve to hold smaller data sets, along with a broad gallery of functions such as URL handling, testing and model description.

Standards for metadata

For the Django server, which primarily serves as an interface to hold metadata for open science products, the metadata can be freely defined, and will develop with time depending on the requirements of the user communities. Currently, the models used by the server consist of the following three classes:

- **KM3OpenResource** describes each registered data element, independent of the actual type of data. It holds metadata on the publication, the content and a reference link to a description of the content, various optional identifiers like DOI in addition to the km3net identifier (kid), a link to the storage location of the actual data and metadata on data access.

- **KM3ResourceCollection** holds and describes links to several KM3OpenResources via their kid, and adds information on the collection. Resources of different kinds can be combined here, indicating e.g. that they belong to the same research object, the same analysis, etc.

- **KM3ResourceStream** used to group resources of the same type in a collection that can be extended by new resources over time. These resources are automatically updated from the defined URLs.

Data conversion and upload

Format conversion

The conversion of the data to be uploaded to the server or simply registered as KM3OpenResource is described in Section 3.2.

Publication procedure

The publication procedure involves upload of the data to the server, and adding the resource description to the database. If the data product does not yet have a kid assigned, a kid is added to the resource. With this step, the data is made publicly available.

Additional features

The ODC will also be used to store metadata independently of the publication of the data. It includes e.g. a kid search endpoint to allow drawing the metadata on any digital object registered with the server. This will allow the user to collect information e.g. on storage location of files and their descriptions. It serves as an interim solution until a full workflow management scheme has been developed.

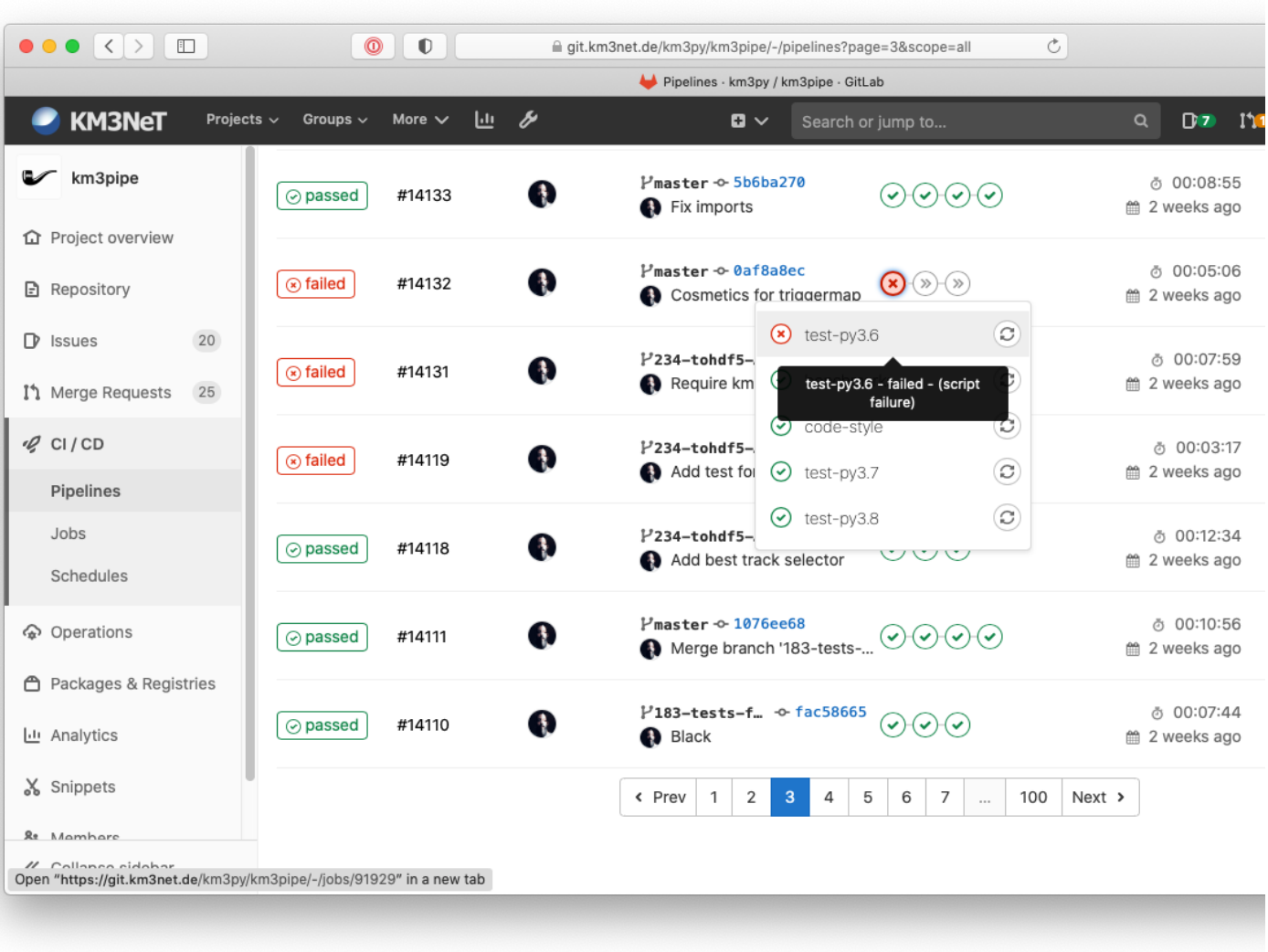

Figure 7: CI Pipelines of the km3pipe projects showing a failing test in a Python 3.6 environment.

3.3.3. Gitlab & Docker server

Gitlab

KM3NeT uses a self-hosted Gitlab instance as the main platform to develop and discuss software, analysis tools, papers and other collaborative creations. Gitlab offers professional and advanced features to keep track of development history and its rich feature set allows for exchanging and archiving thoughts and ideas easily. The continuous integration (CI) that is part of the Gitlab distribution, proves to be a powerful automation tool and is utilised to generate consistently up-to-date test reports, documentation and software releases in a transparent way. The CI pipeline is triggered every time when changes are pushed to a project. Each job runs in an isolated Docker container, making it fully reproducible.

In case of test reports, for example, failing tests are signalled in merge requests, preventing changes that broke them to be applied accidentally. The documentation is built in a dedicated pipeline job and is published through the web upon successful generation. A tight integration of the documentation into the software projects is mandatory and highly improves its up-to-dateness.

The KM3NeT Gitlab server is accessible to the public but only projects which are marked as *global* are visible to a visitor without a KM3NeT account. External users can download the projects and all the public branches, access the issues, documentation and Wiki. They, however, are not allowed to collaborate, i.e. to comment or contribute in any way. To circumvent this problem, open source projects are mirrored to an as yet unofficial GitHub group (https://GitHub.com/KM3NeT), where everyone with a GitHub account is allowed to interact.

Docker

Due to the huge variety of operating systems, programming languages and frameworks, the number of possible system configurations has grown rapidly in the past decades. Operating-system-level virtualisation is one of the most successful techniques to tackle this problem and allows conserving environments, making them interoperable and reproducible in an almost system-agnostic way. KM3NeT utilises Docker for this task, which is the most popular containerisation solution with high interoperability. Docker containers run with negligible performance overhead and create an isolated environment in a fully reproducible manner, regardless of the host system, as long as Docker itself is supported (Linux, macOS and Windows). These containers are used in the Gitlab CI to run test suites in many different configurations. Python based projects, for example, can easily be tested under different Python versions in this way.

3.4. Registries and archiving

Registering the KM3NeT open science products with external platforms and assigning global identifiers is a key to findability of the data. As several KM3NeT servers (VO server, Open Data Centre, Gitlab server) currently hold key data and software, products from these servers must be linked to larger platforms, ideally through semi-automatic integration. Also, after termination of the KM3NeT project, these products must be transferable from KM3NeT-hosted platforms to external repositories that ensure a sustained lifetime of the digital products.

3.4.1. Virtual Observatory registry of registries

Registry structure

In the VO, the KM3NeT server is registered as a registry of resources, i.e. of the datasets and services offered by the server. The IVOA Registry of Registries (RofR) is a service maintained by IVOA that provides a mechanism for IVOA-compliant registries to learn about each other, being itself a compliant publishing registry which contains copies of the resource descriptions for all IVOA registries. When a resource metadata harvester harvests from these publishing registries, it can discover all published VO resources around the world.

Identifiers

The KM3NeT VO server is registered with the following metadata to the RofR:

- name: KM3NeT Open Data Registry
- IVOA Identifier: `ivo://km3net.org/__system__/services/registry`
- OAI service endpoint: http://vo.km3net.de/oai.xml

With this registration, KM3NeT data in the VO is fully findable within the Virtual Observatory, and each resource is identifiable through the naming of the individual endpoint of the service within the registry.

Archiving

Various national organisations in the KM3NeT member states offer long-term repositories, to which the KM3NeT data sets might be transferred after the end of KM3NeT operation. As an example, the GAVO Data Centre provided by *Zentrum für Astronomie Heidelberg* on behalf of the GAVO already hosts other ANTARES data sets and is provider of the DaCHS software used in the KM3NeT VO server, so transfer for archiving would, at least from today's perspective, be easy to achieve.

3.4.2. DataCite and Zenodo

Universal citation of data and digital objects is generally handled through assignment of a Digital Object Identifier (DOI), a persistent identifier used to identify objects uniquely, which is standardised by the International Organization for Standardization (ISO). The VO supports the integration of a DOI in their resource metadata, but does not provide the authority to assign DOIs, as the organisation assigning a DOI generally also has to host the data to ensure the longevity of the data resource.

This leads to the dilemma that the KM3NeT Collaboration would have to become a member organization of e.g. DataCite, a global non-profit organisation that enables the assignment of DOIs for its member organizations. This would enable the collaboration to operate a repository and assign DOIs, or the data would have to be copied or mirrored to another repository in order to obtain a DOI. While this issue is not yet resolved and will be investigated further also in the context of the ESCAPE project, the second option has been chosen for the small amount of data that is currently provided in this demonstrator. As hosting repository, Zenodo was chosen as well-established data repository in the physics community.

Registry structure

Zenodo was initiated by the OpenAIRE project and is hosted by CERN to offer a repository for research funded by the European Commission. It hosts publications, audio and visual media, software and datasets. It allows to create “communities” to group various resources, and has been used to create a KM3NeT community. Uploads can be managed through an API as well as through a web interface and standardised metadata is required.

Identifiers

Zenodo assigns DOIs to resources on upload if a DOI cannot be provided, making the dataset also easily citable. For this demonstrator, the event sample from the KM3NeT ORCA use case will be registered with Zenodo.

Archiving

As Zenodo offers long-term support of its resources and the data is already mirrored to the repository, archiving of the data can easily be accomplished.

3.4.3. Software integration

GitHub, Zenodo and PyPI

With GitHub, a major platform exists for software development that also allows for easy interaction between software developers on various projects. Open KM3NeT software is mirrored to GitHub, making the software findable for software developers. For software grouping, a KM3NeT Collection has also been established here.

For Python-based software, easy installation via the pip package installer is integrated in the software packages. This installer links to the Python Package Index (PyPI), a repository of software for the Python programming language. Here, a KM3NeT user account has been established to group and administrate the software.

Identifiers

Neither GitHub nor PyPI make software citable in the strict sense, as they do not assign DOIs. However, Zenodo allows the integration of GitHub repositories to their platform. Therefore, the combination of mirroring software to GitHub, making it installable via PyPI and registering it with Zenodo, makes the software accessible for community development, easily installable for users and citable in a scientific context.

Archiving

Both PyPI and Zenodo copy the relevant software to their platforms, copies of the source code are stored at multiple sites and make archiving easy and safe.

3.5. Documentation and tutorials

3.5.1. Documentation

The use of inline code documentation such as Doxygen or sphinx to create detailed references and documentation of the full source code is a requirement for the KM3NeT software, see Section 2.5.2. In addition, nomenclature of classes, functions and variables must follow a standardised format to ensure a consistent code style.

In addition to the inline code documentation, each software package is accompanied by documentation served through webpages hosted on the Gitlab instance, which includes a quick start and a basic user manual. In the open science context, example workflows are made available as Jupyter notebooks either provided alongside the software or downloadable from the gallery provided in the OSP.

Like the Open Science Portal, dedicated Gitlab projects have been introduced for the documentation of central aspects of the Open Science System like licensing, data models and quality control (internal). This approach allows an easier interlinking between the various components of the Open Science System and flexible handling of documentation.

3.5.2. Virtual Education Centre

The educational material of the KM3NeT Collaboration is provided through the Virtual Education Centre. While the first prototype was based on an Indico service, this was found insufficient for the aims of a Virtual Education Centre during user tests. There was, for

example, no video streaming service and no feedback mechanism available. The new Virtual Education Centre consists of a server running the WordPress content management system. The LifterLMS plug-in has been installed in WordPress, providing a specialised education management system.

The courses in the Virtual Education Centre are divided in two categories. They are either protected or freely available. The courses belonging to the first category are accessible exclusively for KM3NeT members, who are able to enrol by authenticating with their KM3NeT user identity. These courses are oriented mainly to newcomers making their first steps in the collaboration. The courses of the second category are openly and freely accessible by all users, containing information on how to use the open data provided by the collaboration and aimed to provide an introduction to scientific research with the KM3NeT detectors.

The education management system provided by the LifterLMS plug-in offers the possibility to set up the sequencing of instructions in hierarchical categories like courses, sections and lessons. The sequence in which the courses have been set up is, in principle, a reference program for face-to-face training meetings. In addition, it allows the users to follow the instructions in a self-paced study program independent of dedicated meetings.

3.6. Accessing data

3.6.1. The Open Science Portal

Independent from the technical accessibility of the data, the user needs to have the possibility to gain an overview over the Open Science System and on how to proceed to access the data. To this end, a webpage using the sphinx documentation software has been set up as knowledge base for the KM3NeT Open Science System. This Open Science Portal (OSP) contains descriptions of the experiment, science cases and data generation, and points to the various services, tutorials and example use cases. This portal serves as central access point and landing page for external scientists intending to use the data. Therefore, the various platforms link to the portal for further instructions, and will also serve as starting point for the system testing of the Open Science System, see Section 5.2.

3.6.2. Python environment

KM3NeT develops open source Python software for accessing and working with data taken by the detector, produced in simulations or in other analysis pipelines such as event reconstructions, and a number of other data types like metadata, provenance history and environmental data. The software is following the Semantic Versioning 2.0 conventions and releases are automatically triggered on the Gitlab CI by annotated Git tags. These releases (including alpha and beta releases) are uploaded to the publicly accessible Python Package Index, which is the main repository of software for the Python programming language. The installation of these packages is as simple as executing the command: `pip install PACKAGE_NAME`. Additionally, the packages can also be installed directly from the Gitlab repositories, for example in case of experimental branches.

Preferred Python packages

The general philosophy behind all Python packages is to make commonly used open source scientific tools, libraries and frameworks available for use in KM3NeT. While the common

base is built on NumPy, the de facto standard for scientific, numerical computing in Python, other popular packages from the SciPy stack are heavily used. Examples are *matplotlib* to create publication quality plots, *pandas* which is used to work with tabular data, *astropy* for astronomical calculations or *numba* for high-performance low level optimisations, or *gammapy* for common analyses with CTA.

Preferred formats for interoperability

The output format is preferably CSV and JSON to maximise interoperability. For larger or more complex datasets, two additional formats are supported. HDF5 which is a widely used data format in science and accessible in many popular computer languages, is used to store data of all data levels, including uncalibrated low-level data and high-level reconstruction summaries. Additionally and mainly for astronomical data, the FITS data format is considered if required due to its high popularity among astronomers.

Python interface to KM3NeT data

In addition to offering services and data through the KM3NeT Open Data Centre, the openkm3 python client was developed to directly use open data in python from local computers and within e.g. Jupyter notebooks. It interlinks with the ODC REST-API and allows to query metadata of the resources and collections. In addition to that, it offers functionality to interpret the data according to its KM3NeT type description (ktype), e.g. returning tables in a required format. These interface options will be expanded according to the requirements of data integrated to the ODC.

In addition to that, basic functions relevant for astrophysics are offered in the km3astro package. As development of the python environment is an ongoing process, the number of packages offered for KM3NeT data interpretation will surely grow in the future.

3.7. Further development of the Open Science System

Currently, KM3NeT participates in the ESCAPE (European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures) project, integrating their open science products to the European Open Science Cloud. In the framework of this cooperation, various issues for KM3NeT in the open science context will be picked up and their solutions developed further.

3.7.1. Development of Virtual Observatory standards

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. 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 standardisation. This process, however, will draw guidance from the VO expertise and framework.

3.7.2. Data citation and access

Integration of large data sets, authorisation for users to access these data sets, and finally providing a scheme for usage of computing resources for processing large data sets are issues that are addressed in the context of ESCAPE. This will be especially relevant for future full releases of KM3NeT datasets, for which downloading to a personal computer is no longer a feasible option.

In this context, also issues of data citation, registration and integration of data handling for the wider science community will be addressed.

3.7.3. Software licensing

In the ESCAPE context, an interest group has been formed addressing the issue of software and data licensing, and especially the interoperability of various software licenses. Especially if community development of software is to be pursued, copyright issues can become complex and at some point even damaging for the further development of a project. Although a working solution for KM3NeT regarding licenses has been found for the time being, results from this working group might lead to an adaption of the licensing policy in the future.

3.7.4. Workflow management

As described in Section 2.2, two software solutions are to be integrated to ensure solid documentation and managed execution of the KM3NeT data processing chain. To this end, another working group in the ESCAPE context is testing the use of the DIRAC interware, a management system for distributed, data driven job execution across several HPC centres, and a better description of complex workflows through a workflow description language is investigated, namely the Common Workflow Language (CWL).

3.7.5. EOSC Integration

Further integration is already under way, as currently the example data is integrated into the ESAP (ESFRI Science Analysis Platform) of the ESCAPE project to allow data sharing within this emerging environment. Also, software is provided within the OSSR (Open-source scientific Software and Service Repository) of ESCAPE, and experience shared in the development of the various platforms.

4. Example programs and use cases

4.1. ANTARES 2007-2017 point source analysis

4.1.1. Use case description

One of the primary goals of ANTARES is the identification of astrophysical neutrino sources, with clusters of events at given coordinates in the sky as signatures. The ANTARES data

used for this search are therefore a set of reconstructed neutrino arrival directions (in equatorial coordinates: right ascension RA, declination dec). The significance of a neutrino excess from a given sky position must be assessed over the expected background fluctuations using, for instance, the Feldman-Cousins statistics Phys.Rev.D57:3873-3889,1998. The background mostly comes from atmospheric neutrinos: neutrinos originating from cosmic ray interactions in the Earth's atmosphere. Their arrival directions are isotropical and at the ANTARES detector, uniform in right ascension and with a distribution in declination that depends on the detector geographical position (latitude).

This use case is to inspect a sample of neutrino arrival directions in equatorial coordinates (RA, dec), evaluate the expected background rate for a user-selected sky position, and finally assess the significance of a cluster defined as all arrival directions that fall inside a given radius, selected by the user and indicated here as 'region of interest' (RoI).

The background is evaluated in a declination band with width equal to the RoI diameter. The observed signal-like events are those falling inside the RoI. Feldman-Cousins statistics is applied to determine the significance of this observation accounting for the fluctuations expected in the background. The aim of the analysis is to set an upper limit, at a given confidence level, on the presence of a source with a given energy spectral index at a given sky position.

An upper limit on the number of events observed is not a very useful quantity for the community outside ANTARES, because it still contains the detector response. The upper limit on the number of events is therefore turned into an upper limit on the flux emitted by the neutrino source, which is the relevant information for drawing conclusions on a neutrino source model. To unfold the effect of the detector response, the acceptance of the detector is computed and provided alongside the ANTARES arrival direction data.

4.1.2. Data set

Two ANTARES data sets have been already published using the services of the German Astrophysical Virtual Observatory (GAVO). The most recent public data sample of the 2007-2017 point source search is available through the ANTARES website. However, it is not findable through the VO and therefore does not match the FAIR criteria. Including ANTARES data in the development of future VO data types increases the chance for a long-term availability of high-quality ANTARES data. Furthermore, the KM3NeT VO server has been registered to the VO and protocols have been tested using the ANTARES 2007-2017 sample.

The provided data set includes:

- The **full event list** of 2007-2017 selected astrophysics neutrino candidates, provided through the VO server.
- Supplementary distributions from simulations provided via the Open Data Centre including:
  - the **detector acceptance** for a given source spectral index and declination,
  - the interpolated **distribution of background events** for a given declination and region of interest,
  - the detector **effective area** for an E^{-2} source spectrum in three different zenith bands.

4.2. KM3NeT use case

4.2.1. Analysis of an atmospheric muon-dominated dataset

The KM3NeT use case contains a dataset with triggered events reconstructed with 4 Detection Units of KM3NeT/ORCA. It is therefore dominated by atmospheric muon events (several Hz, compared to the atmospheric neutrino rate of ~mHz) reaching the detector from above. The recorded and reconstructed dataset is provided on the KM3NeT VO server in HDF5 format for download and analysed with Jupyter notebooks.

4.2.2. Run selection and online sample

ORCA events are pre-selected by imposing requirements on variables stored in the internal database:

- 'PHYSICS' runs (i.e. no calibration runs, etc.),
- long runs (>2 h duration),
- less than 100 s recorded live-time missing with respect to the run duration in the database,
- runs with fairly low optical rates (a 'High Rate Veto' (HRV) variable is used, indicating the time-averaged fraction of photo-sensors with too high rates (20 kHz) due to optical background noise. Runs with HRV fraction < 0.2 are selected).

4.2.3. Processing to metadata enriched HDF5 format

For demonstration purposes, a range of data-taking runs meeting the run selection criteria above is published on the KM3NeT VO server.

The HDF5 dataset contains all events satisfying all event trigger conditions of KM3NeT/ORCA. The data sample currently contains events reconstructed with the track reconstruction algorithm JGandalf. In addition to the basic event features needed for analysis (*time*, *zenith*, *azimuth*, *energy)*, it contains also other features provided by the reconstruction algorithm: an estimate for the *angular error*, the reconstruction *quality* and the number of signal hits used in the reconstruction. More sophisticated high-level variables needed for selecting a pure neutrino sample are not available for the time being, but will be added in the future.

The dataset was converted from the internally used aanet format to HDF5 format, and enriched with metadata describing the provenance and contents of the file.

4.2.4. Analysis examples

Requirements

Jupyter notebooks are based on python3 and require packages installable via pip.

Analysis examples 1-4: pip installable packages (astropy, numpy, matplotlib, openkm3)

Analysis example 4 only:

- Aladin: https://aladin.u-strasbg.fr/java/nph-aladin.pl?frame=downloading

- TOPCAT: http://www.star.bris.ac.uk/~mbt/topcat/#install .

Description of KM3NeT/ORCA use cases

The Jupyter notebooks for the KM3NeT use case are available on the Open Science Portal.

Four analysis examples are provided as KM3NeT use-cases:

A01 In the first example the HDF5 data file is retrieved from the KM3NeT server, the continuity of data-taking and the event distribution in local coordinates is analysed.

A02 In the second example, the use of the reconstructed direction but also the other quality parameters contained in the HDF5 file is outlined in order to select 'interesting' events. Here 'interesting' events are those with an enhanced probability to be neutrinos.

A03 The provided event sample is converted and visualised in galactic coordinates.

A04 The HDF5 dataset is retrieved and celestial coordinates are added. The event sample is then used to search for events coincident in space and time with a Gravitational Wave (GW) event.

5. Tests of the Open Science System

Based on this testing tutorial, the following testing scheme has been implemented for the Open Science System. It ranges from automatic testing at the software development level to manual testing by a test user to verify the system usability and acceptance.

5.1. Automated software testing

5.1.1. Unit testing

Unit testing tests the core functionality of code parts and functions. Unit tests are part of each software project used in the KM3NeT Open Science System that is not third-party software. Test evaluation is automatic and part of the continuous integration pipeline in Gitlab, and tests are implemented by the programmer. Publication of the software and integration into the Open Science System requires a high degree of coverage of the overall code for automatic testing. In the current setup, the workflow tool km3pipe, the Django-based Open Data Centre, the astropy derivative km3astro, the metadata tool kmeta and the python client openkm3 all implement the "unit testing".

5.1.2. Integration testing

The software design is tested in integration testing as part of the automated procedure. Test cases of the software include reading test data, performing complex functions of the software and testing the construction of the software by performing tests on various subsets and groups. The KM3NeT software also includes integration testing as part of the automated test cases.

5.1.3. System testing

The interoperability of software components is evaluated in system testing. Workflows relevant for testing are defined also as example workflows for documentation and provided e.g. as Jupyter notebooks, but they also serve to test the functioning of the full workflow. Part of this testing is also performed manually, as new test cases can be defined during the acceptance testing which lead to new system test workflows.

5.2. Manual testing of the system

In addition to the automated testing, a manual testing procedure has been performed on the full system by an external test user. In this testing procedure, the following steps were performed:

- Introduction to the Open Science System for the test user by pointing to the documentation to the Open Science Portal.
- Definition of test cases by the tester, including example work flows and additional testing of partial functions.
- Testing and reporting on the testing by the tester through a standard test protocol.
- Implementation of changes and recommendations from the test results.

The tests were implemented at the time of writing this deliverable. Documentation of the testing procedure can be made available on demand. Similar manual test will be performed when introducing major architectural changes or new types of open data to the Open Science System.

5.3. Test of the Virtual Education Centre

A webinar (online workshop) has been held in order to test the open courses. The participants were external scientists as well as KM3NeT members. They were asked to have watched the first set of the educational videos before the webinar, in order to install the relative programs and to get familiar with them. The webinar was coordinated by a KM3NeT member. The participants watched the second set of educational videos at the same time, and then some additional time was given by the coordinator for a hands-on practise with a task assigned for each video. Positive Feedback was provided by the participants at the end of the workshop in form of an open feedback request.

6. Conclusion

In this report, all current parts of the Open Science System in KM3NeT are presented and their use has been tested and documented. In the Open Research Data Pilot, the general architecture has been established for sharing of data and software. The integration of the KM3NeT open data procedures and plans for other European projects, especially the ESCAPE project contributing to the EOSC, will ensure the further development and integration of KM3NeT's open science products beyond the pilot stage.