Funded by
the European Union

TOWARDS FULL IMPLEMENTATION OF THE KM3NeT RESEARCH INFRASTRUCTURE

KM3NeT-INFRADEV2 – HORIZON – 101079679

Report on the results of the review of the KM3NeT DMP by external experts

KM3NeT-INFRADEV2 GRANT AGREEMENT DELIVERABLE: D4.1	
Document identifier:	KM3NeT-INFRADEV2-WP4-D4.1-v.final
Date:	31/03/2024
Work package:	WP4 Data Management and Open Science
Lead partner:	FAU
Document status:	FINAL
Dissemination level:	PUBLIC
Document link:	https://www.km3net.org/km3net-eu-projects/km3net-infradev2/infradev2-outputs/

ABSTRACT

A review of the Data Management Plan (DMP) and options for future computing strategies by external experts was arranged close to the project start. A detailed list of recommendations and suggestions was issued by the reviewers and is being taken into account in updating the DMP. This deliverable is the report of the DMP review panel.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the REA can be held responsible for them.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner and WP	Date
From	Jutta Schnabel	FAU, WP4	21/02/2024
Author(s)	Mieke Bouwhuis Vincent Cecchini Kay Graf Jutta Schnabel	NWO-I, WP4 CSIC, WP4 FAU, WP4 FAU, WP4	29/02/2024
Reviewed by	Yuri Shitov Richard Randriatoamanana		22/03/2024
Approved by	Paschal Coyle KM3NeT IB	CNRS, WP1	31/03/2024

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	21/02/2024	1st draft	Jutta Schnabel, FAU
2	29/02/2024	Final draft ready for review	
3	22/03/2024	Review completed	Yuri Shitov, CVUT Richard Randriatoamanana, CNRS
4	25/03/2024	Draft ready for IB submission	Victoria Ciarlet Thaon, CNRS
5	29/03/2024	Corrections following IB review	Jutta Schnabel, FAU

IV. APPLICATION AREA

This document is a formal deliverable of the Grant Agreement of the project, applicable to all members of the KM3NeT-INFRADEV2 project, beneficiaries and third parties, as well as its collaborating projects.

V. TERMINOLOGY

AAI	Authentication and Authorization Infrastructure
AP	Action Point
CC	Computing Center
CERN	*Conseil Européen pour la Recherche Nucléaire* (European Organization for Nuclear Research)
CNAF	National Centre for the Research and Development in INFN Information and Communication Technologies
CPU	Central Processing Unit
CSA	Coordination and Support Action
CTA	Cherenkov Telescope Array
DL	Data Level
DMP	Data Management Plan
DU	Detection Unit
EGI	European Grid Infrastructure
EOSC	European Open Science Cloud
ESCAPE	European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures
GCN	General Coordinates Network
HPC	High-Performance computing
IRODS	Integrated Rule-Oriented Data System
IVOA	International Virtual Observatory Alliance
KM3NeT	Cubic Kilometre (km3) Neutrino Telescope
LCG	LHC Computing Grid
LHC	Large Hadron Collider
LSST	Large Synoptic Survey Telescope
LUPM	Laboratoire Univers et Particules de Montpellier
MoU	Memorandum of Understanding
RAM	Random-Access Memory
RI	Research Infrastructure
TCO	Total Cost of Ownership
VO	Virtual Observatory
WLCG	Worldwide LHC Computing Grid
WP	Work Package
ZTF	Zwicky Transient Facility

VI. LIST OF ANNEXES

VII. PROJECT SUMMARY

The Kilometre Cube Neutrino Telescope (KM3NeT) is a large Research Infrastructure (RI) comprising a network of deep-sea neutrino telescopes in the Mediterranean Sea with user ports for Earth and sea science instrumentation. During the EU-funded Design Study (2006-2010) and Preparatory Phase (2008-2012), a cost-effective technology was developed, deep-sea sites were selected and the Collaboration was formed in 2013. This proposal constitutes a second INFRADEV project dedicated to KM3NeT in order to implement an efficient framework for mass production of KM3NeT components, accelerate completion of its construction and provide a sustainable solution for the operation of the RI during ten or more years. Following the appearance of KM3NeT on the 2016 ESFRI Roadmap and in line with the recommendations of the Assessment Expert Group, this project addresses the Coordination and Support Actions (CSA) to prepare a legal entity for KM3NeT, accelerate its implementation, establish open access to the RI and its data and ensure its sustainability by implementing an environment-friendly operation mode and evaluating the Collaboration socio-economic impact.

VIII. EXECUTIVE SUMMARY

This document summarises the KM3NeT DMP review process that took place within WP4 of the KM3NeT-INFRADEV2 project, in 2023. It describes the organisation of a review by a panel of external experts, its outcomes and the resulting action points for the update of the DMP. Some additional action points including changes in data processing or technical implementations are also mentioned as results of this review process.

IX. TABLE OF CONTENTS

1. Introduction

The Data Management Plan (DMP) of KM3NeT had been written during the former INFRADEV call (KM3NeT INFRADEV – H2020 – 739560[1]) and provided in 2020. In the meantime, the strategies and implementation of data management in KM3NeT have developed further in close exchange with partners in the context of EOSC-related initiatives and multi-messenger astronomy. In order to streamline these developments and implement a concise data management strategy for the full runtime of the KM3NeT detectors, the DMP was reviewed and presented to external experts in the first part of the INFRADEV2 project. This deliverable summarises the preparation, conduction and outcome of the review, as well as the action points drawn from this process for the future implementation of data management in KM3NeT.

2. Organization of the DMP review

Review panel expert selection and communication

Candidates to serve as experts on the review panel were collected by inquiring for recommendations from the current computing centres of KM3NeT, colleagues from EOSC-related projects and contacting known experts in the field in Q2/2023. In the selection of suitable candidates, it was ensured that both experts for HPC infrastructures as well as scientific computing in the area of high-energy astroparticle physics were represented.

It was possible to set up the review panel consisting of the following experts:

- Arrabito, Luisa: Software engineer at LUPM CNRS/IN2P3 for CTA
- Bouvet, David: Technical manager for LCG France, CNRS/IN2P3
- Espinal, Xavier: Senior applied physicist at CERN
- Litmaath, Maarten: WLCG Operations Coordination co-chair, CERN

Prepared material and review panel meeting

The DMP was updated during the first half of 2023 to version 2.1 to include major changes from the status at the end of 2020. Main alterations included the transition to Grid computing in data management and the addition of extended chapters on the open science and multi-messenger alert systems. Also, management aspects were brought in alignment with the current status of considerations towards a legal entity in KM3NeT.

The DMP v2.1 was circulated to the collaboration, and a dedicated internal meeting was held for all collaboration members (Open Science Forum, 20/09/2023) for discussions about the

[1] see https://cordis.europa.eu/project/id/739560

current data management strategy and to gather input for the upcoming review panel meeting.

The DMP v2.1 was provided to the reviewers by 24/07/2023, accompanied by partially internal additional information containing extended documentation for the online alert system, the data quality plan, information on data levels and modelling and the open science policy.

The review panel meeting was held on 27/09/2023[2], with part of the day dedicated to a walk-through presentation of the DMP content, the second part used for open discussions. The full minutes of the meeting are available at the event page.

3. Outcome of the DMP review

Feedback by the experts

In addition to the direct oral feedback and outcome of discussions during the review panel meeting, the experts provided a condensed feedback of major comments within a week after the meeting, see Annex 1.

Action points derived from the DMP review

From the written feedback provided by the panellists and from the minutes from the full review panel day, a list of action points was derived which will be implemented during the remaining time of the project. These action points were addressed to the various stakeholders, including the KM3NeT management team, the INFRADEV2 work package on the legal entity (WP2) and the computing and software working group and will be followed up by the members of the WP4 team. They partially included updates to the DMP, which is provided in a revised version below as Annex 2, and organisational and management changes and suggestions for the future data management strategy, which will be followed up during the remainder of the project and implemented according to the considerations by the KM3NeT Collaboration. For these considerations, the relative benefit, technical feasibility and availability of resources in the Collaboration will be taken into account.

To add all these changes and outcome of the adopted change processes, including those requiring a longer implementation time, a completely revised version of the DMP will again be provided at the end of the INFRADEV2 project.

[2] Event on indico: https://indico.cern.ch/event/1304459/

4. Impact on the work package goals in INFRADEV2

All recommendations and suggestions by the reviewers, derived from the written feedback and from the minutes of the meeting, will be taken into consideration. All action points are either derived from the minutes of the meeting or from the written feedback by the reviewers.

Revising the DMP

These changes reflect points that were mainly requested for clarification and restructuring the DMP which have been addressed in the updated DMP in Annex 2:

- Revise the exact needs for data storage and processing at the shore stations and to add information about current agreements, including shore station links to the computing centres,
- Include human resources for computing duties in the DMP, including for the online alert system,
- Document backup strategies (sites, volume, scheduling) and risk assessment applied at the contributing computing centres, including on the database, and information about computing resource needs (RAM, CPU, storage),
- Include the ANTARES data in the DMP and add information about simulation needs in the future and data load from acoustic data,
- Include citizen science in data management and make the hosting of the open science system clearer.

Additional action points

All recommendations and suggestions by the reviewers, derived from the written feedback and from the minutes of the meeting, will be taken into consideration. The following list summarises the main action points that have been already identified and will be taken up during the remainder of the project, which mainly include change processes or recommendations for technical implementations:

- Improve the data processing efficiency starting from the bottlenecks of the data processing chain,
- Produce data sets for common high-level data analysis to reduce reprocessing overhead at the analysis level,
- Clarify user support and the contribution of KM3NeT members to computing duties,
- Create a legal agreement with the main computing centres and consider becoming an observer to the LHCONE and WLCG collaboration,

- Implement the outcomes of the ESCAPE project[3] if possible, and check the use of other CERN-related software for KM3NeT,
- Set up monitoring of services and resources,
- Review the current status of AAI in KM3NeT, investigate possibly unmet needs and corrective actions with the goal of smooth operation in a fully Distributed Computing environment,
- Review the current status of Oracle licensing and cost outlooks; explore possible free alternatives with full TCO plan including expertise for maintenance and transition cost/personpower.

5. Conclusion

The review of the DMP could be carried out as planned. With the conducting of the review, the data management strategy for KM3NeT received constructive feedback on the goals and for the implementation of a sound data management system for KM3NeT. The results of the review not only helped to cut the written DMP clearer, but also gave rise to several helpful developments which will be followed up during the remainder of the project duration. On the other hand, the review showed that work in this regard is already well under way and therefore confirmed to a large extent the project path laid out in this work package.

[3] https://projectescape.eu

X. ANNEXES

Annex 1 : Review panel Recommendations

KM3Net DMP Reviewers High Level Summary of Inputs

First of all, we would like to acknowledge the high quality of the Data Management Plan and the additional documentation on Open Science, Data quality and Online Processing.

Please find below a collection of topics / items as the main guidelines and recommendations from the review board.

The reviewers suggest / recommend to:

- Check consistency of raw data rates and general data volumes across the different sections in the DMP document.
- Beyond the estimation of the expected data volumes for the different categories of data, for each data level and/or category, provide information about the number of replicas, the number of versions and the target storage (disk and/or tape).
- Estimate the requirements in terms of memory usage for the various processing stages in relation with the CPU consumption provided throughout the DMP document.
- The chapter of the DMP related to roles and human resources should also cover an estimation of the personpower needed for the processing of the low-level data (DL0 through DL2): data processing and simulation operations, software maintenance and IT service administration (e.g. Rucio, Dirac, etc.).
- Agree on an AAI framework and deploy it as a first building block to adopt a Distributed Computing model.
- Consider the possibility to replace Oracle with an open-source product as a provider for the database (e.g. MySQL, MariaDB or rather PostgreSQL).
- Ensure that open science resources (hardware, services and human effort) are provisioned separately from resources available to KM3NeT members and do not interfere with the activities of the latter.
- Migrate out from iRODS to a general purpose Grid Storage service (Dirac) that integrates seamlessly with Data Management tools, ie. Rucio and FTS.
- Investigate if independent, concurrent analyses of large quantities of the same data by different users might benefit from a framework in which users can participate in order to have their analysis jobs combined in a way that allows such data to be downloaded or streamed only once, thus allowing potentially enormous savings on I/O.

- Set up an MoU or an equivalent agreement with the collaboration regarding involvement in computing (personpower) to participate in operations, maintenance, user support and central computing duties (eg. processing campaigns, etc.).
- In the context of the transition to a Distributed Computing model, consider for the participating sites only those which commit to a minimum level of resources. For storage an indicative value could be at least 1 PB, for computing at least 1000 cores.
- Set up an MoU or an equivalent agreement between the KM3Net collaboration and EGI and/or a participating institute regarding the service level of the DIRAC service(s) required by the collaboration.
- Evaluate the participation in and the usage of the WLCG overlay network “LHCONE”.
- Evaluate the participation in WLCG as an observer. The following experiments: DUNE, Belle-II, JUNO and Virgo, have already taken the observer role in WLCG in the past few years.

Annex 2 : KM3NeT Data Management Plan v2.2

KM3NeT Data management plan

Authors: The KM3NeT Collaboration (M. Bouwhuis, V. Cecchini, K. Graf, U. Katz, F. Salesa Greus, J. Schnabel, F. Vazquez de Sola Fernandez)

Abstract

This document presents an update of the data management plan which was created as an output of the first INFRADEV project. It contains information on the data to be collected, generated and processed; the applied standards and methodologies; the open data access standards to be employed; and the implementation of data preservation and curation. The data management plan is based on FAIR data management guidelines and will be updated during the KM3NeT project lifetime whenever new decisions and implementations of relevant points have been addressed.

Recipients: The KM3NeT Collaboration

This document was prepared with funding by the European Union from the KM3NeT INFRADEV – H2020 – 739560 project and the KM3NeT-INFRADEV2 – HORIZON – 101079679 projects

Document Status

Revision	Date	Comment	Reviewed by	Approved by
v 2.2	29/02/2024	Update including feedback from review panel of INFRADEV2		

Revision History

Revision	Date	Description
v 2.1	24/07/2023	Update for review panel in KM3NeT-INFRADEV 2 project
v 2.0	31/10/2020	Update at the end of the KM3NeT-INFRADEV Project
v 1.4	22/06/2017	Deliverable D4.1 first version at beginning of INFRADEV
v 1.3	12/06/2017	Version for endorsement of KM3NeT-INFRADEV PMB and KM3NeT IB
v 1.2	07/06/2017	Updated with feedback of WP4 members
v 1.1	01/05/2017	Initial draft derived from the KM3NeT computing model

I. Terminology

AAI	=	Authentication and authorization infrastructure
API	=	Application Programming Interface
ARCA	=	Astroparticle Research with Cosmics in the Abyss (KM3NeT neutrino astroparticle physics telescope)
ASTERICS	=	Astronomy ESFRI & Research Infrastructure Cluster (EU H2020 project)
BB	=	Building Block
CVMFS	=	CernVM File System
DIRAC	=	Software Framework for distributed computing
DL	=	Data Level
DMP	=	Data Management Plan
DU	=	Detection Unit
DOI	=	Data Object Identifier
EOSC	=	European Open Science Cloud
ESCAPE	=	European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (EU H2020 project)
ESFRI	=	European Strategy Forum on Research Infrastructures
FAIR	=	Findable – Accessible – Interoperable – Reusable
GCN	=	Gneral Coordinates Network
GEDE-RDA	=	Group of European Data Experts in Research Data Alliance (RDA)
GNN	=	Global Neutrino Network
HDF5	=	Hierarchical Data Format
HPC	=	High-Performance Computing
HPSS	=	High Performance Storage System
HTC	=	High-Throughput Computing
IAM	=	Identity and Access Management
IB	=	Institutional Board
INFRADEV	=	Infrastructure Development, Horizon Europe Program Optimising the European Research Infrastructures Landscape
iRODS	=	Integrated Rule-Oriented Data System
IVOA	=	International Virtual Observatory Alliance
JSON	=	Javascript Object Notation
LHCONE	=	Large Hadron Collider Open Network Environment
MC	=	Monte Carlo
ODC	=	Open Data Center
ORCA	=	Oscillation Research with Cosmics in the Abyss (KM3NeT neutrino particle physics detector)
OSSR	=	Open-source Scientific Software and Service Repository
PID	=	Persistent identifier
PMB	=	Project Management Board
PMT	=	Photo-Multiplier Tube
RI	=	Research Infrastructure
VO	=	Virtual Observatory (astronomical dataset context) Virtual Organisation (Grid computing context)
VODF	=	Very high energy Open Data Format
VRE	=	Virtual Research Environment
WP	=	Work Package

II. List of figures

III. List of tables

IV. Project Summary

KM3NeT is a large Research Infrastructure that will consist of a network of deep-sea neutrino telescopes in the Mediterranean Sea with user ports for earth and sea sciences. Following the appearance of KM3NeT 2.0 on the ESFRI roadmap 2016 and in line with the recommendations of the Assessment Expert Group in 2013, the KM3NeT-INFRADEV project addresses the Coordination and Support Actions (CSA) to prepare a legal entity and appropriate services for KM3NeT, thereby providing a sustainable solution for the operation of the Research Infrastructure during ten (or more) years. In the KM3NeT-INFRADEV2 project, the objectives comprise, amongst others, the update of the Data Management Plan in the dedicated work package 4 containing also Open Science and Multimessenger Alert handling.

The Data Management Plan is a working document that has been and will be updated through all phases of the KM3NeT project. The current version reflects the status at the end of the first phase of the KM3NeT-INFRADEV2 project. It summarises the KM3NeT data and computing model as well as KM3NeT plans for data management for open science. The DMP will serve as a basis for the implementation of data management during the remainder of the KM3NeT-INFRADEV2 project.

V. Table of Contents

1. Introduction

KM3NeT is a large Research Infrastructure (RI) that will consist of a network of deep-sea neutrino detectors in the Mediterranean Sea with user ports for Earth and Sea sciences. The main science objectives, a description of the technology and a summary of the costs are presented in the KM3NeT 2.0 Letter of Intent [1].

KM3NeT will open a new window on our Universe, but also forward the research into the properties of neutrinos. With the ARCA telescope, KM3NeT scientists will search for neutrinos from distant astrophysical sources such as supernovae, gamma-ray bursts or active galactic nuclei. Using the exact same technology, the ORCA detector will provide data of unprecedented quality on neutrino oscillations, exploiting neutrinos generated in the Earth's atmosphere. Arrays of thousands of optical sensors will detect the faint light in the deep sea from charged particles originating from collisions of the neutrinos with atomic nuclei. The facility will also house instrumentation for Earth and Sea sciences for long-term and on-line monitoring of the deep-sea environment and the sea bottom at a depth of several kilometres. KM3NeT is designed such that all data is sent to shore and is processed there by an online computer cluster (“all-data-to-shore”).

The KM3NeT Collaboration has developed a data policy plan reflecting the research, educational and outreach goals of the facility. For a certain embargo time after data taking, the processing and exploitation of the data is granted to the collaboration members as a return for constructing, maintaining and operating the facility. During this phase, each collaboration member has full access rights to all data, software and know-how. The collaboration commits itself to process the data during the embargo phase so as to generate high-quality calibrated and reconstructed event data suited for a wider user community. This data will be made publicly available after the embargo time under an Open Science policy on web-based services and will not only allow the public to validate the scientific results presented by the collaboration but also allow for individual analyses.

The prompt dissemination of scientific or methodological results achieved during the embargo time is in the responsibility of the KM3NeT collaboration. The scientific responsibility and the publication rights for results derived from public data are with the scientists performing the corresponding analyses. The KM3NeT Collaboration offers analysis support to external analysers on their request, and after scrutinising the validity of the respective analyses. In this case, both the external scientists and the KM3NeT collaboration will author the resulting publications.

The KM3NeT RI is constructed in a phased and distributed approach, where the basic element is one *Building Block (BB)* that comprises a logical and technical sub-unit of the respective detectors. BBs have a spatial distribution that guarantees that the physics sensitivity grows approximately linearly with the BB size and thus partitioning is possible without penalty in the physics outcome. Table 1 details the phases of the KM3NeT project. Current construction sites are in the Mediterranean deep-sea South of Toulon (KM3NeT-Fr, ORCA) and East of Sicily (KM3NeT-It, ARCA).

Phase	Detector Layout	No. of DUs	Start of construction	Full Detector
Phase 2.0	2 BBs ARCA/ 1 BB ORCA	345 DUs	2016	2028/2028
Reference	1 BB	115 DUs		

Table 1 Phases of the KM3NeT project
(planning status: 07/2023)

This document details an intermediate version of the KM3NeT Data Management Plan (DMP). It contains information on the data to be collected, generated and processed, the applied standards and methodologies, the open science standards to be employed, and the implementation of data preservation and curation. The DMP content has been based on the "Guidelines onFAIR Data Management in Horizon 2020" [2] and was updated during the project whenever new decisions and implementations of the relevant points were addressed by the collaboration. It has been reworked during the INFRADEV2-project following a review by experts and will be updated again with the inclusion of changes requested in the review at the end of the project. New adaptations will be necessary when the KM3NeT legal entity is established.

One of the aims of the KM3NeT data management plan is to provide a sustainable basis for the development and utilisation of e-Infrastructure commons. KM3NeT will therefore contribute to the development of standards and services in the e-Infrastructures both in the specific field of research (astroparticle physics and neutrino physics) and in a wider context. The guidelines for these developments are recommendations by expert groups (ASPERA, e-IRG, RDA, W3C or similar[4]) and best-practice approaches. KM3NeT colleagues play active roles in various groups, like the Group of European Data Experts[5] and the EOSC Architecture Working Group[6], and have contributed in consortia like the ASTERICS H2020 project[7] and ESCAPE H2020 project[8], becoming the ESCAPE open collaboration. In addition, KM3NeT has established contacts with European-wide e-Infrastructure providers, like EGI[9] and GEANT[10]. The collaboration also studies the possibility of cloud computing, especially in the context of the EOSC[11].

Machine learning techniques and the associated computing model are more and more employed in the scientific community. Within the KM3NeT collaboration, there are efforts on employing machine learning techniques to classify different types of neutrino interactions and to reconstruct their characteristics, as well as generative approaches to simulation of detector response. This work is not covered in the current document but will in all probability become part of an updated version.

[4] ASPERA: https://www.aspera-eu.org/, e-IRG: https://e-irg.eu/, Research Data Alliance: https://www.rd-alliance.org/, W3C: https://www.w3.org/

[5] GEDE-RDA: https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda

[6] EAWG: https://www.eoscsecretariat.eu/working-groups/architecture-working-group

[7] See also KM3NeT-INFRADEV deliverable D4.2 at [5] for the liaison of KM3NeT and the ASTERICS project

[8] https://projectescape.eu

[9] EGI: https://www.egi.eu

[10] GEANT: https://geant.org

[11] European Open Science Cloud: https://eoscpilot.eu

2. KM3NeT data

2.1. KM3NeT data characteristics

The purpose of the data, collected with the RI and generated in simulations, is fundamental research in the fields of particle physics, astrophysics and Earth and Sea sciences. The detector, currently under construction, has a modular design, and data taking is ongoing with the detection units that are deployed. A detection unit (DU) is a string containing 18 detection modules that contain the primary sensors. The relevant data encompasses data products at different abstraction levels. In order to provide science-ready data, data processing in the following steps is required (see Table 2 for a summary):

- data acquisition and simulation
- calibration
- quasi-online reconstruction
- offline reconstruction
- physics analysis

The KM3NeT computing model is based on the LHC computing model [3]. The general data processing concept consists of a layered system, commonly referred to as Tier structure, shown in Figure 1.

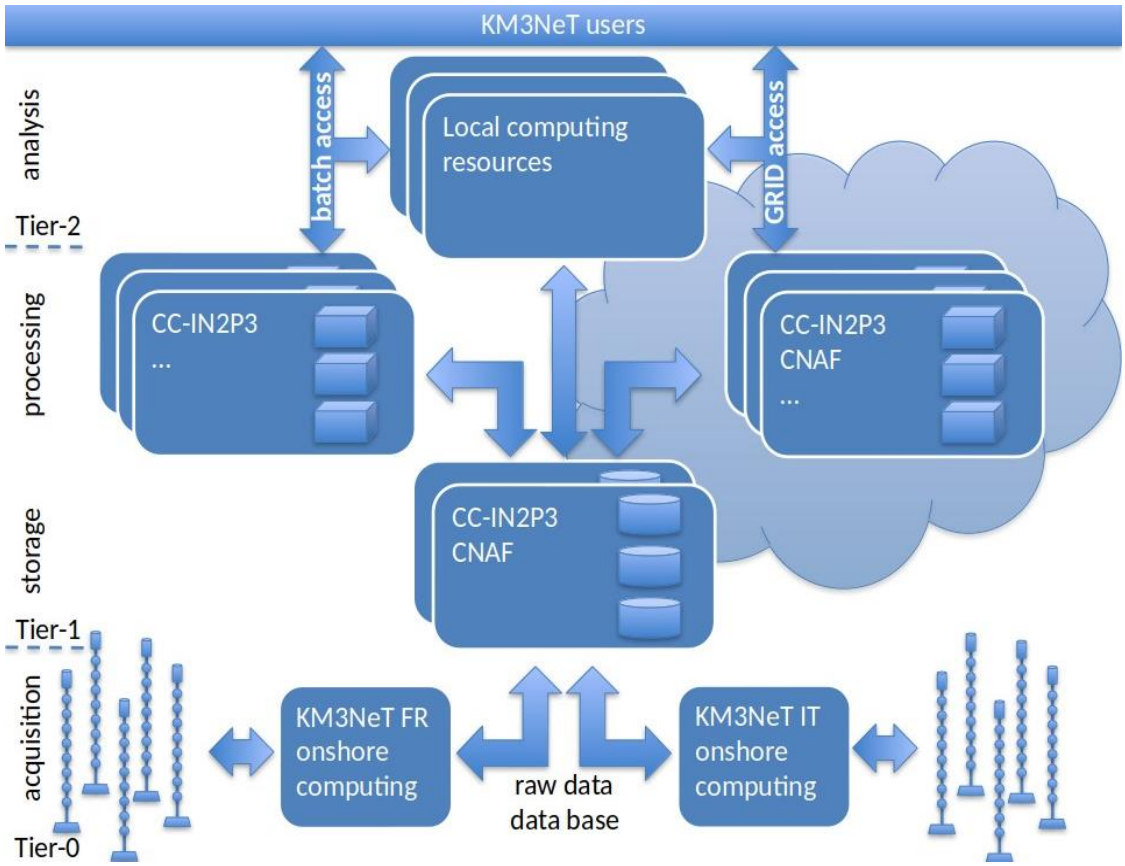

Figure 1 The KM3NeT Tier structure

2.2. Tier approach

KM3NeT adopts a general Tier approach with 3 levels (Tier-0, 1 and 2) that is described in the following. Due to adopted techniques for flexible software and data distribution, the classification is not strictly followed in day-to-day operation, where tasks of the different tiers can be mixed between computing resources. For a clear and concise description, it was opted to assign all computing tasks to fixed Tier levels, nonetheless.

At the Tier-0 level, all data that are detected by the primary sensors – the photo-multiplier tubes– are sent to the shore stations. These data are referred to as Level 0 (L0) data. The L0 data are filtered by processes in the on-shore data acquisition system. The shore stations thus operate as the Tier-0 sites. The computing power and storage resources at the shore stations are relatively modest compared to the LHC Tier-0 sites: the output data rate is about 100 Mb/s per DU, and will eventually amount to about 10 Gb/s per BB. Currently, at about 10% of the final detector configuration, each shore station is equipped with (currently) 2 nodes[12] dedicated to Level-0 data handling and filtering, and 5 TB of disk storage. The required number of CPU cores for real time processing is provided in [1], and amounts to 50 per shore station for the complete infrastructure. The output of the Level-0 filtering is referred to as the raw data, and is daily copied from the shore stations to permanent storage at two Tier-1 computer sites. After verification of successful data transfer to safe storage, the raw data is deleted from the storage in the shore stations. Currently, at about 10% of the final detector configuration, the shore stations can accommodate the raw data covering about 100 days. The shore stations have a dedicated network link of a few kilometres to the general internet. The bandwidth of these links is currently 1Gb/s. This is the limiting bandwidth in the network from the shore stations to the computer centres, but is sufficient for the current detector configuration. The bandwidth of these dedicated links are foreseen to be upgraded to 10 Gb/s in the coming period.
For fast reaction to and sending out of alerts to the community (cf. Section Online data processing), data are processed (direction and energy reconstruction as well as event type classification) in real time at the Tier-0 level already.

At the Tier-1 level, the calibration sets are generated, and the raw data are further processed. This processing step is commonly referred to as reconstruction. Both the calibration and data processing are performed at Tier-1 sites. The inputs for the calibration steps are partly retrieved from a central Oracle database system, and partly from the raw data. The output of the calibration is, as a temporary solution, stored in an archive at the KM3NeT GitLab instance, and is used as input for the processing of the data. During data processing, the raw data are corrected for the calibration, and various models are fit to the data to determine the energy and direction of the neutrino candidates. Also, at the Tier-1 level, Monte Carlo simulations are made to assess the detector efficiency and systematics. The simulated data are generated and processed alongside the processing of the raw data. The results of the processing of the raw data as well as the simulated data are stored on storage elements at the Tier-1 sites.

At the Tier-2 level, the processed data are input to the higher-level analyses. This primarily takes place at the local computing clusters of the KM3NeT partner institutes.

[12] 256GB RAM, 32Core-64 Threads (2 physical CPUs, each 16core/32T)

2.3. Data levels and formats

Data Level	Tier	Computing Facility	Processing steps, Access
DL 0	Tier-0	at shore station computer farm	Detector response and status, event simulation Internal
DL 1	Tier-1	computing centres (HPC, HTC)	Data calibration Internal
DL 2	Tier-1	computing centres (HPC, HTC)	Full reconstruction Internal/Open
DL 3	Tier-2	local computing clusters, personal computers	Science-ready data Open
DL 4	Tier-2	local computing clusters, personal computers	Binned science data Open
DL 5	Tier-2	local computing clusters, personal computers	Advanced science data Open
DL 6	Tier-2	local computing clusters, personal computers	High-level catalogue data Open

Table 2 Data processing steps and locations, user access

For the further description of data models and formats used in KM3NeT, the concept of data levels (DL) is used, which refers to the different processing stages of the data products, as listed in Table 2. As different processing steps are related to their respective computing environment, one can map these data levels roughly to the tiers.

Data Level 0 covers the genesis of the data at the detector site (or the respective simulation) and together with the initial processing steps through calibration (DL1) and full reconstruction (DL2) is detailed in the Chapter 3.2 Data types.

From this data level onwards, data level description follows the nomenclature introduced by the VODF initiative[13]. The full data sets from DL2 have undergone a selection governed by a quality assessment procedure, and a selection reducing background events. For simulation, the detector response is provided as instrument response functions to facilitate the assessment of the experimental data. More details on the data entities are available in the data format description[14].

Also, the online data processing, described in Chapter 3.6 Online data processing, can be described along the data levels. An overview over the main data entities per data level and the main processing steps is given in Figure 2.

[13] The VODF format, ICRC 2023 (see supplementary material)

[14] See “Data levels and Data models in KM3NeT” in the supplementary material

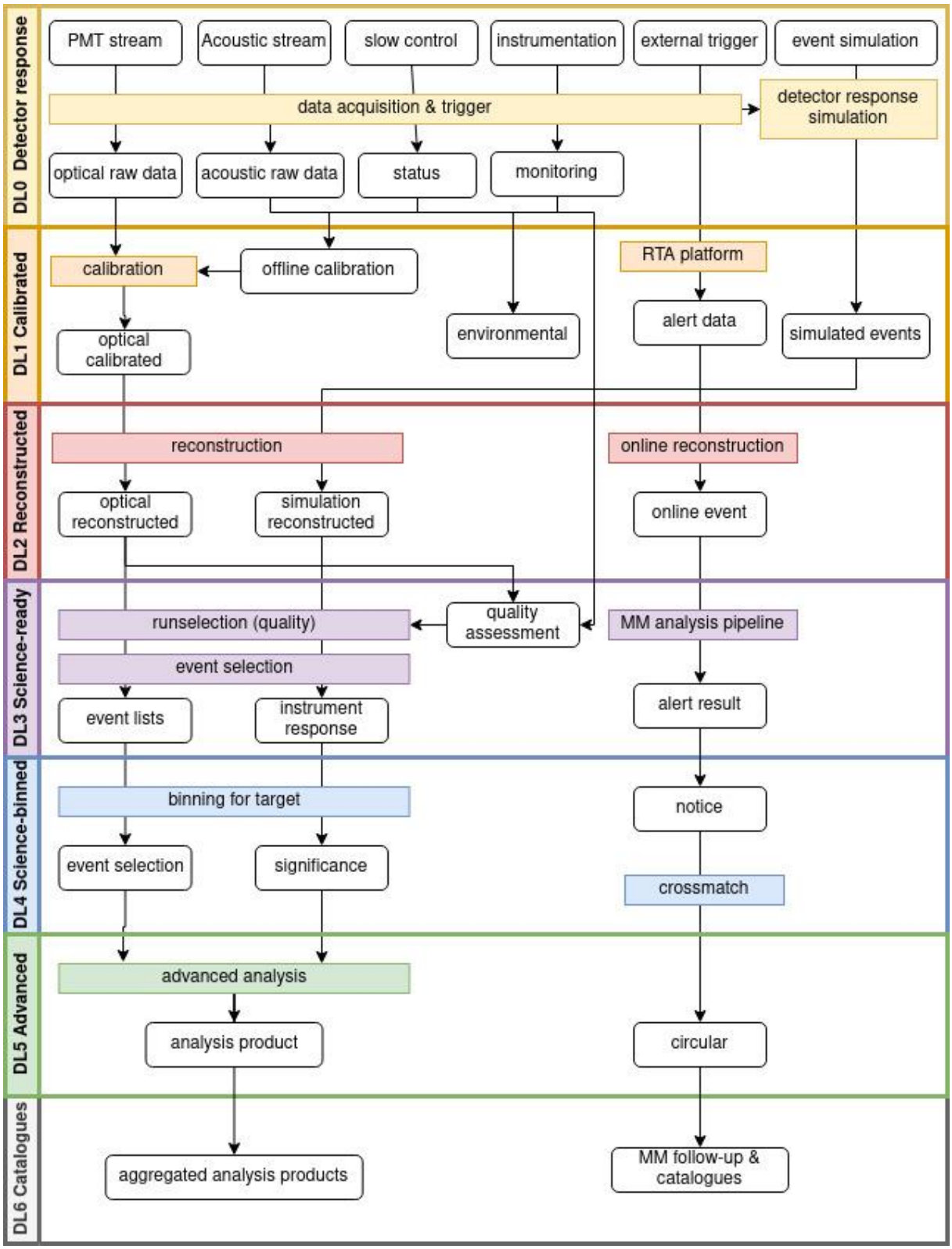

Figure 2 Overview over data entities per data levels

3. Data storage and processing

3.1. Data types

Experimental data

All data from the primary sensors arrive from the deep sea at the shore stations where they are processed real time. This Tier-0 processing is continuous (24/7), and is organised in data taking runs which cover a few hours (typically three hours). The Tier-0 data processing reduces the data stream by a factor of about 10^4. The PMT data are converted to events that are stored in ROOT files[15]. Each ROOT file contains the events from a single data taking run (15%), data for PMT calibration purposes (25%), summary data for realistic simulation of the environmental conditions (24%), and so-called "supernova data", for studying low-energy neutrinos (35%). Experimental data also include the output of an online reconstruction system, which operates as a low-latency alerting system for the multi-messenger community. The online reconstruction data make up about 5% of the experimental data, and are stored in ROOT files and JSON files[16]. The output files from the ORCA (ARCA) detector are temporarily stored on the local disk storage system in the ORCA (ARCA) shore station. Each night, all data files – from ORCA as well as ARCA– are copied to two computer centres: CC-IN2P3[17] and CNAF[18]. Data that are copied to CC-IN2P3 are ingested into the iRODS system[19] via an iRODS client that runs at both shore stations. Once the data are ingested in iRODS they are staged to disk, after which they are copied to tape on an HPSS within a day. Data is copied from the shore stations to CNAF with the WebDAV protocol via the CNAF StoRM WebDAV service. A checksum error is applied (sha256) after the copy procedure to verify a successful copy. Only when the data copy was successful, the data is deleted from the disks at the shore station.

The output of the full detector of KM3NeT Phase 2.0 will eventually amount to 5 MB/s per BB that needs to be accessible on permanent storage. For the three building blocks in Phase 2.0, this amounts to about 520 TB per year of raw data on permanent storage.

The raw data are processed for the higher level analyses. The output of this processing is also part of the experimental data. This output is, independent of the Tier-1 site where the data are processed, ingested into iRODS at CC-IN2P3 on the HPSS system. The processed data are available on storage for the analyses, and consist of calibration and reconstruction information, as well as metadata for reproducibility. The data processing reduces the raw data by a factor of about 2.5, and adds to the experimental data. As improved calibration sets and/or improved reconstruction algorithms are likely to become available, the data will undergo more than one data processing pass. The data from multiple passes will remain available for the analyses. The overall rate of fully calibrated processed data amounts to less than 2 MB/s per BB for one pass.

[15] https://root.cern/

[16] https://www.json.org

[17] CC-IN2P3: https://cc.in2p3.fr/

[18] CNAF: https://www.cnaf.infn.it/en/

[19] https://irods.org/

Simulation data

To assess the detector efficiency and systematics, Monte Carlo (MC) simulations are processed. Standard MC simulations are performed alongside the processing of the experimental data, such that the actual background rates and other environmental conditions are taken into account. This allows for a realistic representation of the environmental and running conditions during each run, and provides a time-dependent simulation of the detector response. The MC data are generated at the detector level and are subjected to the same processing as the PMT data. The simulated data are also stored in ROOT files, and are stored in the same event data format as the measured experimental data.

As the data processing of the experimental data and the MC simulations are integrated, the generation and the processing of the MC data also takes place at the Tier-1 sites. The simulated data are stored on the permanent storage via iRODS on HPSS at CC-IN2P3. Also the output of the simulation processing at Tier-1 sites other than CC-IN2P3 are manually ingested into iRODS at CC-IN2P3.

The simulation data that are generated for each data taking run, in Table 3 indicated as standard, are the basic simulations, and exceed the data volume of the experimental data by a factor of 1.5. In addition, custom simulation data are generated that will significantly exceed the measured data in volume since large statistics are required for precise analyses. In Table 3, these simulation data are indicated as custom. A special case of MC simulation consists of cosmic ray interactions in the atmosphere above the detector[20]. In addition to the detector systematics, systematics in the cosmic ray flux, the atmosphere and the interactions are also studied. The cost-benefit of these studies will still need to be evaluated and in-kind contributions are sought specifically for this purpose. The custom simulations are not structurally carried out in this phase. KM3NeT is currently in the process of organising computing in general. In the computing strategy that is being developed, the different physics groups will submit a specific request for custom simulations, which can then be considered for resource allocation and scheduling. The implementation of this strategy will provide an estimate for the required resources for custom simulations.

Acoustic data

The acoustic data arrive from the deep sea at the shore stations where they are processed in real time. Like the data from the primary sensors, these data arrive at the shore stations in a continuous data stream (24/7). The amount of acoustic data that is produced depends on the number of acoustic sensors in the detector, the number of acoustic emitters, and the frequency at which they ping. The number of acoustic emitters that will be deployed, and their operational settings, can vary. These specifications will be clarified by the calibration working group.

The produced acoustic data are currently stored in the central Oracle database at CC-IN2P3, however a new mechanism is being implemented that writes these data to files. After each sea operation where new DUs are deployed, these data are used to perform a pre-calibration of the new detector geometry. The results of this pre-calibration are currently stored in a central GitLab archive, hosted by ECAP[21]. This archive is a temporary solution and will, in the transition to distributed processing, be replaced by the Rucio, a data management tool (see section Transition to distributed data processing). The GitLab archive is expected to be easily portable to Rucio. In the meantime, essential features that are part of Rucio can now be provided by the GitLab archive. For example, it allows for efficient data processing as it is accessible from the different Tier-1 sites. It also offers the possibility to provide files in the

[20] https://www.iap.kit.edu/corsika/

[21] https://ecap.nat.fau.de/

required format that directly serves as input for the data processing. In terms of efficiency, this is preferred over querying a relational database, since the data processing is file-based. During the processing of the experimental data, these data are retrieved to perform the real-time position and orientation calibration. Like Rucio, the GitLab archive allows for a sophisticated versioning mechanism for easy reproducibility and traceability. In addition, it allows for an efficient storage of the calibration data in case the calibration does not change significantly over different runs.

In addition to filtered acoustic data, occasionally raw acoustic data are stored for marine and sea sciences. This will be done for only a selection of modules each time, as the output rate of acoustic data is significant. These data are copied from the shore stations to two computer centres, on permanent storage: CC-IN2P3 and CNAF, in the same way as the raw experimental data are copied. The strategy for these raw acoustic data dumps has not been determined yet, and hence the volume of these raw acoustic data are not included in Table 3.

Condition data

The condition data consist of logging information, slow control information, and data from secondary detector sensors and calibration devices. After a data taking run has ended, currently every three hours, these data are ingested from the shore stations into the central Oracle database in CC-IN2P3. These data are retrieved from the database at different stages and for different purposes. The logging data is occasionally consulted for debugging purposes. The slow control data is retrieved for offline monitoring. The data from the calibration devices are retrieved for generating the calibration sets for the data processing of the experimental data.

The retrieval of the condition data from the database is done with data stream services based on SQL commands. Slow control data and data from secondary detector sensors are the input for the monitoring tools. The calibration data are, prior to the data processing of the raw data, transferred to the GitLab archive in a format that is required for the data processing. The data volume of the condition data amounts to about 200 GB per year per DU. This data volume scales linearly with the number of DUs, and will thus amount to about 25 TB per year for ORCA, and 50 TB per year for ARCA.

3.2. Data storage at the Tier-1 level

For the data storage at the Tier-1 level, distributed storage elements at different sites are used, as indicated in Figure 1. These sites are mainly computer centres in the countries of the different KM3NeT partners (see Section 3.5). All of these storage sites offer disk storage, some offer tape storage as well. As part of the transition to the grid (see Section 3.5) a strategy is being developed for automatic and distributed data processing, which will, depending on data lifetimes and the data processing sites, lead to a compliant data strategy in terms of organisation of the data in data sets, and on which storage element they will be stored.

The data at the Tier-1 level can roughly be divided into four types: experimental data, acoustic data, simulation data, and condition data. The experimental data correspond to the raw data, as well as the processed raw data. The acoustic data, output by the data acquisition system, come from a separate data channel, and are stored separately from the experimental data. The acoustic data are used for calibration purposes. The simulation data are generated alongside the processing of the experimental data, at the Tier-1 sites. These data are required to assess the detector efficiency and systematics. The condition data are generated at all stages: during detector construction, detector operation, data

processing, and data analyses. These are needed throughout the lifetime of the experiment and include information on detector construction, calibration, logging, configuration, and environmental and run conditions. These data are used in all layers of the Tier structure.

The storage consists of tape systems, disk systems, and the storage system of an Oracle database. The Oracle database is hosted by CC-IN2P3. CC-IN2P3 provides a daily backup of the database which includes DataGuard and a standby database. A replica exists at the KM3NeT partner institute at Salerno (IT) that works on the file exports from the CC-IN2P3 instance, instead of using direct database links. Table 3 indicates an estimate of the data volumes for each of the aforementioned data types, per DU per year. These numbers are based on the results obtained with a 10% detector size of the complete infrastructure. As completed DUs are deployed and become operational, an estimate of the data volumes can be made for future detector configurations by scaling with the corresponding number of operational DUs. This is shown in Figure 3 and Figure 4, for ORCA and ARCA respectively.

data type		**data volume in TB per DU per year**	
experimental data	raw	1.50	disk tape
	processed per pass	0.60	
simulation data	standard	2.25	disk tape
	custom	4.50	
acoustic data	filtered	0.05	Oracle database
condition data		0.20	Oracle database

Table 3 The data volumes of the different KM3NeT data types at the Tier-1 level per DU per year

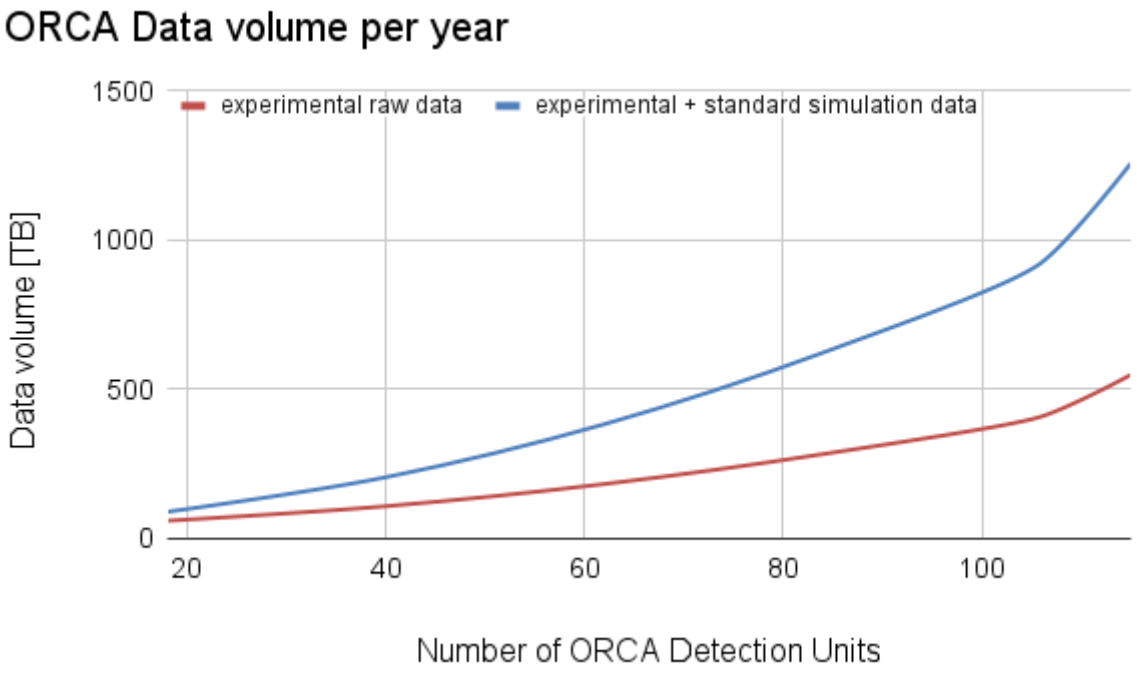

Figure 3: The ORCA data volume per year as a function of operational detection units. The red line indicates the raw experimental data, and the blue line indicates the sum of the raw experimental data and the processed and simulation data for one pass.

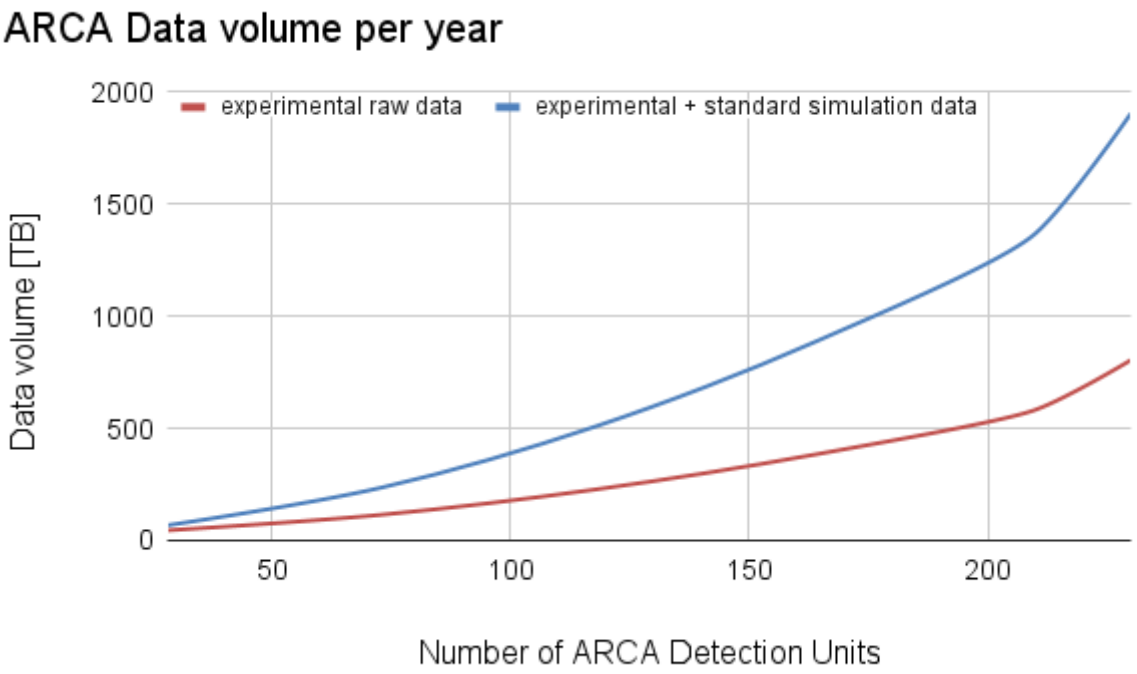

Figure 4: The ARCA data volume per year as a function of operational detection units. The red line indicates the raw experimental data, and the blue line indicates the sum of the raw experimental data and the processed and simulation data for one pass.

Not included in Figure 3 and Figure 4 are the data that are stored in the database, the custom simulation data, and the ANTARES legacy data. Although the data volume of the database is modest compared to the raw data volume, the calibration data that are currently stored in the database will eventually be written to disk, to reduce the load on the database. This includes the acoustic data and part of the condition data. The needs for the custom simulation data in terms of storage are only roughly estimated, and need to be evaluated in more detail (See section “Simulation data”). The ANTARES legacy data are preserved for future internal and external use and are now under the care of the KM3NeT collaboration. The data volume of the ANTARES legacy data amounts to O(50 TB) of experimental raw data, and O(1 PB) of processed and simulation data.

Not included in Table 3 are the data for the purpose of research by associated KM3NeT partners, such as data from bio cameras, seismometers, and radioactivity metres. These data arrive at the shore stations where they are directly separated from the Tier-0 data processing system, and passed to the equipment of the associated partners. No KM3NeT compute or storage resources are used for this data.

3.3. Processing

The Tier-1 compute resources required for KM3NeT are primarily used for data processing. For a detector size corresponding to about 5% of a BB, the full processing of one ORCA (ARCA) data taking run, covering three hours, requires about 200 (125) CPU hours, single core. This processing includes the processing of the raw data, and the complete generation and data processing of the standard simulated data covering that run. An indication for the generation of the calibration is provided in Appendix A. The quoted processing time does not include the generation of the custom simulations.

The processing time of ORCA data is currently dominated by the algorithm reconstructing a shower-like signal in the detector. This is one of four reconstruction algorithms, and takes up about 75% of the processing time. For ARCA data, the processing time is currently dominated by the generation of the neutrinos and the simulation of their interactions. This takes up about 75% of the processing time.

The required processing time for a single data taking run will scale linearly with the event rate in the detector and the size of the events. For the mentioned detector size, the event rate and event size correspond to about 10 Hz and 5 kB, respectively. For a complete BB, consisting of 115 DUs, the event rate is expected to be 100 Hz, and the event size is expected to increase to about 25 kB.

3.4. Data lifetimes

Ultimately, the data processing procedures are expected to operate with a fixed latency of a few hours or less. This includes the generation and processing of the simulated data, which is in KM3NeT done in parallel to the processing of the raw data. The inputs to the data processing are the calibration sets which are the output of a sequence of calibration procedures. The time it takes to perform each calibration step is shorter than the coverage of the data to which the specific calibration applies. All calibration sets are stored in files that are, as a temporary solution, pushed to the GitLab archive. For a full description of the calibration procedures, see Appendix A.

During the construction and commissioning phase of the experiment, the processing of the data is organised in batches of multiple data taking runs (*O*(3000)). The mass processing of these data is foreseen to be initiated twice per year. This is done to be able to regularly take into account improved versions of the software and calibration that will become available relatively frequently during these phases. It will also allow for fixing issues that are detected through data taking and low-level analysis, to transition to stable and high quality data taking.

3.5. Transition to distributed data processing

In view of the expanding detector over the next few years, the collaboration is moving to distributed data processing in order to be able to continue to process the foreseen influx of data. This includes the transition to distributed data. Recently the collaboration has implemented all data processing workflows in the Snakemake[22] workflow management system. This makes the workflows easily portable to different computer clusters. In addition to the two sites from which the data are available, CC-IN2P3 and CNAF, data processing can now also take place at the local clusters of KM3NeT partner institutes like ECAP, Nikhef, and Demokritos. The data are first copied by the workflow from the storage to the cluster, unless the workflow is running at CNAF (from where the data are directly accessible). For all other Tier-1 sites, the data are retrieved from HPSS at CC-IN2P3 via iRODS. After data processing, the data products are ingested into the central storage at CC-IN2P3 via iRODS, to make them available to the collaboration. Several local clusters hosted by KM3NeT partners, such as ECAP, Viper[23] and Demokritos[24] will remain available to KM3NeT for local batch processing.

The next step is the transition to the Grid, to ensure the availability of sufficient computing power, and to reduce dependence on specific sites and thus provide more flexibility. A dedicated VO km3net.org exists to authenticate for Grid computing. Within the INFRADEV2 project, the collaboration has started to use the Dirac[25] interware that offers the possibility for the extension to Grid resources. A multi-VO instance of Dirac, hosted by EGI, is available to KM3NeT. This service has been available since 2014, and is planned to remain operational for a long time, thus providing the availability to KM3NeT for the longer term. Several KM3NeT partners have allocated grid resources to KM3NeT, compute as well as storage, and include CC-IN2P3, CNAF, ReCaS Naples, SURF, Nikhef and CPPM. The transition to distributed data processing goes hand in hand with the transition to distributed data. Within the ESCAPE project, the use of the Rucio[26] data management system was investigated for KM3NeT, and also demonstrated on the basis of a number of use cases. After this experience, KM3NeT is now setting up a KM3NeT Rucio instance, which is carried out within a funded project of the Dutch eScience organisation. With the availability of Rucio, the data will be available in a fully automatic way from different data processing sites. In combination with Dirac, it will also allow the execution of the data processing closer to the physical location of the data. The file based GitLab archive, that is now temporarily used for the storage of the calibration sets, will also be transferred to Rucio.

[22] https://snakemake.readthedocs.io/en/stable/

[23] https://hpc.wordpress.hull.ac.uk/

[24] https://www.demokritos.gr/

[25] https://dirac.readthedocs.io/en/latest/index.html

[26] https://rucio.cern.ch/

As an user identity service, KM3NeT will use INDIGO IAM[27] that is hosted by CNAF. This service will manage the access to the grid tools such as Dirac and Rucio, and also acts as token issuer, to cover the upcoming transition to tokens.

For the data transfer between the different storage elements, KM3NeT is using an FTS service hosted by Cern. In the coming period KM3NeT will join the LHCONE[28] community, and make use of the LHCONE network to assure a fast and reliable network at the Tier-1 level.

The KM3NeT software is available as Apptainer[29] containers and distributed via a CVMFS[30] server. These are currently used by the Snakemake workflows, and will also be used when running the workflows in a Grid context.

3.6. Online data processing

With the aim of participating in multi-messenger astronomy observations, KM3NeT should be able to both receive external alerts from other observatories and to trigger its own alerts. In order to perform astronomy in real time, the KM3NeT data have to be handled online, with latencies of a few minutes, demanding a high level of automatization[31].

Analysis Online Pipeline

To achieve this, the KM3NeT Collaboration has developed an online analysis platform that processes all the KM3NeT events in real-time, and monitor the external alerts[32]. Following the data acquisition, the online analysis platform must be able to reconstruct the events, using the standard KM3NeT reconstruction algorithms (with an online default calibration), classify the events according to their topology (track or shower) with a rate of about ~10 Hz for the current detectors and ~100-200 Hz for the full detectors, for both ARCA and ORCA, and distinguish between neutrinos and atmospheric muons.

The outputs of the reconstruction are stored in both ROOT and JSON format (first stored in the shore stations, then copied to CC-Lyon for both detectors once per day, ARCA additionally saves them to CNAF). Currently, the amount of data produced for ARCA is: ~ 225 MB in ROOT files and ~120 MB JSON files. For ORCA there are ~ 400 MB ROOT files and ~360 MB JSON files. In a long-term plan, with the full detectors, the order of magnitude the data volume could reach 5 GB/day (very preliminary estimation).

The KM3NeT Online Analysis Platform (KOAP) fulfils two purposes: After filtering the external alerts, the KOAP performs a quick search for coincidences in time/space around these alerts, and reports whether or not a neutrino counterpart has been detected. The time windows used to search for coincidences range from a few seconds to days, depending on the type of alert and/or targeted source.

[27] https://indigo-iam.github.io/v/v1.8.2/

[28] LHCONE

[29] https://apptainer.org/

[30] https://cernvm.cern.ch

[31] For further details, see the “Addendum to the Online Alert System”, supplementary material

[32] RTA Software Description, see supplementary material

When notable events are detected in KM3NeT, the KOAP should trigger the sending alert system. Those steps are illustrated in Figure 5.

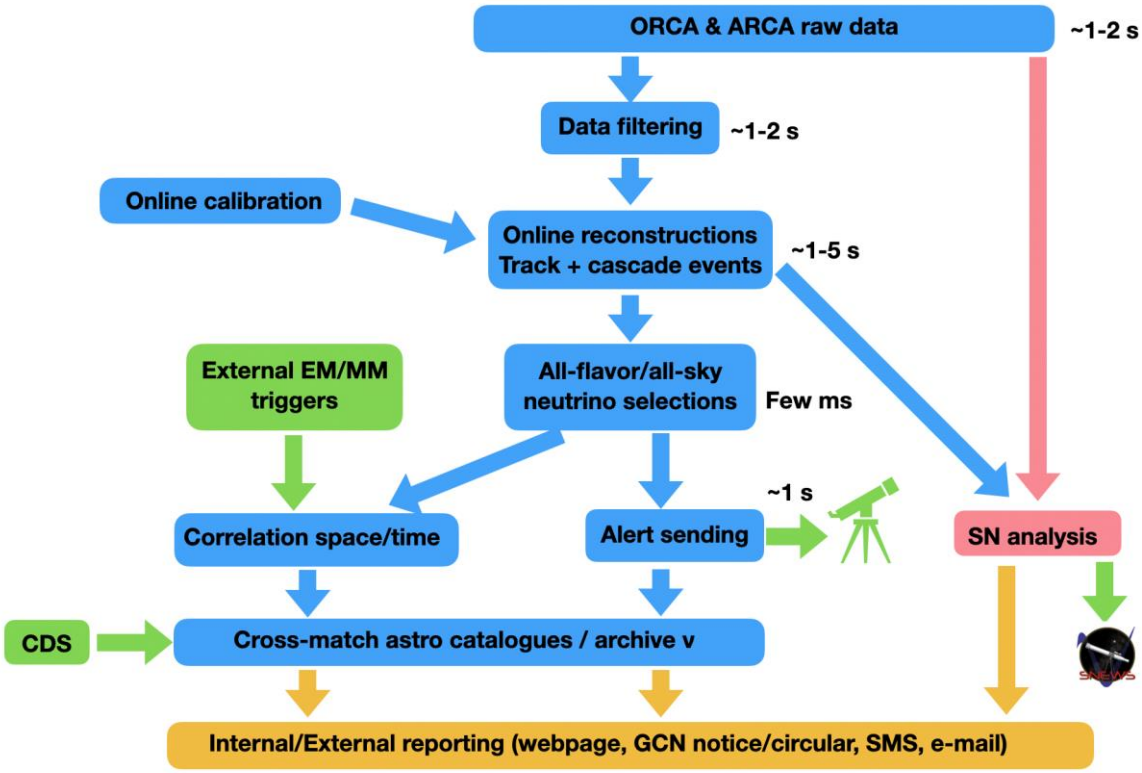

Figure 5 The KM3NeT Online Analysis Pipeline

The implementation of such a kind of data processing follows a multi-stage roadmap. The first phase consists of being ready for external trigger follow-ups. This step already requires the ability to receive and filter the external alerts, to collect the corresponding KM3NeT data and to run an online analysis. The second step will be to send an alert if KM3NeT detects an event likely to come from an astrophysical source of interest.

The development of the KOAP has gathered about ten full-time workers for nearly two years. Its maintenance and future upgrade would require at least the same amount of personnel. However, to cover the wide scope of application offered by the online analysis, supplementary human resources could be used.
As for the technical resources, the shore stations are currently using the equivalent of one machine (32 core, 128 GB of RAM and 4 To of storage) per junction box for real-time data processing. The analysis would require two additional machines, thus the order of magnitude of the needs for the two shore-stations is somewhere between 10 and 15 machine equivalents to the one described.

External alerts handling

The external alerts are received as Notices mainly distributed through the GCN (General Coordinate Network)[33] broker. GCN Notices are real-time, machine-readable alerts that are submitted by

[33] https://gcn.gsfc.nasa.gov/

participating facilities and redistributed publicly. Currently, the notices are received with the VOEvent format[34] through the GCN classic broker. However, GCN is transitioning to a more modern system. This modernization started by distributing the notices with a cluster of Kafka brokers[35] in the cloud and ultimately it will result in a unified schema of alerts with a JSON format. There will be several external alert brokers: microquasar (own KM3NeT development), FINK[36] for the ZTF/LSST transients[37], TNS for the fast radio bursts or other optical transients and fermi flare's advocate, among others.

The alert handling pipeline includes the following steps, which are independent of the notice format. First, we subscribe only to the alerts sent by the observatories of interest, via the GCN broker. Then, when an alert is received, it is filtered according to the role of the alert (test or observation), and to the nature and visibility of the source. As a result of this filter, the alert will get a label (EXCLUDED or VALID).

The content of each received alert is written to a file at the common analysis platform (temporarily stored in the antorcamm2 machine), and stored in a specific folder regarding the "excluded" or "valid" status. Tens of alerts are received per day, each resulting in a file of a few kB (<5 kB). Additionally, if the notice is valid a JSON message, containing the information required by the online analysis pipeline, is generated and added to the KM3NeT online analysis scheduler (a SQLite database used to orchestrate the analysis).

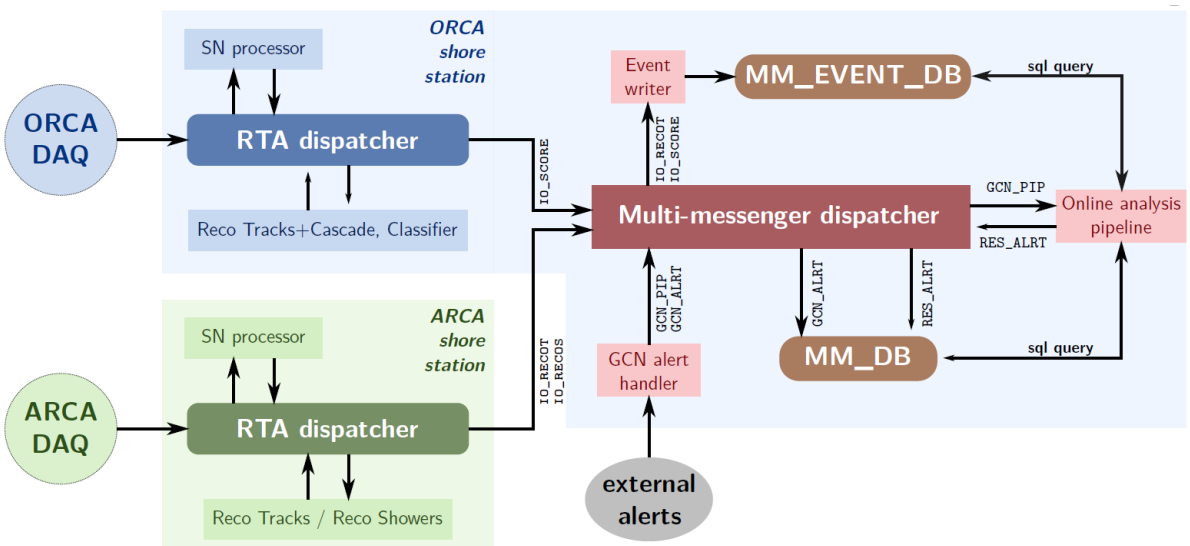

Figure 6 Online data processing and external alert handling

[34] https://www.ivoa.net/documents/VOEvent/
[35] https://kafka.apache.org/
[36] https://fink-broker.org/
[37] https://lasair-ztf.lsst.ac.uk

4. Data sharing

The dissemination of data is addressed by Open Science in KM3NeT. Following a dedicated policy, a system of platforms and servers is developed to grant access to KM3NeT data for scientific use. The following section gives an overview over the digital products that will be considered part of open science in KM3NeT, the interfaces and user access for the products. The system described below exists in a prototype version [4] and development of interfaces procedures and interfaces are addressed by the Open Science Committee in KM3NeT, which has been charged to implement the Open Science Policy.

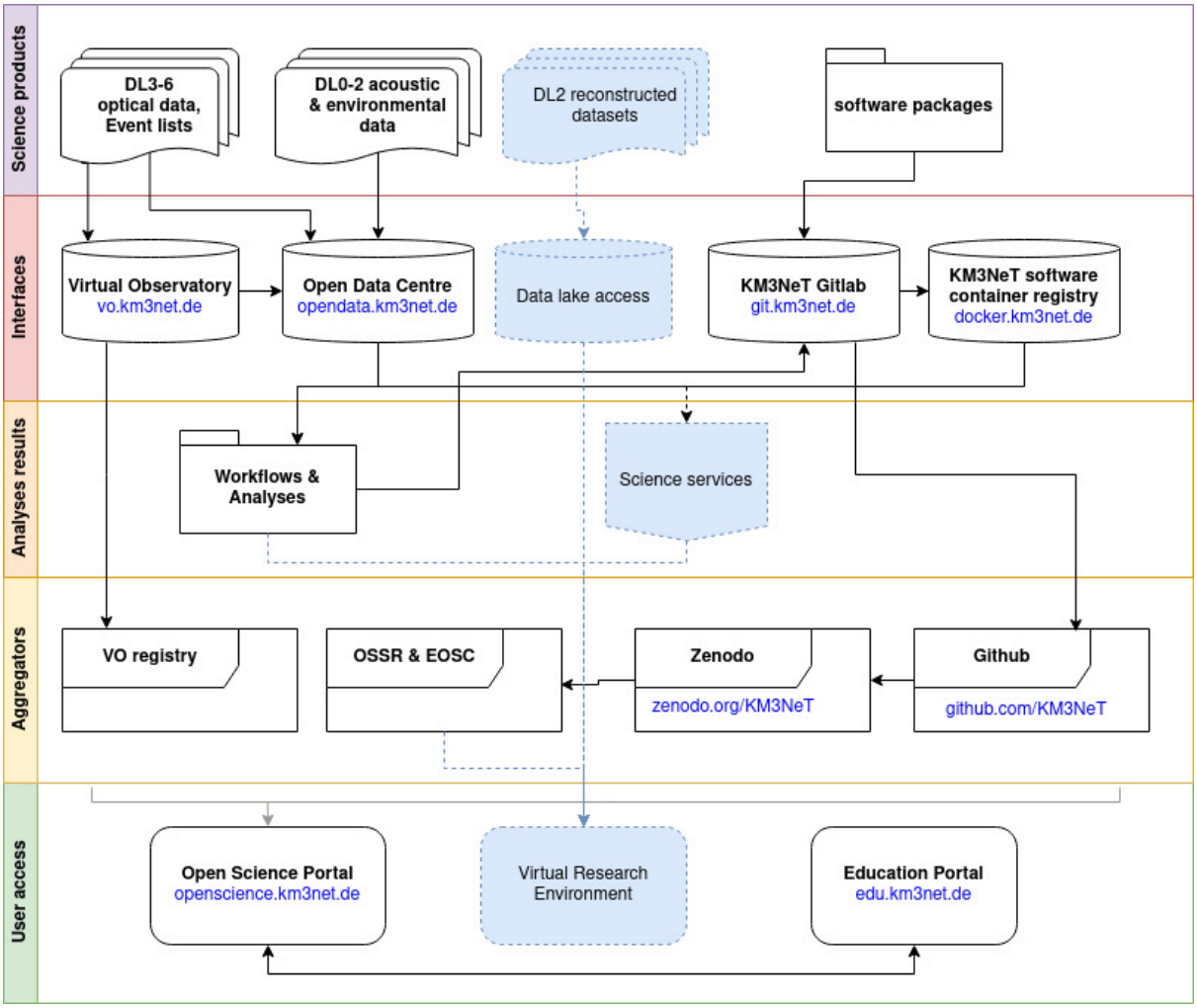

Figure 7 Components of the KM3NeT Open Science system, in blue the potential extensions

4.1. Products for Open Science

Data sets

Access to data regarding fairness is strongly correlated with the possibility to make large datasets and computing resources available for an outside user. Therefore, optical reconstructed data sets at DL2 form the basis for an event selection which leads to DL3 data products, which will be made public, see Science Products in Figure 4. However, using the DL2 data necessitates an authenticated access scheme which might develop in the context of the EOSC, but which cannot be provided by the KM3NeT collaboration alone. For the time being, DL3 data and beyond can therefore be fully treated as having to follow the FAIR principles, while at DL2 quality control and selection mechanisms have to be established to lead to well-defined and usable derived data sets in the following levels.

Name	Data model/File format	Volume/Frequency (per year)	Storage/Access
DL0-2 Environmental data	various	various	ODC (link)
DL3 Event lists astrophysics candidates	Tables (VODF event list) /FITS	O(1-10MB) / several	VO
DL3 Event lists neutrino samples	Tables (KM3NeT DST) /HDF5	O(10-100MB) / several	ODC
DL3 Instrument Response	VODF IRFs / FITS	O(MB), according to event lists / several	ODC/service
DL4 Event selection	VODF & others	O(10-100MB) / per analysis	ODC (link)
DL4 Significance	VODF & others	O(1-10MB) / per analysis	ODC (link)
DL5 Analysis product	VODF & others	O(1-10MB) / per analysis	ODC (link)
DL6 Aggregated analysis product	VODF & others	O(10-100MB)	ODC (link)

Table 4 Data intended for publication

In Table 4, an overview is given over the expected volume of public data excluding full DL2 data sets. Here, acoustic and environmental data from DL0-2 is not further detailed as data formats and providing infrastructure might vary widely in the future, depending on the community needs that cannot be foreseen at the moment. However, the data will be searchable through the Open Data Center, at least through linkage.

Event lists at DL3 level will be made both accessible through the VO server and as multi-purpose data sets in a well described and widely used data format like HDF5[38], as the needs of communities like in neutrino research or machine learning are not as well defined, which makes aiming for generality a sensible approach.

[38] https://www.hdfgroup.org/solutions/hdf5

Last but not least, objects from DL4-DL6 will be further defined in cooperation with the VODF initiative. However, making the products also findable through the ODC is planned.

Open Software

KM3NeT is developing software as open source, and will offer packages supporting the reading and processing of open data. This software is stored at the KM3NeT Gitlab instance, and will also be made available through Github to enable community collaboration. Automatic build of software containers, provided through the KM3NeT docker registry[39], also enables easy providing and use of dedicated software environments.

Analysis workflows

Scientific work consists of analysis scripts which can require diverse processing steps. KM3NeT aims to offer a suitable environment in which the execution of complex workflows is possible. As a first approach, software containers with an installable software environment serve to enable the reproduction of full analysis workflows.

4.2. Interfaces and storage

The KM3NeT Open Science system contains several servers providing access to data sets and software. Here, the Open Data Center (ODC)[40] serves as a central broker to host metadata about available data sets even if not hosting the data itself. Community-specific interfaces and a software repository complement the various access options for KM3NeT data.

Open Data Center

The Open Data Center serves both as a data server as well as a centralised registry for all open science products by KM3NeT. Through a dedicated API, the ODC can be queried for related entries arranged as collections, and used to search and download public data sets either from the ODC directly or from linked resources. Metadata to the entries is developed in order to accommodate a large variety of file types and content.

Virtual Observatory server

The Virtual Observatory server[41], running the DaCHS software[42], hosts neutrino event lists and is planned to be used for datasets in the context of astrophysics. Currently, extensions of the Virtual Observatory interface for high-energy particle astronomy are discussed and will be used to provide the datasets to the VO through dedicated interfaces. Datasets are usually stored directly on the server.

[39] https://docker.km3net.de

[40] KM3NeT Open Data Center: opendata.km3net.de running on Django: https://www.djangoproject.com/

[41] KM3NeT Virtual Observatory: vo.km3net.de

[42] GAVO DaCHS: https://dachs-doc.readthedocs.io/index.html

Gitlab and Software container repository

KM3NeT uses a self-hosted GitLab instance as the main platform to develop and discuss software, analysis tools, papers and other private or collaborative creations. The continuous integration (CI) that is part of the GitLab distribution is utilised to generate consistently up-to-date test reports, documentation and software releases. Docker containers are stored at the KM3NeT docker registry. All KM3NeT software used by the data processing workflows is available as Apptainer containers and are published via the CVMFS server hosted by RAL[43]. The access to gitlab projects is public for those projects having been explicitly published, internal or private projects require access through the KM3NeT collaboration login.

Data lake access

In the context of the ESCAPE collaboration, access to the data lake infrastructure using Rucio and the ESCAPE AAI are being developed[44]. In the future, access through the ESCAPE data lake for KM3NeT might make it possible to also provide large data sets on the DL2 level.

4.3. Aggregators

External aggregators are used to increase the findability of KM3NeT data through major platforms or community-specific interfaces.

Mirrored repositories

With Github, a major platform exists for software development that also allows easy interaction between software developers on various projects. Open KM3NeT software is mirrored from the KM3NeT Gitlab instance to Github, making the software findable for software developers and open for community development. For grouping of the software, a KM3NeT collection has also been established here. This group also contains analysis-related repositories to provide executable scientific workflows.

Other interfaces might be used, depending on the type of software. For python-based software, easy installation via the pip package installer is integrated to 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.

Driven by the need to assign DOIs to digital products and ensure archiving of main research results, Zenodo was chosen as a well-established data repository in the physics community to serve these needs. Repositories from Github can be mirrored here for main releases. Also, Zenodo is used as a basis for exposing repositories to the OSSR software repository of ESCAPE[45]. Therefore, a community for KM3NeT was created at Zenodo for this purpose[46].

[43] https://www.gridpp.ac.uk/wiki/RAL_Tier1_CVMFS

[44] ESCAPE DIOS: https://projectescape.eu/services/data-infrastructure-open-science-dios

[45] https://projectescape.eu/services/open-source-scientific-software-and-service-repository-ossr

[46] https://zenodo.org/communities/km3net

ESCAPE repositories

Material that is provided in the ESCAPE context is made findable through the OSSR of the ESCAPE project. This interface is built on Zenodo and a dedicated API that also facilitates the integration of repositories, notebooks and software containers in the Virtual Research Environment of EOSC, see below. Therefore, open software and repositories are planned to be exposed additionally through the ESCAPE interfaces by registering them in the OSSR.

The VO registry

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[47] 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. The KM3NeT VO server is registered to the RofR as "KM3NeT Open Data Registry" under km3net.org.

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.

4.4. User environment

The current user environment is built on the assumption that the user can run all relevant software on their personal computer and store data locally. Therefore, the user platform is restricted to documentation on how to build a local analysis environment and use KM3NeT data. However, if the Virtual Research Environment of ESCAPE is adopted, the user could also create a personal workspace in a dedicated server environment in the VRE, reducing the preconditions for access to the use of a web browser.

The user group here can be assumed to be more diverse than only the researchers of the associated science community. Providing easy-to-follow science examples from open data can be helpful already for high school science projects, and can e.g. be offered through the education portal in KM3NeT. With increasing difficulty, this can be followed by material relevant for undergraduate projects and master classes up to graduate school environments or data challenges for the advanced science enthusiast. Using common data sets and infrastructure also intended for the use of researchers increases synergies while at the same time lowering the threshold for external scientists using open data by increasing the amount of available documentation and introduction material.

Open Science Portal

The Open Science Portal[48] serves as a documentation platform for Open Science in KM3NeT and an access point for information about open data and the Open Science System. Being built from Gitlab pages, relevant links to the other parts of the system as well as information about policy and standards used by KM3NeT can be found here.

[47] https://ivoa.net/

[48] Open Science Portal: openscience.km3net.de

Education Center

The KM3NeT Education Center[49] hosts courses for KM3NeT members with KM3NeT authentication as well as public courses that introduce the use of the open science interfaces. Here, Wordpress is used for content management, including extensions for online course management[50]. The courses can serve also as basic material for outreach and education workshops as well as starting point and introduction to the use of KM3NeT data for researchers.

Virtual Research Environment

The Virtual Research environment[51] (VRE) and the ESAP analysis framework[52] of the ESCAPE collaboration offer a self-hosted platform to easily integrate workflows, data lake access and software container use. Being currently under development in the EOSC-Future project[53], the KM3NeT collaboration will evaluate the introduction of the VRE or ESAP as a tool to facilitate access to KM3NeT open data.

5. Documentation and curation

5.1. Data models and metadata

Creating data models

Open Science products are collected in the ODC, which lies at the border of the internal and public data regime in KM3NeT. Data published via the KM3NeT Open Data Center (ODC) is annotated as KM3OpenResource, which includes basic metadata for resource content, accessibility and identification. As resources can be provided either as part of a collection, e.g. data set or multiple resources related to an analysis, or as part of a stream of similar objects, e.g. of alert data, resources are grouped in the server as KM3ResourceCollection or KM3ResourceStream to facilitate findability. For the various resources included here, the data models will be documented while they are included in the ODC. This documentation will also be made available through the Open Science Portal.

Metadata approach

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

[49] Education Center: edu.km3net.de

[50] Wordpress: https://wordpress.org/ and the LifterLMS plugin: https://lifterlms.com/docs

[51] VRE home: https://escape2020.pages.in2p3.fr/virtual-environment/home/

[52] ESAP home: https://sdc-dev.astron.nl/esap-gui/

[53] https://eoscfuture.eu

Naming and identifiers

Naming conventions are relevant both in early data processing as well as in the identification of data sets in the open data regime. Beyond this, the identification of individual particle events to facilitate cross matching through various data sets is standardised.

The convention of the file names at the Tier-1 data processing level has been regulated for easier findability. The file names indicate what the data products contain. They also indicate the abstraction level of the data products. The overall structure of the file name consists of the following categories: {coverage}.{contents}.{version}.{extension}[54].

An ordering schema for class definitions and content descriptors helps in the interpretation of a specific digital object. To this end, the ktype and kid have been introduced and are applied in the Open Data Center.

- The kid is a unique identifier which follows the uuid schema. The uuid is ideally assigned at the generation of the digital object where possible 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 and has been implemented for all open science products.

- The ktype serves as a content descriptor and is defined as a string with a controlled vocabulary of words separated by ".", starting with "km3.". 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}

At the data aggregation level, an identifier has to be introduced to uniquely identify a particle detection in one of the KM3NeT detectors. The internal KM3NeT event identifier is defined as km3.{detector_id}.{run_id}.{frame_index}.{trigger_counter} with the entries referring to the initial data taking of the event.

5.2. Quality assurance

The processes involved in the KM3NeT data processing chain can be grouped into a few main categories: data acquisition, detector calibration, event reconstruction, simulations and finally scientific analyses based on the data processed in the previous categories. Implementing a complete and consistent set of data quality control procedures includes the setting of data quality criteria which should be initiated at the highest level of the data processing chain, and propagated towards the lowest levels. For each of the aforementioned categories there exists a working group within the KM3NeT collaboration implementing the steps and establishing quality criteria.

For the final setting of quality flags for full runs, the data quality plan[55] has been established to identify problematic data taking periods and define quality criteria to declare data fit for scientific analysis depending on the intended use case.

[54] See "File naming convention in data processing context", supplementary material

[55] Data quality plan, see supplementary material

For the release of data and scripts in the context of analyses, i.e. from DL3 onwards, a review process has been established. For the release of multi-purpose data sets, this review process will still have to be established.

5.3. Development strategy

The Open Science System is currently under development, starting from the prototype that has been described so far. In order to develop the system further, steps are taken minding the following principles:

1. Working along use cases: the functionalities and interfaces are improved along the requirements of current analyses and latest data releases specifically targeted to advance the use of the system. In this context, the ANTARES collaboration has agreed to pass the legacy data from the experiment, which concluded in 2022, to KM3NeT to ensure future use and accessibility. It will therefore serve as example data to test and develop the KM3NeT data management and interfaces.

2. Standardisation in common working groups: For the development of common metadata and formats, KM3NeT is actively involved in initiatives like the VODF and ESCAPE collaboration, and participates in the formation of a VHE interest group in the IVOA.

3. Offer issues for open development: The services and data products are offered in such a way that issues can be raised on either github or another feedback channel to directly incorporate the user experience and requests from the community.

4. Benefit for internal processes in KM3NeT: As working with FAIR principles increases transparency and data use also in the collaboration, policies and standards are developed such that their adoption by the researcher early in the analysis process benefits the overall research setup in the collaboration.

5.4. Long-term storage and archiving

As can be seen in Table 1, the construction of the KM3NeT detector will continue up to 2028. After completion, it is foreseen to operate the detector for about 15 more years. Also, after data taking ends, the data need to be accessible for about another 10 years to finalise the analyses. The raw data, as well as all data in the Oracle database, will remain available at least for the lifetime of the experiment. In principle, the calibration sets, the data products from data processing, and the simulation data are reproducible. The fully calibrated data products are stored for further analysis and are made available internally to the scientists for detailed scientific analyses. Data are reprocessed when updated calibration sets or improved data processing algorithms become available.

In addition, data used in theses will be stored in the partner institutes under the national or local regulations – this is the responsibility of the respective KM3NeT institutes.

For data offered through the KM3NeT interfaces, the availability of entries will be assured by automated testing. In case of superseding versions, the old versions will be kept available in so far as they are relevant for the reproducibility of related science products. Where deletion of data sets is required due to e.g. storage capacities, higher-level or newer products will be linked from an archiving note to ensure the best possible replacement for the entry.

In Zenodo, long-term archiving is ensured as former versions of an entry are not deleted. Here, material that is intended for long-term preservation will be mirrored in a dedicated community.

For the development of an archiving strategy, KM3NeT currently deliberates ingesting the ANTARES legacy data[56] into the KM3NeT data system. This would preserve the research results of the ANTARES collaboration and could serve as a test case for archive access and integration of similar data sets in the KM3NeT Open Science system.

6. Data Governance

6.1. Data security

In general, data handled in the KM3NeT project does not contain sensitive data of any kind. Personal data and authentication are separated from the actual data and handled via central authentication and authorisation methods based on SAML[57] and Google OAuth 2.0[58]. Backup of scientific data is ensured to a feasible extent through the distributed nature of the tier-processing structure as outlined above, which allows recovery of all relevant data sets through the reproducibility measures employed.

6.2. Open Science Policy

The Open Science Committee was established to set into action the Open Science Policy of KM3NeT. It works in parallel to the Publication Committee and the Conference Committee and in close cooperation with the PC. It sets, develops and maintains procedures for open science, by implementing a publication procedure for data and software following the corresponding procedures for publications. This includes setting up requirements, defining standards of quality, and implementing review processes. The Open Science Committee develops guidelines for the publication of scientific products in the spirit of the Open Science Policy[59].

[56] The ANTARES collaboration: https://antares.in2p3.fr/data/

[57] https://en.wikipedia.org/wiki/SAML_2.0

[58] Google OAuth: https://developers.google.com/identity/protocols/oauth2

[59] Open Science Committee Guidelines: https://open-data.pages.km3net.de/osc-guidelines/

6.3. Licensing

The licensing has been defined in a dedicated licencing policy[60]. The lowest possible obstacle for data re-use will be implemented, while making sure that the quality of the data analysis can be controlled and the rights and obligations of all parties are granted and documented according to the relevant legal standards and best-practice procedures.

In short, a permissive licence with attribution, no share-alike and no warranty will be applied to all scientific open products of KM3NeT.

6.4. Embargo time and access

The first exploitation of the data is granted to the collaboration members as they build, maintain and operate the facility. It is also required to produce high-quality data for public dissemination. High-level data (event information including quality information) will be published as data releases after an embargo period (of two calendar years) after data taking, starting with the completion of the detectors. Currently, this targets multi-purpose data sets at the DL3 level. Data relating to specific analyses for DL3-6 can be published alongside a dedicated publication or as a specific science service and can therefore be exempt from the embargo at the consideration of the KM3NeT collaboration.

The KM3NeT collaboration aims at providing a unified approach to data access for all members. To this end, a single identity provider will be aimed for and chosen coherently with the developments for Grid authentication.

6.5. Publication and Authorship

Assigning DOIs

The VO does support the integration of a DOI in their resource metadata, but does not provide an 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. While this issue is not yet resolved and will be investigated further also in the context of the EOSC, currently, Zenodo was chosen as a hosting repository to assign DOIs.

Authorship

Authorship rules are part of the publication strategy in KM3NeT and currently under development in the context of the establishment of a legal entity.

[60] Licencing policy for KM3NeT, https://open-data.pages.km3net.de/licensing/

6.6. Ethical and environmental aspects

Ethical aspects are covered in KM3NeT-INFRADEV WP5 *Societal role, societal impact*, see [5]. We anticipate no additional aspects that are not covered there.

A strategy to reduce the carbon footprint of the KM3NeT computing is under development, focusing on both recommendations to the individual users as well as criteria to be deliberated with the computing centres contributing to the KM3NeT computing.

7. Data management Responsibilities and Resources

7.1. Maintenance

The maintenance of the open science system requires dedicated workforce but will also be set up such that the effort should be kept minimal. Therefore, the servers will be run with a dedicated testing suite to check the upload of material for accuracy regarding metadata, uptime monitoring will be established and a versioning and archiving procedure established to keep data sets to a current status and archive or delete outdated content.

Also, the continuous feedback for new requirements and bug reports will have to be addressed by dedicated maintenance and development efforts. Therefore, maintenance of the open science system has to be addressed with dedicated human resources, see below.

7.2. Roles and human resources

Strategy and governance

Currently the governance of the KM3NeT collaboration is regulated through a Memorandum of Understanding (MoU). The management team includes the Physics and Software manager who is also the prime responsible for the data management and delegated tasks and partial responsibilities to the data processing experts. The future governance of KM3NeT will be performed through a legal entity which will be established by the Institute Board and the concrete implementation of which is prepared in the Infradev2 WP2.

The primary development and implementation of the Open Science Policy is addressed by the Open Science Committee and the related working groups.

Open Science Committee and working group members perform reviews and initialise developments according to the current needs of the collaboration. They participate in the community collaborations mentioned above.

System maintenance

The maintenance of the Open Science System requires dedicated human resources for system updates, backups, general server maintenance and facilitating the providing of data by the collaboration members. Currently, the long-term maintenance relies on the in-kind contribution of the hosting institute (ECAP). For the maintenance of the data processing environment, commitments by the Computing Clusters and funding from the INFRADEV project are available as stated below.

System development

The open science system development at this point includes the definition of data models, test implementation of the VRE, providing of converters between different data formats and documentation of the system. For related projects, members and their responsible PIs apply for grants for the further development of the system. These efforts are generally seen as in-kind contributions to the collaboration efforts.

Providing content

Content providers are

- researchers in KM3NeT performing their own analysis and providing the outcome in the open science framework. Their contribution will be covered in the individual science projects.
- The KM3NeT collaboration, generating (multi-purpose) data sets as the results of the generation of neutrino data. Here, data taking processing efforts performed by the collaboration result in the generation of the public data sets. A specialised service group within the collaboration will process the data from low-level to high-level and will provide data-related services (including documentation and support on data handling) to the collaboration and partners. At this point, no specific funding is provided to perform data processing.
- KM3NeT members, providing tutorials and teaching material, either for the use by scientists or for outreach activities. Currently, there is no specific funding for these activities.
- The ANTARES collaboration, offering ANTARES legacy data as test case and for archiving.

For the continuous maintenance and curation of the data management system and user support, independent from further developments and high-level data analysis, the following estimate for person power to maintain the fully running system can be made.

Task	Description	Number of FTE
Maintaining Workflow Management System	Monitoring, interaction with Computing Sites, resource allocation	2
Maintaining Data Management System	Site interaction, troubleshooting	1
General User support Grid	User training, troubleshooting, support of code migration	2
Curating Open Science System	External & internal user support, data curation	1

Curating Online Alert System	Monitoring, Organization of Shifts	1
Maintenance of Main Offline Software Frameworks and Algorithms	Support for full software lifecycle	8
Maintaining Central KM3NeT IT Services	Service administration, Updates, Monitoring, User Support	2

Table 5 Estimate for Human Resources

7.3. Allocation of resources

The commitment of long-term and sustainable computing resources for KM3NeT, including both processing and storage needs, has been made by the KM3NeT partner institutes responsible for the two large computing centres CC-IN2P3 and CNAF, CNRS and INFN respectively. An agreement exists between the KM3NeT partner Nikhef and their national e-Infrastructure organisation SURF for the usage of Grid compute resources up to 2025.

All resource needs are negotiated on a yearly basis between the KM3NeT collaboration and the computing centres, or responsible research organisations. For Tier-1 centres, this is by request of the contact persons from the computing centres to the computing coordinator of KM3NeT. For Tier-2 centres, the agreements are based on negotiations between the KM3NeT IB representatives and KM3NeT computing working group directly.

Formalised agreements - in the form of a memorandum of understanding - are foreseen within the context of setting up a legal entity for KM3NeT. The MoU between the funding institutions for Tier-1 computing and the KM3NeT legal entity will ensure the provisioning of the resources in the future. The computing needs at the shore station for Tier-0 will be covered by Service Level Agreements in the framework of the legal entity. To ensure the use of a well-established computing infrastructure, the KM3NeT collaboration is exploring the option to become an observer to the Worldwide LHC Computing Grid[61].

[61] https://wlcg.web.cern.ch/

VI. Appendix

A. Calibration procedures

The different calibration steps (that include pre-calibration, time calibration, PMT calibration, position calibration, and orientation calibration) are performed at different stages and with different frequencies. The pre-calibration of the positions within a new detector configuration is performed once after each deployment. Deployments take place about three times per year, for ORCA as well as ARCA. This pre-calibration typically takes a few processing days, and takes place at a Tier-1 site. The input for this pre-calibration is retrieved from the Oracle database. The output of this procedure is stored in ROOT files, and are used in the data processing of all data for the corresponding detector configuration.

The initial time calibration is performed in the laboratory for each DU before deployment. The data of these measurements are manually ingested into iRODS and stored on HPSS at CC-IN2P3. The results of the calibration procedure are stored in the central Oracle database, as they are used by the real-time data filter during operation. Immediately after deployment, subsequent time calibration measurements are performed using the first real-time processing of the data. This is done to transfer the time calibration to all other PMTs using a designated procedure using real data. This procedure takes a few hours to perform. During operation of the DU, the time calibration is continuously monitored, and occasionally updated, for example when the high voltage of the PMTs are tuned. Typically, the time calibration is stable during a few weeks of data taking, An updated time calibration takes a few minutes to perform per data taking run. The time calibration data are ingested into the GitLab archive in ASCII format, to be used for data processing.

The calibration of the gain, the gain spread, and the efficiency of the PMTs is performed for a set of data taking runs, each currently covering three hours. The data required for this calibration reside in the raw data files. This calibration is performed prior to the data processing of a run and takes a few minutes to perform per run. The output is stored in ASCII files in the GitLab archive.

The typical update frequencies of the PMT positions and orientations are typically 10^{-2} Hz. These are taken from the pre-calibration set, available from the GitLab archive, and are dynamically taken into account in the data processing to correct the geometry of the detector. As this is a dynamical procedure, the data that are needed to continuously update the PMT positions and orientations are kept in RAM.

VII. References

[1]	S. Adrián-Martínez et al., “Letter of Intent for KM3NeT 2.0,” *Journal of Physics G: Nuclear and Particle Physics,* vol. 43 (8), p. 084001, 2016.
[2]	European Commission - Directorate-General for Research & Innovation, “H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020,” v3,2016.
[3]	J. Knobloch et al., LHC Computing Grid, Geneva: CERN, 2005.
[4]	J. Schnabel, T. Gal, Z. Aly, “The KM3NeT Open Science System”, 2021, arXiv:2101.06751,
[5]	KM3NeT Collaboration, “Deliverables of the KM3NeT-INFRADEV Project,” [Online]. Available: https://www.km3net.org/km3net-infradev/project-outputs/. [Accessed 18 11 2020].