KM3NeT INFRADEV – H2020 – 739560

KM3NeT Data Management Plan

KM3NeT-INFRADEV GA DELIVERABLE: D4.1

Document identifier:	**KM3NeT-InfraDev-WP4-D4.1_v1.4**
Date:	**22 June 2017**
Work package:	**WP4**
Lead partner:	**FAU**
Document status:	**Endorsed by PMB and KM3NeT IB**
Dissemination level:	**Public**
Document link:	

Abstract

This document details a first version of KM3NeT data management plan, at the start of the H2020 KM3NeT-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 project whenever new decisions and implementations of the relevant points have been addressed by the collaboration.

I. Copyright notice

II. Delivery slip

	Name	Partner/WP	Date
Author(s)	K. Graf and U. Katz	FAU / WP4	12/06/2017
Approved by	PMB and KM3NeT IB		14/06/2017

III. Document log

Issue	Date	Comment	Author/Partner
1.1	01/05/2017	Initial draft derived from the KM3NeT computing model	K. Graf/FAU
1.2	07/06/2017	Updated with feedback of WP4 members	K. Graf, U. Katz/FAU
1.3	12/06/2017	Version for endorsement of KM3NeT-InfraDev PMB and KM3NeT IB	K. Graf, U. Katz/FAU
1.4	04/08/2017	Adapted to template for deliverables, added document information	K. Graf, U. Katz/FAU

IV. Application area

This document is a formal deliverable for the GA of the project, applicable to all members of the KM3NeT INFRADEV project, beneficiaries and third parties, as well as its collaborating projects.

V. Terminology

ARCA = Astroparticle Research with Cosmics in the Abyss (KM3NeT neutrino astroparticle physics telescope)
BB = Building Block
DMP = Data Management Plan
DOI = Data Object Identifier
FAIR = Findable – accessible – interoperable – reusable
GEDE-RDA = Group of European Data Experts in Research Data Alliance (RDA)

GNN	=	Global Neutrino Network
IB	=	Institutional Board
MC	=	Monte Carlo
ORCA	=	Oscillation Research with Cosmics in the Abyss (KM3NeT neutrino particle physics detector)
PID	=	Persistent identifier
PMB	=	Project Management Board
PMT	=	photo-multiplier tube
RI	=	Research Infrastructure

VI. List of figures

VII. List of tables

VIII. Project summary

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

IX. Executive summary

The Data Management Plan is a working document to be updated throughout the project. The current version reflects the status at the beginning of the KM3NeT-InfraDev project. It summarises the KM3NeT data and computing models and requirements and summarises the KM3NeT plans for data management under particular consideration of making data findable, accessible, interoperable and reusable (FAIR).

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 depth of several kilometres (2). KM3NeT is designed such that all data are sent to shore and processed there by an online computer farm (“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 (e.g. two years) 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. These data will be made publicly available after the embargo time under an open-access policy on a web-based service and will not only allow the public to validate the scientific results presented by the collaboration but also 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 is with the scientists performing the corresponding analyses. The KM3NeT Collaboration offers analysis support to external analysers on their request, and after scrutinising the validity of the respective analyses. In this case, both the external scientists and the KM3NeT collaboration will author the resulting publications.

The KM3NeT RI is constructed in a phased and distributed approach, where the basic element is one *Building Block (BB)* that comprises logical and technical sub-units of the respective detectors. BBs have a dimension that guarantees that the physics sensitivity grows 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) and East of Sicily (KM3NeT-It). A future Phase 3 with additional 4 ARCA BBs is under consideration but not yet in the planning phase.

Table 1 Phases of the KM3NeT project (planning status: 05/2017)

Phase	Detector Layout	No. of DUs	Start of construction	Full Detector
Phase 1	approx. ¼ BB	31 DUs	2015	2018
Phase 2.0	2 BBs ARCA/ 1 BB ORCA	345 DUs	2017	2022
Reference	1 BB	115 DUs		

This document details a first version of KM3NeT data management plan (DMP), at the start of the H2020 KM3NeT-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 DMP is based on the "Guidelines on FAIR Data Management in Horizon 2020 (3)" and will be updated during the project whenever new decisions and implementations of the relevant points have been addressed by the collaboration.

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 research field and in general. The guidelines for these developments are recommendations by expert groups (ASPERA (4), e-IRG (5), RDA (6), W3 (7) or similar) and a best-practice approaches. KM3NeT colleagues play active roles in various groups, like the Group of European Data Experts (GEDE-RDA (8)) and consortia like the ASTERICS H2020 project (9). In addition, KM3NeT has established contacts with European-wide e-Infrastructure providers, like EGI and EUDAT. The collaboration will also study the possibilities of cloud computing, especially in the context of the EOSC (10).

2. KM3NeT data

2.1. KM3NeT computing model

The DMP is a fundamental part of the KM3NeT computing model (data distribution and data processing system), which is based on the LHC computing model (11). The general data processing concept consists of a hierarchical system, commonly referred to as Tier structure (see Figure 1).

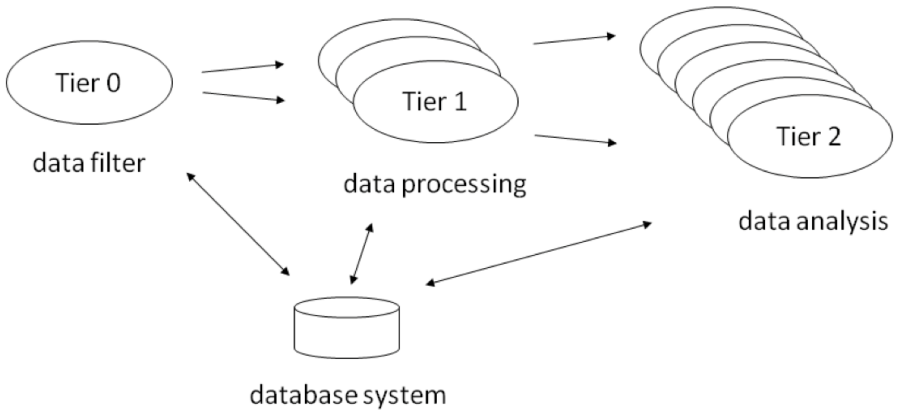

Figure 1 Data Distribution Model

At the first layer (Tier 0), farms of computers filter the full raw data from the deep sea in real-time and selects signal pattern consistent with scientifically relevant information ("events"). These computer farms are located in the shore station at each detector site. The main requirement is the reduction of the overall data rate by about three orders of magnitudes (from 5 GB/s to 5 MB/s for one BB). The output data are temporarily stored on a persistent medium and distributed with fixed latency (typically less than few hours) to various computers centres. For detector monitoring and alert handling, specific monitoring data are recorded and selected events are reconstructed using fast algorithms with a low latency of the order of seconds to minutes (quasi-online reconstruction).

In the second layer (Tier 1), the filtered data are processed. Various models for hit patterns are fitted to the data (usually referred to as reconstruction) and non-reconstructable events are discarded. In order to achieve an optimal pointing resolution of the detector, the reconstruction step requires a detector calibration using auxiliary data to determine the time offsets, efficiencies and positions of the photo-multiplier tubes (PMTs). The results of the calibration and quality identifiers are stored in a central database system. The typical update frequencies of the PMT positions and time offsets are 10^{-2} Hz and 10^{-5} Hz, respectively. The corresponding data rate amounts to less than 1MB/s. An automated calibration procedure is needed to process the filtered data in a timely manner. Ultimately, the calibration and the reconstruction procedures are expected to operate with a fixed latency of a few hours or less. The fully calibrated reconstruction results are stored for further analysis and are made available to the scientists for detailed scientific analyses. The overall rate of reconstructed data amounts to less than 1 MB/s per BB.

To assess the detector efficiency and systematics, dedicated Monte Carlo (MC) simulations are processed. Due to the changing running conditions of the detector in the deep-sea environment, time-dependent simulation data sets are required. These can be implemented optimally in a run-by-run simulation strategy, where runs are sufficiently small time intervals of data taking with stable conditions. The detector response is simulated individually for these periods, allowing for a direct comparison of experiment and simulation. Due to the construction of the detector in independent, nearly identical BBs, simulations can be performed for the fixed geometry of two types of building blocks (ARCA and ORCA). The MC data are generated at the raw-data level and are subjected to the same filter and reconstruction processing, and are stored in the same event data format as the real data. The MC data will significantly exceed the real data in volume since large statistics are required

for precise analyses, and also since the “true MC information” must be stored together with the reconstruction results.

The analyses of the reconstructed and the MC data are primarily processed at the local computing clusters of the partner institutes (Tier 2).

The KM3NeT computing model is schematically shown in **Figure 2 Detailed Computing Model**).

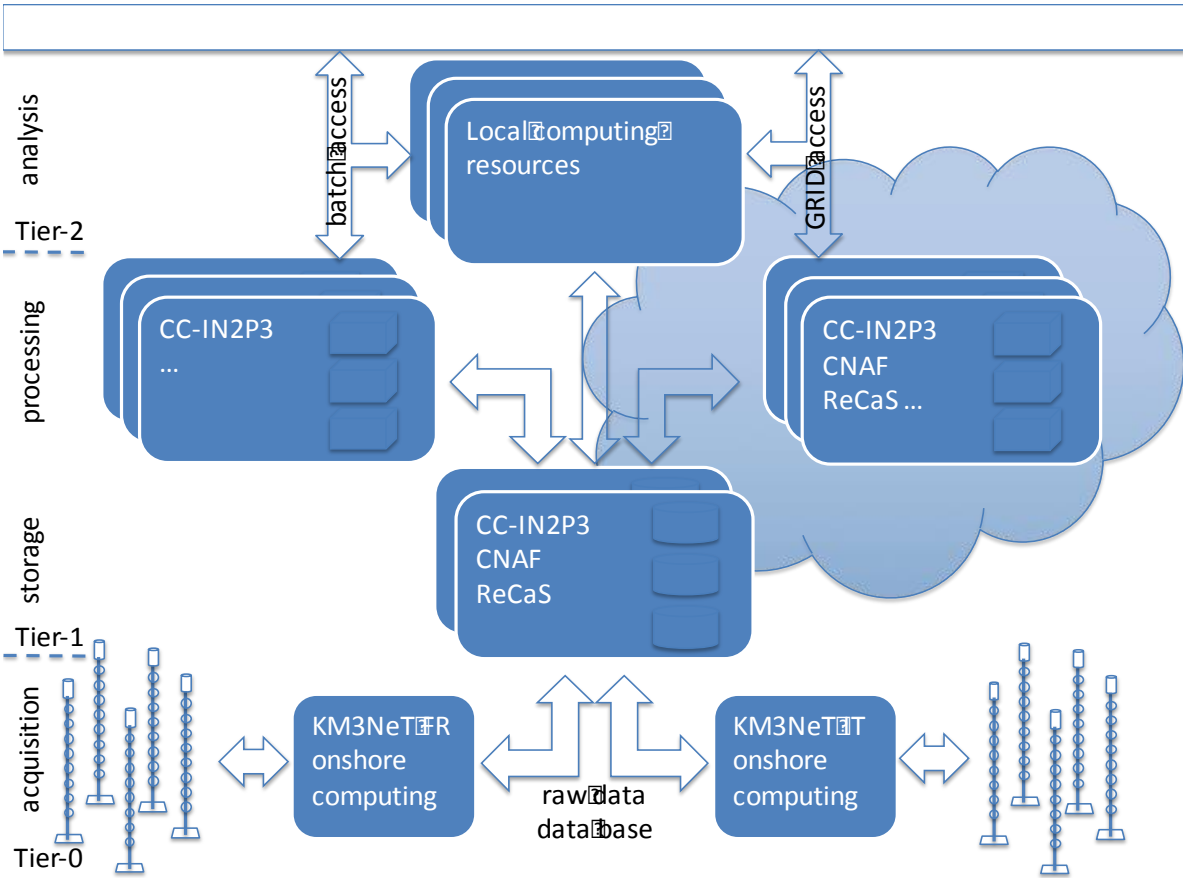

Figure 2 Detailed Computing Model

2.2. KM3NeT data characteristics and processing

Purpose of the data collection/generation

The purpose of the data collected with the RI and generated in simulations is basic research in the fields of particle physics, astrophysics and Earth and Sea sciences. The data collection encompasses data products at different abstraction levels:

- raw data (including metadata) from the detectors themselves;
- monitoring data;
- calibrated data: adding detector status and calibration to the raw data;

- reconstructed data;
- simulation data;
- analysis results.

All data products include metadata, such as processing information, data quality identifiers or uncertainties. As described above, these data streams are required for the scientific exploitation of the RI.

Relation to the objectives of the project

The objective of the project is to forward scientific exploration of neutrino astrophysics and neutrino particle physics. This requires scientific analysis on the event level, and on similar data abstraction levels for the Earth and Sea sciences. In order to provide these data, data processing in the following steps is required (see Table 2 for a summary):

- calibration;
- quasi-online reconstruction;
- offline reconstruction;
- simulation;
- physics analysis.

Types and formats of data generated/collected

The primary data are stored in the ROOT format (12) – interfaces and converters to other standard formats (e.g. HDF5 (13) as proposed as standard format by the ASTERICS project) are available or being developed (e.g. for the FITs format). Metadata is provided by a central database in customisable format – primary in standard ASCII text format.

Table 2 Data processing steps and locations, user access

Tier	Computing Facility	Processing steps
Tier-0	at detector site	triggering, online-calibration, quasi-online reconstruction
Tier-1	computing centres	calibration and reconstruction, simulation
Tier-2	local computing clusters	simulation and analysis

Is existing data being re-used?

Raw data is re-used in later data processing cycles for adding additional information and implementing knowledge gained from detector operation and calibration, and from data analysis. High-level analyses will use existing data from e.g. astronomical catalogues, astrophysical observation (mainly in the multi-messenger approach of astrophysics) and cooperating experiments (ANTARES, IceCube, VIRGO/LIGO, …).

Origin of the data

Data is produced directly in the detectors, in test stands in laboratories of the collaboration and in the e-infrastructures employed by the KM3NeT collaboration (see Figure 2).

3. FAIR data usage

The methodology and implementation of FAIR (findable, accessible, interoperable and reusable) data usage is the major deliverable of the KM3NeT-InfraDev WP4 Open Data Access; preliminary considerations will be reported below. One of the aims of the KM3NeT data management plan is to provide a sound 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 research field and in general. KM3NeT will adapt the open science concept by providing access to high-level data and data analysis tools.

3.1. Making data findable, including provisions for metadata

Discoverability of data (metadata provision)

High-level data will be distributed with open-access via a web-based interface – the implementation, e.g. in form of Virtual Observatories, will be defined. All metadata necessary to uniquely identify data sets will be provided.

Identifiability of data and use of persistent and unique identifiers such as Digital Object Identifiers

Currently universal data identifiers (DOIs or PIDs): KM3NeT is actively involved in the definition of standards via the GEDE-RDA group and GNN. Prerequisites are persistency, uniqueness and reach – including a versioning of data, as the KM3NeT data sets will grow over time.

Internally the data are identified uniquely from their run numbers and processing version.

Naming conventions used

The data made publicly available will be in data sets with a naming convention following their origin (detector or physical origin) as well as the processing version. Internally, the following information is encoded in the naming: run (link to time of generation), detector and version of data processing and in metadata hosted via the central database.

Approach towards search keyword

In addition to general keywords, such as “neutrino”, “astrophysics”, etc., analysis- and data-set related keywords will be stored in metadata that will – depending on the service chosen for publication – be freely searchable.

Approach for clear versioning

Versioning will be in a MAJOR.MINOR release approach. There will be new versions for the publicly available data sets when:

- new data is added to a set, e.g. due to additional simulation runs or further detector data acquisition – this will increase the minor version number;
- the underlying data structure or processing chain has changed without backwards compatibility – this will result in an incremented major version number.

Standards for metadata creation

Well-defined standards are currently under development. The metadata is generated via the central KM3NeT database, so a flexible approach is possible, optimised to the needs of the service used for data dissemination.

3.2. Making data openly accessible

The KM3NeT collaboration has developed measures to ensure the reproducibility and usability of all scientific results over the full lifetime of the project and in addition for a period of about 10 years after shutdown. Low-level data (i.e. raw data recorded by the experiment) and high-level data will be stored in parallel at central places. A central software repository, central software builds and operation system images are provided and continuously maintained until the end of the experiment.

Which data will be made openly available? If some data is kept closed provide rationale for doing so

As explained above, data access will be reserved to the KM3NeT collaboration for a certain embargo period after data taking. This is required as a return for construction, operation and maintenance of the RI, but also for the generation of high-quality high-level data suited for open access dissemination.

As a first stage of dissemination, high-level data will be made publicly available. At a later stage of the experiment, providing additional, lower-level data and the related documentation and application program examples can be considered to extend the scope and quality of data analyses by external users. The analyses of low-level data (i.e. raw data) need an expertise and resources that are only available to the collaborators and cannot be transferred to open access with a reasonable effort – making these data openly available would thus not increase the general knowledge.

How will the data be made available?

The data will be openly available using a web interface (the simplest solution being a web page). More standardised implementations like Virtual Observatories are under investigation. The decision on the actual implementation is part of WP4 of the KM3NeT-InfraDev project.

Several backups will be distributed on the central computing centres of the collaboration, to guarantee persistence and failure safety.

What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

This is under study as well in WP4. The obstacle for accessing the data should be as low as possible. This requires providing sample programs, an environment where they can be executed and access the data, and documentation on data contents and formats, as well as on the software. Documentation is an important sub-task of WP4 and will be provided in form of documents, via trainings and through in-code documentation (which is required by the KM3NeT software quality plan).

Where are the data and associated metadata, documentation and code deposited?

They will be deposited on the same web-interface making the data available, with the same security and safety measures.

How will access be provided in case there are any restrictions?

No restrictions are currently planned or considered desirable for open-access data. A licensing procedure regulating rights and obligations of data users and data providers is under discussion. Access to lower level data can be provided, with full support by the collaboration, if a relevant result has been achieved with the openly available data. The rules and regulations for this case, as well as the licensing, are part of the KM3NeT-InfraDev program.

3.3. Making data interoperable

Interoperability of data: Data and metadata vocabularies, standards, and methodologies

Interoperability will most likely focus on multi-messenger analyses in the astrophysics community. The data will be provided in a standard format that allows for analysis with widely used tools in that community (HDF5, FITS (14) or ROOT formats). Including the data in a Virtual Observatory will ensure interoperability.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Not applicable.

3.4. Increase data re-use (through clarifying licenses)

How will the data be licenced to permit the widest reuse possible?

The licensing will be part of the KM3NeT-InfraDev package. 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.

When will the data will be made available for re-use; why and for what period a data embargo is needed?

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 after an embargo period (e.g. two years).

Is the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

All openly available data – as defined above – will be usable by third parties after the embargo time and also at least 10 years after the end of the project.

Data quality assurance processes

The data is processed by a dedicated group of experts in a central queue. A data quality group is responsible to assess the quality of all data. This process is to be automated to the maximal possible extent in the KM3NeT-InfraDev project in order to reduce latencies in data analysis and alerts. The openly available data will be produced in the same processing chain as the official data productions leading to KM3NeT publications, thus data quality control is an integral part of the KM3NeT QA/QC procedures.

Length of time for which the data will remain re-usable

At least 10 years after the end of the KM3NeT RI operation.

4. Allocation of resources

The storage and computing needs of the KM3NeT project are highly advanced. The KM3NeT collaboration has developed a data management plan and a corresponding computing model to answer these needs. The latter is based on the LHC computing model utilising a hierarchical data processing system with different layers (tiers). Data are stored on two main storage centres (CCIN2P3-Lyon, CNRS and CNAF, INFN). These data centres are fully interfaced with the major European e-Infrastructures, including GRID-facilities (INFN, HellasGRID and ReCaS provide resources to KM3NeT). The main node for processing of the neutrino telescope data is the computer centre in Lyon (CCIN2P3-Lyon). A corresponding long-term and sustainable commitment has already been made by CNRS, which is consistent with the needs for long-term preservation of the data. 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. WAN (GRID) access tools (e.g. xrootd, iRODS, and gridFTP) provide the access to high-level data for the collaboration. The analysis of these data will be pursued at the local e-Infrastructures of the involved institutes (both local and national). The chosen data formats allow

for the use of common data analysis tools (e.g. the ROOT data analysis framework) and for integration into e-Infrastructure common services.

Estimate of the costs for making your data FAIR

The cost of making KM3NeT data FAIR depends primarily on the implementation of the open data access methods that will be established in KM3NeT-InfraDev WP4. Based on preliminary estimates those costs will be covered by the general collaboration funds as well as in-kind-contributions of the partner institutes.

Responsibilities for data management

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 RI governance will be established by an ERIC, which will be set up during the KM3NeT-InfraDev project.

Costs and potential value of long term preservation

The central services are mainly funded through CNRS and INFN that have pledged resources of their main computing centres to the project. Additional storage space and its management are provided by the partner institutes (e.g. INFN has provided 500 TB of disk space for KM3NeT at the ReCaS GRID infrastructure).

In addition to the major storage, networking and computing resources provided by the partner institutions and their computing centres, grid resources have been pledged and will be used by KM3NeT (ReCaS, HellasGRID). These will provide significant resources to be used for specialised tasks (as e.g. for special simulation needs). The major resources, however, will be provided by the partners. External services are employed to integrate the KM3NeT e-Infrastructure into the European context of the GRID – in the fields of data management, security and access; services will be implemented in collaboration with EGI.

Data preservation beyond the RI operation time will be important since the recorded data is unique and relates to time-dependent astrophysical and marine processes (i.e. it cannot be recollected if lost). Technology and methods of data preservation will be decided in due time (i.e. a few years before the end of operation). Only then a realistic cost estimate will be possible.

5. Data security

Address data recovery as well as secure storage and transfer of sensitive data

In general, data handled in the KM3NeT project does not contain sensitive data of any kind. Personal data and authentication is separated from the actual data and handled via central authentication and authorisation methods.

6. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Ethical aspects are covered in KM3NeT-InfraDev WP5 *Societal role, societal impact*. We anticipate no additional aspects that are not covered in WP5.

7. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

All data leading to publications will be centrally stored for a period of 10 years, in addition data used in theses will be stored in the partner institutes.

References

1. *Letter of Intent for KM3NeT 2.0.* **Adrián-Martínez, S., et al.** 2016, Journal of Physics G: Nuclear and Particle Physics, Vol. 43 (8), p. 084001. arXiv:1601.07459 [astro-ph.IM]. DOI: 10.1088/0954-3899/43/8/084001.

2. **The KM3NeT Collaboration.** KM3NeT Homepage. [Online] 2017. [Cited: 29th May 2017.] https://www.km3net.org.

3. **European Commission - Directorate-General for Research & Innovation.** *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020.* v3,2016. available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

4. **AStroParticle ERAnet - ASPERA.** ASPERA homepage. [Online] http://www.aspera-eu.org.

5. **e-Infrastructure Reflection Group - e-IRG.** e-IRG homepage. [Online] http://e-irg.eu.

6. **RESEARCH DATA ALLIANCE - RDA.** RDA homepage. [Online] https://www.rd-alliance.org.

7. **World Wide Web Consortium - W3C.** W3C homepage. [Online] https://www.w3.org.

8. **Group of European Data Experts in RDA - GEDE-RDA.** GEDE homepage . [Online] https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda.

9. **ASTERICS.** ASTERICS homepage. [Online] https://www.asterics2020.eu.

10. **European Open Science Cloud.** EOSC for Research Pilot Homepage. [Online] https://eoscpilot.eu.

11. **Knobloch, J., et al.** *LHC Computing Grid.* Geneva : CERN, 2005. LCG-TDR-001, CERN-LHCC-2005-024. ISBN 92-9083-253-3.

12. **CERN.** ROOT Data Analysis Framework homepage. [Online] https://root.cern.ch.

13. **The HDF Group.** HDF5 support page. [Online] https://support.hdfgroup.org/HDF5/.

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