|  |
| --- |
| **SUBMERSE.EU** |
| **Data Management Plan (Version 1.0)** |
| **Deliverable D1.2** |
| **A blue and white background with lines and a bird  Description automatically generated with medium confidence** |
| **Hannah Mihai (DeiC);Rene Belsø (DeiC);Angelo Strollo (GFZ);Panagiotis Douris (GRNET);Kostas Koumantaros (GRNET);Chris Atherton (GÉANT);Olaf Schjelderup (Sikt);Lars Fischer (NORDUnet)**  12-4-2023 |

|  |
| --- |
|  |

31-11-2023

Deliverable 1.2 Data Management Plan (Version 1.0)

|  |  |
| --- | --- |
| Contractual Date: | 30-11-2023 |
| Actual Date: | 04-12-2023 |
| Grant Agreement No.: | 101095055 |
| Work Package: | WP1 |
| Task Item: | Task 4 |
| Nature of Deliverable: | DMP (Data Management Plan) |
| Status: | Submitted Draft, not yet approved by the EC |
| Dissemination Level: | PU (Public) |
| Document link: | https://doi.org/10.5281/zenodo.11065792 |
| Lead Partner: | DeiC |
| Authors: | Hannah Mihai (DeiC); Rene Belsø (DeiC); Angelo Strollo (GFZ); Panagiotis Douris (GRNET); Kostas Koumantaros (GRNET); Chris Atherton (GÉANT); Olaf Schjelderup (Sikt); Lars Fischer (NORDUnet); |
| Internal Reviewers: | Frederik Tilmann (GFZ) |
|  |  |
|  |  |

Table of Contents

[Executive Summary 1](#_Toc151653371)

[1 Introduction 2](#_Toc151653372)

[2 SUBMERSE Data Collection Sites and Use-cases 4](#_Toc151653373)

[3 SUBMERSE Data Summary 5](#_Toc151653374)

[4 SUBMERSE Data Policy 6](#_Toc151653375)

[4.1 Policy Partners – EPOS, EMSO and Others 7](#_Toc151653376)

[4.2 Policy perspectives – Domain Specific vs. Generic, and Restricted vs. Open 8](#_Toc151653377)

[5 FAIR Data in the SUBMERSE project and beyond 10](#_Toc151653378)

[5.1 Making data Findable (F) 11](#_Toc151653379)

[5.2 Making data Accessible (A) 12](#_Toc151653380)

[5.3 Making data Interoperable (I) 14](#_Toc151653381)

[5.4 Making data Reusable (R) 15](#_Toc151653382)

[6 Other research outputs 17](#_Toc151653383)

[7 Allocation of resources 18](#_Toc151653384)

[8 Data Security 19](#_Toc151653385)

[9 SUBMERSE Ethics 20](#_Toc151653386)

[11 SUBMERSE outlook - final remarks 21](#_Toc151653387)

[Appendix A - Scientific Use Cases Surveys 22](#_Toc151653388)

[Appendix B - Data flow 29](#_Toc151653389)

[SUBMERSE Glossary 30](#_Toc151653390)

[References, Footnotes & Endnotes 31](#_Toc151653391)

# Executive Summary

The data management procedures that are developed and implemented during the SUBMERSE project are presented in this document as a first version 1, in that it is an integral part of the SUBMERSE project to devise an optimal management for the very complex SUBMERSE data – continuously streamed, in vast quantities, from several locations, with several bottom-of-ocean sensors, potentially providing highly sensitive data. The data initially unintelligible, is to be repeatedly filtered and processed, ending up significantly reduced in research community repositories, as ***F***indable, ***A***ccessible, ***I***nteroperable and ***R***eusable (FAIR) data. I.e. FAIR compliant Distributed Acoustic Sensor (DAS)[[1]](#endnote-2) and State of Polarisation (SOP)[[2]](#endnote-3) data, of high value to a number of research domains.

This Data Management Plan (DMP) is based on the Horizon Europe Data Management Template. In this version, we describe the framework for data acquisition, scrubbing, storage, and publication to the best of our understanding six months into the project, where part of the project is concerned with defining many of the details now noted as decisions to be taken. When any significant changes to the Data Management practices within the SUBMERSE project occur, an updated version of the DMP will be published.

This document describes extensively the importance of open and FAIR data to the SUBMERSE project rational, simultaneously taking into account the potentially military dual-use nature of the raw data being generated, which therefore cannot be immediately made openly available. It is a clear objective for SUBMERSE to develop an instrument for scientific applications. The DMP is crucial in ensuring that scientific applications are impeded as little as possible by security restrictions. The plan for the development of automated scrubbing mechanisms and long-term storage of – a small part of – the streamed data, are described. Even though a separate task (T1.3) is dedicated to guidelines around ethics and security we also discuss them in this DMP version, since the DMP is very much dependant on security agreements with national authorities, impacting the overall instrument architecture and hence the DMP.

The ethics and security guidelines from T1.3 will be integrated into a final version of the DMP near the end of the project, in dialogue with the project’s *Security Advisory Committee*, as well as planed *National Data Access Committees.*

# Introduction

The SUBMERSE project is a response to the call *Next generation of scientific instrumentation, tools, and methods* (HORIZON-INFRA-2022-TECH-01). The project seeks to develop a new research instrument integrating already existing infrastructures, specifically ocean submarine fibre-optic communication cables, within the National Research and Education Networks (NREN)[[3]](#endnote-4), and the data distribution facilities of the European Plate Observing System (EPOS-ERIC)[[4]](#endnote-5) and Copernicus[[5]](#endnote-6) Marine Service. Through collaboration between three European nations (Norway, Greece, and Portugal), as well as Pan-European and Pan-American institutions, the instrument aims to disseminate FAIR[[6]](#endnote-7) & [[7]](#endnote-8) compliant Distributed Acoustic Sensor (DAS)[[8]](#endnote-9) and State of Polarisation (SOP)[[9]](#endnote-10) data, gathered from already installed, live, telecommunications submarine optical fibres, in a temporally consistent and geographically dispersed way. The instrument will deliver innovative scientific instrumentation, methods and tools as well as advance the involved research infrastructures by developing new state of the art technology, services, and long-term data sets of unique scientific value.

The SUBMERSE project will enable continuous recording of these types of data for research offered by NRENs for the first time, at three data collections sites, thus supporting diverse academic, industrial, and governmental user communities. Through the integration with existing data infrastructures (EPOS-ERIC and Copernicus Marine Services) research communities in solid earth and marine science (European Multidisciplinary Seafloor and water column Observatory, EMSO[[10]](#endnote-11)) as well as government agencies and industry can be provided with entirely new types of data sets not available to them before, which will lead to new areas of research.

A key challenge, and indeed a success criterion, in the SUBMERSE project (addressed further in this WP and in WP1.3; Ethics and Security) is how this distributed scientific instrument can deal with the potentially sensitive data the DAS instrument generates on marine activity, which may have military implications, of interest to national security agencies. Hence, the DMP must devise mechanisms, dataflows, and procedures that:

1. either filter out (by scrubbing, spatial and temporal decimation or frequency/wavenumber filtering) the sensitive data, before any further data analysis and processing for any other usages – scientific, commercial, or governmental.
2. or screens researchers and/or software algorithms allowing them to access even the sensitive data, prior to initial data scrubbing.

Other types of data produced in this project are Polarimeter and SOP. These have low resolution, and pose less of a security concern.

The former a) can be done in the sensing hardware (interrogators) or at a security approved national computer centre, but in any case, must be done (almost) in real-time, due to inevitable data-overflow from the instruments overwhelming data stream. The later b) can be done by a vetting process connected to a secure computer centre. Both alternatives need further clarification in dialogue with national security agencies, via the project’s *Security Advisory Committee*, as well as national Data Access Committees, that are in contact with research stakeholder communities. Selected stakeholders, e.g. national agencies, might have privileged access even to sensitive data. Different procedures have different research strategic, organisational, and economic implications that the DMP must uncover and consider.

In addition, it is a challenge to efficiently handle the data continuously streamed from the SUBMERSE instruments, in amounts that eventually will overflow data storage capacities – possibly from 1 to 10 terabytes per day per site, increasing drastically with increased data resolution (e.g., gauge length and sampling rate) and how many sensors are added. Therefore, streamed data must be processed on-the-fly in real-time. Though limited data buffering is planned, storing all the data is economically untenable. The minor fraction of the full sampling rate data from all channels, deemed of potential interest by stakeholders and as approved by independent research committees, must be extracted from the live data stream in real-time, or from temporal buffer, before decimating the continuous data in space and time to become small enough for permanent archival, while retaining as much signal of interest as possible, before discarding the full resolution data. Depending on the use case, higher level data products might be derived from the full resolution data at secure national computing centre, or even on the interrogator itself. Independent information about some phenomenon of interest might come from outside sources, which could be possible to investigate from subtle but important imprint on SUBMERSE data, buffered for a limited period. Hence, the challenge is knowing which data with (seemingly) no value to throw away, as early as possible. Development of the selection and scrubbing methods needed is closely connected to the scientific work done to understand the information that is gathered from the streamed raw data.

Firstly, one needs trusted scientific competence – e.g., geophysics – with necessary security clearance to decide, quantify and qualify what kind of information is present in the recorded data.

Secondly, one needs large temporary secure data storage capacity for raw data to train AI-models to do auto-classification, as well as for buffering data for approximately three months, in case researcher want to go back to the raw data, based on findings from the data in repositories. If such buffered data is of specific interest, then a more long-term storage solution for that raw data is needed.

Thirdly, empirical, or AI-based auto classification methods being developed within the project are key to reduce the total amount of raw data potentially in need of storage and/or post dissemination to other research communities. Though data-buffering allows for time to analyse data, one must still prevent growth of the buffer beyond an economically reasonable size, Therefore, automating the process down to seconds of analysis on the real-time data stream is a primary goal. Furthermore, a need for speed in data processing is driven from early warning applications of SUBMERSE data – like underground eruption, earthquake, tsunami alerts. We aim for achieving a real-time instrument, with a minimal need for longer term storage capacity.

# SUBMERSE Data Collection Sites and Use-cases

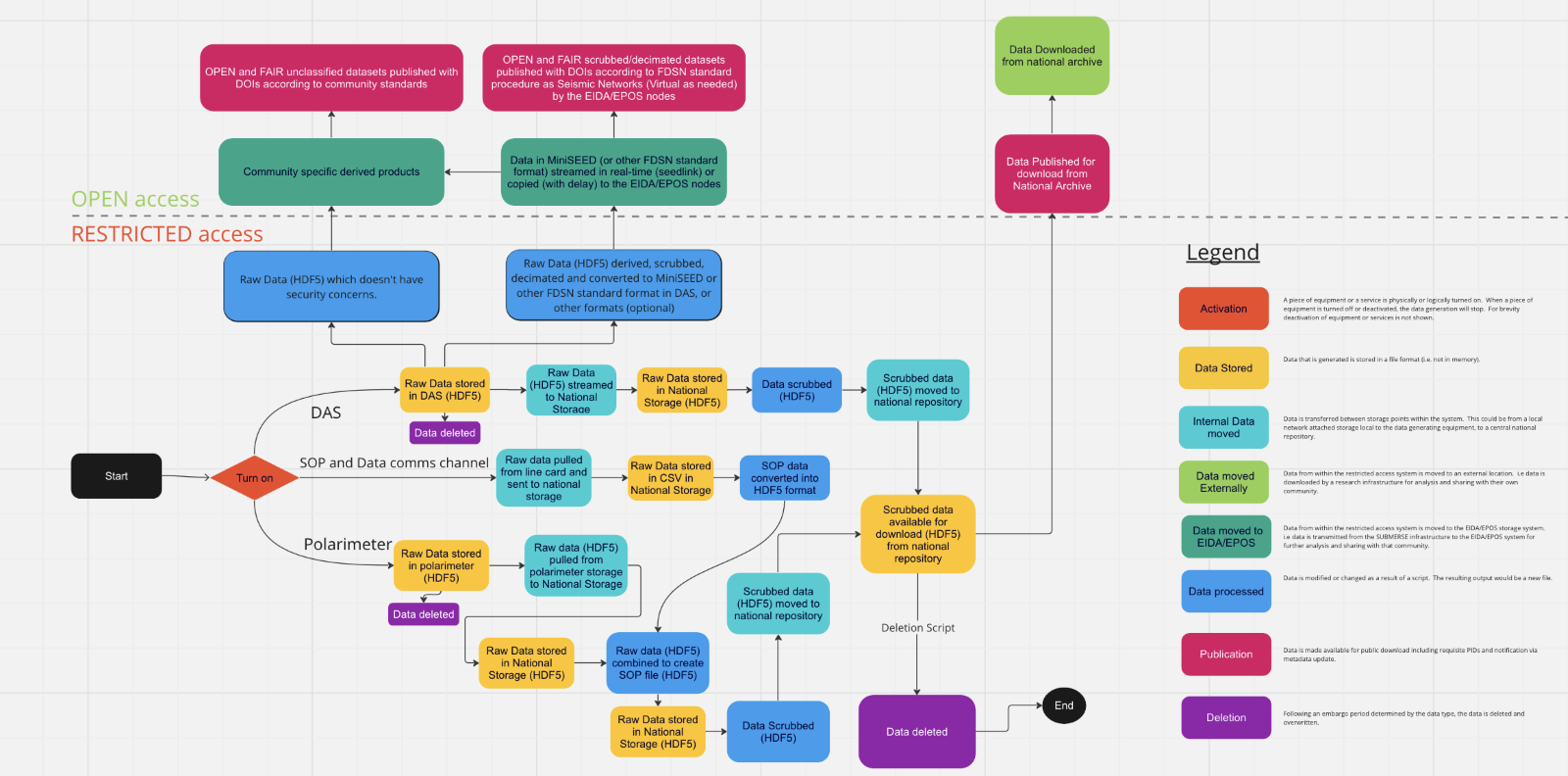
The project will collect data from sites in three countries (Norway, Portugal, and Greece) made accessible through the respective National Research and Education Networks (NREN) where at least three different technologies (DAS, SOP, Polarimeter and possibly SOP-OTDR) will be tested and used. Table 2.1 describes the data collection sites in terms of data types being acquired, use cases planned (see Appendix A) and national contact points.

The different national sites will have to follow different national security policies on which the data filtering (scrubbing) methods will be tailored during the SUBMERSE project under the guidance of national security advisors and stakeholders.

Table 2.1: Data collection sites in each country, with details on collected data types and intended use-cases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Country | Test sites | National Coordinator | Data Types | Use-cases |
| Greece | Preveza | GRNET | DAS  SOP  Polarimeter | NOA and HCMR   * Seismology * Volcanology * Oceanography * Tsunami monitoring |
| Portugal | Sines | INESCTEC | DAS  SOP  Polarimeter  (SOP-OTDR) | University of Lisbon, IPMA and UaH   * Oceanography * Seismology * DAS |
| Portugal | Madeira | INESCTEC | DAS | GFZ, University of Lisbon, IPMA and UaH   * Oceanography * Seismology * DAS |
| Norway (Svalbard) | Longyearbyen | Sikt | DAS  SOP  Polarimeter | NTNU and Bergen University   * Marine Biology * Oceanography * Seismology * DAS |
| Norway (Svalbard) | Ny-Ålesund | Sikt | DAS  SOP  Polarimeter | NTNU and Bergen University   * Marine Biology * Oceanography * Seismology * DAS |
| Brazil | Fortaleza | RedClara | SOP |  |

# SUBMERSE Data Summary

The SUBMERSE project is dealing with a large variety of data types. From measurement instruments, raw data is endlessly produced and streamed through a complex dataflow, manipulating, analysing, discarding, and storing the data along the ways, thus producing modified data along the way, i.e. data is filtered, cleaned (scrubbed), reduced, discarded, annotated, tagged and stored – temporarily and/or in longer-term repositories.

Furthermore, higher level products (e.g., arrival times of seismic waves or whale vocalisation, information on wave heights, earthquake locations, whale tracking, etc.), which need to be disseminated.

Figure 3.1: Data flow; see Appendix B for full size image.

Planning the optimal dataflow is essential for optimal use of the project instrumentation, and thereby for the many envisioned research agendas. It is an evolutionary dynamic process to design an ever more optimal dataflow. Decisions as to what and when to process data, why, how, and what to scrub, analyse, store, discard etc. is an ongoing process of improvement by incorporating results of developments within the project or from outside, with the aim of developing a world-class research instrument. An additional complexity is the fact that the SUBMERSE project is working with not one, but several geographically distributed instrument sites, as described in section 2. The data flow is best known for the DAS measurements (see full size figure in Appendix B). Therefore, we will use this as the baseline to describe the data flow as we understand it so far and use this as a template to construct the data flow for the other measuring technologies, which, however, are associated with acquisition rates that are orders of magnitudes less compared to DAS. It is very likely that, over the course of the project, we will discover limitations and opportunities for the collected data and how the data is processed and stored. The Data Lifecycle task will continuously monitor and adapt the DMP accordingly.

The SUBMERSE consortium will ensure that all the data collected, processed, and produced by the project follows well established standards with respect to security and ethics, respecting all relevant regulations at national and the European level. All data handled within the Consortium is collected, processed, and maintained lawfully, fairly, and transparently with respect to its subjects, as well as kept safe from unwanted observation and intervention.

Regional test site data are all scientific data, products, and software is produced by the three collaborating regional test sites. These data and products are generated by three technologies in use within the project: DAS, SOP and Polarimeter[[11]](#endnote-12).

# SUBMERSE Data Policy

The SUBMERSE project Data Management Plan (DMP) needs a clear data policy, to be agreed upon in collaboration between the SUBMERSE project, its Security Advisory Committee, and its user communities. It is too early in the project to settle most of these policy questions, which must be elaborated in later versions of the SUBMERSE DMP

The SUBMERSE project works with data in various states of movement and storage, many of which need policy clarification:

* A ***dataflow*** (stream) designates data in real-time (speed of light) movement, from its point of origin (a sensor) or from a data processing point (computing, data analytics, scrubbing etc). Streamed data might or might not contain sensitive data, depending on where scrubbing occurs.
* **Buffer storage** is for catching the stream temporarily, either due to fluctuations in data processing capacity, or due to potential later scientific utilization. The buffer is *First In First Out (FIFO)* and at present moment envisioned to be able to hold at least three months of streamed data, in the order of 10 terabytes per day per site, increasing drastically with increased data resolution (e.g., gauge length and sampling rate, and how many sensors are added), before it must be deleted. Buffered data has identification tags carried downstream, through various scrubbing/processing, into repositories, where communities’ specific researchers, within the buffered data’s lifespan, might find cause to request the original data from the buffer. Buffered data might or might not contain sensitive data, depending on where scrubbing occurs. Needless to say, data value (research utilization) declines after scrubbing for sensitive data, albeit possibly a necessary fate.
* **Storage** isdesignating data that is moved from the temporary buffer storage to other storage, due to research communities flagged interest, based on findings from the scrubbed/processed data in repositories at the end of the stream. Again, this stored data might or might not contain sensitive data, depending on where scrubbing occurs. Unless the security-sensitive part of these higher resolution datasets is essential to the research output, these data should be added metadata and submitted to community repositories. Exceptionally, e.g., when there is security sensitive data involved that cannot be scrubbed without making the research output irreproducible, they must be stored raw in protected national computer and data centers.
* **Repository data** is FAIR data with community metadata, as elaborated in the section "*FAIR Data in the SUBMERSE project and beyond*". The FAIR data is taken over, curated, indexed, and distributed by the community. It is this data that constitutes SUBMERSE products, where the SUBMERSE project has inclusion of several initial use cases: DAS; Marine Biology; Oceanography; Oceanography; Seismology; Volcanology. Further use cases such as tsunami detection can be envisaged but must still be feasibility-tested.

At its point of creation, the data is owned by the owner of the sensor instrument creating the data – i.e., the NREN’s in collaboration with the research Primary Investigator (PI) – currently not clearly defined and dependent on national policies. The SUBMERSE project is building a research instrument acquiring scientifically interesting data that is made available to other researchers openly (after scrubbing).

Initially a simple security scrubbing is aimed for – e.g., remove certain frequencies, or combinations of time periods and channel numbers. The responsibility for what must be removed lies with the national security agencies, via their presence in the project Security Advisory Committee. Since this is nationally dependent, different solutions work for different set-ups.

The SUBMERSE data distribution policy takes its point of departure from SUBMERSE data being an Open Access data offering to any interested research community, private or public organization. Data distribution policies may vary from open and real-time via embargoed to scrubbed data, depending on project development stage, as well as the specific use case at a specific site/technology, and specific clearance levels obtained from national security agencies. As an Open Access data offering, the SUBMERSE project will expect any interested party to present its own understanding (value proposition) of the SUBMERSE data streams, and its optimal utilization for their research agenda. In so doing, it is a SUBMERSE data distribution policy to demand interested parties do their own data processing, and have their own repository plans, including adding of metadata, negotiating with their community specific repositories and submission of data – either scrubbed by SUBMERSE or by the research community themselves.

The SUBMERSE project will, however, cooperate intimately with specific communities, namely within Geo-INQUIRE[[12]](#endnote-13), EPOS and user community Copernicus Marine Services, obtaining a robust array of SUBMERSE data products – useful FAIR datasets that will be part of these communities’ community specific repositories, for further research utilization. SUBMERSE will reach out to other ERICs, notably EMSO[[13]](#endnote-14) to reach additional user groups. Engagement of other communities will in principle be at the initiative of these communities and their ability to engage their community specific repositories, where the community links to solid earth and marine researchers through EPOS, Copernicus Marine Service and possibly EMSO, will serve as blueprint examples.

Hence, SUBMERSE plans to leverage on the data management policies and practises for DAS-data, developed within the Geo-INQUIRE EU project, which does this in coordination with the long-term sustainable European infrastructures EPOS and EMSO[[14]](#endnote-15) ERIC and in coordination with international domain-specific coordination bodies.

## Policy Partners – EPOS, EMSO and Others

The Data Management strategy planned within Geo-INQUIRE for DAS data follows the basic principles below:

1. datasets within the European Integrated Data Archive (EIDA)[[15]](#endnote-16) are curated in accordance with the EPOS data policy[[16]](#endnote-17) in synergy with local policies of the EIDA nodes[[17]](#endnote-18);
2. all datasets are attributed a Digital Object Identifier (DOI) according to a standard global procedure[[18]](#endnote-19);
3. when possible persistent identification of instruments is also used (Handle Identifiers);
4. data are licensed with CC-BY when possible, as defined in the EPOS data policy, PIs having ownership and make final decision on license;
5. data and metadata standards are not yet formalized. Geo-INQUIRE plans to foster the international discussion within the international Federation of Digital Seismograph Networks (FDSN)[[19]](#endnote-20)towards proposals and adoptions;
6. metadata are initially included as stationXML[[20]](#endnote-21) and mapped also to the recently provided guidelines from the DAS-RCN[[21]](#endnote-22) group;
7. data (after scrubbing) will be down sampled to MiniSEED[[22]](#endnote-23) and/or other seismological standards using existing tools (e.g. Quinteros 2021) and distributed via EIDA nodes with standard FDSN web services;
8. raw data – planned to be buffered for at least three months, then deleted – remains with the national facilities in original format (e.g. HDF5) which are linked and described, when possible, from the DOI landing page;
9. access to data can be embargoed using the EIDA AAI system which is based on eduGAIN[[23]](#endnote-24) with the federated identity providers dealing with identity data. EIDA services are only receiving a token digitally signed with encrypted user information; PIs are responsible to provide list of allowed users to the EIDA node operators;
10. Quality Assurance applied and web services with quality metrics are available for users.
11. The SUBMERSE Data Management Plan is produced and adopted within the project, alongside the development and adoption of standard metadata and data formats.

For the remaining two data types, SOP and Polarimeter data, SUBMERSE will foster adoption of a similar data management policy, as developed for DAS – where feasible and possible. When standard formats for data and metadata are not yet sufficiently established, SUBMERSE will foster adoption of FAIR data management policies, which are iteratively evolving during the SUBMERSE project, and future updates to this DMP.

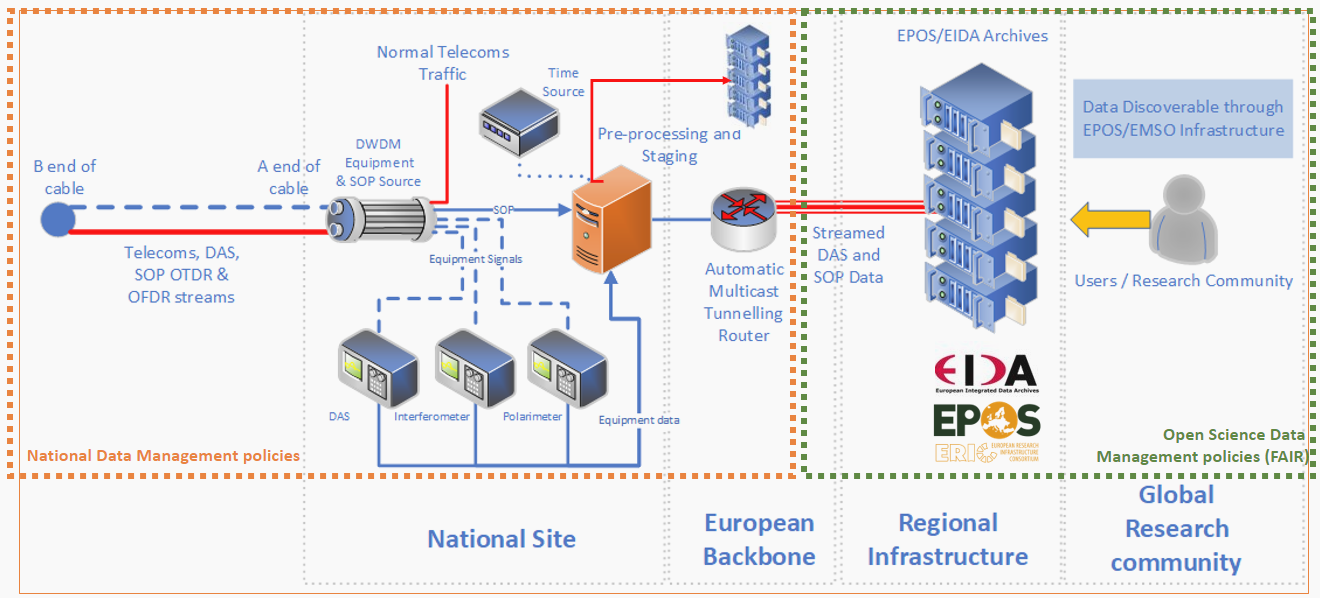


Figure 4.1 schematic view of the data for DAS, SOP, Interferometer. Note the separation between the (a) National Data Management policies (left part) from the (b) Open Science Data management policies (FAIR).

For all data types produced in the project, SUBMERSE data management policies will go through two steps as schematically represented in Figure 4.1.

In **(a)** the data cannot be made openly available but is restricted by national data management policies in which raw data are protected from the public, and most researchers. In this step the data will be pre-processed and scrubbed according to the national storage facility rules and in some cases already established practises before it will be streamed into repositories.

Data in repositories **(b)** is, for SUBMERSE selected use-cases, to be at the disposal of research communities, and follow open science and FAIR data management principles. I.e., chosen use case repositories take over data responsibility, once the raw data is pre-processed. In the case of DAS this is envisioned to be the EIDA Distributed Data Archive, which is the European Infrastructure for seismic waveform data in EPOS. SOP and Polarimeter data could be published in the same repositories – pending further talks with EPOS and EIDA. Other future communities will need to resolve data processing, metadata, and repository questions themselves, though the SUBMERSE community has a keen interest in assisting with data access and DMP advice. It should be noted that EIDA repositories, though being international and domain specific, are in fact national repositories (as per EIDA’s distributed nature).

## Policy perspectives – Domain Specific vs. Generic, and Restricted vs. Open

SUBMERSE is currently in continued dialogue with EPOS and EIDA on submitting data to their different community repositories, depending on the processing state. Specifically, regarding continuous open waveform data, i.e., after scrubbing, decimation, and enrichment with metadata for permanent archival. As EIDA is a federated archive, the national EIDA nodes of Norway (U Bergen), Greece (NOA) and Portugal (under construction at IPMA) would be responsible for this archival task. If there are any issues, e.g., EIDA nodes not being ready in time, other interested EIDA nodes, e.g., GFZ[[24]](#endnote-25) can possibly take over.

Other downstream data products (oceanographic parameters, data on marine mammal movements, earthquake detections) will be forwarded to appropriate community-specific computer and data centers, e.g., Copernicus Marine Services or EMSC (European Mediterranean Seismological Centre). I.e., data flow from each national SUBMERSE site can, after being processed/scrubbed, also be submitted to other national or international repositories. These repositories can be domain specific or generic. The trade-off here is between domain specific advantages (e.g., metadata capabilities, API, machine accessibility, and exposure to relevant communities) on the one hand, and national autonomy and control (flexibility, scalability, ownership) on the other. Whichever repository solutions SUBMERSE user communities decides upon, the repository must be able to efficiently handle:

* On-the-fly (real time) submission of data, via an appropriate API, from a processed (scrubbed) data-stream. This includes automated filling-in of metadata templates and minting of persistent identifiers (PIDs)
* A periodically changing data-form and -content, due to changes in instrumentation and data processing. This entails periodically updating metadata templates and API submission and retrieval script.

In addition to uploading scrubbed/processed data to EIDA and others domain specific repositories, it can also be considered to seek increased utilization of the raw data. I.e., the non-scrubbed, sensitive data from the FIFO data buffer, or the saved-on-request data storage, could be securely passed on to national high security computer and data centers. Vetted researchers and/or approved software could be allowed access to the sensitive data. For evaluating the feasibility of such alternatives within the SUBMERSE project, the responsible compute and storage centers would be under the auspice of Sikt (Norway); INESTEC (Portugal); GRNET (Greece) and in dialogue with the SUBMERSE project’s *Security Advisory Committee*.

# FAIR Data in the SUBMERSE project and beyond

The FAIR data concept has been broadly accepted by the SUBMERSE project as something to be pursued in the interest of the SUBMERSE research community itself, in that it strengthens the understanding, control and utilization of its data, thereby supporting its own research and development agenda, in that by doing so it gets better science out of acquired data.

It is also acknowledged that FAIR SUBMERSE data is in the interest of a vast array of other research communities, that to varying extent utilize parts of the data the SUBMERSE instrumentation provides – now and in the foreseeable future. This FAIR SUBMERSE data support for other research communities is indeed part of the ethos of FAIR data, as it is also part of the EU Directive[[25]](#endnote-26) on open data and the re-use of public sector information. However, the SUBMERSE project makes no secret out of the fact that usage by and appreciation from other research communities is a welcome legitimisation of the SUBMERSE task at hand – delivering first class data generating instruments and infrastructure.

Most scientific communities seek agreement on FAIR guidelines – i.e., for increased Findability, Accessibility, Interoperability and Reusability of data (e.g., GO FAIR foundation, Research Data Alliance, European Open Science Cloud) for data access by researchers, as well as machine accessibility. We will not here elaborate further on the merits of FAIR data and machine accessibility, but only point to the origin of the FAIR[[26]](#endnote-27) data principles, and the GO FAIR[[27]](#endnote-28) interpretation guidelines. However, we clarify some important points:

* **FAIR data does not mean, nor implies Open Data**. Data that is closed and subject to restrictions can and should possibly just as well also be FAIR. The SUBMERSE data is to some extent closed, and restricted, but nevertheless FAIR, in that it possesses Findability, Accessibility, Interoperability and Reusability – albeit sometimes only for the selected few.
* **Not all data can or should be FAIR**. Making data FAIR implies work effort, which in many cases, is not worth the effort. Data is abundant and one can quickly drown in unintelligible data, a point at hand in the SUBMERSE project aiming at identifying the data that indeed has value to somebody – albeit fluctuating over time and research discipline. The SUBMERSE raw data that is continuously streamed in vast quantities, needs to be processed (on the fly, or just afterwards from current buffer) to have meaning. Most data are then discarded, due to processing, storage, and economic restraints.
* **Data is not either FAIR, or not FAIR.** One speaks of degrees of *FAIRness*, based on a case-by-case, research-intersubjective rational for pursuing FAIR data in the first place. Researchers continuously evaluate the extent it is feasible and possible to make the available data also FAIR data, and to which extent. We speak about *FAIRness*, and when we aim to make data FAIR, we work within a *FAIRification process.*
* **Data *FAIRness*, and** **the *FAIRification process* is continuously changing**. Based on utilisation estimates, technical and economic limitations, and the scientific understanding of the day, data, data FAIRness, and the FAIRification process evolves. Only data which is deemed “*valuable at the time, in light of resource costs*” is to be kept, submitted to repositories, and given Persistent Identifiers (PID) – ready for publications or utilisation by others.

Generic FAIR guidelines/rules exist, but it is up to the specific community to agree on their own *FAIRification* rational, realistic levels of ambition, set rules, standards, and guidelines regarding the various types of data.

In this document we lay out the data workflows and data specifications at this current point in time. We aim to manage the specific SUBMERSE data[[28]](#endnote-29) as best we can – only until we have a better plan.

[](https://deic.dk/en/data-management/GO-FAIR/The-Danish-GO-FAIR-Office)Figure 5.1: FAIRification process[[29]](#endnote-30)

By embracing the FAIR data concept and by applying an evolving *FAIRification Process,* valuabledata is securely, easily, and fast to be accessible for a maximum number of researchers – security constraint taken into consideration.

Defining what is or is not valuable data, in the light of resource costs and research communities shifting focus, is an integral part of the SUBMERSE project. Storing data in FAIR repositories presumes deciding on several sub-steps in a *FAIRification Process*, as outlined in Figure 5.1. For this we need the research communities to agree on data types, metadata, protocols, API’s etc. It’s a process of consensus-building and convergence. An example of such is Geo-INQUIRE[[30]](#endnote-31), pursuing for the integration of data, and applying specific tools[[31]](#endnote-32) across communities.

In the following sections, we will describe the details and implications of the numerus decisions the *FAIRification Process* entails – in the version 1 of the SUBMERSE DMP starting with the various sub-principles of the overall FAIR principles. In later versions of the SUMBERSE DMP we will turn more to the *FAIRification Process* and investigate possible and feasible FAIR implementation profiles for real time machine accessibility.

Some of the FAIR sub-principles are already clear, whereas others are pending crucial data management decisions, that are to be worked out in the duration of the SUBMERSE project, since the technology and its use-cases are still new and under construction.

In the following we address the individual parts of the FAIR Principles[[32]](#endnote-33), as outlined by GO FAIR.

## Making data Findable (F)

The workflow for the collected data is complicated, in that data is initially unintelligible, streamed endlessly, in overwhelming amounts. Making it all FAIR is neither technically or economically doable, feasible, nor scientifically useful. The SUBMERSE project is about analysing (processing) data on-the-fly (or very soon thereafter via buffering) and defining data that is worthy of submission to a repository, and hence in need of being referenced or searched, and found at a later stage.

**F1: (meta)data are assigned a globally unique and persistent identifier**

A dataset obtains findability via a unique persistent identifier (PID). It is necessary to isolate where – at which stage in the workflow – the data in question can become a data product and deemed worthy of *findability* in the FAIR data context. Here a PID is assigned to the dataset, as one crucial metadata element among many.

One of the repository options for SUBMERSE use-cases is the repository connected to the Copernicus[[33]](#endnote-34), an Earth observation component of the European Union’s Space programme, which however, is not automatically assigning PIDs to the uploaded datasets. The Data Lifecycle task (T1.4) will work towards a solution that fits both the requirement and needs from the connected research community as well as the FAIR requirements of SUBMERSE regarding PID’s and metadata.

**F2: data are described with rich metadata (defined by R1 below)**

Rich metadata is an essential tool to increase findability of datasets. This includes generic metadata, as well as domain specific metadata. Data "worthiness" implies adding well defined metadata before submitting data to a repository, as opposed to moving it down stream for further processing or deleting it. The data in question and its "worthiness" will evolve over time and research domains utilisation of SUBMERSE. So will also its metadata, through standardisation, consistency, and interoperability, which are always centre stage. DataCite's Metadata Schema[[34]](#endnote-35) will be the point of departure. Additional enrichments will evolve, in dialogue with community peers.

For DAS data, there is already a metadata standard existing[[35]](#endnote-36). T1.4 aims to work towards a correlating a metadata standard for SOP data collection. The metadata schema will include instrument specific metadata as well as information on space and time. The metadata will be assigned in different formats (XML and NetCDF[[36]](#endnote-37)), to satisfy the standards from the different research communities.

**F3: metadata clearly and explicitly include the identifier of the data it describes**

Since the metadata and the dataset, they describe likely will be separate files, a mutually linked association between a metadata file and the dataset must be made explicit by mentioning the dataset’s globally unique and persistent identifier (PID) in the metadata itself. This PID originates from F1 above.

**F4: (meta)data are registered or indexed in a searchable resource**

Metadata and datasets are indexed and searchable directly in the repositories search engine immediately after publishing. Furter proliferation of indexed (meta)data is aimed for, but rests with the repository at hand. Hence this FAIR principle and the choice of repositories is of strategic importance to optimal usage and impact of SUBMERSE data.

## Making data Accessible (A)

SUBMERSE will make use of several community specific repositories, taking into consideration the need to restrict access and/or filter data for certain bandwidths. These choices will be determined over the course of the project, regarding sensitive data in dialogue with national security agencies.

SUBMERSE, dealing with large amounts of data, will aim for having its metadata machine accessible, to allow for automated queries through APIs, following community standards. However, again regarding sensitive data, this will be in dialogue with national security agencies.

A FAIRification process will be established at each point in the data flow where data is worthy of submission to a repository. Appropriate arrangements with the identified repositories will take place regarding automated data submission.

**A1: (meta)data are retrievable by their identifier using a standardized communication protocol**

Metadata for individual records as well as files, file collections (datasets) must be harvestable using low-barrier mechanism for repository interoperability, e.g. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[[37]](#endnote-38). Metadata must also be retrievable through the public REST API. Again, choice of repositories is of strategic importance to optimal representation, usage and impact of SUBMERSE data.

Data will either be deposited in the already existing community specific repositories or in national repositories (see chapter 4, Submerse Data Policy). It will be a part of the workflow created within the project to analyse possibilities and best practice for individual repositories, community specific or national. Repositories must ensure that SUBMERSE data is assigned an identifier (PID) that resolves to a digital object, that is machine accessible.

Metadata, as opposed to data, can be made more openly available and e.g. licenced under a public domain dedication CC0. i.e. access to metadata is less (but possibly still) sensitive, than access to data.

Over the course of the project a standardized metadata schema will be created in collaboration with the technical and domain experts. The work will build up on previous work, e.g. already existing metadata schemas for DAS measurements. The same could be created for SOP measurements.

**A1.1: the protocol is open, free, and universally implementable**

In continuation of point A1, OAI-PMH and REST will be evaluated since they are open, free, and universal protocols for information retrieval on the web.

SUBMERSE data will be made available *as open as possible, and as closed as necessary*, taking into consideration necessary restrictions to be settled in dialogue with national security agencies. To provide data for e.g., tsunami-early-warning-systems, the project works towards releasing relevant data openly and (very) fast, through appropriate channels, which yet need to be identified.

A data embargo might be considered regarding restricted data, in dialogue with national security agencies. However, this is unlikely because it does not scale well due to the constant inflow of large amounts of new data. The data flow must become 100% automated from data’s point of origin to the repository, including all metadata, filtering, PID, licencing and possible access restrictions etc. I.e., human analysis/evaluation of data cannot happen other than being implemented in software that filters and sorts in real-time. Talks with the national security services of Norway, Greece and Portugal have been started up, possibly resulting in a standing Data Access Committee.

**A1.2: the protocol allows for an authentication and authorization procedure, where necessary**

All metadata might not be publicly accessible nor licensed under public domain. Most data will be open, though authorization will be necessary for access to other data, while some data will not be accessible even through authorization, short of via national security agencies.

Access level will be identified beyond the dichotomy of fully open and highly restricted. Intermediate access level might include institutional access via federated access models like eduGAIN[[38]](#endnote-39) providing an efficient, flexible way for participating federations, and their affiliated users and services, to connect to – some parts of – SUBMERSE data.

**A2: metadata are accessible, even when the data are no longer available**

Data and metadata will by default be retained for the lifetime of the repository. However, reassessment cannot be excluded, since one must be open to possible future budget restraints e.g. due to altered evaluation of the utility of SUBMERSE data.

Metadata takes up considerably less space than the data itself and can be kept available for a longer time independent of data amounts, cost dynamics and repositories policy and funding schemes. The potential value of metadata for data that no longer exist is unclear. Repositories do not delete metadata records, since there are PID’s resolving to that metadata, prompting an explanation of data absence.

SUBMERSE data must be long-term accessible through free and standardized access protocols, which is to be decided along with evaluation of repository options.

## Making data Interoperable (I)

Researchers across a multitude of disciplines should be able to exchange and interpret each other’s data. When applied to computers, it means that data should be readable for machines, without human interference or specialised translators or mappings. Through repository standards, open agreed-upon protocols, metadata vocabulary, the SUBMERSE project refers to a set of terms and concepts used to describe and organize the specific SUBMERSE data, resources, and other content.

**I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.**

The SUBMERSE project – like within all other research areas – has an intersubjective language and vocabulary. The language used must have a formal specification, i.e., the semantics and syntax must be consistent and defined in a precise way. The language must be shared and accessible so others – although distant research domains might need to learn the language. Especially the metadata language must be designed to be used across many research domains in many scenarios.

However, the SUBMERSE data will be utilising, thus dependant on, the repository choices to be made. Metadata templates might be lacking, and vocabulary might be inappropriate. Analysis of repositories will take this into consideration, and possible approaches will be compared, e.g. the use of JSON[[39]](#endnote-40) Schema as internal representation of metadata and the possibility of export to other popular formats such as Dublin Core[[40]](#endnote-41) or MARCXML[[41]](#endnote-42).

**I2: (meta)data use vocabularies that follow FAIR principles**

An important part of data interoperability lies in metadata and vocabulary standards, as well as file formats. The SUBMERSE project aims towards reusing existing vocabulary standards for generic metadata (Possibly Dublin Core for descriptive metadata and Prov-O[[42]](#endnote-43) for provenance metadata). There is a metadata standard available for DAS data acquisition: DAS-RCN[[43]](#endnote-44), which is recognized by the community. For the SUBMERSE project, we will be working on a comparable metadata standard for SOP data. These metadata standard will be discussed in the light of repository options.

When describing data or metadata SUBMERSE uses vocabularies that are adequate to represent SUBMERSE data content. Analysis will be conducted securing that used vocabulary is well defined and generally accepted – i.e. metadata is in itself also FAIR, so that machines can find, access, interoperate and reuse data. SUBMERSE aims to compile a carefully selected list of standardized terms and phrases used to index and retrieve SUBMERSE data – i.e. a *controlled vocabulary*.

The SUBMERSE community will seek to define the required FAIRness level of the vocabularies used. Preferably, the vocabulary and its terms/concepts have globally unique and persistent identifiers (F1) that can be resolved by all, providing a formal, accessible, shared and broadly applicable language for SUBMERSE knowledge representation (I1). Sources for certain terms in open, external vocabularies will be investigated.

**I3: (meta)data include qualified references to other (meta)data**

A good data model is the point of departure and significant time will be spent on revisiting and revising the data model – tiding up language and adding references, cross-references and meaningful links.

It will be specified how one dataset builds on another dataset, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. The scientific links between the datasets need to be described and clear also to researchers not part of the SUBMERSE community. Metadata will be defined and qualified by a resolvable URL.

Since the data is getting streamed to different domain specific repositories the SUBMERSE project needs to account for – if not resolve – the potentially different standards used by the research communities. In the first part of the project, focus will be on three different use-cases, that have their own metadata standards:

* Seismology (FDSN StationXML Schema[[44]](#endnote-45), metadata collected in StationXML or datalessSEED)
* Physical Oceanography (Copernicus Marine In-Situ NetCDF format[[45]](#endnote-46), metadata collected in NetCDF)
* Marine Biology

## Making data Reusable (R)

The SUBMERSE project will aim for publishing the workflows on data scrubbing via the project’s communication channels. The explanation of the data will be facilitated via the ontologies for DAS and SOP data and follow community standards.

Not all data will be made publicly available, due to the restrictions imposed by national security agencies. Such data might never be made publicly available, or there will be an embargo period imposed on the data. Talks have begun with the national security agencies of Norway, Greece and Portugal, and these talks will expand and iteratively conclude over the course of the project. Apart from this, all other data will be made openly available as fast as possible, though intermediary access models between fully open and highly restricted might evolve, possibly based on policy decisions from the SUBMERSE project Security Advisory Committee in cooperation with a standing Data Access Committee.

The responsible domain specific repositories will be keeping the data alive following their standard data storage policies. Since this project is aiming to build a new infrastructure for data collection, we are counting on new projects and in the long-term more sustainable funding opportunities for expanding data collection and storage.

**R1: (meta)data are richly described with a plurality of accurate and relevant attributes**

It will be much easier to find and reuse data if there are lots of metadata and surrounding descriptions attached to the data. Hence this FAIR principle is very much related to F2 (rich metadata) but focuses on the ability of machine or human to evaluate usefulness of data. In addition to the repository indexing enough metadata to allow for discovery (publishing), it must also describe the context under which the data was generated, including protocols, instrument descriptions, manufacturer, model number of machines or sensor that created the data. SUBMERSE will seek to be as generous as possible in providing metadata, even including information that may seem irrelevant.

However, it is recognised that SUBMERSE data discovery depends on repositories choice, in that several already existing domain specific repositories have their policies/limitations, like national generic repositories also do.

**R1.1: (meta)data are released with a clear and accessible data usage license**

Interoperability (I) is typically about technical or linguistic interoperability, but regarding reuse (R) of digital assets, we need to look at legal interoperability. Fortunately, this is a straightforward uncomplicated FAIR principle for SUBMERSE. Regarding the data provide openly, there are no legal constrains. The data that is restricted is subject to national laws governing espionage, treason etc., which is not something the SUBMERSE project will need to deal with, short of possibly being addressed by a possible Data Access Committee.

**R1.2: (meta)data are associated with detailed provenance**

Provenance of SUBMERSE data is about documentation and tracking of its origin, ownership, history, changes made to the data and its origin, who to acknowledge and cite. Related to this is the specifics of origin (sensors) but also data-processing, in that data is "marked", transformed by these specifics – introducing a scientifically valid debate about potential bias – of all sorts.

To show openness to these challenges, ideally, metadata will be included that can show acknowledgment of such potential challenges – workflows, hardware and software details, elaborations on scrubbing filtering process etc. – all described in a machine-readable format.

**R1.3: (meta)data meet domain-relevant community standards**

With metadata and repositories there is an ingrown dilemma between the generic and domain specific. While the domain specific is indeed preferable, it might not be up to par. For example, uploading data to the Copernicus[[46]](#endnote-47) repository does not automatically assign PIDs to the uploaded datasets. In contrast generic repositories, while often better than domain specific, they do not have the community standards (e.g. metadata vocabulary) and best practices for data archiving and sharing, which the domain specific often do have. One repository might be less formal, or even of lower quality, but nevertheless, publishing (meta)data in a manner that increases its use(ability).

The SUBMERSE project will seek to strike a balance, based on investigation, analyses, and strategic data management choices.

# Other research outputs

During the SUBMERSE project, other types of outputs will be made available, other than the standard instrumentation datasets made available for users. This type of data is categorized as “Other research outputs”. Due to the nature of the project, anticipated outputs, other than SUBMERSE data, could include journal articles, software and scripts used to scrub and pre-process the data, etc. These outputs relate to FAIR principle *R1.2: (meta)data are associated with detailed provenance*. It is an acknowledged challenge to make the data available to a wider audience, but nevertheless understood to be necessary to ensure maximization of the project’s impact.

For example, the European Open Science (EOSC) marketplace defines a simple taxonomy of research products categories[[47]](#endnote-48), which are aligned in many cases with the Resource Type vocabulary[[48]](#endnote-49) as proposed by the OpenAire guidelines for institutional repositories. In particular, the high-level categories include publications, datasets, software, services, data sources, interoperability guidelines, training, bundles, and other. The latter is further divided, among others, into collections, interactive resources, physical objects, and research objects. Research objects, aka FAIR digital objects, are semantically rich aggregations of data, methods and people in scientific investigations supporting the implementation of FAIR guiding principles and the systemic change of science practices to Open science. They describe and associate all this content together in a machine-readable mechanism so that it can be easily shared and exchanged. Hence, in line with the European Commission priorities and Open Science policy, research objects may be a relevant type of object for SUBMERSE project, particularly as open Science is concerned not only with making data FAIR but also other research outputs.

All other research outputs created within the SUBMERSE project will be made FAIR following the FAIR principles as described in chapter 5.

# Allocation of resources

The needed resources in the SUBMERSE project are distributed unequally over the course of the project. In the beginning, a large amount of data storage will be needed to train the newly developed AI models for scrubbing sensitive data. This will not be necessary at a later stage of the SUBMERSE instrument development as the AI gets better at instantly filtering/reducing the data amounts as well as improving data handling in real-time data flow. The short-term storage of the large amount of raw data as well as the long-term storage of raw data linked to published datasets will be handled by the national infrastructures (e.g. Sigma2 in Norway).

A collection of raw data containing examples of sensitive and insensitive measurements will allow us to propose a class of algorithms that will potentially be able to classify and filter this data. Based on this, we will be able to estimate processing time, latency and what computational resources are required.

Once the initial scrubbing algorithms for sensitive data are in place, raw data will still be buffered for three months, in case researchers express interest. Again, national infrastructures will provide the storage capacity for this buffer.

The expected amount of created raw data is very varying depending on many factors like gauge length and sampling rate and how many sensors are added for every site. We currently estimate in the order of 1TB to 10TB, per day, (per sensor) per location. For >3 months of buffer storage capacity totalling something in the order of anything from hundreds of terabytes to petabytes of storage per location – initially large, later lower as AI gets trained.

The long-term preservation is only planned for down-sampled datasets that are made FAIR and considered useful by researchers. The amount of data planned for long-term storage will be considerable smaller than what is collected daily as raw data. This storage will be provided by the domain specific repositories identified by the SUBMERSE project, as well as repositories of other research domains that may show interest in SUBMERSE data later.

# Data Security

Data Security is of exceptionally high importance in the SUBMERSE project. Even though the data may be dual use, the project’s aim is purely scientific. All sensitive data must be scrubbed from the raw data before it is made openly and FAIR available. To avoid any security risks, the raw data will remain in the respective country, where it will be scrubbed from any sensitive information before it will be sent to the domain specific repositories. The raw data will be stored securely for three months in a storage facility provided by the national NREN before deletion, as a natural part of the SUBMERSE data flow.

SUBMERSE has a separate task (T1.3 Ethics and Security) dedicated to guide the development of security policies that will ensure that no sensitive data s released or exposed. In addition, a security advisory board of the project gives guidance.

A backup is only provided for the scrubbed FAIRified data sets that will be preserved in the domain specific repositories for long term storage following their own policies.

# SUBMERSE Ethics

Ethics covers the whole range of the project execution, encompassing all phases of data lifecycle encountered, including but not limited to the creation, collection, processing, storage, use, sharing and dissemination, archiving, and destruction in full compliance with the SUBMERSE data management plan (DMP) and data policy of the project as well as with applicable legislation and frameworks in force. Moreover, towards the direction of meeting the goals associated with the successful adoption and application of the required ethics framework, measures will be in place to ensure the following: (i) effective data security, (ii) data protection & privacy, (iii) proper mitigation actions, and (iv) detailed, applicable remediation steps.

Furthermore, the strategy adopted and applied, starting from the identification, adoption, and use of the relevant ethical and legal requirements across the project. Indicatively, the methodology is a well-documented, tested, and certified one will be based, and build upon the structure in ISO27001, as outlined therein. Indicatively, the measures include, among others, the prevention of any attempt to gain unauthorised access to the systems and data, in close collaboration with and liaison with the activities associated with the definition and setup of the DMP. Towards the purpose of ensuring the security and protection of data and privacy, several activities will be in place as well, such as (i) the creation and maintenance of a risk register, (ii) the development of processes and procedures that will contribute to the achievement of the security objectives set, (iii) regular audits, preferably both scheduled and unannounced ones to ensure the developed safeguards are adopted and enforced, (iv) regular checks and evaluation as to the sufficiency of the aforementioned safeguards, and the (v) preparation and delivery of a periodic, half-yearly report, detailing the state of the security measures.

The CIA (confidentiality, integrity, availability) triad of information security will be taken into serious consideration. Finally, it is worth highlighting that a separate task, indicatively T1.3, is dedicated to ethics and security within the context of the SUBMERSE project and that the DMP task will closely collaborate with that task.

# SUBMERSE outlook - final remarks

We have described in this deliverable the data planning and policies for the datasets after scrubbing/down-sampling, thereby producing data "products", that can be submitted into well-established exiting data infrastructures (e.g. repositories) for the communities involved (e.g. EPOS). This is the case for the DAS datasets which are already used within this community. Further SUBMERSE "products" are envisioned.

Data Management policies for the raw sensitive data within the national facilities context are being developed within the project and will be included in the final version of the DMP at the end of the project.

The Security Advisory Board of this project will be included in further discussion of the security issues related to data collection in the SUBMERSE project. The final version of the DMP is also expected to include refined data management policies for the public SOP, Polarimeter and SOP-OTDR which are being produced and piloted within the project.

A next step preparing the next version of the SUBMERSE DMP is elaboration on the data security implications that limit the SUBMERSE instrument and therefore guides the evolution of the DMP.

This will be done in consultation with the Security Advisory Board.

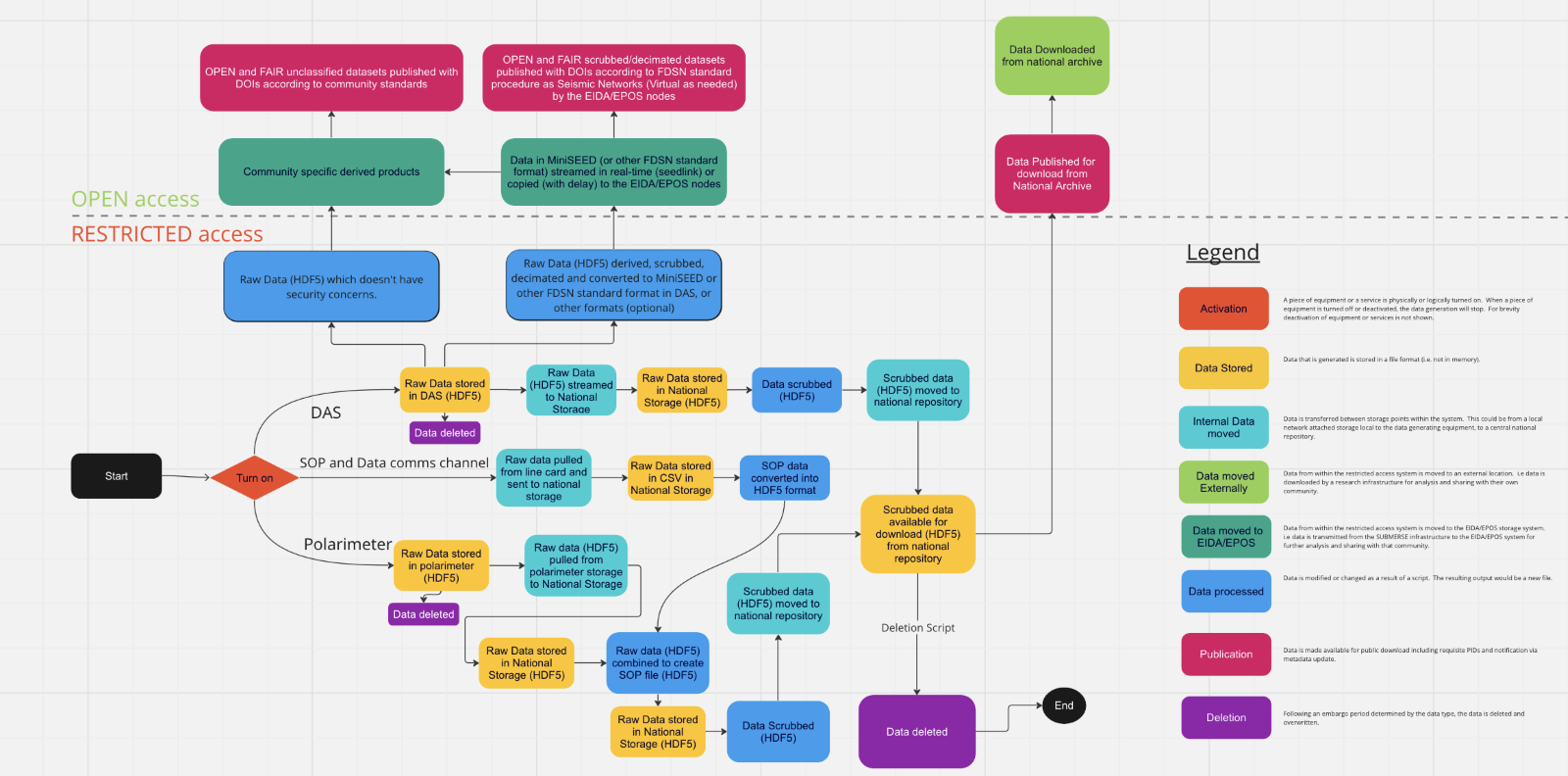
# Appendix A - Scientific Use Cases Surveys

Table A.1: Scientific Use Cases Surveys (answers received September-October 2023)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Marine Biology** | **Seismology** | **Physical Oceanography** |
| **Contact** | Martin Landrø, NTNU | Christos Evangelidis, NOA | Athanasia Papapostolou, HCMR |
| **1.** If your community is a part of the SUBMERSE project, to what site is the use case connected? | Ny Ålesund, Svalbard | As a geographical site data from SUBMERSE can densify existing seismic networks in the area especially in the offshore regions between Greece and Italy or nearby underwater volcanic edifices in the Aegean | HCMR is part of the SUBMERSE project, leading Task 3.3 on oceanographic applications. The Greek site will be used to develop the methods for evaluating and or validating the DAS derived oceanographic parameters (e.g. surface ocean waves properties). The end goal, however, is to produce methods that are applicable to any DAS site. |
| **2** What file formats do you handle your data today? |  | MiniSEED for seismic data in SDS (SeisComp Data Structure). Many DAS data rae handled and distributed in HDF5 |  |
| **2.a)** Are you using the direct formats produced by the instruments or are they converted? | Yes, hdf5 | Mostly yes since seismic instruments produce MiniSEED | HCMR, through the POSEIDON system, operates moored oceanographic buoys that could be used within the SUBMERSE context. The raw data from the buoy instruments are transmitted in real time and converted from the instruments’ data formats (proprietary encrypted format provided by the buoy’s manufacturer) to NetCDF files in real time as well. |
| **2.b)** Are there any plans in your community to replace the current file format with another.  If so, which file format(s)? | It is more a question of standardizing file format based on hdf5. | For DAS a new more compressed format is needed. Perhaps HD5 | This data flow is very well established and even in the raw data formats change in the future, i.e. if new instrumentation is deployed, the provision of NetCDF files in real time is highly unlikely to change, at least during the lifetime of SUBMERSE. |
| **3** What is your current scientific workflow? |  |  |  |
| **3.a)** Is the data getting changed during the workflow (format, is data discarded...)? | No | Up to now all seismic data are stored. | The raw data transmitted from the POSEIDON buoys are converted from the instrument’s data format to NetCDF format and then go through an automated real time quality control following the Copernicus Marine In Situ guidelines (<https://catalogue.marine.copernicus.eu/documents/QUID/CMEMS-INS-QUID-013-030-036.pdf>). Each data point is given a quality control flag following the Copernicus Marine in Situ conventions (QC=0: No QC was performed, QC=1: Good data, QC=2: Probably good data, QC=3: Bad data potentially correctable, QC=4: Bad data, QC=5: Value changed, QC=6: Value below detection, QC=7: nominal value, QC=8: Interpolated value, QC=9: Missing value), thus no data is discarded, it rather characterized as good or bad. |
| **3.b)** To what repository is your data uploaded? | Local disks at the university. | European Integrated Data Archive (<https://www.orfeus-eu.org/data/eida/>) | The data is uploaded to our data base and pushed directly to the “Mediterranean Sea- In-Situ Near Real Time Observations” Copernicus Marine Service product (<https://doi.org/10.48670/moi-00044>) |
| **3.c)** Will data (and/or metadata) automatically get a PID, license? | No | All seismic networks have a doi. PID for data is only attributed from some data centers for experimental EOSC projects | No |
| **3.d)** What are data access conditions to the data? | Restricted, all data needs to be checked prior to given a wider access. | Mostly open data. For some experiments and temporary deployments some embargo period may be possible | Open source free access data through the subscription to either the HCMR POSEIDON data base or the Copernicus Marine Service Database |
| **4** How fast do you usually get access to the data (almost real time, daily, monthly, quarterly)? | We can view data in real time, but massive amounts of data are transferred to local disks on a daily or weekly basis. | Almost real time |  |
| **4.a)** What would be the perfect timeframe for this? |  |  | Answered in 2-3. |
| **5** How much data gets passed through the workflow? | Approximately 4 TB per day. |  | All received data from the POSEIDON moored buoys gets passed through the workflow, this equals to 1GB/day |
| **6** How do you handle metadata? | We use the metadata setup provided by the provider of the interrogator. |  |  |
| **6.a)** Is there a standard metadata template? | yes | YES. Station XML (<https://www.fdsn.org/xml/station/>) | The data and metadata are formatted into netCDF (Network Common Data Form) files, following globally used format conventions and standards by OceanSites and Copernicus.    Format Information    Copernicus Marine In Situ TAC - physical parameters    Copernicus Marine In Situ NetCDF format manual    Standards that the above documentation is based on:    <http://cfconventions.org/>    <https://www.unidata.ucar.edu/software/udunits/>    <https://en.wikipedia.org/wiki/ISO_8601> |
| **6.b)** How is it collected? | It is not collected, the operator of the experiments has to fill in. | Each station operator is responsible to handle the correct metadata to the EIDA node that distribute this station | Answered in a |
| **6.c)** In what format is it collected? | Header information part of the hdf5-forma | StationXML or datalessSEED | Answered in a |
| **6.d)** What metadata fields are mandatory? | That varies, but as a minimum a short description of the experimental set up. | Many! Almost all fields in StationXML file or datalessSEED | Position (latitude, longitude), time, depth and/or pressure of measurements, the institution and the principal investigator responsible for the operation of the station, distribution statement, citation |
| **6.e)** Regarding security, the exact location of the sensor might not be feasible to make openly available. How accurate do you need to know latitude and longitude of the sensor position? | For the Svalbard case researchers that needs to know the exact position get access to this information. | It depends on the study. Roughly tens of meters accuracy are relatively ok (1-20m?) | The accuracy with which the latitude and longitude need to be defined is at least within 1-5 km, in order to use the output of the Mediterranean Sea Waves Reanalysis Model of the Copernicus Marine Service (<https://doi.org/10.25423/cmcc/medsea_multiyear_wav_006_012>) to complement the development of the methods for the oceanographic use case. |

# Appendix B - Data flow

Figure B.1: Data flow for DAS, SOP and Polarimeter data of data generating instruments.



# SUBMERSE Glossary

|  |  |
| --- | --- |
| **Acronym** | Definition |
| DAS | Distributed Acoustic Sensing |
| DMP | Data Management Plan |
| DOI | Digital Object Identifier |
| EIDA | European Integrated Data Archive |
| EMSO | The European Multidisciplinary Seafloor and water column Observatory |
| EOSC | European Open Science Cloud |
| EPOS | The European Plate Observing System |
| EU | European Union |
| FAIR | Findable Accessible Interoperable Reusable |
| FDSN | International Federation of Digital Seismograph Networks |
| NetCDF | Network Common Data Form |
| NREN | National Research and Education Network |
| OAI-PMH | Open Archives Initiative Protocol for Metadata Harvesting |
| OTDR | Optical Time-Domain Reflectometer |
| PID | Persistent Identifier |
| RCN | Research Coordination Network |
| REST-API | Restful Application Programming Interface |
| SOP | State of Polarisation |
| SUBMERSE | SUBMarine cablEs for ReSearch and Exploration |
| XML | Extensible Markup Language |

# References, Footnotes & Endnotes

1. Distributed Acoustic Sensor (DAS): xx [↑](#endnote-ref-2)
2. State of Polarisation (SOP): xx [↑](#endnote-ref-3)
3. National research and education network: <https://en.wikipedia.org/wiki/National_research_and_education_network> [↑](#endnote-ref-4)
4. European Plate Observing System: <https://www.epos-eu.org/about-epos> [↑](#endnote-ref-5)
5. Copernicus Marine Service: <https://marine.copernicus.eu/> [↑](#endnote-ref-6)
6. The FAIR Guiding Principles for scientific data management and stewardship: <https://www.nature.com/articles/sdata201618> [↑](#endnote-ref-7)
7. The GO FAIR Principles: <https://www.go-fair.org/fair-principles/> [↑](#endnote-ref-8)
8. Distributed Acoustic Sensor (DAS): xx [↑](#endnote-ref-9)
9. State of Polarisation (SOP): xx [↑](#endnote-ref-10)
10. Marine Science (EMSO ERIC): <https://emso.eu/> [↑](#endnote-ref-11)
11. Polarimeter: <https://en.wikipedia.org/wiki/Polarimeter> [↑](#endnote-ref-12)
12. Geo-INQUIRE: <https://www.geo-inquire.eu/> [↑](#endnote-ref-13)
13. EMSO ERIC: <https://emso.eu/> [↑](#endnote-ref-14)
14. EMSO ERIC: <https://emso.eu/> [↑](#endnote-ref-15)
15. EIDA Distributed Data Archive: Strollo et al. 2021 and <http://www.orfeus-eu.org/data/eida/> [↑](#endnote-ref-16)
16. EPOS data policy: <https://www.epos-eu.org/data-policy-2018> [↑](#endnote-ref-17)
17. EIDA node operators: <https://geofon.gfz-potsdam.de/eas/privacy.html> [↑](#endnote-ref-18)
18. Data set example: <http://doi.org/10.14470/4A7563971328> (Jousset et al. 2022) [↑](#endnote-ref-19)
19. Digital Seismograph Networks (FDSN): <https://www.fdsn.org/webservices/fdsnws-dataselect-1.1.pdf> [↑](#endnote-ref-20)
20. stationXML: <https://www.fdsn.org/xml/station> [↑](#endnote-ref-21)
21. DAS-RCN group: <https://das-metadata.gitbook.io/das-metadata-standard-by-das-rcn/> [↑](#endnote-ref-22)
22. miniSEED: <http://docs.fdsn.org/projects/miniseed3/en/latest/> [↑](#endnote-ref-23)
23. eduGAIN: <https://edugain.org/> [↑](#endnote-ref-24)
24. German Research Center for Geosciences (GFZ): <https://geofon.gfz-potsdam.de/contact/> [↑](#endnote-ref-25)
25. Directive on open data and the re-use of public sector information: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024> [↑](#endnote-ref-26)
26. The FAIR Guiding Principles for scientific data management and stewardship: <https://www.nature.com/articles/sdata201618> [↑](#endnote-ref-27)
27. The GO FAIR Principles: <https://www.go-fair.org/fair-principles/> [↑](#endnote-ref-28)
28. I.e., data from Ocean-Bottom Distributed Acoustic Sensors (DAS) and State of Polarisation (SoP); (xx And 1 or 2 more?) [↑](#endnote-ref-29)
29. FAIRification process: <https://deic.dk/en/data-management/GO-FAIR/The-Danish-GO-FAIR-Office> [↑](#endnote-ref-30)
30. Geo-INQUIRE: <https://www.geo-inquire.eu/about/about-geo-inquire> [↑](#endnote-ref-31)
31. E.g. the F-UJI tool: <https://www.f-uji.net/> and FAIR implementation profiles [↑](#endnote-ref-32)
32. the FAIR Principles: <https://www.go-fair.org/fair-principles/> [↑](#endnote-ref-33)
33. Copernicus: <https://www.copernicus.eu/en/about-copernicus> [↑](#endnote-ref-34)
34. DataCite Metadata Schema: <https://schema.datacite.org/> [↑](#endnote-ref-35)
35. DAS metadata standard homepage (DAS-RCN): <https://das-metadata.gitbook.io/das-metadata-standard-by-das-rcn/> [↑](#endnote-ref-36)
36. The Network Common Data Form (NetCDF): <https://nsidc.org/data/user-resources/help-center/what-netcdf#:~:text=NetCDF%20(network%20Common%20Data%20Form,format%20for%20representing%20scientific%20data>. [↑](#endnote-ref-37)
37. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): <https://www.openarchives.org/pmh/> [↑](#endnote-ref-38)
38. eduGAIN: <https://edugain.org/about-edugain/what-is-edugain/> [↑](#endnote-ref-39)
39. JSON Schema: <https://json-schema.org/> [↑](#endnote-ref-40)
40. Dublin Core: <https://www.dublincore.org/> [↑](#endnote-ref-41)
41. MARCXML: https://www.loc.gov/marc/marcxml.html [↑](#endnote-ref-42)
42. Prov-O: https://www.w3.org/TR/prov-o/ [↑](#endnote-ref-43)
43. DAS-RCN: <https://das-metadata.gitbook.io/das-metadata-standard-by-das-rcn/> [↑](#endnote-ref-44)
44. FDSN StationXML Schema: <https://www.fdsn.org/xml/station/> [↑](#endnote-ref-45)
45. Copernicus Marine In Situ NetCDF format: <https://archimer.ifremer.fr/doc/00488/59938/> [↑](#endnote-ref-46)
46. Copernicus: <https://www.copernicus.eu/en/about-copernicus> [↑](#endnote-ref-47)
47. EOSC Marketplace Resources: <https://search.marketplace.eosc-portal.eu/> [↑](#endnote-ref-48)
48. COAR Controlled vocabularies for Repositories: <https://vocabularies.coar-repositories.org/resource_types> [↑](#endnote-ref-49)