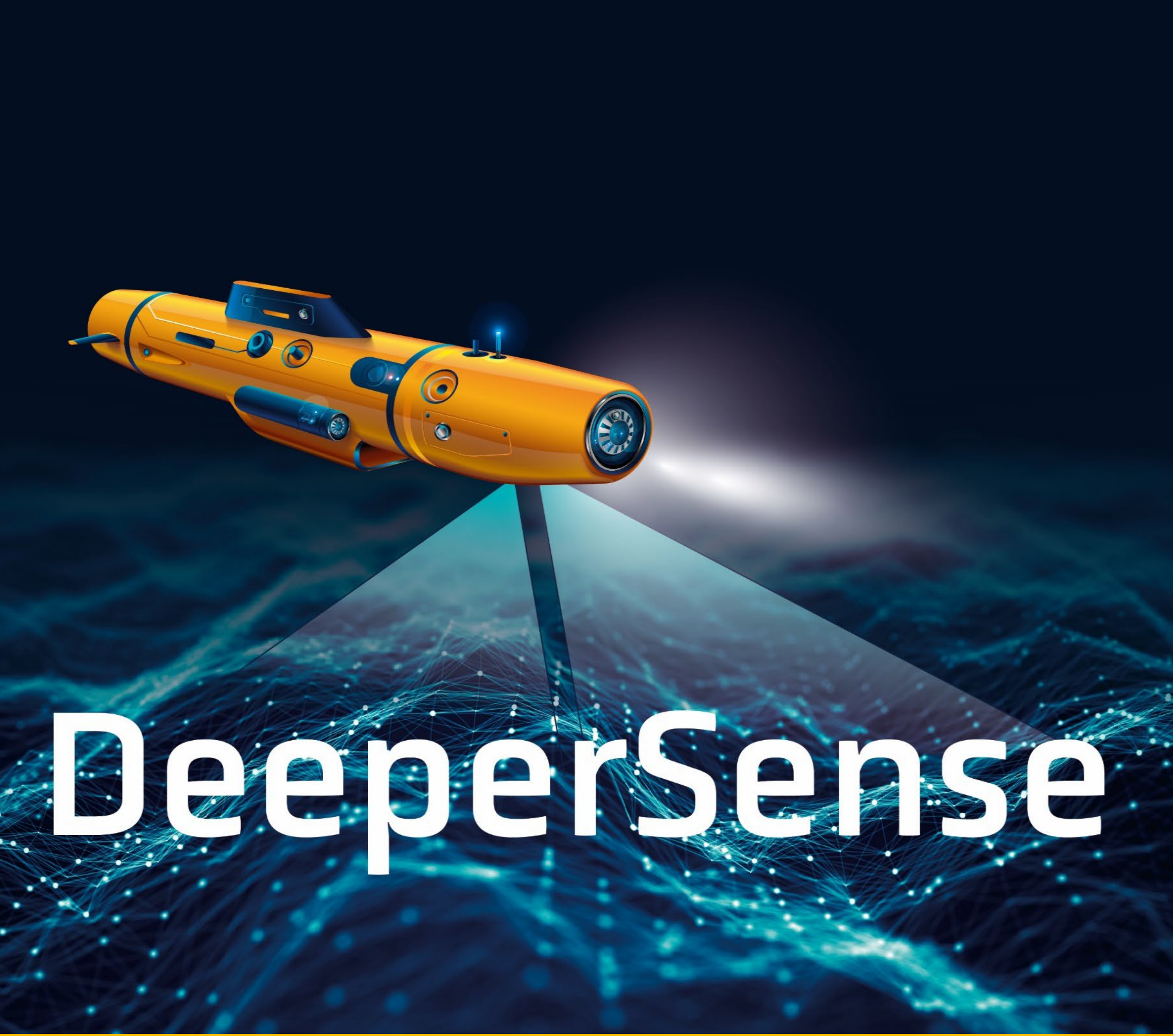


Training Data Concept

Deliverable number:	D2.4
Version:	1.4
Due date:	30/6/21
Nature:	Report
Dissemination Level:	Confidential/Public
Work Package	2
Lead Beneficiary:	UH
Contributing Beneficiaries	ALL

Document History

Version	Date	Author	Description
Version 1.0	15.05.2021	Tali Treibitz	First Draft
Version 1.1	10.06.2021	Bilal Whebe	Added relevant info
Version 1.2	15.06.2021	Nuno Gracias	Added relevant info
Version 1.3	29.06.2021	Nuno Gracias	Final review
Version 1.4	02.07.2021	Thomas Vögele	Peer-review and QC

Project Coordinator

Organisation:	DFKI, Research Department: Robotics Innovation Center
Responsible Person:	Dr. Thomas Vögele
Address:	Robert-Hooke Str. 1
Phone:	+49 17845 4130
e-mail:	thomas.voegele@dfki.de

Consortium

Participant name	Short name	Country
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	DFKI	Germany
Universitat de Girona	UdG	Spain
University of Haifa	UH	Israel
Kraken Robotik GmbH	KRA	Germany
Bundesministerium des Inneren	THW	Germany
Israel Nature and National Parks Protection Authority	INPA	Israel
Tecno Ambiente SL	TA	Spain

Copyright

Table of Contents

List of Figures:

List of Tables

1. Executive Summary

As high-quality and high-volume training data are the key to the successful application of any ML algorithm, this document comes to detail concepts for the generation and collection of training data for the 3 algorithms that will be developed in this project. Training data will be derived mainly from three main sources:

Legacy Data: Both the DeeperSense end-users and research partners have done extensive underwater missions in the past. Available data from those missions will be examined and transformed into training data for the algorithms, if appropriate. In addition, if suitable, datasets published by other groups will be considered.

Real-World Data: The bulk of training data will be newly created during the project. Each group will gather it in the lab, available test tanks and/or in controlled and natural outdoor environments.

Synthetic Training Data: For all three algorithms, additional synthetic training data will be created using simulators such as the Stonefish simulator (Cieślak, 2019), developed at UDG. This tool supports intuitive simulations of marine robots in realistic scenarios.

The document details the training data generation plan for each use case. It also provides a set of rules intended to facilitate the management of the training data. Such rules are guided by the FAIR data principles, and address the topics of data formatting including naming and annotation, and data publication.

2. Use Case 1: Hybrid AUV for Diver Safety Monitoring (SONAVision)

2.1. Legacy Data

For the first usecase, several legacy and online available datasets will be used for initial development of the SONAVision algorithm. Although these datasets do not include any sonar or camera images of divers, they can still be used to perform similar multi-modal translation tasks or to evaluate self-supervised learning methods.

2.1.1. Marine Debris FLS Watertank Dataset

The Marine Debris dataset is an open dataset that consists of 1868 full sonar images of several marine debris objects. It was captured using a Forward-Looking Sonar (FLS), the ARIS Explorer 3000, at 3.0 MHz frequency.

Classes included are: bottle, can, chain, drink carton, hook, propeller, shampoo bottle, standing bottle, tire, and valve. Bounding boxes are annotated in all images, with annotations stored in a JSON file.

Full sonar images projected in cartesian coordinates are available as PNG files. Image crops for each class are also included, which can be easily used with keras' ImageDataGenerator or other frameworks. These crops vary with size, so they need to be normalized before training a model with them.

This dataset was generated under the Marie Curie ITN program “Robocademy” FP7-PEOPLE-2013-ITN-608096, at the Ocean Systems Lab Water Tank (Heriot-Watt University). The dataset is publicly available and can be found under the following github repository https://github.com/mvaldenegro/marine-debris-fls-datasets/releases/tag/watertank-v1.0. A few examples of this dataset can be shown in Fig. 1. We will use this dataset for self-supervised pre-training of neural network models.

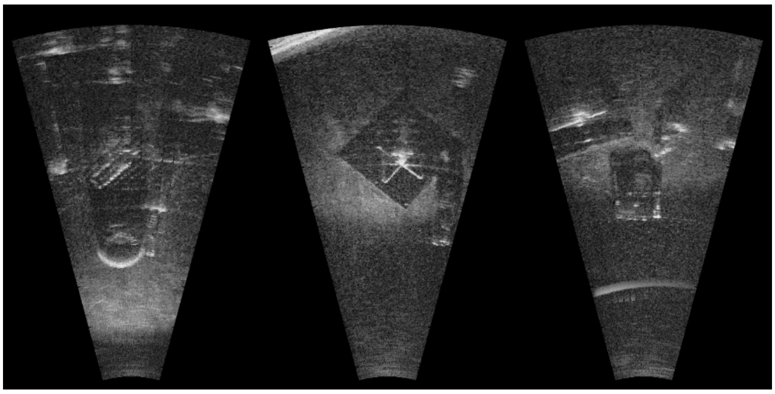
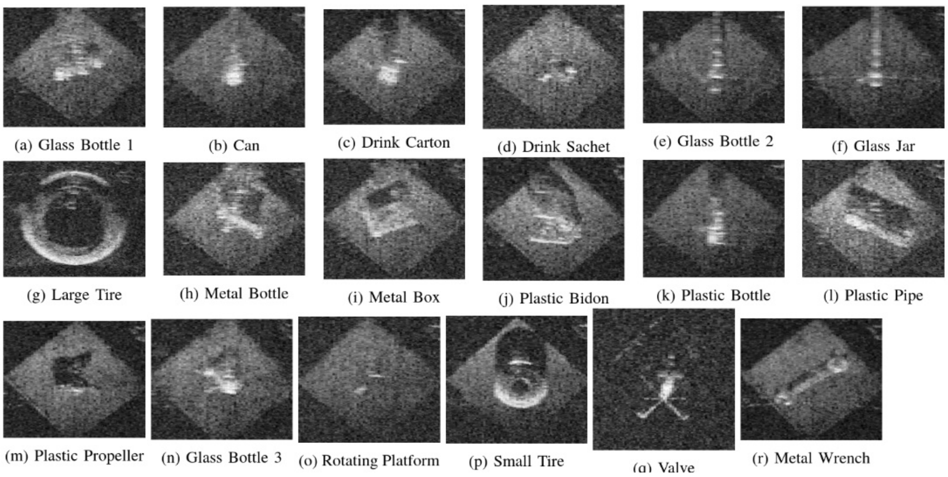


Figure 1. Examples of the Marine Debris dataset.

2.1.2. Marine Debris FLS Turntable Dataset

This dataset is similar to the Marine Debris Watertank dataset, but the capturing setup is different. It was captured by placing a rotating turntable underwater, and keeping the sonar sensor fixed. Multiple views of each object are captured, as the turntable rotates from –180 to 180 degrees in the Z axis. Classes can be seen in Fig. 2.



(a) Glass Bottle 1
(b) Can
(c) Drink Carton
(d) Drink Sachet
(e) Glass Bottle 2
(f) Glass Jar
(g) Large Tire
(h) Metal Bottle
(i) Metal Box
(j) Plastic Bidon
(k) Plastic Bottle
(l) Plastic Pipe
(m) Plastic Propeller
(n) Glass Bottle 3
(o) Rotating Platform
(p) Small Tire
(q) Valve
(r) Metal Wrench

Figure 2. Examples of the Marine Debris FLS Turntable Dataset

The dataset is available in the same address as the Watertank dataset. This dataset will be used for pre-training convolutional neural network models for transfer learning, and for evaluation of self-supervised learning methods.

2.1.3. Other online available datasets

The following are open-source publicly available datasets for preliminary testing of the developed algorithms. None of the partners in DeeperSense consortium were involved in collecting or generating these datasets. Both datasets mentioned below will be used for preliminary evaluation of the SONAVision algorithm as they consist of pairs of images (acoustic and optical) that can be directly integrated into image-to-image translation tasks.

2.1.3.1. Fish Monitoring Dataset

The Fish Monitoring dataset [Terayama et al., 2019] is an open-source dataset consisting of pairs of camera and FLS images which captures schools of sardine fish in an aquaculture tank at the Saikai National Park Kujukushima Aquarium, Nagasaki, Japan.

This dataset was collected using an underwater optical camera (GoPro HERO4) placed at the bottom of the tank, and an imaging sonar (ARIS EXPLORER 1800) from the side of the tank. A sample of this dataset can be shown in Fig. 3. This dataset is publicly available and could be found under the following link http://www.tsudalab.org/files/camera_sonar_dataset.zip .

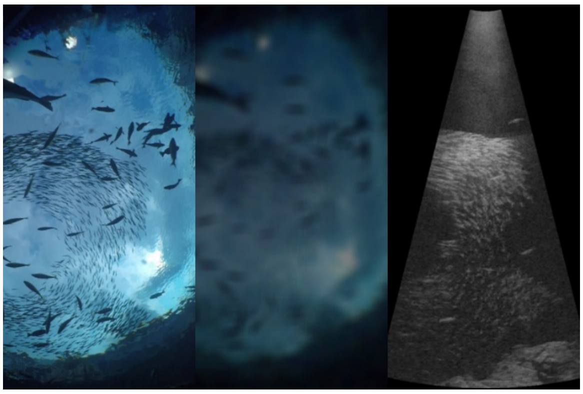


Figure 3. Example of the fish monitoring dataset where the left image shows the ground truth image recorded during day light, the middle image shows the darkened visual image to simulate nighttime, and the right image showing the corresponding sonar image.

2.1.3.2. Facades Dataset

The Facades dataset consists of 506 building facades and their corresponding segmentation images. This dataset is obtained from the original Pix2Pix [Isola et al., 2017] datasets available at UC Berkeley's official directory: https://people.eecs.berkeley.edu/~tinghuiz/projects/pix2pix/datasets/. A few examples of this dataset are shown in Fig. 4.

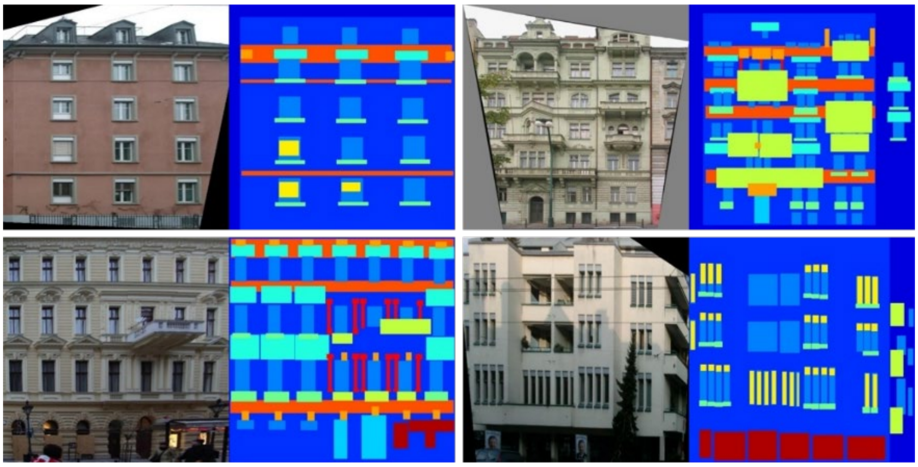


Figure 4: Example of the facades datasets showing several images of buildings facades with their corresponding segmentation image.

2.2. Real-World Data

To create the necessary training data for UC1, a number of experiments will be set up at the Maritime Exploration facility of DFKI in collaboration with KRA and THW. In this section, a plan detailing the logistics of this experimental setup is presented.

2.2.1. Acquisition Platforms

In the training data collection phase, a sensor rig will be constructed by DFKI, where different sensors could be easily attached to. The rationale behind constructing such rig is to have a mobile floating platform that is easily accessible and deployable, where any kind of sensor could be mounted onto with a desired configuration. This would also make the placement of sensors easily reconfigurable and facilitate achieving an optimal viewpoint configuration between the different visual and acoustics sensors, while capturing footage of the divers.

The sensor rig will consist of a floating platform made up of several floating pontoons, which are held together by an aluminum-profiles structure, as shown in Figure 5. The figure shows a CAD model of the platform to be used to record sonar and camera images of underwater objects. The left image corresponds to the top view and the right image corresponds to the bottom view, showing how a sensor will be placed in order to record required training data. At the bottom side of this structure, a perforated plate is attached that allows any kind of sensor to be mounted to. A splash-proof box will be mounted on top of the platform, which will house all the necessary electronics and computers for logging and synchronizing the data.

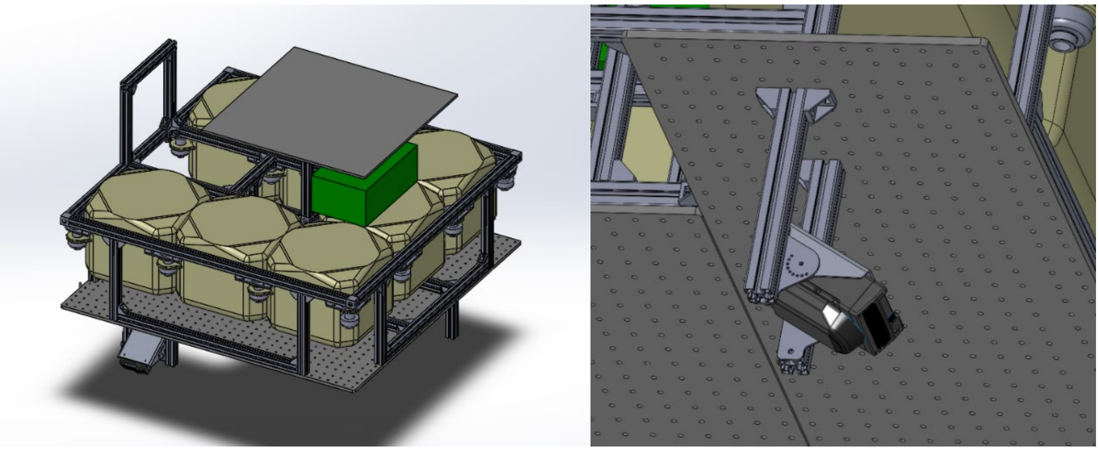


Figure 5: CAD model of platform to record sonar and camera images of underwater objects.

A specific frame will be developed to hold the Seavision bottles that houses the RGB lasers and camera. As described in D2.2, this sensor can be used as a 3D laser scanner as well as a color camera which can deliver dense full color 3D point cloud images of subsea infrastructure. The technical details of the SeaVision bottles can be found in D2.2. This will have the possibility to change the geometric configuration of Seavision bottles in order to test different mounting options. Seavision's frame will be compatible with the sensor rig aforementioned, but it will also allow to be deployed from the basin walls, since the sensor resolution would be affected when scanning from the floating structure, due to the distance to the floor. To synchronize SeaVision and the other sensors, the data will be logged using standard Unix timestamp (with resolution of a millisecond).

2.3. Locations

2.3.1. DFKI's Maritime Exploration Hall

The collection of the training data for Use Case 1 will mainly take place at the Maritime Exploration Hall[1] at DFKI in Bremen, Germany. The infrastructure comprises of a large water basin with the dimensions of 23 m x 19 m x 8 m, that contains 3.4 million liters of saltwater (18 g of salt per liter) (Fig. 6). Two crane systems (12.5 t and 250 kgs) are installed over and to the side of the basin that can be used to immerse systems and test objects into the basin.

The large basin was selected to be the main facility where most of the training data for Usecase 1 will be collected. This choice was made due to the clear visibility conditions in the basin that will allow the visual sensors to produce good quality data that can be used as labels in the training process.

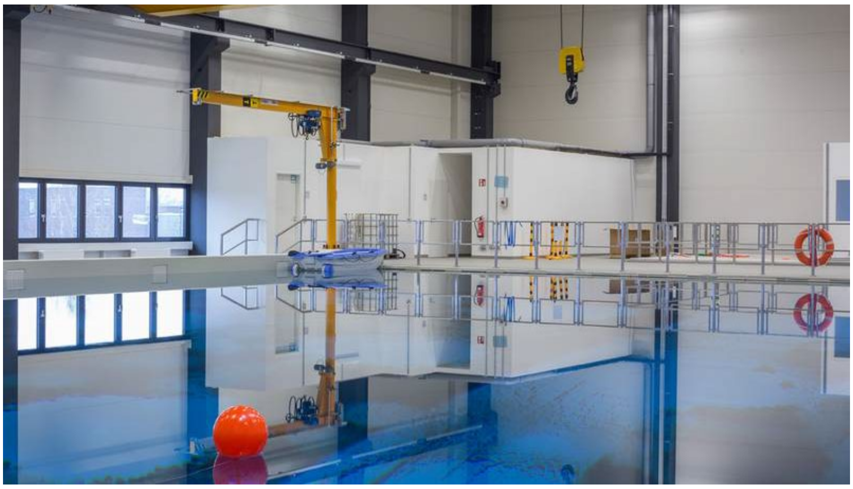


Figure 6: The water pool at the Maritime Exploration labs of DFKI in Bremen, Germany.

2.3.2. Scenarios

To recreate the scenery of Use Case 1, a mock-up environment will be setup in the water basin where the divers can perform their regular underwater tasks (Fig. 7). In the training data collection phase, the divers will only perform activities that do not create fine particulate matter or induce turbidity in the water, such as manipulating objects, drilling or cutting wood, etc. Fiducial markers will also be added to the scene to ease the co-calibration between different sensors.

Based on the sensor pairing concept that was presented in D2.2, a multitude of cameras and sonars will be attached onto the floating platform to record the divers' activities. Additionally, there will be the option to also mount different sensors onto the walls of the basic suspended from aluminum profile structures. The SeaVision sensor will be mounted on a frame which can be either be attached to the sensor rig or deployed on the basin

[1] https://robotik.dfki-bremen.de/en/research/research-facilities/maritime-exploration-hall.html

walls when required to be closer to the scene. When mounted onto the basin wall, a sonar sensor can be attached to the frame so both sensors can see the scene from the same perspective.

The training data will be collected across 4 multi-day sessions. During these sessions, the divers from THW will perform their designated tasks using real equipment to generate the necessary training data. The data collection sessions will be mainly hosted at the premises of DFKI, where the water basin described above will be the main testing environment. Additionally, optional data collection sessions might be hosted at a lake in the vicinity of DFKI, if necessary.

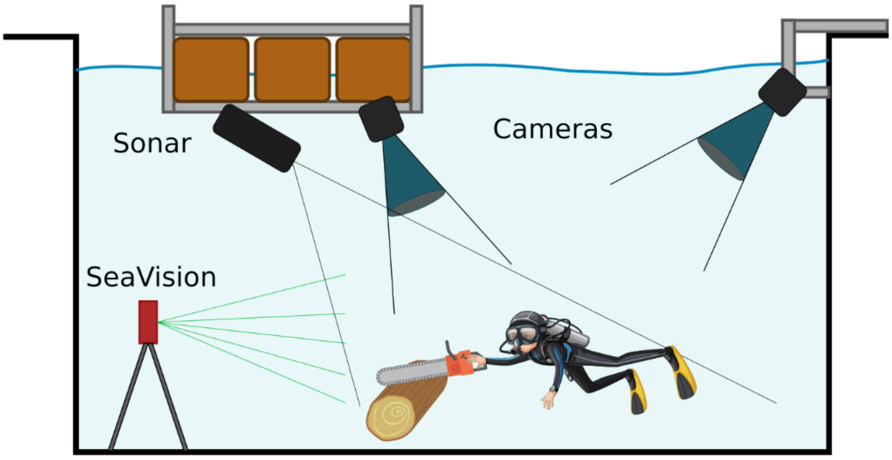


Figure 7: A schematic showing the scenario for collecting the training data for Use Case 1.

2.4. Synthetic Training Data

The Stonefish simulator [Cieślak, 2019] was selected for UC1 due to 2 main features. Firstly, it provides a realistic visual rendering of an underwater environment including effects such as light absorption and scattering, as well as suspended particulate matter known as marine snow. The second feature is the incorporation of visual sensors such as color and depth cameras and well as a simulation of a forward-looking sonar.

For the usecase of diver monitoring, a simple environment will be setup using the Stonefish simulator where human-shaped 3D meshes will be created to simulate divers as well as meshes of various objects that divers may use or operate on (Fig. 8).

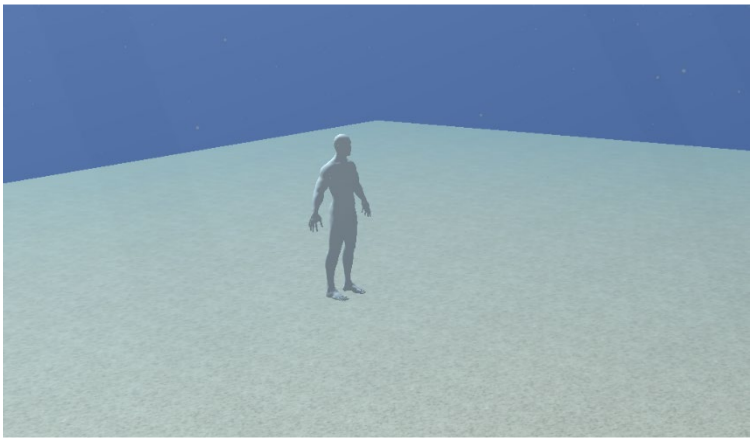


Figure 8: A simple example of a simulation environment for diver monitoring.

3. Use Case 2: Surveying and Monitoring Complex Benthic Environments (EagleEye)

3.1. Legacy Data

This algorithm requires images acquired by a co-located forward-looking camera (FLC) and forward-looking sonar (FLS). Such dataset is not available. Because of the scarcity of available relevant data we will try to use other datasets as much as possible.

3.1.1. Marine Debris FLS Watertank Dataset

This is the dataset described for UC1, in section 1. As in UC1, this dataset might be of use for some steps in the initial training.

3.1.2. Crossview USA (CVUSA)

This is a large dataset containing millions of pairs of ground-level and aerial/satellite images from across the United States (http://mvrl.cs.uky.edu/datasets/cvusa/). An example is shown in Fig. 9. Although the images were all acquired using a camera, they contain different viewpoints: front and top, similar to our case and it might be useful for some steps in the training.

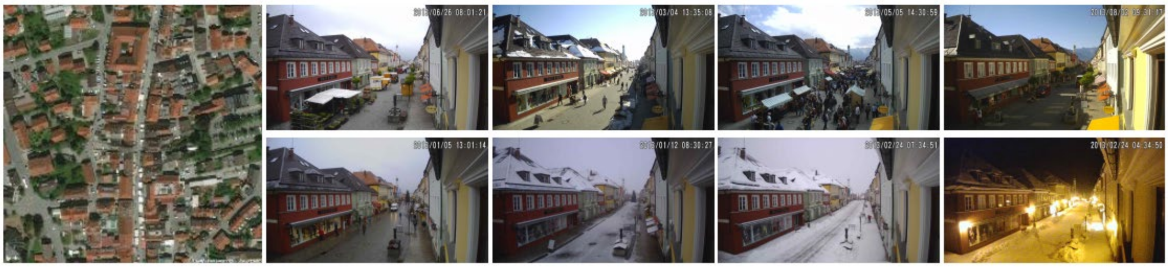


Figure 9. Examples from the Crossview USA dataset.

3.2. Real-World Data

3.2.1. Acquisition Platforms

Real-world data will be gathered with the UH SPARUS II AUV (named ALICE) that is equipped with co-located FLS-FLC sensors (FIG. 10 left). In the AUV, a full frame camera is installed as the FLC, and a BLUEVIEW M900 as FLS. In addition, BlueRobotics ROV (Fig. 10 right) which will be equipped with an FLS that was purchased for this project (Blueprint M1200d). As described in D2.2, this FLS has a particularly high resolution, which makes it suitable for working with images. The camera on the ROV was upgraded to an IDS UI-3260CP with a sensor of 1/1.2" with auto-focus.

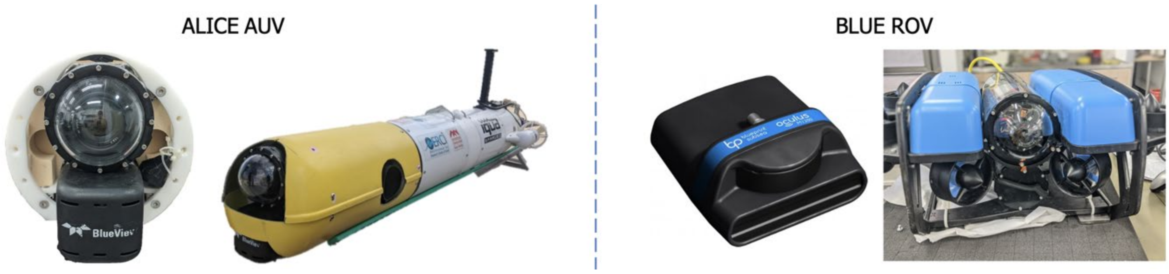


Figure 10. Acquisition platforms of UH.

3.2.2. Locations

Data will be gathered in three main locations, as shown in Figure 11:

- In the UH pool.
- In the Mediterranean off the coast of Haifa.
- In the Red Sea.

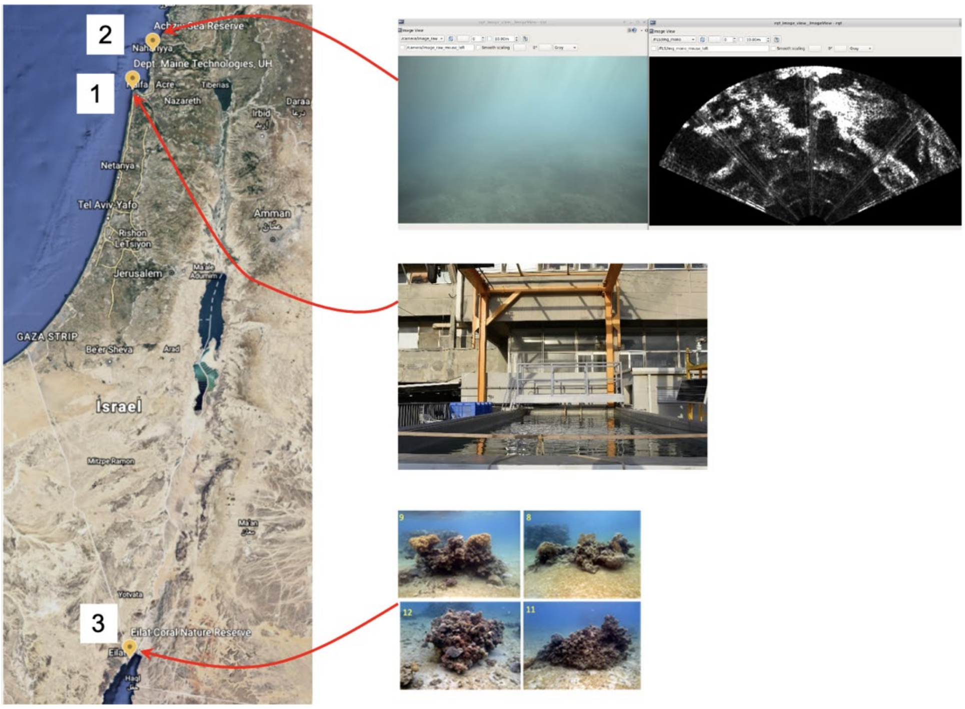


Figure 11. Acquisition locations.

3.2.3. Scenarios

We will collect scenes in a controlled environment with mock-up obstacles. Examples of planned mock-up obstacles include barrels and walls. In addition, image of real-world scenes will also be acquired, such as rocky reefs in the Mediterranean (e.g., Fig. 12), coral reefs in the red sea and man-made structures (e.g., piers) in both locations.

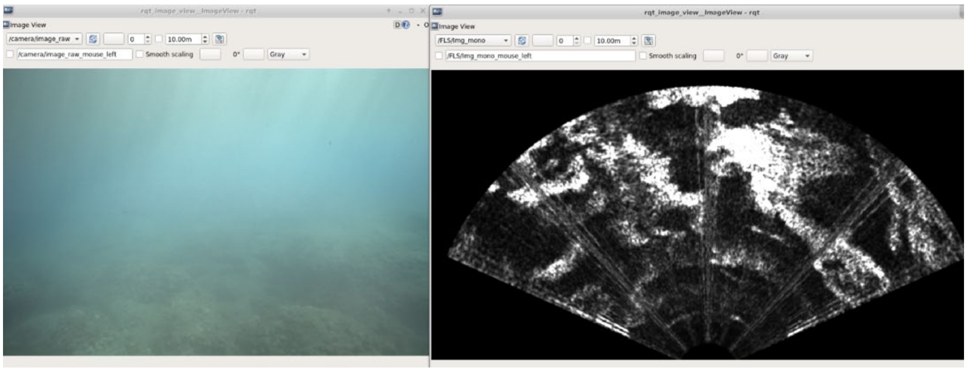


Figure 12. An example FLS-FLC set from a rocky reef in the Mediterranean.

3.3. Synthetic Training Data

3.3.1. Simulation Environment and Parameters

We are using the Stonefish simulator (Cieślak, 2019), as a tool to create synthetic training dataThe simulator requires the knowledge of the extrinsics parameters between sensors (relative positions and orientations) as input to the simulation. We will input the calibrated extrinsics of both our platforms.

The implemented FLS simulator currently does not have frequency input, i.e. the ability to generate acoustic images at different frequencies. We will consider changing the FLS resolution in post-processing. We will determine its FOV according to the sensors we are going to use.

For the camera images, we will control the following parameters: Field of view (FOV), resolution and exposure. Noise will be added in post-processing on top of the simulation to increase the level of realism. For the optical properties of the water in the simulator, we will use the Jerlov water types I-1C.

For scenes we will use the 3D models we already have (https://sketchfab.com/Marine_Imaging_Lab). An example is given in Fig. 13.

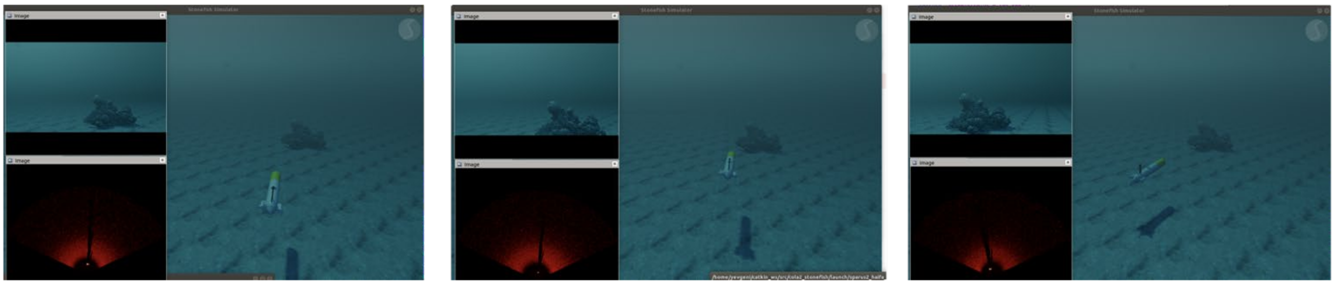


Figure 13. An example scene from an FLS-FLC simulation.

3.3.2. Simulation Output

The output of the simulation for every viewpoint (each data point in the dataset) will include the optical image, the FLS image, and a depth map of the scene from that viewpoint.

4. Use Case 3: Sub-Sea multisensor bottom mapping and interpretation for geophysics (SmartSeafloorScan)

4.1. Legacy Data

The datasets available for this UC are those provided mostly by TecnoAmbiente SL. Among these, four relevant datasets were acquired from geophysical surveys before the beginning of the project. The datasets consist of sonar data captured with a Klein 3000 Side Scan Sonar, optical images from ROV videos, and the interpretation generated by an expert in the form of a thematic map.

Three datasets were acquired within the Mediterranean Sea at a maximum depth of 80m and one in deeper waters in the Canary Islands, in the locations marked in Fig. 14.

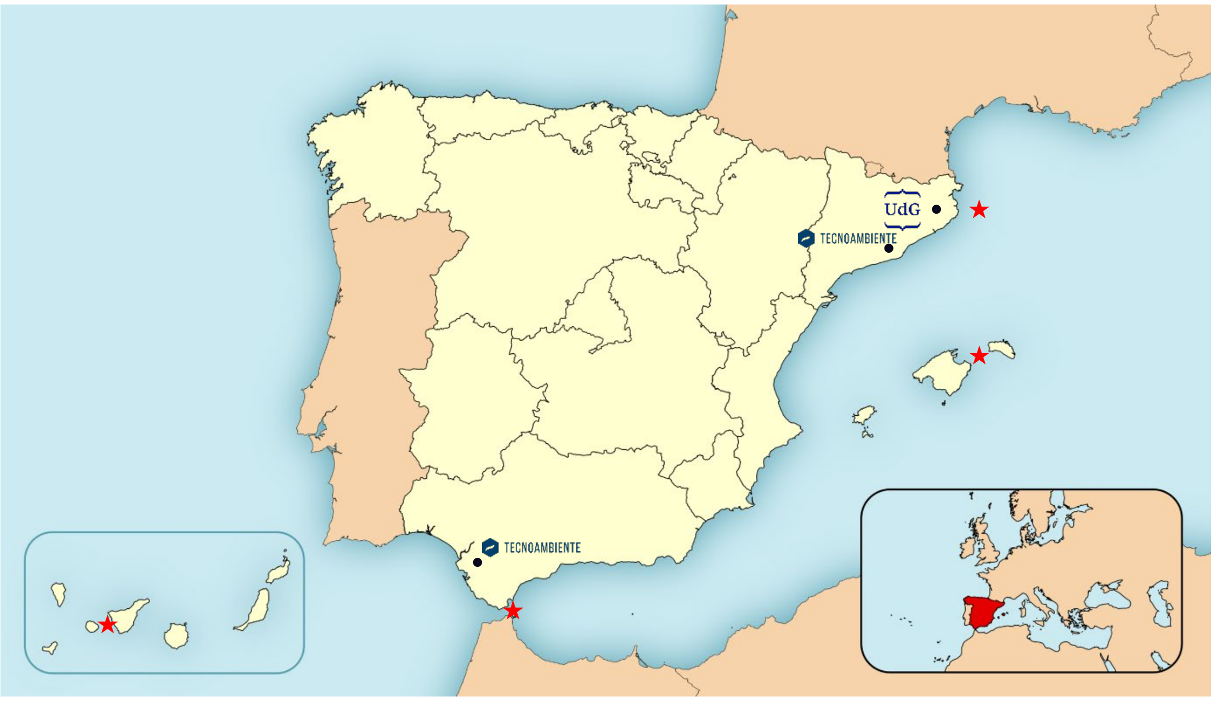
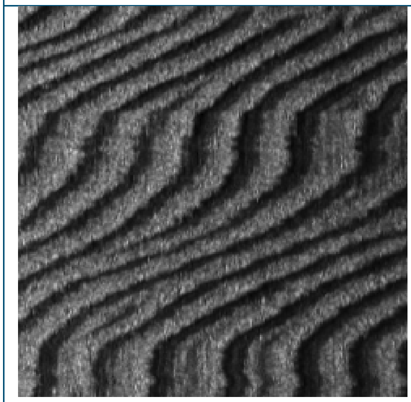
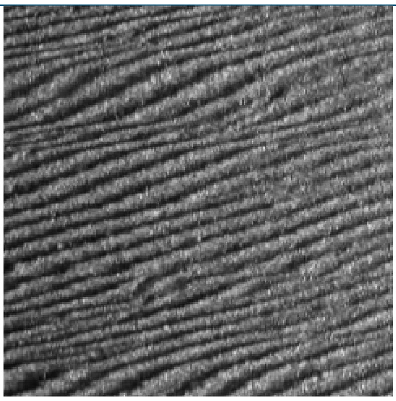
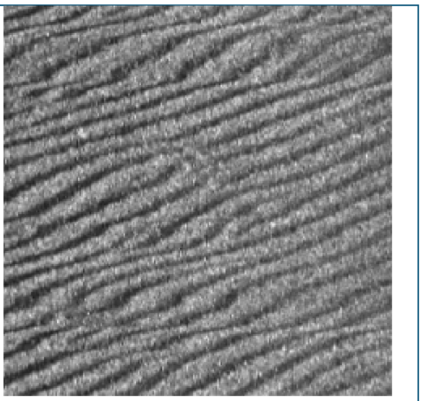


Figure 14. Locations where legacy data was collected marked as red stars.

Among the available datasets, one (referred to as *Project A*) is currently being explored and used for creating adequate training data. This dataset was acquired in the Balearic Islands, along the Menorca channel and comprises sidescan sonar data and short video transects.

There is a difference between the amount of existing data in sonar and in optical formats. While the long sonar transects cover extensions of kilometers with swath width of more than 40m per side (80m in total), the ROV videos transects are considerably shorter (ranging from 65m to 310m) and have images footprints covering 9 to 25m^2.

Three categories of sediments were identified, namely Sand Ripples, Silt and Rock, covering approximately 50%, 36% and 14% of the total area respectively. Examples of the same are depicted in Fig. 15.

Examples of sidescan patches of different and easily distinguishable bottom types		
Sand Ripples		
		

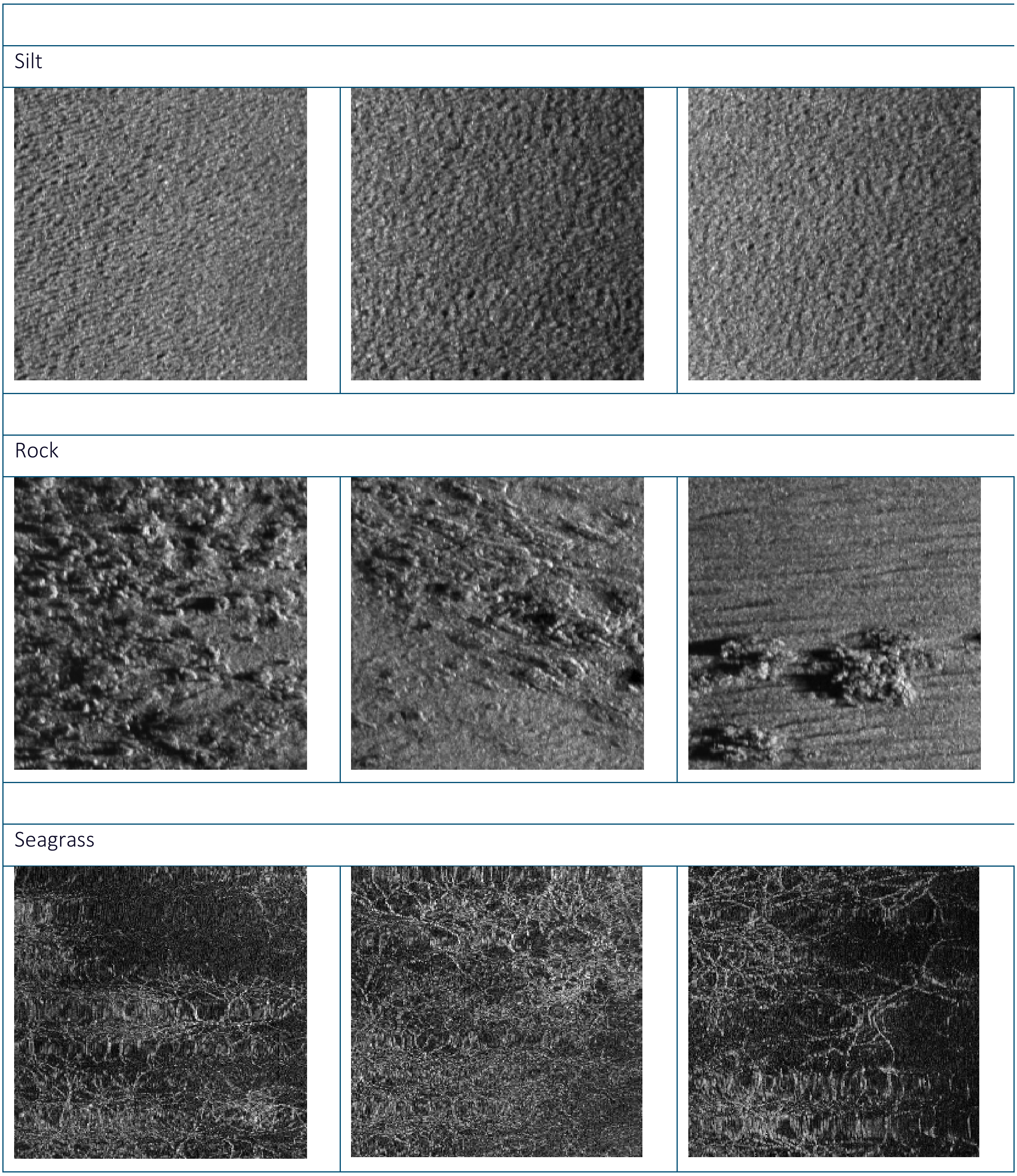
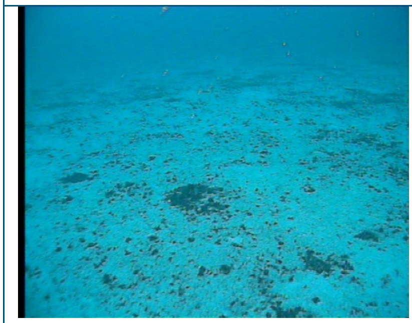
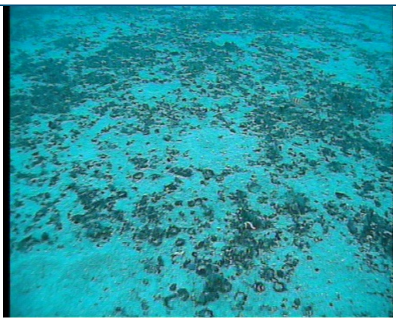
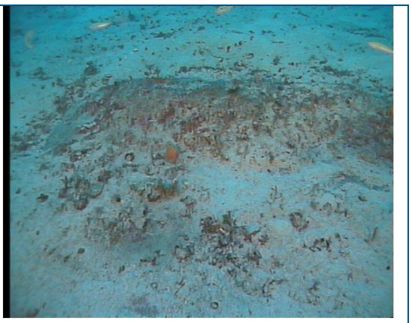
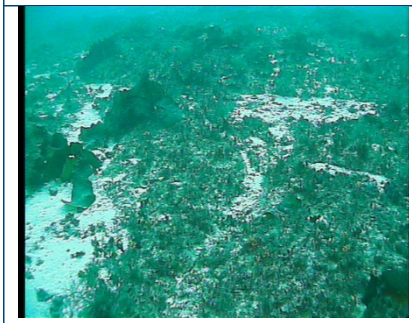
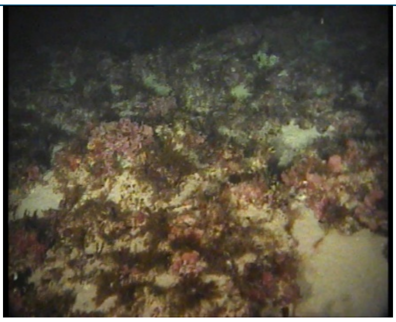
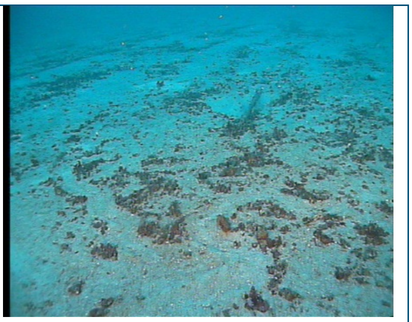
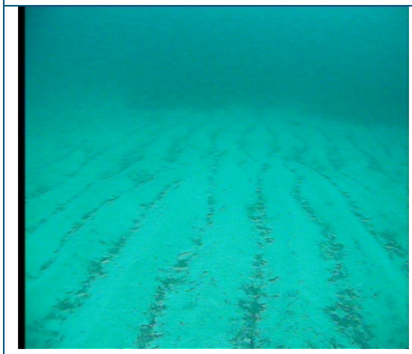
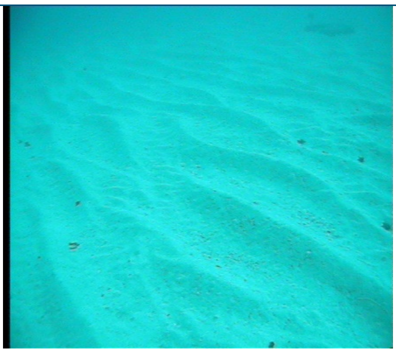
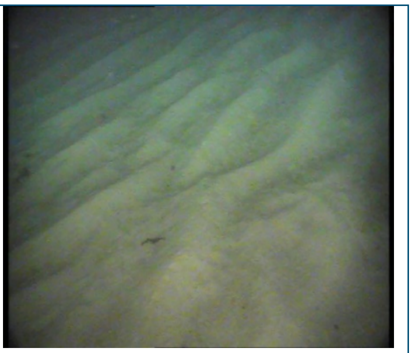
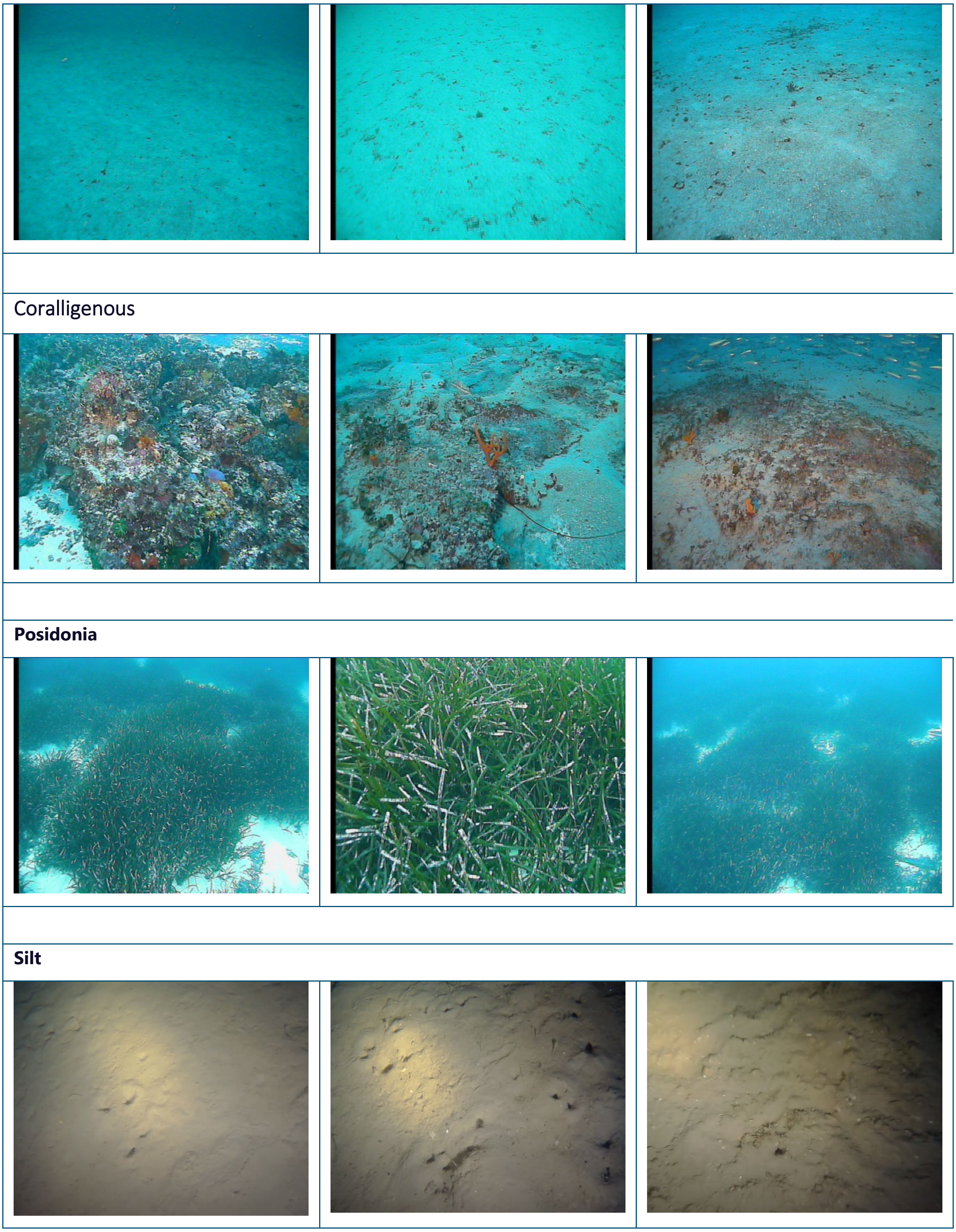


Figure 15. Examples of sidescan patches.

There is a strong difference on the amount of sonar and optical data. While the sonar transects cover large extensions of several kilometers with swath widths of more than 40m per side, the ROV videos transects are considerably shorter (ranging from 65m to 310m in length) and have images footprints covering only 9 to 25m^{2}of the seafloor in each frame.

Although not part of the original dataset, manual annotations for the video sequences have been created, under the following categories: Detrital, Maerl ,Sand , Sand Ripples, Coralligenous, Posidonia, and Silt. Examples of such classes are given in the following figure.

Examples of video frames with predominant examples of different classes		
Detrital		
		
Maerl		
		
Sandripples		
		
Sand		



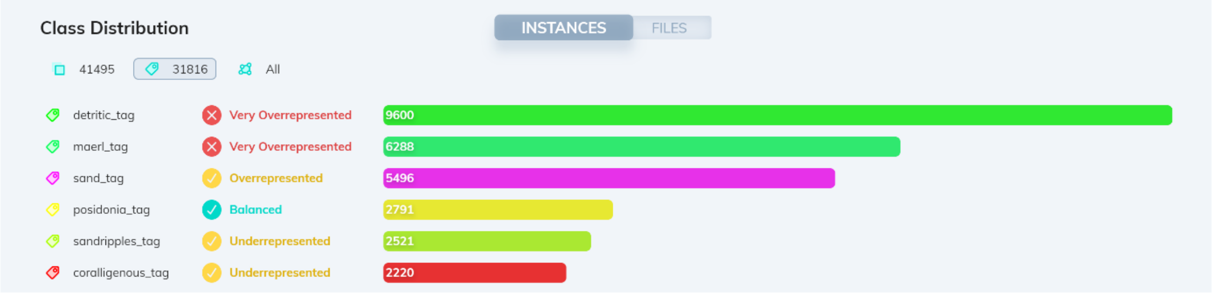
Coralligenous

Posidonia

Silt

Figure 16. Examples of video frames.

The class distributions, ordered by decreasing order of representativeness, are shown in Fig. 17.



detritic_tag	☒ Very Overrepresented	9600
maerl_tag	☒ Very Overrepresented	6288
sand_tag	✓ Overrepresented	5496
posidonia_tag	✓ Balanced	2791
sandripples_tag	✓ Underrepresented	2521
coralligenous_tag	✓ Underrepresented	2220

Figure 17. Class distribution of the annotated video

An important shortcoming of the existing datasets is that the acoustic and optical data were collected in separate surveys. In fact, the goal of the video acquisition was to provide a visual validation of the bottom types on certain locations, and not to do an extensive optical survey. As such there is not an easy way to co-register the two types of data. Furthermore, there is also a strong difference in the distribution of the observed bottom types. These shortcomings are intended to be solved during the data acquisition effort planned for DeeperSense during years 1 and 2 of the project.

4.2. Real-World Data

4.2.1. Acquisition Platforms

The acquisition will be based both on the use of sidescan sonar and optical camera equipment operated from three main platforms:

- deployments from boats operated by Tecnoambiente as part of their surveying activities,
- deployments from the UdG boat *Sextant,* for specifically targeted data collection efforts,
- deployments using the Girona1000 AUV as the sensor platform, for acquiring data under conditions similar to those intended for the final demos.

While the videos in the legacy datasets have been recorded using a frontal oblique position, a zenithal (down-looking) configuration for the camera orientation will be preferred, given that it allows for a better alignment with the sidescan data.

4.2.2. Locations

For logistical simplicity, data acquisition will prioritize the areas near Costa Brava in the north east of Spain. This area is near the UdG and is frequently surveyed by Tecnoambiente. Furthermore, there are some locations for which interpretation is already available, such as the ones reported by Ricart[2016], based on sidescan data acquired with a Tritech StarFish 990F. In 2021 and 2022, Tecnoambiente will carry out the acquisition of side scan sonar data and video along the coast of Catalonia from -5m to -50m of depth.

4.2.3. Scenarios

The targeted scenarios will serve two purposes. The first is to provide data with better class balancing than those of the legacy datasets. As such, sites where there is larger abundance of seagrasses will be prioritized, as well as sites where there are frequent changes in dominant classes and clear transition zones. The second purpose is to identify areas that are suitable for the final demonstrations that will involve the use of the Girona 1000 AUV. Towards this goal, part of the data acquisition planned for the second year of the project will be executed using the Girona1000 AUV as a platform to deploy both the sidescan sonar and the camera system. In this sense the acquisition setup is intended to be close to the one planned from the demo, with the exception that no processing will be done by the AUV at that stage, and no reactive behaviors will be yet implemented.

4.3. Synthetic Training Data

The structural complexity of natural environments translates into acoustic and optical images which are also of high complexity and spatial variability. As such there are no readily available tools to create synthetic opto-acoustic data that would be able to capture such complexity in a realistic way. Nonetheless we are investigating the feasibility of using existing sonar simulators, both commercial products and research-oriented implementations. Furthermore, as part of the planned effort for the implementation and testing of the learning algorithms, we will use (and further develop) techniques for data augmentation. Although not entirely synthetic, data augmentation techniques aim at generating new training data from representative example, by applying geometry and photometric transformations.

5. Data Management

Data management is concerned with the logistics of data formatting and data publication. It ensures that the project results are reproducible, and that the generated data is re-usable, both between the three different use cases of the project, and for the interested public.

Not all data that will be produced during the project may be publishable. This is discussed below. With respect to publishable data, data management is committed to the FAIR data principles [Wilkinson et al. 2016]. This includes the production of appropriate high-quality metadata.

Whether a dataset will be published may be unknown at the time of its creation. To facilitate the transition from private to public status, all datasets shall be held to the same standards.

5.1. FAIR Data Principles

The FAIR data principles state that data should be findable, accessible, interoperable, and re-usable. The following explanation provides high-level requirements for the implementation of the FAIR principles. It is copied verbatim from [Wilkinson et al. 2016], except for the emphasis.

To be Findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata (defined by R1 below).
- F3. metadata clearly and explicitly include the identifier of the data it describes [sic].

- F4. (meta)data are registered or indexed in a searchable resource.

To be Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
  - A1.1 the protocol is open, free, and universally implementable.
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:

- R1. meta(data) have a plurality of accurate and relevant attributes.
  - R1.1. (meta)data are released with a clear and accessible data usage license.
  - R1.2. (meta)data are associated with detailed provenance.
  - R1.3. (meta)data meet domain-relevant community standards.

5.2. Data Formatting

All datasets shall be bundled, named, annotated, and encoded consistently.

Not all formatting rules can be specified definitively in this document, because they depend in part on properties of the data that will emerge during data collection and data processing. **The partners will designate a person responsible for continuous update, internal dissemination, and controlling of the formatting rules.**

5.2.1. File bundling

File bundling is illustrated by the following file tree layout.

```
sensors_raw/
    [YYYY-MM-DD-HH-MM-SS]_[TASK_DESCRIPTION]/
        data/
            ... [DATA_FILES]
        metadata/
            metadata.yaml
            README.md
            ... [ADDITIONAL_FILES, e.g. plots, stats, ...]
sensors_processed/
    [YYYY-MM-DD-HH-MM-SS]_[TASK_DESCRIPTION]/
        data/
            ... [DATA_FILES]
        metadata/
            metadata.yaml
            README.md
            ... [ADDITIONAL_FILES, e.g. plots, stats, ...]
models/
    [YYYY-MM-DD-HH-MM-SS]_[TASK_DESCRIPTION]/
        data/
            ... [DATA_FILES]
        metadata/
            metadata.yaml
            README.md
            ... [ADDITIONAL_FILES, e.g. plots, stats, ...]
```

The first layer (sensors_raw/, ...) is debatable and may evolve. The idea is that multiple data bundles might be grouped by a shared purpose.[2]

The second, third, and fourth layers are important and should be handled consistently throughout the project. The idea is that each data file can be attributed to a specific task, and that each task is accomplished at a certain time. All data files that belong to the same task are bundled together and are accompanied by all their metadata. Each task is prefixed by the time stamp of its accomplishment to facilitate retrieval.

Each metadata subdirectory contains two mandatory files: a machine-readable metadata.yaml and a human-readable README.md, which are discussed below. Additional metadata files, like plots, statistics, etc., may serve as material for the data management tool to be produced in task T4.5. Their formatting and naming will be standardized when necessary. **Data producers are encouraged to add further files to the metadata subdirectory that may support the understanding and re-use of the data.**

5.2.2. Names

All directory and file names are composed only of the following characters:

- Lower case ASCII letters: [a-z]
- Digits: [0-9]
- Underscore, Dash, Dot: [_-.]

Any other characters are invalid. Examples for invalid characters are:

- Upper case letters: [A-Z]
- Non-ASCII letters: [äאñ]
- White space: [ ]
- Any other interpunction: e.g. [/\\:]
- Symbols: e.g. [$#%]
- ... (This list is not exhaustive.)

[2] Modeling such purposes as file system directories is a bad idea if the purposes are not mutually exclusive. Alternatively or additionally, we could (in the context of deliverable D4.2. / task T4.5) implement an abstraction layer based on metadata, and restrict the first layer to a few select purposes.

Each file name ends with an extension that indicates its content type, e.g. .csv, .bag, .yaml, .h5, .md, .png, .m, .py, .mp4, ...

The names data, metadata, metadata.yaml and README.md are reserved.

The naming of the top-level data bundle directory ([YYYY-MM-DD-HH-MM-SS]_[TASK_DESCRIPTION]) needs discussion: What is the content of the task description? Which datetime items are contained in the time stamp?

5.2.3. Annotation

5.2.3.1. METADATA.YAML

The file metadata.yaml is a machine-readable collection of all information necessary for effective re-use of a given data bundle, in compliance with the FAIR data principles outlined above. The YAML[3] formatting standard allows human-friendly editing. Since it is also machine-readable, it supports automated processing for indexing and publication.

The content of the file shall be compatible with relevant existing metadata standards[4]. Below is an illustrative example to convey the general idea of the format. As mentioned above, the specification will be continually updated as necessary during data collection and processing. **All data producers are encouraged to extend or adjust the metadata format as they see fit.** Standardization and possible resolution of conflicts will be managed by the person in charge of formatting development and controlling.

Example for metadata.yaml:

```
Title (required): ...
Date of recording (required): YYYY-MM-DD
Public domain (required): YES/NO
Project link (required): TBD
Related datasets (optional):
    - ...
Authors (required):
    -
        Family Name (required): ...
        Given Name (required): ...
        Affiliation (required): ...
        Email (optional): ...
    - ...
Contributors (optional):
```

[3] **https://yaml.org**

[4] **e.g. https://zenodo.org, https://earthdata.nasa.gov/esdis/eso/standards-and-references/iso-19115, https://www.emodnet-seabedhabitats.eu/access-data/launch-map-viewer/, http://ide.cime.es/geonetwork/srv/search?type=dataset, https://github.com/geonetwork/core-geonetwork, https://datasetsearch.research.google.com/**

```
    -
        Family Name (required): ...
        Given Name (required): ...
        Affiliation (required): ...
        Email (optional): ...
    - ...
Capturing device (required):
    Sensors (required):
        -
            Sensor type (required): ... (e.g., RGB camera XYZ, Side Scan Sonar ABC, GPS)
            TODO: Further sensor specification items
        - ...
    TODO: Further items to specify the capturing device
Location of recording (required):
    Location name (required): ...
    Bounding box (required):
        -
            Latitude (required): ...
            Longitude (required): ...
        - ...
    - TODO: Further items to specify the location
Habitats (optional):
    - ...
Biological species (optional):
    - ...
TODO: Further top-level categories
```

5.2.3.2. README.MD

The file README.md is a human-readable description of the given data bundle in Markdown[5] syntax. It complements the metadata.yaml by giving room for a coherent free-form presentation of introductory remarks, overview material (including e.g., graphics, timeline of events (although this should probably rather be covered by a dedicated data file)), or any other relevant aspects not covered by the metadata.yaml (e.g., peculiar circumstances during data collection, intended purpose of the data bundle, relation to existing data bundles, and others).

The Zenodo data publishing platform has a required "description" field that may be populated with (parts of) the README.md.

[5] **https://learnxinyminutes.com/docs/markdown**

5.2.4. Encoding

The partners will develop and observe collective standards for the encoding of data files, organized by sensor modality.

5.3. Data Publication

Due to rules imposed by publishing providers, **data that is published cannot be changed after publishing**.

Not all data that is produced during the project will be relevant to the public, and for some data, there may be legitimate reasons to keep the data private.

The partners will develop and observe collective criteria for classifying data bundles as either private or public. Data bundles will be treated as private by default, unless explicitly flagged as public in the metadata.yaml by the respective authors.

The partners will designate a person or persons responsible for the handling of all data publications.

The partners will evaluate different publishing platforms and select one that is suitable for their needs. The current top contender is Zenodo[6].

6. References

Cieślak, P. (2019, June). Stonefish: An Advanced Open-Source Simulation Tool Designed for Marine Robotics, With a ROS Interface. In *OCEANS 2019-Marseille* (pp. 1-6). IEEE.

Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

Terayama, K., Shin, K., Mizuno, K., & Tsuda, K. (2019). Integration of sonar and optical camera images using deep neural network for fish monitoring. *Aquacultural Engineering*, *86*, 102000.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Ricart AM (2016). Insights into seascape ecology: Landscape patterns as drivers in coastal marine ecosystems. PhD Thesis, Universitat de Barcelona.

[6] **https://zenodo.org**

DeeperSense Consortium