



D1.4 Final Data Management Plan

05/09/2025

Abstract

This report specifies how research data were collected, processed, monitored and catalogued during the project lifetime. For each dataset, it describes the type of data and their origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.



**Funded by
the European Union**

iImagine receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101058625.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, which cannot be held responsible for them.

Document Description

D1.4 Final Data Management Plan			
Work Package 1			
Due date	31/08/2025	Actual delivery date:	05/09/2025
Nature of document	Report	Version	1.0
Dissemination level	Public		
Lead Partner	EGI		
Authors	Andrea Anzanello (EGI), Tjerk Krijger (MARIS), Dick Schaap (MARIS), Carolin Leluschko (DFKI), Jean-Olivier Irisson (SU), Madeleine Walker (SU), Vanessa Tosello (Ifremer), Catherine Borremans (Ifremer), Damian Smyth (MI), Enoc Martínez (UPC), Igor Atake (CMCC), Wout Decrop (VLIZ), Jesús Soriano-González (SOCIB), Martin Laviale (UL-LIEC)		
Reviewers	Malgorzata Krakowian (EGI), Matteo Agati (EGI)		
Public link	https://zenodo.org/records/17055683		
Keywords	Data management, Use cases		

Revision History

Issue	Item	Comments	Author/Reviewer
V O.1	Draft version	Draft version and main contents ready for comments	Andrea Anzanello
V O.2	Revised version	Review of the structure and general information	Dick Schaap, Valentin Kozlov
V O.3	Revised version	Update of the data management plan for each Use Case	Tjerk Krijger, Dick Schaap, Carolin Leluschko, Jean-Olivier Irisson, Madeleine Walker, Vanessa Tosello, Catherine Borremans, Damian Smyth, Enoc Martínez, Igor Atake, Wout Decrop, Jesús Soriano-González, Martin Laviale
VO.4	Revised version	Quality review	Malgorzata Krakowian, Matteo Agati
VO.5	Revised version	Finalisation	Gergely Sipos
V 1.0	Submitted version		

Copyright and license info

This material by Parties of the iImagine Consortium is licensed under a [Creative Commons Attribution 4.0 International License](#).

Table of Contents

I.	Introduction.....	7
II.	Publication and metadata guidelines for iMagine training datasets on Zenodo.....	8
	Overview	8
	List of training datasets on Zenodo	8
	Zenodo submission guidance	9
	Description.....	10
	Domain specific fields for annotation of Biodiversity related training data sets.....	11
	Additions to Zenodo submission form	12
III.	Data management plans per Use case	14
	Use Case 1. Marine litter assessment.....	14
	Data Summary	14
	Findability.....	15
	Accessibility	16
	Data Interoperability.....	18
	Data Sharing and Re-use.....	18
	Ethics.....	19
	Data Security	19
	Use Case 2 ZooProcess.....	20
	Data Summary	20
	Findability.....	22
	Accessibility	23
	Data Interoperability.....	25
	Data Sharing and Re-use.....	26
	Ethics.....	28
	Data Security	28
	Use Case 3 Marine ecosystem monitoring.....	28
	Findability.....	30
	Accessibility	31
	Data Interoperability.....	35
	Data Sharing and Re-use.....	35
	Ethics.....	37

Data Security	38
Use Case 4 Oil spill detection	38
Data Summary	39
Findability.....	40
Accessibility	40
Data Interoperability.....	42
Data Sharing and Re-use.....	42
Ethics.....	43
Data Security	43
Use Case 5 Flowcam plankton identification	44
Data Summary	44
Findability.....	46
Accessibility	47
Data Interoperability.....	49
Data Sharing and Re-use.....	49
Ethics.....	50
Data Security	51
Use Case 6 Analysis of underwater noise spectrograms	51
Data Summary	51
Findability.....	53
Accessibility	54
Data Interoperability.....	56
Data Sharing and Re-use.....	56
Ethics.....	57
Data Security	58
Use Case 7 Beach monitoring	58
Data Summary	58
Findability.....	60
Accessibility	60
Data Interoperability.....	62
Data Sharing and Re-use.....	63
Ethics.....	65

Data Security	65
Use Case 8 Freshwater diatoms identification	66
Data Summary	66
Findability.....	68
Accessibility.....	69
Data Interoperability.....	71
Data Sharing and Re-use.....	71
Ethics.....	72
Data Security	72
IV. Conclusion	74

List of Figures

Figure 1 The domain specific fields, "Scientific name", "Scientific name ID" and "Taxon rank" to annotate Biodiversity related training datasets.	12
Figure 2 Using the PO2 Vocabulary in the Zenodo input form to annotate training datasets with parameters.....	13
Figure 3 Using the search tool to find the PO2 parameters and identifying the "Preferred label" to be used in the "Keywords and subjects" element in the Zenodo form.	13
Figure 4 Option to include organisations related to the creator or contributor using EDMO as a source.	13

I. Introduction

This deliverable is the final version of the Data Management Plan (DMP) for the iImagine project, funded by the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101058625. It builds upon and updates Deliverable D1.2, reflecting the developments that occurred over the course of the project.

The DMP is structured according to the Horizon Europe DMP template and adheres to the FAIR principles, ensuring that data is Findable, Accessible, Interoperable, and Reusable. It also addresses the Open Access requirements established by Horizon Europe.

Scientific data has been used and generated across eight use cases (UCs). For each use case, Section II outlines the status of pre-existing datasets and describes the data management approaches adopted for both existing and newly produced data. The plan provides details on the type and origin of data, associated metadata standards, data sharing strategies and target audiences, as well as the adopted archival and preservation methods.

This final version of the DMP captures the evolution of data management practices throughout the project, including the integration of newly generated datasets and adjustments in handling previously existing data to reflect changes in project needs and implementation.

II. Publication and metadata guidelines for iImagine training datasets on Zenodo

Overview

To make the project-related training datasets available following the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, a publication strategy has been developed using Zenodo as the primary archive. Zenodo was chosen as it offers:

- Long-term perspective
- DOI is assigned to every publication
- Clear versioning of publications (including datasets)
- 50GB is available by default, more on demand
- Tracks and gives download and view statistics
- Dataset publications can be assigned to the already existing [iImagine Project community folder](#)

However, Zenodo is a general repository for all kinds of EU results in many domains and not dedicated specifically to data, AI/ML, or Aquatic Sciences. Therefore, iImagine has been in touch with Zenodo (CERN) for adding extra elements to their metadata template to better fit the iImagine ML training datasets. As a result of this cooperation, the following improvements and guidance have been made available to the iImagine Use Cases for submission of their training datasets to Zenodo:

- Zenodo has implemented a look-up of the PO2 vocabulary (NERC Vocabulary Server) for marine parameter annotation;
- Zenodo has enabled the use of EDMO (European Directory of Marine Organisations) codes to identify organisations;
- iImagine has provided guidance for usage of the “Domain specific fields” section in the Zenodo form for annotation of Biodiversity related training data sets;
- iImagine has provided guidance for structuring the content information in the description field of the Zenodo input form;
- iImagine has provided guidance for how to publish the training datasets and which metadata fields to be included.

List of training datasets on Zenodo

As of now, 21 training datasets are published under the **iImagine community** on Zenodo:

1. [Smartbay Marine Types Object Detection Training dataset](#)
2. [iMAGINE UC4 – Segmented oil spills](#)

3. [Dataset for publication: Usefulness of synthetic datasets for diatom automatic detection using a deep-learning approach](#)
4. [AI-based fish detections at OBSEA Underwater Observatory](#)
5. [Labeled Images at OBSEA for Object Detection Algorithms](#)
6. [AI-based fish detections at Slagreef biotop deployed near the OBSEA Underwater Observatory](#)
7. [AI-based fish detections at OBSEA Underwater Observatory during a dolphin carcass experiment, May–July 2024](#)
8. [LifeWatch observatory data: phytoplankton annotated training set by FlowCam imaging in the Belgian Part of the North Sea](#)
9. [Nephrops \(Nephrops norvegicus\) Burrow object detection simple training dataset from Irish Underwater TV surveys](#)
10. [Underwater images from OBSEA fish detection training dataset \(YOLO\)](#)
11. [Deep-sea observatories images labeled by citizen for object detection algorithms](#)
12. [Smartbay Marine Species Object Detection Training dataset](#)
13. [Segmentation masks of ZooScan images focusing on images with several objects separated by a human operator](#)
14. [AI-based fish detections at OBSEA Underwater Observatory during a dolphin carcass experiment, March–July 2023](#)
15. [BWILD: Beach seagrass Wrack Identification Labelled Dataset](#)
16. [SCLabels: Labelled rectified RGB images from the Spanish CoastSnap network](#)
17. [RipAID: Rip current Annotated Image Dataset](#)
18. [EyeOnWater training dataset for assessing the inclusion of water images](#)
19. [LifeWatch observatory data: phytoplankton annotated image library by FlowCam imaging for the Belgian part of the North Sea](#)
20. [Beach-imaging derived beach wracks](#)
21. [Dataset for publication: Long-term effects of the herbicide glyphosate and its main metabolite \(aminomethylphosphonic acid\) on the growth, chlorophyll a and morphology of freshwater benthic diatoms](#)

Zenodo submission guidance

The Zenodo submission forms contains many different elements, in the context of the iImagine project, related to training datasets, the following elements should be focused on and filled:

1. Login on Zenodo and navigate to new upload¹
2. Select the 'iImagine project' under 'select a community'
3. Create a DOI if not already available

¹ <https://zenodo.org/uploads/new>

4. Set resource type to Dataset
5. Give a relevant title
6. Publication date: In case your upload was already published elsewhere, use the data of the first publication
7. Creator:
 - a. Can add multiple persons
 - b. Can add multiple organizations
 - i. Now possible to add EDMO organization (see section below)
8. Description: Add image classification activities following the guidance below
9. Licenses: Licenses available for data and software, can also add custom license; preference would be CC-BY-4.0.
10. Contributors can include persons, or organizations (same options as for creators)
11. Keywords and subjects:
 - a. Here possible to add PO2 parameter (see section below)
12. Dates: Multiple dates can be added with:
 - a. Types: Accepted, available, collected, copyrighted, created, issued, other, submitted, updated, valid, withdrawn.
 - b. + description
13. Funding
14. Software: Can include links to software applied on the training data set.
15. Domain specific fields
 - a. Can include here Taxonomic information if relevant (see section below)

Description

Add the activities under “Additional Description” with the “type” set to “Technical Info” and the activity names (e.g., Data Preprocessing) used as a heading. This way, the activities can be clearly outlined, and, if domain-specific fields are introduced in the future, these entries can be extracted and adapted easily.

Suggested headers are outlined below:

- Training dataset (header 1)
 - Brief explanation of the dataset including its purpose, contents (amount of images) and other relevant information, i.e. in what context the training data set is used.
- Data preprocessing (header 2)
 - Details about any preprocessing steps applied to the data, such as for example augmentation.
- Data splitting (header 2)
 - Explain how the training dataset is split into training, validation and prediction.

- Data labelling (header 2)
 - Describe the classes and labels used. If images are annotated, describe the type (bounding boxes, segmentation masks).
- Parameters (header 2)
 - The option to include here also information (next to the Keywords) on the parameters from the training dataset that might not be covered by the NVS. It is recommended here to include links to vocabularies.
- Data sources (header 2)
 - Instrument/gear, sensors, website, database, conditions under which images were captured.
- Data quality (header 2)
 - Information about the quality and reliability of the data, including any known limitations or sources of error.
- Data resizing (header 2)
 - Image resizing to other dimensions of images in pixels (width x height)
- Spatial coverage (header 2)
 - Spatial coverage: geographic extent covered by the data
- Contact information (header 2)
 - Who to contact for questions about the training dataset or the application related to it.

Domain specific fields for annotation of Biodiversity related training data sets

For Biodiversity related training datasets it was investigated how to annotate them with Taxonomic information. The idea is to implement the WORMS catalogue as a look-up page for Taxonomic information in the Zenodo input form. However, at this point of writing, the implementation is planned for the end of this year. In order to annotate the Biodiversity training datasets at this point it is chosen to add the relevant information manually to the Zenodo input form.

This can be done by using the “Domain specific fields” in the Zenodo form and:

1. Adding the following DarwinCore metadata fields:
 - a. Scientific Name²
 - b. Scientific NameID³
 - c. TaxonRank⁴
2. Writing the appropriate content in each field, using the WORMS catalog search interface to find your relevant taxon name(s).

² <https://dwc.tdwg.org/terms/#dwc:scientificName>

³ <https://dwc.tdwg.org/terms/#dwc:scientificNameID>

⁴ <https://dwc.tdwg.org/terms/#dwc:taxonRank>

You need to give the most specific name you have (preferably species) for the scientificName together with the ID and rank, which gives all the info needed. You can find your organism in WORMS using:

- the quick search⁵
- the advanced search⁶

For example (see image⁷), for solea solea after adding the three “Domain specific fields” in your Zenodo form, you would put for scientificName “Solea solea (Linnaeus, 1758)”, for the scientificNameID “127160” (see AphiaID) and for the taxonRank “Species”.

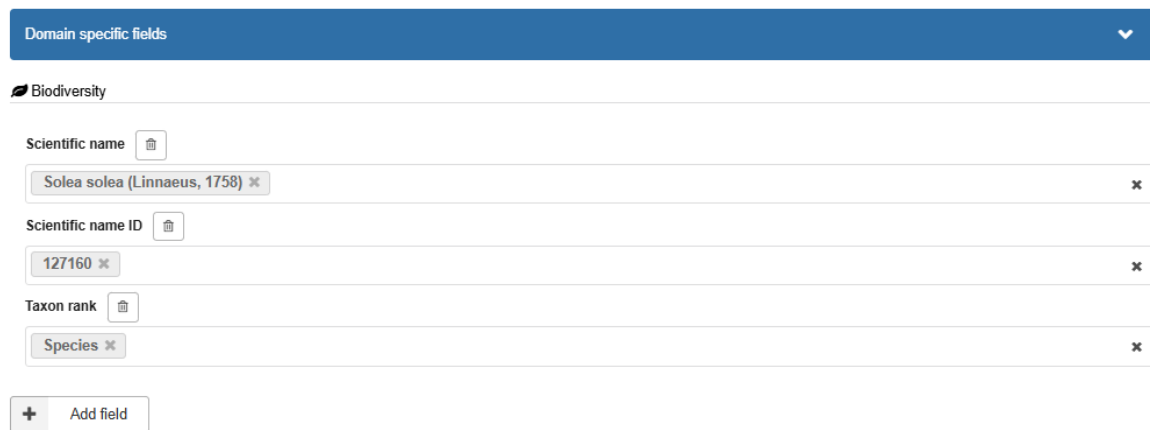


Figure 1 The domain specific fields, “Scientific name”, “Scientific name ID” and “Taxon rank” to annotate Biodiversity related training datasets.

If you already have a list of names you can use the taxon match service to automatically match your species list or taxon list with WoRMS⁸:

Additions to Zenodo submission form

PO2 vocabulary (NERC Vocabulary Server) for parameter annotation

Zenodo now allows the use of PO2 vocabulary terms from the NERC Vocabulary Server (NVS) in the “Keywords and Subjects” section.

How to add PO2 parameters in Zenodo:

- Go to the “Keywords and Subjects” section of the Zenodo form.
- Start typing the “Preferred Label” of the parameter from the PO2 list.

⁵ <https://www.marinespecies.org/aphia.php?p=search>

⁶ <https://www.marinespecies.org/aphia.php?p=search&adv=1>

⁷ <https://www.marinespecies.org/aphia.php?p=taxdetails&id=127160>

⁸ <https://www.marinespecies.org/aphia.php?p=match>

- Select the correct entry from the dropdown.

Keywords and subjects

EyeOnWater x ocean colo x

Add ocean colo

(NVS-P02) Ocean colour and earth-leaving visible waveband spectral radiation

Figure 2 Using the PO2 Vocabulary in the Zenodo input form to annotate training datasets with parameters.

How to find your PO2 parameter:

- Full PO2 vocabulary list⁹:
- Search tool to find your parameter¹⁰:

PO2 (SEADATANET PARAMETER DISCOVERY VOCABULARY)

SEARCH

Free search

ConceptID

Preferred label

Alt label

Entrytermidmod from

Entrytermidmod to

ocean

SEARCH RESET

OVERVIEW

Overview | Export subset of list | Export full list | New query | Found 1 | Current | Previous | Next

ConceptID	Preferred label	Alt label	Definition	Modified
R410	Ocean colour and earth-leaving visible waveband spectral radiation	Ocean_clr	Parameters associated with the measurement of ocean colour, including visible waveband satellite spectral radiometer measurements, their calibration measurements (spectral water-leaving radiance and reflectance) and chart colour estimates.	6/5/2006 15:11:44

Overview | Export subset of list | Export full list | New query | Found 1 | Current | Previous | Next

Figure 3 Using the search tool to find the PO2 parameters and identifying the “Preferred label” to be used in the “Keywords and subjects” element in the Zenodo form.

EDMO codes to identify organisations

Zenodo now supports the use of EDMO (European Directory of Marine Organisations) codes to identify research institutions, agencies, and organisations.

Where to add EDMO codes in Zenodo:

- Go to the “Creator” and “Contributor” sections in the Zenodo input form.
- When selecting an organisation, you can now choose the “source: EDMO”.

Creators

Kruger, Tjerk (MARIS B.V.) Data manager

Marine Information Service

Add creator

Add creator

Person Organization

maris

Marine Information Service (MARIS)
Noodorp, Netherlands Source: EDMO

Maritime Research Institute Netherlands (MARIN)
Wageningen, Netherlands — Nonprofit Source: ROR (Preferred)

Marine Design & Research Institute of China (MARIC)
Shanghai, China — Facility Source: ROR (Preferred)

Matis ohf, Icelandic Food and Biotech R&D (Matis)
Reykjavik, Iceland Source: EDMO

Marine Autonomous Robotic System (MARS)
Southampton, United Kingdom Source: EDMO

Marine Animal Rescue Society (MARS)
Miami, United States — Nonprofit Source: ROR (Preferred)

Mobilis, Aménagement, Transports, Risques et Société (MATRIS)
Cergy-Pontoise, France — Facility Source: ROR (Preferred)

Contributors

Thijssen, Pieter (MARIS B.V.) Data manager

Marine Information Service Data manager

Add contributor

Add contributor

Person Organization

mari

Marine Information Service (MARIS)
Noodorp, Netherlands Source: EDMO

Marine Design & Research Institute of China (MARIC)
Shanghai, China — Facility Source: ROR (Preferred)

Marine Renewable Energy Ireland (MaREI) (MaREI)
Ringaskiddy, Ireland Source: EDMO

Maritime Research Institute Netherlands (MARIN)
Wageningen, Netherlands — Nonprofit Source: ROR (Preferred)

Mikocheni Agricultural Research Institute (MARI)
Dar es Salaam, Tanzania — Facility Source: ROR (Preferred)

Marine & Risk Consultants Ltd (MARICO Marine)
Southampton, United Kingdom Source: EDMO

Marine Autonomous Robotic System (MARS)
Southampton, United Kingdom Source: EDMO

Figure 4 Option to include organisations related to the creator or contributor using EDMO as a source.

⁹ <https://vocab.nerc.ac.uk/collection/PO2/current/>

¹⁰ https://vocab.seadatanet.org/v_bodc_vocab_v2/search.asp?lib=PO2

III. Data management plans per Use case

iImagine Project has in total 8 use cases. The data management plan of each case is defined in 7 aspects: data summary, findability, accessibility, data interoperability, data sharing & reuse, ethics, and data security.

Use Case 1. Marine litter assessment

The use case developed an operational service at the iImagine platform which automatically analyses RGB drone imagery to detect and to quantify floating litter in seas, rivers and lakes, lying at beaches or shores. The analysing process consists of two steps: plastic litter detection, followed by quantification of the detected litter areas. Both steps are performed by separate classification models. Therefore, two different training datasets exist, consisting of image tiles. The training datasets were created before the iImagine project, which is why not all parts can be made publicly available. Since no new training data was generated during the project, the following information refers to the existing training data. The generated, resulting litter datasets are available to users through different interfaces (MinIO, OSCAR) but are primarily intended for download and local storage.

The developed service provides litter datasets that provide information about the quantity and the composition of the detected litter. During the project, the extend to which standardized litter categories can replace the original litter categories or be mapped to them was investigated. The investigation concluded that the JLIST contains too many detailed categories and cannot be used for an automatic detection on the context of this project. The litter categories used are based on previous projects and were selected by local stakeholders as a first step toward automated environmental monitoring.

Data Summary

Description of the existing data	Plastic Litter Detection: Identification of plastic waste. Plastic Litter Quantification: Identification of plastic items and quantities.
Origin of the existing data	<ul style="list-style-type: none"> • Indonesia • Germany • Vietnam • Cambodia and Myanmar • Philippines

Size of the existing data (Number of images and storage size)	Plastic Litter Detection: 26.147 images (split 70/15/15 as training, validation and test) Plastic Litter Quantification: 38.147 images (split 70/15/15 as training, validation and test)
Repository where existing data is stored	<ul style="list-style-type: none"> • All the data: At DFKI servers • Subset of the data¹¹
Licensing of the existing data	Zenodo Data: Creative Commons Attribution 4.0 International Rest of the Data: Closed
Access to the existing data	Zenodo Data: Open Source Rest of the Data: Closed. Only for DFKI researchers
Will you generate any new data?	No plans yet
Description of the new data.	N/A
Purpose of the new data and its relation to the project objectives	N/A
What is the expected size of the data that you intend to generate?	N/A
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	N/A

Findability

Will data be identified by a persistent identifier?	Published Data yes; unpublished data no
Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	Abstract provide information about dataset

¹¹ <https://zenodo.org/record/4552389>

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?	Plastic Litter, Drone Imagery, Detection, Machine Learning, Aquatic, Cambodia
Will metadata be offered in such a way that it can be harvested and indexed?	No

Accessibility

Will the data be deposited in a trusted repository?	Parts of already done, for other parts deposition currently in discussion
Have you explored appropriate arrangements with the identified repository where your data will be deposited?	Not yet
Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	https://doi.org/10.1088/1748-9326/abbd01
Open to making the data available through EOSC and AI4Europe?	Yes
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	Only the part that has already been published remains publicly available. Other data cannot be made publicly available for legal reasons.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	N/A
Will the data be accessible through a free and standardized access protocol?	N/A
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	No restrictions on use, so the data will still be available after the end of the project.
How will the identity of the person accessing the data be ascertained?	N/A
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	No
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	N/A
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	Unlimited
Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	The github repository contains information in the ReadMe file on how the data can be used for training.

Data Interoperability

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	The ontology used in the result of previous projects with the World Bank.
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?	During the project, it was investigated to what extent the current, less detailed litter categories can be mapped to the JList categories. The results show that it is not directly possible to map the highly detailed JList categories, which were developed and are used for manual surveys, to the categories used.
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	N/A

Data Sharing and Re-use

What are the target groups	AI researchers, Plastic Waste researchers, Remote Sensing and Drone operators
What are the main scientific impacts	Quality proofed dataset of plastic waste types in natural marine environment
What are the key channels or method of data sharing	Repositories
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	Via repositories, readme files with information on methodology

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?	Yes
Will the data produced in the project be usable by third parties, in particular after the end of the project?	yes
Will the provenance of the data be thoroughly documented using the appropriate standards?	yes
Describe all relevant data quality assurance processes.	Imagery is labelled by a pool of students (each image is labelled by at least 2 students) to ensure that false labelling is minimized

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	Drone images may raise personal data issues, as drone-captured images of beaches and other aquatic litter areas may include humans (at-risk humans even like migrants). However, the shared data set only contains training data, which does not contain humans
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?	N/A

Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Data is stored on a server with back-up
Will the data be safely stored in trusted repositories for long term preservation and curation?	Not by the end of the project

Use Case 2 ZooProcess

Use case 2, ZooProcess, is about building a new version of a workflow to process images acquired by the ZooScan and upload them to the EcoTaxa application, for taxonomic classification and sharing. The main improvement compared to the current version of ZooProcess will be that the separation between single and multiple organisms on the same image will be semi-automated through AI tools; this is a fully manual and very time-consuming part currently (25min/scan, amounting to ~35 working days over a year for a typical ZooScan operator). The workflow will be free, open and made available to the community. It will eventually replace the current workflow and be of use to all users of the ZooScan instrument worldwide.

However, **this data management plan concerns only the data that is officially part of the iMagine project**, handled and owned by partner Sorbonne Université. It is not possible to write a data management plan for all data processed through the workflow because it involves many partners, several of them outside of the EU, with various funding sources and data management requirements.

Within this data we own, we distinguish between:

- **Raw data:** the scans, which are large image files, associated with some metadata
- **Processed data:** small images with individual plankton objects, segmented from the scans, which also carry the original metadata + some image features computed from the segments
- **Derived data:** tables of occurrence, concentration and biovolume, per sample and taxon, which are computed by the EcoTaxa application after segmented images have been uploaded and taxonomically identified.

While Sorbonne University hosts many image datasets produced with the ZooScan, within iMagine we focused on two specifically:

- The Tara Oceans data which was used to produce a training set to distinguish between single and multiple plankton organisms on an image
- The time series collected with a Juday-Bogorov plankton net, in the Bay of Villefranche (bi-weekly sampling since 1966), which was processed with the new workflow and serves as an ecological demonstrator of its effectiveness.

Data Summary

Description of the existing data	<i>Tara Oceans data (training set)</i> raw data: images acquired by the ZooScan; processed data: segmented images of organisms from the raw images using the current ZooProcess workflow derived data: occurrences, concentrations
----------------------------------	--

	and biovolume of planktonic organisms; those are not directly relevant to iMagine however.
Origin of the existing data	scientific cruises funded by international research projects
Size of the existing data (Number of images and storage size)	raw data: >300 scans representing 210GB processed data: 270k segmented images, which, together with the metadata represent 48GB of data
Repository where existing data is stored	raw data: on local drives processed data: on local drives and on Ecotaxa derived data: on OBIS and GBIF ¹²
Licensing of the existing data	raw data: no processed + derived data: CC-BY 4.0
Access to the existing data	raw data: no (except locally) processed data: through EcoTaxa derived data: through OBIS, GBIF, IPTs.
Will you generate any new data?	Yes
Description of the new data.	<i>Juday-Bogorov net time series in Villefranche sur mer (1966–today)</i> scans, segmented images, and derived identifications = same as the old one
Purpose of the new data and its relation to the project objectives	Data from a long-term plankton monitoring time series to detect ecosystem changes
What is the expected size of the data that you intend to generate?	raw data: 1013 scans, 700GB processed data: 2.4 million segmented images + metadata, 155GB derived data: 54k occurrences, concentrations, biovolume records
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	training dataset: computer scientists and ecologists versed in computer vision, for training of segmentation algorithms. ecological dataset: scientists, monitoring agencies (ex. those tasked with evaluating

¹² <https://dx.doi.org/10.14284/521>

	the Good Ecosystem Status for the Marine Strategy Framework Directive)
--	--

Findability

Will data be identified by a persistent identifier?	<p>training dataset: is published on SeaNoe with a DOI¹³ and linked to from Zenodo¹⁴ to be in the iMagine community.</p> <p>ecological dataset: will be published on OBIS and GBIF with a DOI (current provisional version on GBIF¹⁵) and contains a permalink back to the original images on EcoTaxa</p>
Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	<p>training dataset: is published using the guidelines elaborated during the project.</p> <p>ecological dataset: published as a in DarwinCore Archive (the industry standard) with controlled terms from the NERC vocabulary server.</p>
Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?	Yes
Will metadata be offered in such a way that it can be harvested and indexed?	<p>raw data: SeaNoe requires basic metadata (location, date, variables, etc.) that facilitate discovery.</p> <p>processed data: metadata can be queried through EcoTaxa's API; location, depth, and date do not follow specific standards but are controlled and consistently queryable.</p> <p>derived data: OBIS and GBIF are easily harvested and indexed.</p> <p>In addition, both the OBIS and EcoTaxa</p>

¹³ <https://doi.org/10.17882/99663>

¹⁴ <https://zenodo.org/records/11108274>

¹⁵ <https://doi.org/10.15468/fkhe7w>

	repositories are also made discoverable and accessible through the Blue-Cloud Data Discovery & Access Service ¹⁶ .
--	---

Accessibility

Will the data be deposited in a trusted repository?	<p>training dataset: published on SeaNoe¹⁷ and linked to from Zenodo¹⁸.</p> <p>ecological dataset: will be published on OBIS and GBIF with a DOI (current provisional version on GBIF¹⁹)</p>
Have you explored appropriate arrangements with the identified repository where your data will be deposited?	Yes, we worked with them in the past
Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	Yes (as an option, which we will choose)
Open to making the data available through EOSC and AI4Europe?	Yes
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	Yes
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in	Not applicable

¹⁶ <https://data.blue-cloud.org>

¹⁷ <https://doi.org/10.17882/99663>

¹⁸ <https://zenodo.org/records/11108274>

¹⁹ <https://doi.org/10.15468/fkhe7w>

mind that research data should be made available as soon as possible.	
Will the data be accessible through a free and standardized access protocol?	<p>raw data: I don't know if SeaNoe is queryable.</p> <p>processed data: EcoTaxa has a standard OpenAPI with a Swagger interface and official R and Python clients.</p> <p>derived data: OBIS has an open API with various clients to query it.</p>
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	Not applicable
How will the identity of the person accessing the data be ascertained?	<p>processed data: EcoTaxa the user will need to be logged in to query the data but the queries/identities will not be logged.</p> <p>derived data: OBIS does not retain the identity of the people querying the data.</p>
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	No
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	Metadata is not licensed differently from the data; the data is often CC-BY rather than CCO, to encourage citation (and this will not change). However, in existing repositories, the metadata is queryable, readable, etc. so maybe it can be considered CCO. It would be OK to licence the metadata as CCO if that does not require an additional submission in a different format elsewhere. We believe the best solution is to distribute the metadata+data through the existing community-accepted and standard-

	following channels/databases.
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	<p>raw data: "Forever" at SeaNoe and Zenodo</p> <p>processed data: No commitment (yet) regarding EcoTaxa but we are working with IFREMER (the next host) to get a time guarantee.</p> <p>derived data: "Forever" at OBIS/GBIF.</p>
Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	No software needed.

Data Interoperability

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	<p>training dataset: the data is formatted as the mask2former model expects, which is pretty standard.</p> <p>ecological dataset: data is distributed as a DarwinCore Archive to OBIS/GBIF, with controlled terms from the NERC vocabulary server. From OBIS and EcoTaxa, data can also be extracted in a simpler table-like format which, although consistent, is not defined as any kind of standard.</p>
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to	Not applicable

allow reusing, refining or extending them?	
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	ecological dataset: Existing data from the same time series is already public. It will be integrated with the data generated through the project.

Data Sharing and Re-use

What are the target groups	<p>training dataset: computer scientists and ecologists versed in computer vision, for training of segmentation algorithms.</p> <p>ecological dataset: scientists, monitoring agencies (ex. those tasked with evaluating the Good Ecosystem Status for the Marine Strategy Framework Directive)</p>
What are the main scientific impacts	For the new ecological dataset in particular: probably the longest or at least one of the longest plankton time series in the world, distributed publicly for anyone to use, for climate-change related studies, ecological monitoring etc. + we will exploit this data through the project.
What are the key channels or method of data sharing	SeaNoe + Zenodo, OBIS, EcoTaxa as explained above
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	<p>training dataset: the dataset organisation and processing is documented in the landing page, following the recommendations defined during the project.</p> <p>ecological dataset: the distributed data is standard and self explanatory (concentrations of organisms per taxon and date); the computation methods for</p>

	<p>this data are documented there²⁰. The original data are images and associated morphological features measured on the images, which are already documented there²¹.</p> <p>The pipeline itself is documented there²² and the new version will be documented and the documentation distributed in the same location. The project also contributed to the translation of that documentation in 5 languages.</p>
Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?	CC-BY
Will the data produced in the project be usable by third parties, in particular after the end of the project?	Yes
Will the provenance of the data be thoroughly documented using the appropriate standards?	Yes, within the DarwinCore Archive files
Describe all relevant data quality assurance processes.	<p>The data acquisition process is quality controlled with many criteria, the checking of which is automatised through an existing software (and are too complex to describe here). The taxonomic sorting will be done automatically for a large part of the new data, and spot checked for a few dates each year. The computation of concentrations is then straightforward.</p>

²⁰ <https://sites.google.com/view/piqv/piqv-manuals/ecotaxaecopart-manuals>

²¹ <https://zenodo.org/records/14704251>

²² <https://zenodo.org/records/13928157>

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	Not applicable
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?	Not applicable

Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	raw + processed data: local backups derived data: the data is stored on and IPT, by GBIF and by OBIS is their responsibility (and I am not sure which backup solutions are in place but I am confident there are some)
Will the data be safely stored in trusted repositories for long term preservation and curation?	Yes, SeaNoe, Zenodo, OBIS and GBIF are all long term, trusted repositories. EcoTaxa should become one depending on IFREMER's commitment.

Use Case 3 Marine ecosystem monitoring

Use case 3 aimed to establish operational services on the iImagine platform for automatic processing of imagery, collected by cameras at EMSO underwater sites, identifying and further analysing interesting images and videos for purposes related to ecosystem monitoring. The three sites involved in the use case are EMSO-Obsea (UPC – SE), EMSO-Azores (Ifremer – FR) and EMSO-SmartBay. Within iImagine the objective is to develop services and use the platform for AI analysis of imagery and identification of biota, while developing standards and best practices that are useful for the management and storage of the video imagery and annotated images at EMSO sites. Finally, it should be possible to develop some guidelines useful best practices to share with other EMSO ERIC sites not involved in the project but also other RIs or ERICs such as LifeWatch or EMBRC, and any other external stakeholder that might be interested in AI and Computer Vision approaches for analysing and managing imagery acquired at underwater observatories.

The EMSO-OBSEA, EMSO-AZORES and EMSO Smartbay Sites maintain their own imagery Archives, smaller subsets of these Archives were used for creating training datasets during the iImagine project. The training datasets used by the EMSO sites for creating the iImagine modules are all stored on Zenodo.

	EMSO-OBSEA	EMSO-AZORES	EMSO-SMARTBAY
Description of the existing data	10 years of pictures from different underwater cameras at the same location. One picture every 30 minutes. Some data has already been manually tagged	12 years of video. 2 min every 6 hours / 365 days.	>6 years of Video 2 min Files 24/7/365 with gaps for observatory shutdowns / maintenance periods
Origin of the existing data	Different underwater cameras deployed at the OBSEA underwater observatory	Different underwater cameras deployed at the EMSO-AZORES underwater observatory	Underwater Kongsberg HD Camera at Smartbay observatory, Galway Atlantaquaria Minka-sdg , iNaturalist research grade dataset (currently collating additional imagery from the iNaturalist dataset)
Size of the existing data (Number of images and storage size)	100 GB	TBs	Terabytes of video and image Archives (Again only small subset will be used as part of the iImagine project)
Repository where existing data is stored	Internal repository	Internal repository	Internal repository
Licensing of the existing data	CC-BY	CC-BY	CC-BY
Access to the existing	Yes	Yes	Yes (At present

data			subset of data is available online < 3years)
Will you generate any new data?	yes	Yes (Data is generated continuously by the platform) Only a very small, curated subset will be used for the iMagine project	Yes (Data is generated continuously by the platform) Only a very small, curated subset will be used for the iMagine project
Description of the new data.	a picture every 30 minutes	Images extracted from videos collected by the observatory	2 min video files generated continuously when the observatory is in operation
Purpose of the new data and its relation to the project objectives	biodiversity estimation	biodiversity estimation and exploration	Training Datasets and Model weights for use in species Object Detection
What is the expected size of the data that you intend to generate?	Depends on the model of the camera, but typically 1 GB per year	~ 20 GB per year	Currently gathering many terabytes per year. (Only a small subset of this data will be used by the iMagine project and stored a training dataset on Zenodo)
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	marine scientists	marine scientists	marine scientists

Findability

	EMSO-OBSEA	EMSO-AZORES	EMSO-SMARTBAY
Will data be identified by a persistent identifier?	Yes, but ongoing work	Yes	Yes, Training datasets created for the iMagine project will have

			persistent identifiers on Zenodo
Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	We are open to add whatever metadata is required.	Yes	Yes
Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?	yes	yes	Yes
Will metadata be offered in such a way that it can be harvested and indexed?	yes	yes	Yes

Accessibility

	EMSO-OBSEA	EMSO-AZORES	EMSO-SMARTBAY
Will the data be deposited in a trusted repository?	yes, probably PANGAEA	yes, Training Datasets created for the iMagine project are stored on SEANOE open data repository and Zenodo.	Yes. Training Datasets created for the iMagine project are stored on Zenodo. Smartbay Marine Species Object Detection Training dataset

			Smartbay Marine Types Object Detection Training dataset Nephrops (Nephrops norvegicus) Burrow object detection simple training dataset from Irish Underwater TV surveys
Have you explored appropriate arrangements with the identified repository where your data will be deposited?	yes	yes	Yes
Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	yes	yes	Yes
Open to making the data available through EOSC and AI4Europe?	yes	yes	Yes
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in	everything open	everything open	All data/imagery used in the project are openly available (published on Zenodo)

multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.			
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	N/A	N/A	N/A
Will the data be accessible through a free and standardized access protocol?	yes	yes	yes
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	no restrictions	no restrictions	no restrictions
How will the identity of the person accessing the data be ascertained?	currently not implemented at OBSEA infrastructure	Identity of Users not tracked on Seanoe and Zenodo	Identity of Users not currently tracked on Smartbay Web Archive. Zenodo gathers its own statistics
Is there a need for a data access committee (e.g. to evaluate/approve	no	no	no

access requests to personal/sensitive data)?			
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	yes	yes	yes
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	depends on the publisher, but the intention is to make data available as long as possible	depends on the publisher, but the intention is to make data available as long as possible	depends on the publisher, but the intention is to make data available as long as possible
Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	currently we do not have any software to be used out-of-the-box, but we are open to share any software required	currently we do not have any software to be used out-of-the-box, but we are open to share any software required	Training datasets use standard image formats no specialist software required. The training datasets are also structured in YOLO format for Model training using the python Ultralytics YOLO package.

Data Interoperability

	EMSO-OBSEA	EMSO-AZORES	EMSO-SMARTBAY
What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	Species classification according to FISHBase	Species classification according to WORMS, instrument according to NVS (NERC Vocabulary Server) LO5 list and parameters according to NVS PO2 list	DarwinCore Metadata Fields using WORMS identifiers for Scientific names/ IDs . Annotation Classes will be described in the Zenodo metadata.
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?	N/A	N/A	N/A Annotation Classes will be described in the Zenodo metadata.
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	N/A	N/A	Yes Where we use other datasets our own or others

Data Sharing and Re-use

	EMSO-OBSEA	EMSO-AZORES	EMSO-SMARTBAY
What are the target groups	Marine scientists	Marine scientists	Marine scientists, Fisheries

			management
What are the main scientific impacts	N/A	<p>Contribute to marine species detection and identification that can be used for:</p> <ul style="list-style-type: none"> -Temporal analyses on species populations - Distribution and evolution of species abundance and biology in time and space 	<p>I Assist scientific projects in detecting Marine Species and assist in developing approaches for the semi-automatic annotation of prawn burrows for fishery management studies</p>
What are the key channels or method of data sharing	N/A	<p>Training Datasets created for the iImagine project will be accessible/downloadable from the Zenodo or Seano platforms</p>	<p>Smartbay uses http based folder access for its data repositories. Training Datasets created for the iImagine project will be accessible/downloadable from the zenodo platform</p>
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	N/A	<p>Zenodo records for training datasets and Model weights will describe the data and include information to validate the data and facilitate re-use</p>	<p>Zenodo records for training datasets and Model weights will describe the data and include information to validate the data and facilitate re-use</p>

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?	Yes, we plan to use CC-BY-3.0	Yes, we plan to use CC-BY-3.0	Yes, we use CC-BY-4.0
Will the data produced in the project be usable by third parties, in particular after the end of the project?	Yes	Yes	Yes
Will the provenance of the data be thoroughly documented using the appropriate standards?	Not currently implemented, but we are open to add it. Might need some support	Not currently implemented, but we are open to add it. Might need some support	We will describe the provenance of the data in our Zenodo records and reference external datasets where appropriate
Describe all relevant data quality assurance processes.	N/A	Annotated images datasets generated by citizens are cleaned through a dedicated pipeline available on github	Currently only raw imagery collected. The training datasets contain user (manually) collated, reviewed and Annotated imagery.

Ethics

	EMSO-OBSEA	EMSO-AZORES	EMSO-SMARTBAY
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context	No	No	No

of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).			
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?	No personal data	No personal data	No personal data

Data Security

	EMSO-OBSEA	EMSO-AZORES	EMSO-SMARTBAY
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	no sensitive data, regular backups in other cloud providers for regular data	no sensitive data, secure storage and backups in internal archiving system	no sensitive data, regular backups on internal hard disks
Will the data be safely stored in trusted repositories for long term preservation and curation?	yes	yes	yes

Use Case 4 Oil spill detection

The growing importance of oil spill models in supporting emergency response brought new user requirements to numerical modellers. Delivering accurate spill forecasts at the desired spatial and temporal resolutions within a short time was proposed as challenge to the community.

Scientific research aimed at increasing the accuracy of meteo-oceanographic forecasts and addressing forecasts in oil spill forecasts found fertile ground in recent years, while at the same time, enhancements in the accuracy of oil spill detection from satellite also increased, allowing oil spill models to be assessed and evaluated with more observations from satellite imagery.

Therefore in UC4 Orbital EOS, University of Trento and CMCC joined efforts to create a database of oil spill observations from 2019–2023 and using this dataset to train a Bayesian Optimization algorithm that is capable of automatically selecting the parameters (such as Horizontal Diffusivity and Wind Drift) of MEDSLIK-II, without the need for an expert user to assess the values inserted for each simulation. The product allowed us to reimagine WITOIL (Where is the Oil) and deploy an instance at the iMagine marketplace.

Data Summary

Description of the existing data	We currently count with Sentinel 1 and 2 pre-processed imagery ready for subsequent oil spill detection.
Origin of the existing data	Sentinel Open Access Hub
Size of the existing data (Number of images and storage size)	(1) Processed Images: 1000s of patches and equal number of modelled patches (2) Preprocessed Sentinel Images. In total O (100) GiB
Repository where existing data is stored	<u>IMAGINE UC4 – Segmented oil spills</u>
Licensing of the existing data	Pre-processed S1 and S2 imagery are Proprietary while raw S1 and S2 are open and free.
Access to the existing data	Internal use only – OrbitalEOS
Will you generate any new data?	yes
Description of the new data.	(1) oil spill detections (shapefile format); (2) oil spill forecasts from the collected database (shapefile format)
Purpose of the new data and its relation to the project objectives	Demonstrate the results that can be obtained from the dataset curated by Orbital EOS and establish the Bayesian Optimization process
What is the expected size of the data that you intend to generate?	O (10) GiB
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	(1) Research institutes interested in improving their oil spill forecasting skills. (2) Businesses operating in the oil spill forecasting field interested in improving their forecasting skills.

Findability

Will data be identified by a persistent identifier?	10.5281/zenodo.11354662
Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	Yes, we will include: spill-identifier, oil-type, sheen_area, thick_area, true_color_area and estimated oil volume when it is possible to analyze it.
Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?	Yes
Will metadata be offered in such a way that it can be harvested and indexed?	Yes

Accessibility

Will the data be deposited in a trusted repository?	Yes (valid for open data only)
Have you explored appropriate arrangements with the identified repository where your data will be deposited?	Yes
Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	Yes
Open to making the data available through EOSC and AI4Europe?	yes
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for	Pre-processed S1 and S2 imagery are Proprietary and will not be made openly available.

specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	Data embargo not foreseen.
Will the data be accessible through a free and standardized access protocol?	Yes
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	OrbitalEOS, UNITN and CMCC Foundation have an internal data sharing strategy already in place.
How will the identity of the person accessing the data be ascertained?	No authentication or registration is needed to get access to the oil spills data (they are openly available in Zenodo).
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	No
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	Yes (valid for open data only)
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	The data will remain findable in the long term according to Zenodo policies
Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	No proprietary software will be required and all open data can be accessed using open and free applications (e.g., QGIS, Python packages)

Data Interoperability

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	Yes. All results will be made available in community-endorsed and interoperable formats such as (1) shapefiles and (2) netCDF4.
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?	N/A
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	Yes

Data Sharing and Re-use

What are the target groups	Researchers and consulting companies working on oil spill forecasting
What are the main scientific impacts	Development of an actually fit-for-purpose oil spill modelling framework capable of delivering answers at the desired spatial scale.
What are the key channels or method of data sharing	EOSC and a THREDDS Catalogue hosted at the University of Trento (available also beyond the project lifetime)
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	User manual, readme files and variable definitions.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?	Yes
Will the data produced in the project be usable by third parties, in particular after the end of the project?	Yes
Will the provenance of the data be thoroughly documented using the appropriate standards?	Yes
Describe all relevant data quality assurance processes.	All the data is extracted by high trained remote-sensing experts who manually ensure the quality of the output

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	No
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Yes

Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Data security standards adopted by CMCC Foundation and OrbitalEOS/AWS Cloud
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes

Use Case 5 Flowcam plankton identification

Use Case 5 focuses on enhancing phytoplankton monitoring by establishing an operational service on the iImagine platform for ingestion, storage, analysis, and processing of FlowCam images. Phytoplankton are key components of marine ecosystems and play a critical role in global biogeochemical cycles. Their efficient monitoring is essential for assessing ecosystem health, forecasting harmful algal blooms, and supporting sustainable marine management.

To achieve this, UC5 leverages high-throughput imaging technologies, such as FlowCam, integrated with artificial intelligence (AI) for image recognition, which has transformed the way phytoplankton monitoring is conducted. This approach enables rapid and accurate species identification, reducing dependence on traditional, manual identification methods. UC5 introduces an automated image classification workflow based on convolutional neural networks (CNNs), designed to improve speed, accuracy, and scalability of phytoplankton identification. Furthermore, UC5 will provide an annotated training dataset and trained classifiers through the iImagine platform, enabling the development and deployment of advanced solutions for marine biodiversity monitoring.

Data Summary

Description of the existing data	<p>A. Biomonitoring result data: scientific data, human validated data, aggregated in taxon densities per sample</p> <p>B. Image library:</p> <ul style="list-style-type: none"> Image data and related metadata bucket: Image data collection: binary image files image metadata collection: metadata on sampling, laboratory processing, image parameters, classifications and predictions Training data split: pointers to images used in model training <p>C. Classifier data: trained models and statistics, scripts</p>
Origin of the existing data	<p>A. Biomonitoring Result data: Collected and generated from sampling since 2017 through funding by the Flemish Government.</p> <p>B. Image library: Collected and generated from sampling since 2017 through funding by the Flemish Government</p> <p>C. Classifier Data: Generated after model training</p>

Size of the existing data (Number of images and storage size)	A. Biomonitoring Result Data (aggregated data, densities of taxa/sample) 19271 documents, 6 982 144 b B. Image library: – Image Data and related Metadata: 1 734312 documents each (3 649 503 333 b metadata, 5 237 752 048 b images) – Training data split: 46 documents, 191 013 524 b C. Classifier data: Model versions and scripts: 69 documents each (299 649 b model metadata, 26 126 148 040 b model data) + few mbs in scripts
Repository where existing data is stored	BioSens mongoDB (internal access only)
Licensing of the existing data	A. Biomonitoring result data: CC-BY B. Image library : – Image data and related metadata: CC-BY – training data split: CC-BY C. Classifier set: CC-BY
Access to the existing data	Image (meta)data, training set and classifier data not available publicly, result data can be accessed via Lifewatch Data Explorer (aggregated taxon densities per sample).
Will you generate any new data?	A. Biomonitoring result data: always growing with new validations though LifeWatch project B. Image library: – Image data and related metadata: yes, always ongoing through LifeWatch project – Training data split: yes, can identify new training sets C. Classifier data: will train new models
Description of the new data.	See first row, continuous biomonitoring; result data and image libraries grow every month. Training data splits and classifier iterations are updated more sporadically.
Purpose of the new data and its relation to the project objectives	Semi-automated, fast and accurate biomonitoring of phytoplankton communities in the BPNS.

What is the expected size of the data that you intend to generate?	A. Biomonitoring result data: over 2 billion images – Training data split: 337,514 images distributed across 95 classes,
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	Environmental Agencies, European Agencies, Aquaculture, Fluid Imaging Technologies (Yarmouth Maine USA), Lifewatch, EMO BON (EMBRIC), Biodiversity researchers.

Findability

Will data be identified by a persistent identifier?	Yes, DOIs will be assigned: 10.5281/zenodo.16679297 https://doi.org/10.14284/710
Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	Yes, we can make metadata records in the IMIS (Integrated Marine Information System) dataset catalogue and link them to the data repository. This system is the dataset catalogue for EurOBIS and EMODnet Biology and it relies on a number of standard vocabularies e.g. WoRMS for Taxonomy and Marine Regions for georeferences. The following fields can be filled in in IMIS: <ul style="list-style-type: none"> • Title, abstract, and description which covers the content of the dataset • Name and contact details for the person responsible for the dataset • A citation for the record • Links to access the associated datasets (see Section 4 for more on what these associated datasets are) • Licence of use • Keywords chosen for the dataset • A listing of the parameters measured

	<ul style="list-style-type: none"> • The spatial, temporal & taxonomic coverage of the dataset • A listing of people who contributed to the dataset • Links to related and child/parent datasets
Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?	Yes
Will metadata be offered in such a way that it can be harvested and indexed?	Yes, metadata in the IMIS system can be harvested.

Accessibility

Will the data be deposited in a trusted repository?	Yes. Data is safely stored in house at VLIZ. VLIZ is an IODE certified and World Data System trusted data centre. Biomonitoring results will also be contributed to EurOBIS.
Have you explored appropriate arrangements with the identified repository where your data will be deposited?	Yes, the work of providing the biomonitoring result data to EurOBIS is ongoing.
Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	Yes, via the IMIS DataCite collaboration you can assign DOIs for every published dataset.
Open to making the data available through EOSC and A4Europe?	Open to opening up the data set to the public.

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	Yes
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	An embargo period is currently not foreseen on making these research data available.
Will the data be accessible through a free and standardized access protocol?	Yes
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	No
How will the identity of the person accessing the data be ascertained?	Currently we have no intention to ascertain the identity of the person accessing the data.
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	No
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	Yes
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	As a certified IODE and WDS data centre, VLIZ has a mandate to keep data available for the long term. IMIS metadata (data discovery) records can persist even if data itself is not available anymore.

Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	Yes
--	-----

Data Interoperability

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	DarwinCore standard will be used.
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?	If so, then yes.
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	Yes, the largest part of the dataset is already collected through other projects (e.g. LifeWatch). Biomonitoring result data are openly available via the rshiny app ²³ .

Data Sharing and Re-use

What are the target groups	Biodiversity researchers, environmental policy agencies, biomonitoring actors, etc.
What are the main scientific impacts	An efficient identification service can have a broad impact in biodiversity science and environmental monitoring.
What are the key channels or method of data sharing	Data and method sharing through: Github, research papers, Lifewatch Data Explorer

²³ <https://rshiny.lifewatch.be/flowcam-data/>

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	A lot of documentation is already available through the LifeWatch data explorer . Additional documentation will be provided in readme files or on github repository. The user can either use the readme files, go through the notebooks or watch the webinar. The metadata itself is explained together with the data records.
Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?	Yes
Will the data produced in the project be usable by third parties, in particular after the end of the project?	Yes
Will the provenance of the data be thoroughly documented using the appropriate standards?	Yes
Describe all relevant data quality assurance processes.	Image predictions are always validated by taxonomic experts.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	No
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?	Yes

Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	There is an advanced back-up and archiving system in place.
Will the data be safely stored in trusted repositories for long term preservation and curation?	Data is safely stored in house at VLIZ. VLIZ is an IODE certified and World Data System trusted data centre.

Use Case 6 Analysis of underwater noise spectrograms

The dataset and model described in this work are part of UC6, which focuses on using passive acoustic monitoring to assess human activity in marine environments. The study integrates 116 days of acoustic recordings from two stations in the Belgian part of the North Sea with Automatic Identification System (AIS) data to label vessel proximity. A convolutional neural network (CNN) was trained to classify short acoustic clips into discrete distance categories, representing the nearest vessel's proximity. This approach demonstrates the feasibility of estimating vessel presence and distance from underwater sound alone, providing a foundation for improved monitoring of sensitive areas such as Marine Protected Areas (MPAs) and offshore infrastructure.

Data Summary

Description of the existing data	Raw Data: 1.5 years of underwater sound from Belgian part of North Sea since 2020 along with metadata. Training Set: Smaller subset of the raw data Spectrogram Set: Sound data converted to image format for analysis
Origin of the existing data	Data collection has been funded by the Flemish Government.
Size of the existing data (Number of images and storage size)	Hard to estimate but couple months of raw recording data
Repository where existing data is stored	BioSensMongoDB on VLIZ server.
Licensing of the existing data	CC-BY, acknowledgement required
Access to the existing data	Internal only at the moment, image spectrogram data will become available under the iImagine project (not raw data)
Will you generate any new data?	Yes

<p>Description of the new data.</p>	<p>This dataset contains 10-second underwater acoustic recordings labeled with Automatic Identification System (AIS) data, collected from the Belgian part of the North Sea (BPNS). The data were gathered to develop machine learning models for vessel activity classification and distance prediction. Hydrophones were deployed at two stations, Gardencity and Grafton, near busy shipping routes, capturing vessel-generated sounds. AIS data was used to annotate these recordings with vessel position, speed, type, and activity. The dataset includes 27 524 labeled audio segments recorded over 116 days and is split into training, validation, and testing subsets</p> <p>This dataset was developed to support research on vessel monitoring using underwater acoustic recordings labeled with AIS data. It was created by researchers from VLIZ (Flanders Marine Institute) and collaborators to improve machine learning models for vessel classification and distance prediction. The dataset consists of 10-second underwater sound recordings, collected at two hydrophone stations, Gardencity and Grafton, located in the Belgian part of the North Sea (BPNS). These stations were strategically placed near major shipping routes to capture vessel-generated noise. Vessel positions, speeds, types, and activities were extracted from AIS-Hub data and linked to each recording, providing labeled data for model training. The recordings were collected over 116 days, resulting in 27 524 labeled audio segments. Each segment is accompanied by metadata, including AIS-derived vessel</p>
-------------------------------------	---

	<p>information and the hydrophone station. The dataset splits (training, validation, and testing) are provided as separate files in the data_split folder to ensure structured and reproducible dataset usage for machine learning applications. This dataset enables research in passive acoustic monitoring, vessel detection, and maritime traffic analysis, offering valuable data for studying human activity at sea.</p> <p>The data does not contain military vessels nor does it give any private information about the vessel (e.g., the MMSI is replaced by an anonymous vessel number).</p>
Purpose of the new data and its relation to the project objectives	Building training set
What is the expected size of the data that you intend to generate?	To be seen
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	Policy Makers including Marine strategy framework directive, spatial planning, Shipping companies, other researchers (bio-acoustic), International Quiet Ocean Experiment

Findability

Will data be identified by a persistent identifier?	Yes, DOI: https://doi.org/10.14284/723
---	---

<p>Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.</p>	<p>Yes, we can make metadata records in the IMIS (Integrated Marine Information System) dataset catalogue and link them to the data repository. This system is the dataset catalogue for EurOBIS and EMODnet Biology and it relies on a number of standard vocabularies e.g. WoRMS for Taxonomy and Marine Regions for georeferences. The following fields can be filled in in IMIS: • Title, abstract, and description which covers the content of the dataset</p> <ul style="list-style-type: none"> • Name and contact details for the person responsible for the dataset • A citation for the record • Links to access the associated datasets (see Section 4 for more on what these associated datasets are) • Licence of use • Keywords chosen for the dataset • A listing of the parameters measured • The spatial, temporal & taxonomic coverage of the dataset • A listing of people who contributed to the dataset • Links to related and child/parent datasets
<p>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</p>	<p>Yes</p>
<p>Will metadata be offered in such a way that it can be harvested and indexed?</p>	<p>Yes</p>

Accessibility

<p>Will the data be deposited in a trusted repository?</p>	<p>Yes, eventually</p>
<p>Have you explored appropriate arrangements with the identified repository where your data will be</p>	<p>Not by the end of the project</p>

deposited?	
Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	Yes, via the IMIS DataCite collaboration you can assign DOIs for every published dataset.
Open to making the data available through EOSC and AI4Europe?	Yes, the spectrograms can be made open eventually
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	Yes
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	An embargo period is currently not foreseen on making these research data available.
Will the data be accessible through a free and standardized access protocol?	Yes
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	N/A
How will the identity of the person accessing the data be ascertained?	Currently we have no intention to ascertain the identity of the person accessing the data.
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	No
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant	Yes

Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	As a certified IODE and WDS data centre, VLIZ has a mandate to keep data available for the long term. IMIS metadata (data discovery) records can persist even if data itself is not available anymore.
Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	Yes

Data Interoperability

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	We have looked in to these standards and although limited, we confirm with the ones in our case
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?	Yes
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	Yes

Data Sharing and Re-use

What are the target groups	Biodiversity researchers, environmental policy agencies, biomonitoring actors etc.
----------------------------	--

What are the main scientific impacts	An efficient identification service can have a broad impact in biodiversity science and environmental monitoring.
What are the key channels or method of data sharing	Research papers, software package (github)
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	Read me files, documentation and the published paper
Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?	Yes
Will the data produced in the project be usable by third parties, in particular after the end of the project?	Yes
Will the provenance of the data be thoroughly documented using the appropriate standards?	Yes
Describe all relevant data quality assurance processes.	Annotated spectrograms will be checked by experts

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	To avoid dual usage: The dataset used in this project consists of 10-second processed audio snippets of vessel sounds, predominantly from cargo vessels, and does not include any personally identifiable information, vessel identifiers (such as MMSI), or any data related to military ships. Any raw sensitive data won't go to unauthorised users.
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?	Yes

Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	There is an advanced back-up and archiving system in place at VLIZ.
Will the data be safely stored in trusted repositories for long term preservation and curation?	Data is safely stored in house at VLIZ. VLIZ is an IODE certified and World Data System trusted data centre.

Use Case 7 Beach monitoring

To develop a prototype service to process images from beach video-monitoring systems to automate beach wrack identification, shoreline extraction, and rip currents detection from beach imaging systems.

Data Summary

Description of the existing data	<p>UC7 utilizes oblique images from SIRENA systems and rectified images from CoastSnap Spain.</p> <p>SIRENA: Beach-video monitoring stations operated by the Balearic Islands Coastal Observing and Forecasting System (SOCIB). These stations are installed atop hotels, approximately 30–50 meters high, facing the coastline. Each SIRENA station features 3 to 5 RGB cameras dedicated to beach monitoring, capturing images hourly during daylight.</p> <p>CoastSnap: This is an international initiative where citizens contribute oblique smartphone photos from designated CoastSnap stations. These images are then shared with scientific managers. The images undergo quality control, are spatially co-registered to a target image, and subsequently rectified (georeferenced).</p>
Origin of the existing data	Collected/derived from SIRENA and CoastSnap

Size of the existing data (Number of images and storage size)	Number of images: ~ 1000000; Storage size: ~ 2000 GB
Repository where existing data is stored	SOCIB data repository (is CoreTrust Seal certified)
Licensing of the existing data	CC-BY4.0
Access to the existing data	Open access
Will you generate any new data?	Yes
Description of the new data.	<ul style="list-style-type: none"> • Oblique and rectified RGB images from SIRENA and CoastSnap new acquisitions • Training datasets (i.e., annotated images) • Data products derived from the implementation of AI prototypes (e.g., gridded products, absence/presence databases)
Purpose of the new data and its relation to the project objectives	The primary goal is to establish a foundation for improved monitoring of crucial coastal characteristics, support morphodynamic research, and aid in creating early warning systems for rip currents. This involves expanding datasets for model development and broadening their utility, as well as enhancing data products.
What is the expected size of the data that you intend to generate?	<p>It is not possible to define since:</p> <ul style="list-style-type: none"> • SIRENA images increase hourly • CoastSnap images depend on participation • Training datasets are in continuous development (currently ~ 8000 images). • At least one expected data product per coastal feature (to be completed after iMagine). An example is currently available²⁴.
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	Universities; scientific institute / researchers; local administration: emergency services, coastal surveillance/managers, lifeguards

²⁴ <https://zenodo.org/records/14808058>

Findability

Will data be identified by a persistent identifier?	DOIs have been assigned to training datasets and data products (Zenodo). SOCIB's DOI system will be updated to incorporate persistent identifiers for data sources in the future.
Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	Currently, datasets published in Zenodo include metadata related to authorship (identities and organizations), funding, related works, keywords and subjects (mostly from standard vocabularies), related projects and communities, licenses, and citations. Furthermore, the data product example ²⁵ adheres to CF conventions and incorporates extended, specific metadata such as dimensions, platforms, and methods.
Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?	Yes
Will metadata be offered in such a way that it can be harvested and indexed?	After discoverability is implemented.

Accessibility

Will the data be deposited in a trusted repository?	Yes. <ul style="list-style-type: none"> • SOCIB Data Repository • Zenodo
Have you explored appropriate arrangements with the identified repository where your data will be deposited?	N/A

²⁵ <https://zenodo.org/records/14808058>

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	Yes
Open to making the data available through EOSC and AI4Europe?	Yes
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	Yes
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	N/A
Will the data be accessible through a free and standardized access protocol?	Yes
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	N/A
How will the identity of the person accessing the data be ascertained?	<ul style="list-style-type: none"> • SOCIB: Currently, data access doesn't require identification. We're implementing a registration form that will support different methods, including Google, ORCID, and email. • Zenodo: It does not track the identity of

	individuals accessing datasets.
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	No
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	Yes
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	As long as SOCIB and Zenodo Data Repositories exist (expected: Forever)
Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	N/A. Data sources, training datasets, metadata, and data products use well-known formats (e.g., png, jpeg, txt, json, netCDF).

Data Interoperability

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	Community-endorsed variables and attributes when possible, standards such as CF conventions, and controlled vocabularies such as BODC (British Oceanographic Data Center operating the NERC Vocabulary Server), and EuroSciVoc.
---	---

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?	Yes
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	Yes, if any

Data Sharing and Re-use

What are the target groups	Scientific community, local administration
What are the main scientific impacts	Automation of processes, identification of environmental patterns, development of early warning systems
What are the key channels or method of data sharing	Open access repositories (e.g., SOCIB, Zenodo)
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	<ul style="list-style-type: none"> • Training datasets: Each dataset is supplemented with a README file that enhances its comprehension, elaborating on the context and contents, and providing detailed explanations of technical aspects, annotation process, usage recommendations, potential limitations, and related works. • Modules: Provide code as additional resources, which include Readme files for clarity (e.g., https://dashboard.cloud.imagine-ai.eu/catalog/modules/socib-beach-

	<p><u>wracks-identification</u>)</p> <ul style="list-style-type: none"> • Data products: extended metadata (e.g., global and variable attributes in netCDF) • Research publications are also expected
<p>Will your data be made freely available in the public domain to permit the widest re-use possible?</p> <p>Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?</p>	Yes
<p>Will the data produced in the project be useable by third parties, in particular after the end of the project?</p>	Yes
<p>Will the provenance of the data be thoroughly documented using the appropriate standards?</p>	Yes
<p>Describe all relevant data quality assurance processes.</p>	<ul style="list-style-type: none"> • SIRENA Stations: Robust quality assurance processes are in place to ensure proper functionality, including maintenance, UPS systems, and alert systems with automated status emails, confirming image storage and transmission to the SOCIB server. • CoastSnap Processing: All images undergo supervision by dedicated CoastSnap scientific managers. • Training Datasets: Multiple 'workers' are involved in the development of training datasets to enhance quality. • Data Products: Specific quality control methods are applied to data products, such as filtering predictions that fall outside of predefined spatial or logical boundaries

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	In principle, no.
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?	N/A
Please explain the workflow or protocol how images with human activities captured are processed? What steps are put into practices to meet ethical principles and applicable law defined in MGA-Annex 5	<p>To ensure GDPR compliance and prevent the identification of personal information, such as car registration plates, the following measures are implemented:</p> <ul style="list-style-type: none"> • Low-resolution imagery: This approach makes it impossible to discern fine details that could lead to personal identification. • Blurred areas: Specific regions of images are intentionally blurred to obscure individuals who are in close proximity. • Image selection: Images that feature beachgoers in close-up views are avoided. • Rectification: This process involves 'deforming' images, and makes human identification impossible.

Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	<p>CoreTrust Seal – SOCIB data repository: Back up system, cyber security.</p> <p>Zenodo: comprehensive security infrastructure to protect user data</p>
Will the data be safely stored in trusted repositories for long term preservation and curation?	Yes.

Use Case 8 Freshwater diatoms identification

Diatoms are unicellular microalgae present in all aquatic environments. They are routinely used as bioindicators for the ecological diagnosis of inland waters (rivers, lakes) as part of the implementation of the EU Water Framework Directive (WFD; Directive 2000/60/EC). Diatom taxonomic identification is based on morphological features of their exoskeleton made of silica that can be observed using classical light microscopy (x1000). Moreover, key morphological features such as size and deformations of the exoskeleton are relevant for bioindication but their quantification is not established as a routine task as it is laborious and time-consuming. Using automatic pattern recognition algorithms on microscope images, the use case will develop a prototype diatom-based bioindication service able to identify diatom species but also to quantify key morphological features, leveraging the iImagine AI platform.

A first proof of concept was developed using a synthetic dataset comprising a limited number of diatom images. In order to develop the approach we use the iImagine AI platform and set the following objectives in the project:

- Building an end-to-end detection, classification and trait quantification pipeline, including performance metrics meaningful for diatom experts
- Assembling an extensive quality-controlled dataset for tuning the CNNs
- Deploying the service on the iImagine AI platform

Data Summary

Description of the existing data	<p>Individual (thumbnail) Dataset: Images of individual diatoms for various species (ca. 200 species * 30–150 images per species) or debris (i.e. other objects than diatoms that can be found on a real image).</p> <p>Synthetic Dataset: Virtual raw microscope images generated from the individual dataset (diatom species + debris) synthetically (using a model). Currently 50,000 images (but as much as we need in theory)</p> <p>Real Dataset: Real raw microscope images acquired from field samples. Currently ca 100 images of the same size (2080 x 1042).</p>
----------------------------------	---

Origin of the existing data	<p>Individual Dataset: Gathered from real images (debris) and taxonomic guides (available as open source pdf)</p> <p>Synthetic Dataset: Generated from the individual dataset images by the team (to be used for pre-training the CNNs as an alternative to the lack of annotated real images)</p> <p>Real Dataset: Generated by the team on field samples provided by French Biodiversity Agency–OFB (stakeholder in charge of Water Framework Directive–WFD implementation).</p>
Size of the existing data	<p>Individual Dataset: ca 1 GB (ca 40 kb/thumbnaïl, ca 20,000 thumbnails)</p> <p>Synthetic Dataset: ca 4 GB (ca 250 kb/synthetic image, ca 50,000 images)</p> <p>Real Dataset: ca 1.5 GB (ca 1.1 MB/image, ca 150 images)</p>
Repository where existing data is stored	All the data is stored on the cloud of University of Lorraine (PETA ²⁶) Data used for published papers are stored on open repository (DOREL ²⁷)
Licensing of the existing data	<p>Published data: Open Source Etalab²⁸ (compatible CC–BY)</p> <p>Rest of data: closed</p>
Access to the existing data	for unpublished data: Restricted to the diatom project members (including partnerships external to iMagine). Data used for published works are open access via public repository
Will you generate any new data?	Yes, real dataset will be expended + update of the individual dataset (more images, more species)

²⁶ <https://sme.peta.univ-lorraine.fr/>

²⁷ <https://dorel.univ-lorraine.fr/dataverse/univ-lorraine>

²⁸ <https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf>

Description of the new data.	<p>See the existing data description for the different datasets.</p> <p>Individual Dataset: In total, the dataset represents a collection of 38,719 images (ca. 30kb/image) of various sizes and various scales, representative of 119 genus and 799 species, with 1 to 395 images per category</p> <p>Synthetic dataset: prepared on demand from individual dataset (only the code is stored)</p> <p>Real Datasets: Actual images from field samples that will be generated and annotated. Currently ca 2,900 real images with ca. 10k oriented bounding boxes annotated at genus and species level. A subset of these images are used for segmented mask manually annotation.</p>
Purpose of the new data and its relation to the project objectives	New individual images will be used to train the CNNs. New real images documenting real-life case studies will be used to fine tune + validate the CNNs..
What is the expected size of the data that you intend to generate?	<p>Individual Dataset: ca 1.2 GB (ca 30 kb/thumbnaill, ca 40,000 thumbnails)</p> <p>Synthetic Dataset: ca 4 GB (ca 250 kb/synthetic image, ca 50,000 images)</p> <p>Real Dataset: ca 3.2 GB (ca 1.1 MB/image, ca 2,900 images)</p>
To whom might your (existing and newly generated) data be useful ('data utility'), outside your project?	For education (university students, training of diatom experts)

Findability

Will data be identified by a persistent identifier?	yes for the training sets
---	---------------------------

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	Generic metadata of datasets published in Zenodo (authorship, funding, licenses...). Specific metadata will include annotation categories (taxonomic affiliation) but also size (image, pixel).
Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?	Yes
Will metadata be offered in such a way that it can be harvested and indexed?	Yes

Accessibility

Will the data be deposited in a trusted repository?	Existing datasets datasets were deposited on the academic repository DOREL , connected to the national repository Recherche Data Gouv . New datasets will be published on the dedicated Imagine project repository on zenodo
Open to making the data available through EOSC and AI4Europe?	Yes if relevant
Have you explored appropriate arrangements with the identified repository where your data will be deposited?	currently following zenodo standards and the "Open Science" standard procedure from Univ. Lorraine ²⁹
Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?	yes
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data	all training sets will be made available (useful for training other models).

²⁹ <https://scienceouverte.univ-lorraine.fr/en/home/>

closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	No embargo period is foreseen
Will the data be accessible through a free and standardized access protocol?	N/A
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	N/A
How will the identity of the person accessing the data be ascertained?	N/A
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	N/A
Will metadata be made openly available and licenced under a public domain dedication CCO, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	N/A
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?	Unlimited
Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?	N/A

Data Interoperability

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	Taxonomic names follows WORMS format. Training datasets are released for Model training using the python Ultralytics YOLO package.
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?	Yes (mapping will be provided, annotation categories will be described in the metadata).
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?	N/A

Data Sharing and Re-use

What are the target groups	environmental managers (OFB): diatom experts involved in european WFD implementation working at OFB and private companies (OFB subcontractors) researchers (ecology) teachers (biomonitoring, ecology, AI)
What are the main scientific impacts	release of high quality control training sets (not existing yet)
What are the key channels or method of data sharing	Long term trusted repositories
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	Github documentation, readme files, csv files

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?	yes
Will the data produced in the project be usable by third parties, in particular after the end of the project?	yes
Will the provenance of the data be thoroughly documented using the appropriate standards?	Yes
Describe all relevant data quality assurance processes.	Image annotations were checked by multiple taxonomic experts. Released datasets were prepared following internal standards which were determined after multiple discussions with both domain and AI experts.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	no
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?	N/A

Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	No sensitive data. Data recovery is ensured at the institutional level following an advanced back-up and archiving system in place.
---	---

Will the data be safely stored in trusted repositories for long term preservation and curation?	Yes, according to commitments of the chosen long term trusted repositories (PETA, DOREL, zenodo)
---	--

IV. Conclusion

The Final Data Management Plan of the iImagine project summarises the practices, tools, and policies adopted for handling research data during the project and for ensuring their sustainability beyond its lifetime. Conceived from the start as a dynamic resource, the DMP has been progressively refined to incorporate new requirements, technical adjustments, and project decisions. A major revision was released at Month 21, and the present document closes the process at Month 36, in line with the project schedule.

Revisions were systematically tracked through a dedicated changelog available on the project's Confluence space. Information about updates and decisions was disseminated through the governance structure of iImagine, including Work Package meetings, exchanges with WP leaders, discussions within the Project Management Office, and consultations with the Activity and Service Board.

The plan defines a coherent strategy for research data management, built upon Horizon Europe policies and the FAIR framework. To support this strategy, the project relied on a set of interoperable repositories and collaboration platforms such as OpenAIRE, Zenodo, GitHub, the EGI Document Repository, and Google Drive. These ensured that data could be stored, preserved, and shared in a secure and sustainable way, while also supporting long-term accessibility once the project ends.

Data handling was guided by an assessment of sensitivity and confidentiality. Open access was granted whenever possible, while restricted or ethically sensitive information was kept confidential. All activities respected the principles of informed consent and were carried out in full compliance with the EU General Data Protection Regulation (GDPR).

This final version of the DMP therefore consolidates the project's legacy in terms of data preservation, openness, and reusability. It demonstrates the project's commitment to responsible data governance, ethical standards, and the principles of Open Science, ensuring that iImagine's outputs will remain accessible and valuable well beyond the project duration.