# D1.2 Data Management Plan

28/02/2023


## Abstract

A report that specifies how research data will be 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

iMagine receives funding from the European Union's Horizon Europe resea 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

| Document title | | | |
|---|---|---|---|
| **Work Package 1** | | | |
| Due date | 28/02/2023 | Actual delivery date: | 09/03/2023 |
| Nature of document | [Report] | Version | 1.0 |
| Dissemination level | Public | | |
| Lead Partner | EGI | | |
| Authors | Mandy Y.Lin (EGI) | | |
| Reviewers | DMA Schaap (MARIS); Valentin Kozlov (KIT); Gergely Sipos (EGI) | | |
| Public link | DOI on Zenodo | | |
| Keywords | Use cases | | |

## Revision History

| Issue | Item | Comments | Author/Reviewer |
|---|---|---|---|
| V 0.1 | Draft version | Draft version ready for comments | Mandy Y.Lin |
| V 0.2 | Revised version | Overall proof-reading<br>Including summaries per UC | M. Laviale, Schaap, R. Lagaisse |
| V 0.3 | Revised version | Overall proof-reading | Valentin Kozlov<br>Gergely Sipos |
| V 1.0 | Submitted version | | |

## Copyright and license info

This material by Parties of the iMagine Consortium is licensed under a Creative Commons Attribution 4.0 International License.

# Table of content

# I. Introduction

Deliverable 1.2 is the Data Management Plan (DMP) for the research data used and/or generated within the iMagine project, funded by the European Union's Horizon Europe research and innovation programme under grant agreement No.101058625. The Data Management Plan follows the structure of the Horizon Europe DMP template. It aligns with the FAIR principles, considering data reuse, and addressing the Open Access requirements of Horizon Europe.

Scientific data is used and generated in the project within 8 use cases (UCs). For each use case Section II describes the status of existing data, and the management approaches for both existing and newly generated data. The plan describes the type of data, data origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.

D1.2 Data Management Plan will be a living document updating over the course of the project and the evolution of project activities as new datasets arise, and the management practice of existing datasets need to change.

The final Data Management Plan will be presented as D1.3 Final Data Management Plan by the end of the project.

# II. Data management plans per Use case

iMagine 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 is going to establish an operational service at the iMagine platform for ingestion, storage, analysis and processing of drone images, observing litter floating at surface waters in seas, rivers and lakes, and lying at beaches and shores, delivering standardised classified litter data sets, which are fit for the purpose of environmental management and indicators. The technology is based on the UAV survey from different altitudes and analysing GB drone images with two CNN deep neural networks to get quantification and characterization of observed litter. Approach successfully applied for several countries through World Bank Group and NGOs for providing aquatic litter analyses for local stakeholders and clean-up operations.

## Data Summary

| | |
|---|---|
| Description of the existing data | Plastic Litter Detection: Identification of plastic waste.<br>Plastic Litter Quantification: Identification of plastic items and quantities. |
| Origin of the existing data | • Indonesia: https://www.dfki.de/web/forschung/projekte-publikationen/projekt/aplasticq-idn<br>• Germany: https://www.dfki.de/web/forschung/projekte-publikationen/projekt/aplastic-q-obermaubach<br>• Vietnam: https://www.dfki.de/web/forschung/projekte-publikationen/projekt/aplastic-q-vietnam<br>• Cambodia and Myanmar: https://www.dfki.de/web/forschung/projekte-publikationen/projekt/rs-of-plastic-wb2<br>• Philippines: https://www.dfki.de/web/forschung/projekte-publikationen/projekt/rs-of-plastic-wb3 |
| 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)<br>Plastic Litter Quantification: 38.147 images (split 70/15/15 as training, validation and test) |
| Repository where existing data is stored | All the data: At DFKI servers<br>Subset of the data: https://zenodo.org/record/4552389 |
| Licensing of the existing data | Zenodo Data: Creative Commons Attribution 4.0 International<br>Rest of the Data: Closed |
| Access to the existing data | Zenodo Data: Open Source<br>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 | N/A |

| | |
|---|---|
| data that you intend to generate? | |
| 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 |
| 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 | https://doi.org/10.1088/1748-9326/abbd01 |

| | |
|---|---|
| to a digital object? | |
| 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? | N/AMaybe |
| 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 CC0, 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)? | No |

## 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? | N/A |
| 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 | N/A |

| | |
|---|---|
| data from your project, or datasets from previous research)? | |

## 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? | |
| 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 yet |

## Use Case 2 Zooscan - EcoTaxa pipeline

Use case 2 Zoonscan is about building a workflow to process images acquired by the ZooScan and upload them to the EcoTaxa application for taxonomic classification and sharing. 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 the existing data, which has been processed through the existing workflow and will not have to be reprocessed and the new data that we will process through the new workflow, as part of iMagine**.

## Data Summary

| | |
|---|---|
| Description of the existing data | Raw images acquired by the ZooScan; segmented images of organisms from the raw images using the ZooProcess workflow |
| Origin of the existing data | monitoring actions funded by academic research projects and by the French government, scientific cruises funded by international research projects |
| Size of the existing data (Number of images and storage size) | around 5000 large scans and 20 million segmented images already acquired (no workflow needed); 15TB raw scan data and 0.5TB small images. |
| Repository where existing data is stored | Raw data on local drives; small images are on Ecotaxa https://ecotaxa.obs-vlfr.fr/ |
| Licensing of the existing data | Raw data: no; small: CC-BY |
| Access to the existing data | Raw data: no (except locally); small: through EcoTaxa |
| Will you generate any new data? | Yes |
| Description of the new data. | Raw scans and derived segmented images = 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? | 2~3 milions segmented images = 200G raw images and 2G small images |
| To whom might your (existing and newly generated) data be useful ('data utility'), outside your project? | scientists, monitoring agencies (ex. those tasked with evaluating the Good Ecosystem Status for the Marine Strategy Framework Directive) |

## Findability

| | |
|---|---|
| Will data be identified by a persistent identifier? | The derived ecological data (concentrations of plankton taxa per location/date) will be exported to (Eur)OBIS and have a DOI. The raw data and small images will not be exported but the OBIS |

| | |
|---|---|
| | dataset will contain a link back to the original images |
| 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. | Metadata: at least lat, lon, depth, date+time; Stored in EcoTaxa, exported in DarwinCore Archive format (the industry standard) |
| 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, via OBIS (and somewhat via EcoTaxa's API). Both the EurOBIS and EcoTaxa repositories are also made discoverable and accessible through the Blue-Cloud Data Discovery & Access Service (https://data.blue-cloud.org). |

## Accessibility

| | |
|---|---|
| Will the data be deposited in a trusted repository? | Yes (for the derived table-like data; small the images will be in EcoTaxa; the large one will stay local on backed up drives or within EGI's architecture) |
| 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 mind that research data should be made available as soon as possible. | Not applicable |
| Will the data be accessible through a free and standardized access protocol? | Yes, assuming the querying of OBIS' (or EcoTaxa's) APIs is considered standard. |
| 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? | For OBIS it will not, for EcoTaxa the user will need to be logged in. |
| 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 CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data? | Metadata has not been licensed differently from the data; the data is often CC-BY rather than CC0, to encourage citation (and this will not change). However, in existing repositories, the metadata is queryable, readable, etc. so maybe it can be considered CC0. It would be OK to licence the metadata as CC0 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? | "Forever" at OBIS. No commitment (yet) regarding EcoTaxa. |
| 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? | Data will be distributed in DarwinCore Archive format to OBIS, from EcoTaxa (https://dwc.tdwg.org/terms/)<br>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. |
| 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? | Not applicable |
| Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? | The existing data is from the same time series and is already public. It will be integrated with the data generated through the project. |

## Data Sharing and Re-use

| | |
|---|---|
| What are the target groups | scientists, monitoring agencies (ex. those tasked with evaluating the Good Ecosystem Status for the Marine Strategy Framework Directive) |

| | |
|---|---|
| What are the main scientific impacts | For the new data 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 | 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.)? | The distributed data is self explanatory (concentrations of organisms per taxon and date); the computation methods for this data are documented there https://sites.google.com/view/piqv/piqv-manuals/ecotaxaecopart-manuals . The original data are images and associated morphological features measured on the images, which are already documented and for which the documentation is there https://sites.google.com/view/piqv/softwares/flowcamzooscan.<br>The new pipeline will be documented and the documentation distributed in the same location. |
| 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. | 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. |

## 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)? | Local backup of raw data + small images. The derived data stored 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, OBIS (+EcoTaxa if its long term hosting commitment gets finalised, which should be the case in the coming months) |

## Use Case 3 Marine ecosystem monitoring

Use case 3 aims at establishing an operational and integrated service at the iMagine platform for automatic processing of video imagery, collected by cameras at EMSO underwater sites, identifying and further analysing interesting images 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 iMagine the objective is to use the platform for further developing of AI preselection of interesting images and of AI analysis of selected images for identification of biota, while developing standards for the management and the storage of the video imagery and annotated images from several EMSO sites in databases and setting up an EMSO workflow at the iMagine platform with AI analysis service and connectivity of databases. Finally, it should be possible to develop some guidelines useful to share standard data management practices to use AI analysis pipelines for biota classification and, these should reach 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 this.

| | 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 |
| Size of the existing data (Number of images and storage size) | 100 GB | TBs | > 20GB |
| Repository where existing data is stored | Internal repository | Internal repository | Internal repository |
| Licensing of the existing | CC-BY | CC-BY | CC-BY |

| | | | |
|---|---|---|---|
| data | | | |
| Access to the existing data | Yes | Yes | Yes |
| Will you generate any new data? | yes | yes | yes |
| Description of the new data. | a picture every 30 minutes | | 2 min video file 24/7/365 |
| Purpose of the new data and its relation to the project objectives | biodiversity estimation | biodiversity estimation and exploration | biodiversity estimation, biological observation, engineering observation |
| 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 | ~ 20 GB per year |
| 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 |
| Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case | We are open to add whatever metadata is required. | Yes | Yes |

| | | | |
|---|---|---|---|
| metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. | | | |
| 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, https://www.seanoe.org/ | smartbay.marine.ie/data , http://data.marine.ie/geonetwork/srv/eng/catalog.search#/metadata/ie.marine.data:dataset.3880 |
| 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 10/dzhw 10.20393/dpkq-ez08 |

| | | | |
|---|---|---|---|
| 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 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. | everything open | everything open | everything open |
| 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 currently tracked |
| Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? | no | no | no |
| Will metadata be made openly available and licenced under a public domain dedication CC0, 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 | 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 | Video is in MPEG4 standard - no specialist documentation required |

| | | | |
|---|---|---|---|
| open source code)? | | | |

## 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 | WORMS | WORMS - Seadatanet , Will investigate what are the best practices are and how they are applicable to Smartbay |
| 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 | Will try and use standard vocabularies and are open to defining mappings. |
| 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 |

## Data Sharing and Re-use

| | EMSO-OBSEA | EMSO-AZORES | EMSO-SMARTBAY |
|---|---|---|---|
| What are the target groups | Marine scientists | Marine scientists | Marine scientists |
| What are the main scientific impacts | N/A | N/A | Meet requirements of individual scientific projects |
| What are the key channels or method of data sharing | N/A | N/A | http based folder 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.)? | N/A | N/A | See metadata record - http://data.marine.ie/geonetwork/srv/eng/catalog.search#/metadata/ie.marine.data:dataset.3880 |
| 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 | Not currently implemented, but we are open to add | Not currently implemented, but we are open to add | Not currently implemented, but we are open to add |

| | | | |
|---|---|---|---|
| appropriate standards? | it. Might need some support | it. Might need some support | it. Might need some support |
| Describe all relevant data quality assurance processes. | N/A | N/A | Currently only raw data collected |

## 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 of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA). | No | No | No |
| 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, regular backups in other cloud providers for regular data | 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 grown 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 were proposed as challenges 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. Discoveries on the role played by small-scale-but-hard-to-predict coherent ocean structures (e.g. ocean meanders and filaments) in the oil transport and weathering reshaped our understanding of oil spill forecast uncertainties. Improving the spatial resolution of input meteo-oceanographic fields could, on the one hand, bring results to the spatial scale desired by end-users. On the other hand, fine resolution modelled fields were found to occasionally show higher uncertainties than coarser simulations exactly because of their increased capability to predict smaller scale (but often misplaced) dynamics.

A machine learning algorithm designed to improve the spatial resolution of ocean forcing fields will be developed in the UC 4. The algorithm will rely on modelled ocean fields at multiple spatial resolutions to generate its own ocean fields at higher spatial resolution and, hopefully, with superior accuracy. The impact of the freshly generated fields on oil spill forecast accuracy will be evaluated by extensive comparisons between modelled and (remotely) observed oil spills and later included in the online oil spill forecasting system WITOIL. The current WITOIL architecture will be reviewed to include the innovative product and allow efficient data cataloguing.

### 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 | OrbitalEOS internal storage |
| Licensing of the existing data | Pre-processed S1 and S2 imagery are Proprietary while raw S1 and S2 areopen 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) very high resolution ocean fields (to be generated by ML algorithm) and (3) oil spill forecasts matching detections (netCDF4 format) |
| Purpose of the new data and its relation to the project objectives | Very high resolution ocean fields will be used to perform oil spill simulations and estimate the impacts of detected spills. Oil spill detections will be used to validate oil spill forecasts, estimate uncertainties in impact estimates and improve the oil spill model setup. |
| What is the expected size of the data that you intend to generate? | O (100) 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? | N/A |
| 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 | Yes, we will include: spill-identifier, oil-type, sheen_area. thick_area, true_color_area and estimated oil volume when is possible to analyze it. |

| | |
|---|---|
| do not exist in your discipline, please outline what type of metadata will be created and how. | |
| 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 specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement. | Pre-processed S1 and S2 imagery are Proprietary and will not be made openly available. |
| 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 and CMCC Foundation have an internal data sharing strategy already in place. |
| How will the identity of the person accessing the data be ascertained? | |
| 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 CC0, 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? | To be defined by the EOSC team (?) |
| 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 | (1) Development of a ML-supported algorithm capable of generating high resolution ocean fields with relatively low costs. (2) 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 |
| 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, Use Case results will be made available following |
| Will the data produced in the project be usable by third parties, in particular after | Yes |

| | |
|---|---|
| the end of the project? | |
| 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

| | |
|---|---|
| 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 |

### 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 aims to establish an operational service at the iMagine platform for ingestion, storage, analysis and processing of Flowcam images for determining taxonomic composition of phytoplankton samples. We will make available an annotated training set as well as trained classifiers.

## Data Summary

| | |
|---|---|
| Description of the existing data | A. Biomonitoring result data: scientific data, human validated data, aggregated in taxon densities per sample B. Image library: - Image data and related metadata bucket: Image data collection: binary image files and image metadata collection: metadata on sampling, laboratory processing, image parameters, classifications and predictions - Training data split: pointers to images used in model training C. Classifier data: trained models and statistics, scripts |
| Origin of the existing data | A.Biomonitoring Result data: Collected and generated from sampling since 2017 through funding by the Flemish Government. B. Image library: Collected and generated from sampling since 2017 through funding by the Flemish Government<br>C. Classifier Data: Generated after model training |
| 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)<br>- 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 https://rshiny.lifewatch.be/flowcam-data/ (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<br>- 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: 2327381 b /year B. Image library: - Image data and related metadata: +/- 400 000 images/year, (1370767461 b metadata/year, 585294511 b image data/year)<br>- Training data split: depends on training split C. Classifier data: depends on models trained |
| 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 (EMBRC), Biodiversity researchers. |

## Findability

| | |
|---|---|
| Will data be identified by a persistent identifier? | Yes, DOIs will be assigned. |

| | |
|---|---|
| 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: ● Title, abstract, and description which covers the content of the dataset<br>● Name and contact details for the person responsible for the dataset<br>● A citation for the record<br>● Links to access the associated datasets (see Section 4 for more on what these associated datasets<br>are)<br>● Licence of use<br>● Keywords chosen for the dataset<br>● A listing of the parameters measured<br>● The spatial, temporal & taxonomic coverage of the dataset<br>● A listing of people who contributed to the dataset<br>● 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 AI4Europe? | 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 CC0, 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 at https://rshiny.lifewatch.be/flowcam-data/ |

## Data Sharing and Re-use

| | |
|---|---|
| What are the target groups | Biodiversity researchers, environmental policy agencies, biomonitoring actors, ... |
| 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, 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 (https://rshiny.lifewatch.be/flowcam-data/). Additional documentation will be provided in readme files or on github repository. |
| 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 | No |

| | |
|---|---|
| Description of the Action (DoA). | |
| 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

To develop, using the iMagine platform, a prototype service for processing acoustic underwater recordings for identification and recognition of marine species and other noise types (e.g., offshore piling).

### 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.<br>Training Set: Smaller subset of the raw data<br>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, we will see after all data is processed an uploaded to database |
| 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 iMagine project (not raw data) |
| Will you generate any new data? | Yes |
| Description of the new data. | Continuous monitoring will produce more data |
| 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 https://www.iqoe.org/systems |

## Findability

| | |
|---|---|
| Will data be identified by a persistent identifier? | Yes, DOI |
| 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: ● Title, abstract, and description which covers the content of the dataset<br>● Name and contact details for the person responsible for the dataset<br>● A citation for the record<br>● Links to access the associated datasets (see Section 4 for more on what these associated datasets<br>are) |

| | |
|---|---|
| | ● Licence of use<br>● Keywords chosen for the dataset<br>● A listing of the parameters measured<br>● The spatial, temporal & taxonomic coverage of the dataset<br>● A listing of people who contributed to the dataset<br>● 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 |

## Accessibility

| | |
|---|---|
| Will the data be deposited in a trusted repository? | Yes, eventually |
| 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? | 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 CC0, 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? | We will look into standards and vocabularies for this type of data |
| 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, ... |
| 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.)? | Yes |
| Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed | Yes |

| | |
|---|---|
| using standard reuse licenses, in line with the obligations set out in the Grant Agreement? | |
| 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). | 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 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 detect near-shore morphodynamic processes, particularly, the formation and dismantling events of seagrass beach berms (*Posidonia oceanica*) and rip-currents formation.

## Data Summary

| | |
|---|---|
| Description of the existing data | The beach-video monitoring system of the Balearic Islands Coastal Observing and Forecasting System (SIRENA) consists on video monitoring stations mounted at the top of hotels (elevation~30-50 m) located in front of the coastline (Mallorca Island: 2 active stations, 3 inactive station; Menorca Island: 1 active station). Each station is composed of sets of 3 to 5 RGB cameras dedicated to the monitoring of a particular beach, except the 3 inactive stations which were all dedicated to the monitoring of the same long beach. Every hour - including daylight hours at least - all cameras start recording simultaneously during the first 10 minutes of each hour (at 7.5 fps; ~4500 frames/video), and three different products are computed: 1. Snapshot: Frame of the video at the 5th minute; 2. Timex image: Time exposure image of the full video (average of ~4500 frames); 3. 'Variance' image: Standard deviation of the ~4500 frames. |
| Origin of the existing data | Collected/derived from cameras: snapshots, timex, variance. |
| Size of the existing data (Number of images and storage size) | Number of images: 2533056; Storage size: 3543.424 GB |
| Repository where existing data is stored | SOCIB data repository (is CoreTrust Seal certified (see https://www.coretrustseal.org/about/)) |
| 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. | The SIRENA video-monitoring system captures daily images automatically, so the stored data is increasing every hour. Besides, from the iMAGINE project we would like to generate segmented images (different classes such as sand, water, glint, *Posidonia Oceanica* berms, human, buoy, vessel, tree, vegetation, ...) and datasets of presence/absence of rip currents (most challenging task, both from 'labelling' point of view as well as from 'model development' sight). |
| Purpose of the new data and its relation to the project objectives | The main purposes are to set the basis for further unveiling *Posidonia* sp. berms formation/dismantling mechanisms and coast-protection functions, and developing early warning systems for rip currents. Further applications such as automatic shoreline extraction, beach-occupation metrics, and number of vessels could be derived. |
| What is the expected size of the data that you intend to generate? | One segmented image (or another format) for each 'valid' image-timestamp from the stored data, and a dataset with 'identified rip currents' imagery. It is difficult to measure the size right now. |
| To whom might your (existing and newly generated) data be useful ('data utility'), outside your project? | Universities; scientific institutes / researchers; local administration: emergency services, coastal surveillance/managers, lifeguards |

## Findability

| | |
|---|---|
| Will data be identified by a persistent identifier? | SOCIB is now working on DOI's for all data sets, but the process is still in an early stage. During the next few years, all datasets, including SIRENA imagery, will be standardised and identified with a DOI. |

| | |
|---|---|
| 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. | Current metadata (PNG metadata = textual information chunks?) is related to image properties (width, height,format, depth, channel statistics, date-time created...) and also includes 'capture properties' such as FPS and duration. We need to think about which metadata is useful, which can be removed, and which is lacking. For instance, the name of the images follows the structure: 'Station_TypeOfImage_NumberOfCamera_Date-Time.png' (e.g., clm_m_03_2014-02-01-12-00.png).<br>It would probably be more useful to include station name and date created inside the metadata too (user friendly, interpretable), so discoverability can be implemented in the future in a structured manner. In addition, other metadata such as station positioning (x,y,z) of the cameras could be included (*e.g.* Filtering images by longitude and/or latitude not needing to know the name of the SIRENA station) . |
| Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? | We are fully open to working on Metadata enrichment. |
| 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? | SOCIB Data Repository (see https://www.socib.es/data/data-repository/) |
| Have you explored appropriate arrangements with the identified repository where your data will be deposited? | n/a |

| | |
|---|---|
| Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? | Not yet. First, data is stored in the repository near-real time (currently working), later on a DOI is assigned to the dataset which evolves with time (DOI versioning) (Under implementation; expected for 2023-2024) |
| 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, all data will be made openly available |
| 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? | Accessing data does not require identification yet. Registration will be required in a near future (under implementation) using different methods such as google, orcid, and mail. |

| | |
|---|---|
| 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 CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data? | Yes (openly available and licenced under public domain) |
| 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 Data Repository exists (expected: Forever) |
| 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. Data consists of standard '.png' images. |

## 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? | To be defined. |
| 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 | Yes. |

| | |
|---|---|
| openly publish the generated ontologies or vocabularies to 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)? | Not planned |

## Data Sharing and Re-use

| | |
|---|---|
| What are the target groups | Scientific community, local/regional 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 | Currently, all images are displayed in the Beamon app, which only allows for visualisation (https://apps.socib.es/beamon), but it is planned to include direct download in the form of a 'product cart' from the same web. In the future, all SIRENA images should be also available from the SOCIB API (https://api.socib.es/home/). |
| 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.)? | Readme files including data cleaning/preparation codes and further aspects to be defined, depending on the achievements/processes developed during the imagine project. |
| 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. 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. | The existent quality assurance processes ensure that the SIRENA stations are working properly (maintenance labours, UPS system) and images are being stored and sent to the SOCIB server (alert systems, including auto-mailing of system status) |

## 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 | Low-resolution: not possible to identify personal information , for instance, 'car plates'; GDP compliance, blurred areas in one of the SIRENA Stations to avoid near users identification. The blurred area could be extended and applied to all existing and new images. |

### 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)? | CoreTrust Seal - SOCIB data repository: Back up system, cyber security. |
| Will the data be safely stored in trusted repositories for long term preservation and curation? | Yes. CoreTrust Seal - SOCIB data repository |

## 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 iMagine 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 iMagine 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 he iMagine AI platform

## Data Summary

| | |
|---|---|
| Description of the existing data | 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).<br>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)<br>Real Dataset: Real raw microscope images acquired from field samples. Currently ca 100 images, enough for first training of the CNNs. |
| Origin of the existing data | Individual Dataset: Gathered from real images and taxonomic guides (available as open source pdf)<br>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)<br>Real Dataset: Generated by the team on field samples provided by French Biodiversity Agency-OFB (stakeholder in charge of Water Framework Directive-WFD implementation). |
| Size of the existing data | Individual Dataset: ca 1GB (ca 40kb/thumbnail, ca 20,000 thumbnails)<br>Synthetic Dataset: ca 4GB (ca 250 kb/synthetic image, ca 50,000 images<br>Real Dataset: ca 1.5 GB (ca 1.1 MB/image, ca 150 images) |
| Repository where existing data is stored | All the data is stored on the cloud of University of Lorraine (PETA: https://sme.peta.univ-lorraine.fr/)<br>Data used for published papers are stored on open repository (DOREL, https://dorel.univ-lorraine.fr/dataverse/univ-lorraine) |
| Licensing of the existing data | Published data: Open Source Etalab (compatible CC-BY)<br>https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf |

| | |
|---|---|
| | Rest of data: closed |
| 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) |
| Description of the new data. | See the existing data description for the different datasets.<br>Real Datasets: Actual images from field samples that will be generated and annotated. 1 dataset per case study, < 10 case studies during iMagine project |
| 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? | Individual Dataset: few GBs. Target of 1000 diatom species (from 200; realistically to 300-700) and 50-100 images per species.<br>Synthetic Dataset: not sure yet if it will be necessary if we have access to training sets based on real images<br>Real Dataset (case studies): several GBs, under discussion, will depend on the image acquisition approach (for each sample, either a collection of raw microscope images or a mosaic of raw microscope images aka "gigaslide") |
| 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. | not known/considered yet |
| Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? | not known/considered yet |
| Will metadata be offered in such a way that it can be harvested and indexed? | not known/considered yet |

## Accessibility

| | |
|---|---|
| Will the data be deposited in a trusted repository? | Currently, published datasets will be deposited on the academic repository DOREL (https://dorel.univ-lorraine.fr/dataverse/univ-lorraine), connected to the national repository Recherche Data Gouv (https://recherche.data.gouv.fr/en) |
| 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 the "Open Science" standard procedure from Univ. Lorraine => https://scienceouverte.univ-lorraine.fr/en/home/ |
| 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 | all training sets will be made available (useful for training other models). Not sure yet if other data (case-studies) will be or if it is relevant |

| | |
|---|---|
| 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. | not sure yet |
| 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 CC0, 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? | To be defined along the project |
| 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? | To be defined along the project |
| Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? | To be defined along the project |

## 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)<br>researchers (ecology)<br>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 | 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.)? | repositories, readme 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? | ideally yes |
| Describe all relevant data quality assurance processes. | under discussion |

## 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)? | To be defined along the project |
| Will the data be safely stored in trusted repositories for long term preservation and curation? | yes |