Ref. Ares(2023)1465806 - 28/02/2023

Technical development roadmap for the AI image analysis use cases

iMagine Deliverable D3.1

28/02/2023

Abstract

iMagine is a 36-month-long project to serve aquatic researchers with a suite of high-performance image analysis tools empowered with Artificial Intelligence (AI). The iMagine services leverage the AI Platform of the project. Through eight image analysis use cases of iMagine, the Best practices for future adopters will be formulated. The document describes the methodology and the corresponding analysis of the use cases, their development roadmaps, gathered requirements for the AI Platform, as well as the means of tracking those requirements.

Funded by the European Union

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

Document Description

Document title			
Work Package number			
Due date	28/02/2023	Actual delivery date:	
Nature of document	[Report]	Version	1.0
Dissemination level	Public		
Lead Partner	KIT		
Authors	Valentin Kozlov (KIT)		
Reviewers	Names (institutions)		
Public link	10.5281/zenodo.7684346		
Keywords			

Revision History

Issue	Item	Comments	Author/Reviewer
V 0.1	Draft version	Early draft of the plan	V. Kozlov (KIT)
V 0.2	Revised version	First full draft of the deliverable	V. Kozlov (KIT)
V 0.3	Revised version	Reviewed by the use cases	C. Leluschko (DFKI), J-O. Irisson (SU), E. Martinez (UPC), J-M. Baudet (Ifremer), D. Smyth (MI), P. Gaughan (MI), A. Sepp (CMCC), S. Fiore (CMCC), R. Lagaisse (VLIZ), E. Debusschere (VLIZ), J. Soriano (SOCIB), M. Laviale (UL-LIEC)
V 1.0	Submitted version		
V 1.1	Corrections		

Copyright and license info

Table of content

Figures

List Of Images

Tables

List Of Tables

Document structure

Introduction

The use cases in the iMagine project focus on AI imaging services for the aquatic sciences. They represent major Research Infrastructures (RI) in the marine and inland waters domains, namely LifeWatch, EMBRC, JERICO, EMSO-ERIC, EurOBIS, and SeaDataNet, and relevant EU initiatives such as Copernicus Marine Environmental Monitoring Service (CMEMS) and European Marine Observation and Data network (EMODnet).

Purpose of the document

This document describes the methodology and the corresponding analysis of the use cases in the iMagine project, their development roadmaps, the gathered requirements for the AI Platform, as well as the means to track these requirements.

Scope of the document

The D3.1 document presents the initial analysis with the high-level architecture of services, identified user stories and epics, gaps and bottlenecks, and structures development timelines, which are aligned with the general timeline of the project. It summarises current use cases requirements for the AI Platform for both AI Development and AI Serving installations, which allow to understand an average and median case for an image service in aquatic science. The presented information will be monitored and updated accordingly in the next deliverable D3.2 on M17. The requirements related to data management within the eight use cases can be found in the dedicated deliverable, D1.2.

Structure of the document

The deliverable is organised as follows: we first introduce the methodology to analyse the use cases and their requirements, then summarise that analysis for every use case, which also includes the use cases development timelines. In the end, we outline the main requirements for the AI platform, and how they are monitored in the course of the project.

Methodology used

To analyse the use cases, we applied the methodology inspired by the DSDM Agile project Framework[1] and the course "Agile Meets Design Thinking" from Coursera[2]. The analysis is based on the '*Persona-Epics-User stories*' approach, where we attempt to understand the end-users of use cases and what functionalities of the product they are missing. Their problem description is converted into *user stories* and similar user stories constitute an *Epic*. For capturing that information, we developed a template which was filled by the use cases. It includes a brief description of *Persona*, what he/she *Thinks-Sees-Feels-Does* in the current activity (Fig. M1), what *Problems* he/she has, what is a current alternative solution and how it might be solved in a better way (Fig. M2). This information helps us to construct user stories from the point-of-view of users formulated as : "*As a [persona], I want to [do something] so that I can [realise a reward]*". The similar user stories are combined in Epics. The filled by use cases documents are listed in Technical Development Roadmap[3]. These documents provide us with qualitative information about use cases.

In parallel we collect quantitative information from use cases relevant for every stage of the AI/ML/DL development and serving: *Data sources*, *Data preparation*, *Modelling*, *Model serving*; and *CI/CD*. This information is filled by the use cases in another document[4] to the best of their knowledge. This helps to quantify the requirements for the AI Platform.

The information filled by the use cases in the aforementioned documents was presented by them during the face-to-face Competence Centre Workshop[5] and discussed in person.

Results and the analysis derived from the applied approach are depicted in two next sections: "Summary of the use cases analysis" and "(initial) Requirements for the platform".

[1] https://www.agilebusiness.org/dsdm-project-framework.html

[2] "Agile Meets Design Thinking" course, offered by University of Virginia. https://www.coursera.org/learn/uva-darden-getting-started-agile

[3] Preparation for D3.1 - Technical development roadmap (list of 'Persona-Epic-User stories' documents for each use case): https://docs.google.com/document/d/1UTdWzxuqKfdQyh5kiHTlgaQfspR6GeF1OHzyXwvGenQ

[4] iMagine - Gathering Technical Aspects (Sheet): https://docs.google.com/spreadsheets/d/1a1fdOd95quL-IVSbzy2SlxVXSLFBAHwz3AUBNzDp6uU/

[5] iMagine Competence Centre Workshop in Villefranche, 30-31.01.2023, https://indico.egi.eu/event/5999/

Persona A

Some general info about the persona.

Thinks	*[Persona] thinks [things should be different in a certain way]. This is important because [why?]*
Sees	*[In certain situation], [persona] sees [key observation of importance]. [Repeat, etc]*
Feels	*When [some event], persona feels [emotion]. It's [cause] that makes them feel this way.*
Does	*[Persona] [does activity] [x] times per [period].*

Figure M1 - Persona description table in the template

Problem Scenario + Alternatives Pairs + Value proposition

Problem Scenarios / Jobs-to-be-Done	Current Alternative	Value proposition
[What problems, needs does persona have in the area?]	*[What do they do right now to solve this problem/meet this need?]*	*[What product ideas do the problem scenarios and current alternatives give you?]* *[What component of the AI platform may solve the problem?]*

Figure M2 - Problem description from a persona perspective in the template

Summary of the use cases analysis

There are five operational (UC1-5) and three prototype (UC6-8) image analysis services with image repositories which are highly relevant for the overarching theme 'Healthy oceans, seas, coastal and inland waters'. All of the use cases aim to either enhance operational services or develop new components based on AI/ML/DL by leveraging the iMagine AI platform. The following sections summarise development plans for every use case in terms of envisaged service's high-level architecture integrated with the iMagine AI platform, identified epics, user stories, gaps and bottlenecks, and the development timeline. The latter is aligned with the project's general timeline (Fig. G1), where the main stages are: Development guidelines (already provided by WP4 - AI and Infrastructure Services[6] on M3); Development roadmap (this deliverable); Update of the development

[6] iMagine D4.1 Best practices and guideline for developers and providers of AI-based image analytics services: https://doi.org/10.5281/zenodo.7372358

roadmap on M17; release of mature use cases (UC1-5) at full scale by M24; providing Best practices for image analysis, based on the lessons/experiences gained by mature use cases during the development; and validation of prototype services on M35. In total we identify 35 user stories for all use cases together.

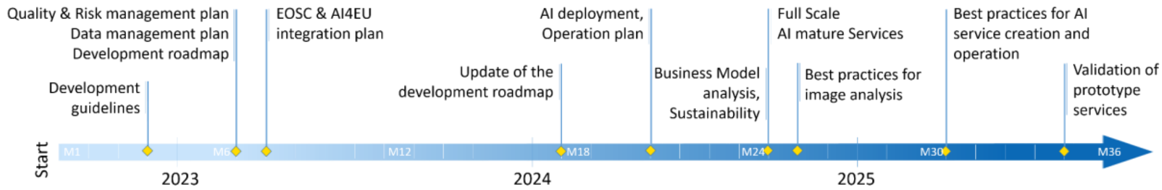

Figure G1 - General timeline of the project

The summary of the current expertise relevant to the AI/ML/DL development and existing training datasets is shown in [Table G1](). Together with the communication channels described in the "Requirements tracking" section, it helps to stimulate the know-how exchange and synergies within the project.

Use Case	Labelling	Training datasets	AI/ML/DL expertise
UC1	Custom tools + Experts + Student Workers	Exists (subset)	Exists In-house (e.g. APLASTIC-Q)
UC2	Zooprocess + EcoTaxa (experts)	Exists	Exists In-house + subcontracted
UC3o	Custom tools + Experts	Exists (2-year dataset of annotated images, subset)	developing
UC3a	DeepSeaSpy + Citizens + Experts	Exists	developing
UC3s	Web-based ML (experts)	Creating	developing
UC4	No need	Creating	Exists in-house
UC5	In-house solution (experts)	Exists	Exists + developing
UC6	Various tools (Raven, Audacity, PlaVA) + experts	Creating	developing
UC7	LabelStudio (experts +	Creating	developing

Use Case	Labelling	Training datasets	AI/ML/DL expertise
	student workers)		
UC8	LabelBox for segmentation (experts, students)	Exists but needs to be extended	Exists in-house

Table G1 - Use cases summary relevant for the AI development

UC1 Marine litter assessment

Use case overview

The use case is going to establish an operational service at the iMagine platform for ingestion, storage, analysis and processing of drone images (see Fig. UC1.1) , 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[7] 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. The training subset is published on Zenodo.

[7] APLASTIC-Q Github: https://github.com/DFKI-NI/APLASTIC-Q; Mattis Wolf et al., 2020 Environ. Res. Lett. 15 114042, 10.1088/1748-9326/abbd01

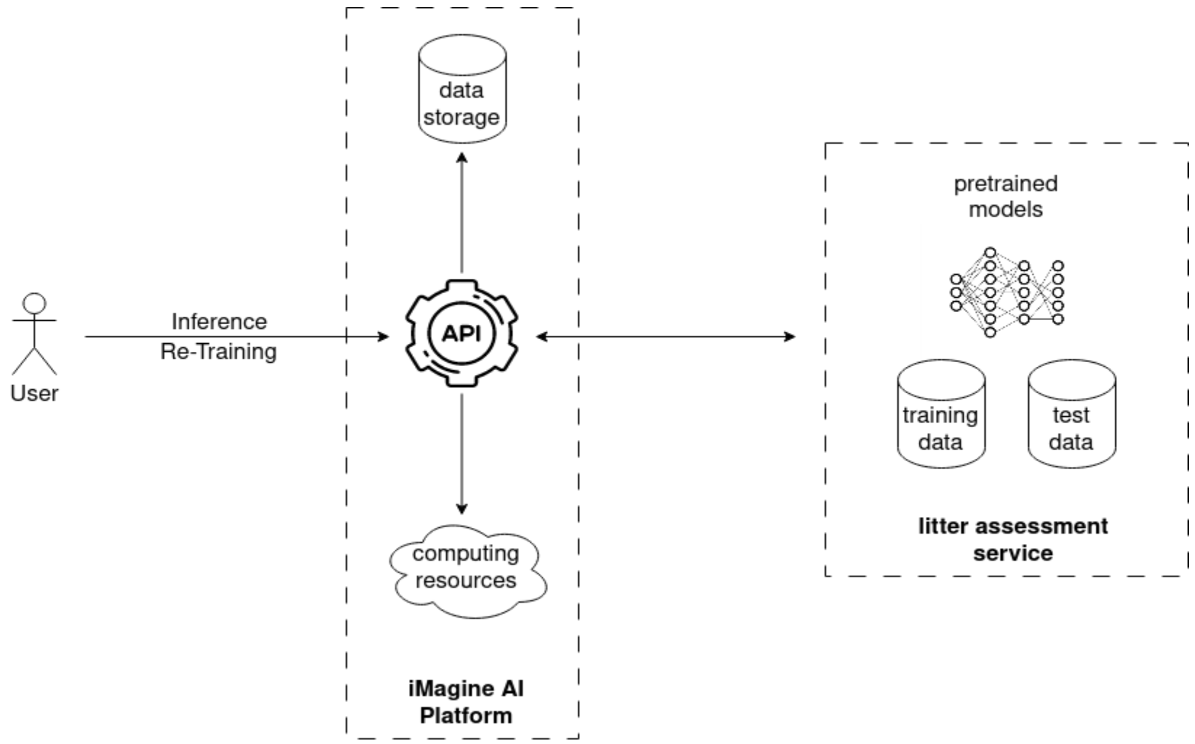

Figure UC1.1 - High-level architecture of the UC1 image service

Epics and User stories

During the process of capturing of user requirements, the follow Epics and User stories were identified:

Epic UC1.E1 "Using the pre-trained models for a quantitative analysis of litter distribution"	
Personas	1. A person working at NGO and performs cleanup missions 2. Environmental manager, establishes strategies for the plastic waste management
UC1.E1.US1	A user wants to perform breakdown of total number of items per litter category, in order to make easier the sorting of litter and/or prepare guidelines for plastic waste management
UC1.E1.US2	A user wants to accomplish breakdown of number of items per input image, so that he/she can plan the cleanup procedures more efficient
UC1.E1.US3	A user wants to receive information about the development over time of the individual waste categories, so that he/she can evaluate and improve the plastic waste management
Epic UC1.E2 "Fitting the pre-trained model to individual data"	

Personas	1. Researcher who wants to retrain and apply the AI model on his/her particular data
UC1.E2.US1	A user wants to retrain the model with individual data, so that the model is optimised for his/her particular case
UC1.E2.US2	A user wants to test the retrained model by using provided test data, in order to verify that the model generalises and performs well

Table UC1.1 - Identified Epics and User stories for UC1

Gap and Bottlenecks analysis

Currently, the service lacks easy usability and there is a number of manually involved steps. The following missing functionalities are identified and going to be added in the course of the project, including the usage of the iMagine platform:

- Easy storage and access for custom data
- User-friendly API
- Ready-to-use environment (e.g. Docker)
- Information about required image processing
- Simple usage of provided test data for retrained model
- Documentation / step-by-step guide

Development roadmap

The development time plan is depicted in the Fig. UC1.2. It starts with the integration of the existing AI model[8] with the iMagine platform and configuration of the API for basic inference. We are going to leverage the data storage available in the platform. We continue with the development of the service to allow authentication and retraining. We extend the service with more data and a broader choice of models and enhance the model output with additional metadata. The documentation and step-by-step guides will be prepared and provided to the users, which also will help for dissemination and exploitation of the service. Once the updated service is available for customers, e.g. via the EOSC Marketplace, we will monitor user feedback and implement it accordingly.

[8] See previous footnote for reference

	2022		2023						2024						2025			
Task \ Project Month	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**	**36**
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
Model integration into iMagine platform (UC1.E1.US1-2)																		
Configure API for basic inference (UC1.E1.US1-2)																		
Allow re-training of the models via API (UC1.E2.US1-2)																		
Extension and provision of additional input data and models (UC1.E2.US1-2)																		
Enhance of model output by additional metadata																		
Documentation / Step-by-step guide																		
Adaption to user feedback																		

Figure UC1.2 - Development timeline for UC1

UC2 Zooscan - EcoTaxa pipeline

Use case overview

The use case provides processing of zooplankton images taken using the Zooscan and aims to establish an operational handling service (Fig. UC2.1) at the iMagine platform that ingests, stores, processes images of marine water samples and uploads the resulting regions of interest to the EcoTaxa platform for later taxonomic identification. The technology, planned for implementation, is based on the processing of grayscale images of 356 megapixels with classical image segmentation and measurement methods, which are further improved through neural network algorithms, in that case instance segmentation. Then, EcoTaxa uses a combination of deep and classic machine learning to predict likely identifications for the uploaded images and a dedicated user interface to validate those.

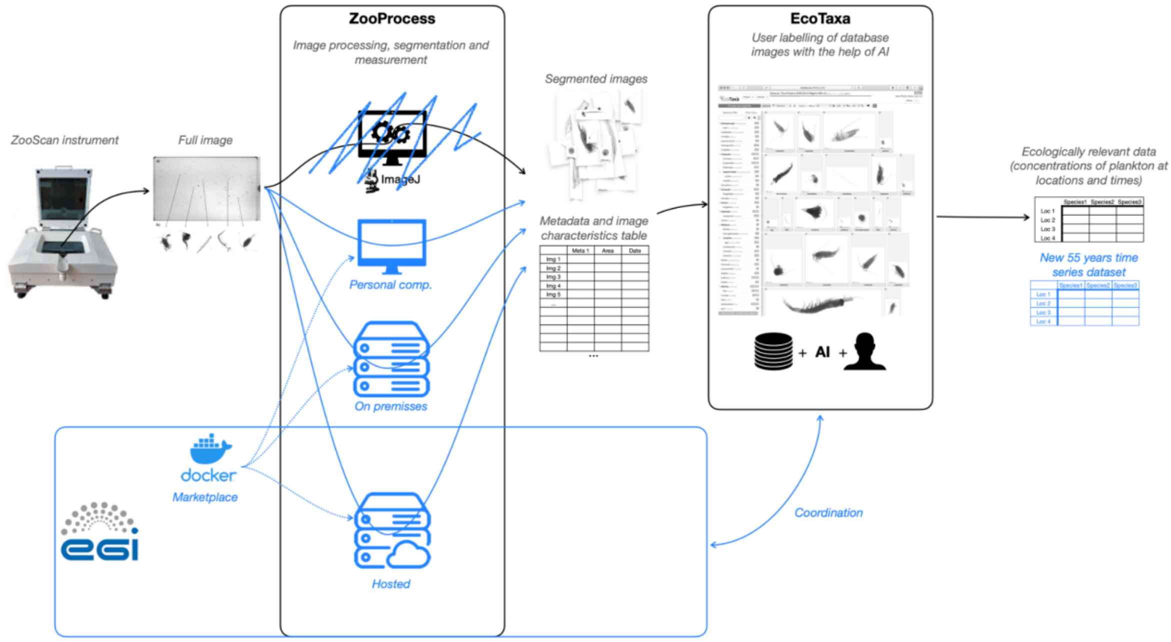

Figure UC2.1 - High-level architecture of the UC2 image service

Epics and User stories

The following Epics and User stories were formulated in the analysis of the use case:

Epic UC2.E1 “Simplification of the plankton samples digitisation”	
Personas	1. A technician using the ZooScan and EcoTaxa
UC2.E1.US1	As a user I want to automate separation of touching organisms in the taken samples so that the digitisation process becomes faster
UC2.E1.US2	As a user I want to import the processed data into EcoTaxa, so that the biological objects can be properly identified
Epic UC2.E2 “Metadata propagation in compliance with DwCA conventions”	
Personas	1. A scientist wanting to analyse a plankton sample and publish the results
UC2.E2.US1	As a researcher I want to automate metadata propagation from the data acquisition to the export from EcoTaxa, so that exported datasets are better compliant with the DwCA conventions

Table UC2.1 - Identified Epics and User stories for UC2

Gap and Bottlenecks analysis

A technician who is responsible for digitising plankton samples, spends several hours a day handling plankton samples, scanning them with the ZooScan, running custom software to process the images, manually correcting some of the automated processing mistakes (in particular separating organisms that touch each other on the processed images, which would lead to incorrect data), importing the resulting images on EcoTaxa and sorting them taxonomically. The process is tedious and little automated.

For publishing and analysing a dataset some metadata is important: volume observed, net used, imaging instrument, imaging settings, etc. There are terms in the controlled BODC vocabularies[9] to document them and those are mentioned in a best practices document[10]. Those should be used in a DarwinCore Archive (DwCA[11]) file to document the data. A researcher, e.g. a plankton ecologist, often has little time to look this metadata up and to create the DwCA file using those conventions. He/she thinks that data processing, management and distribution should be better automated.

Development roadmap

We start with curating the training datasets. Once there is a large enough training dataset, we assess different instance segmentation models to separate the organisms by means of the iMagine platform. In the meantime, we write the specifications for Zooprocess v2, which should reproduce the main features and image processing of the current Zooprocess but be implemented as a client-server web application. In addition, to solve the metadata bottleneck and improve the compliance with the DwCA conventions, we ask the relevant metadata during data acquisition and make sure the pipeline carries that metadata and its BODC mapping all the way to EcoTaxa. The developed and trained model will then be integrated into Zooprocess v2. Finally, EcoTaxa can create the DwCA file with correct identifications and rich metadata. Once the service is ready, it will be deployed and users of the service will get trained.

[9] https://www.bodc.ac.uk/resources/vocabularies/

[10] Martin-Cabrera, P., Perez Perez ,R., Irisson, J-O., et al, (2022) Best practices and recommendations for plankton imaging data management: Ensuring effective data flow towards European data infrastructures. Ostend, Belgium, Flanders Marine Institute, 31pp. DOI: http://dx.doi.org/10.25607/OBP-1742 ,
URI https://repository.oceanbestpractices.org/handle/11329/1917

[11] http://www.eurobis.org/data_formats

	2022		2023						2024						2025			
Task \ Project Month	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**	**36**
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
Machine learning tasks																		
Curate training dataset																		
Train instance segmentation model																		
Integrate model in iMagine platform																		
Software pipeline tasks																		
Write detailed specs of the pipeline																		
Implement general architecture																		
Implement image processing																		
Integrate instance segmentation																		
EcoTaxa improvements																		
Deployment and training																		
Deploy!																		
Explain and train users																		

Figure UC2.2 - Development timeline for UC

UC3 Marine ecosystem monitoring at EMSO sites (OBSEA, Azores, SmartBay)

Use case overview

The use case performs underwater video monitoring and aims to establish 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 of ecosystem monitoring. There are three EMSO sites in the project: EMSO-Obsea (UPC), EMSO-Azores (Ifremer), and EMSO-SmartBay (Marine), with different capacity for analysis and handling of taken data. Therefore, the use case analysis was handled separately for each of the sites and presented in the following. Figures UC3o.1, UC3a.1, and UC3s.1 represent high-level service architecture for every EMSO site in the project.

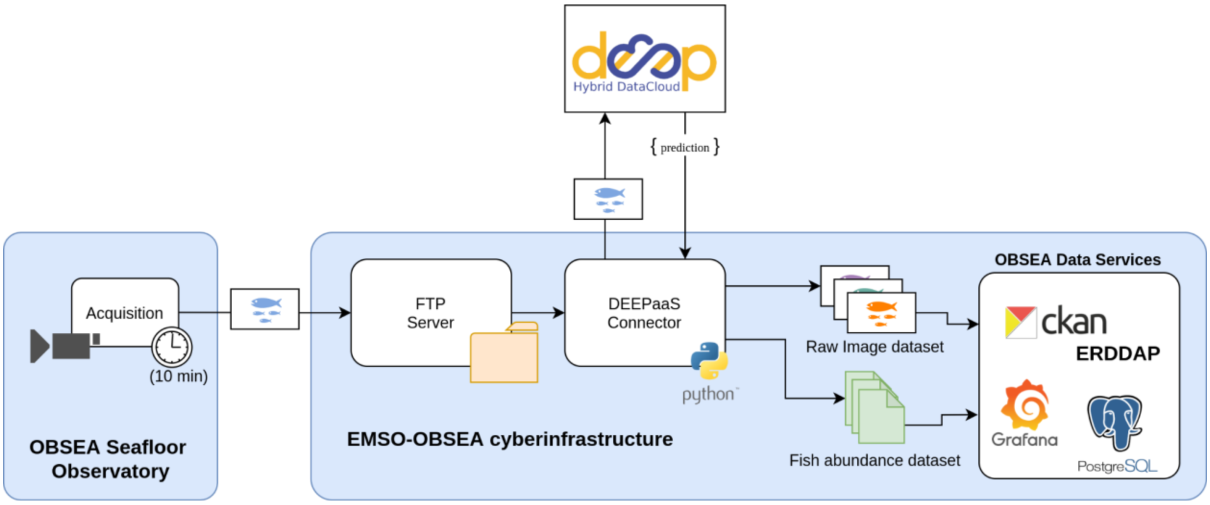

Figure UC3o.1 - High-level architecture of the UC3-Obsea image service

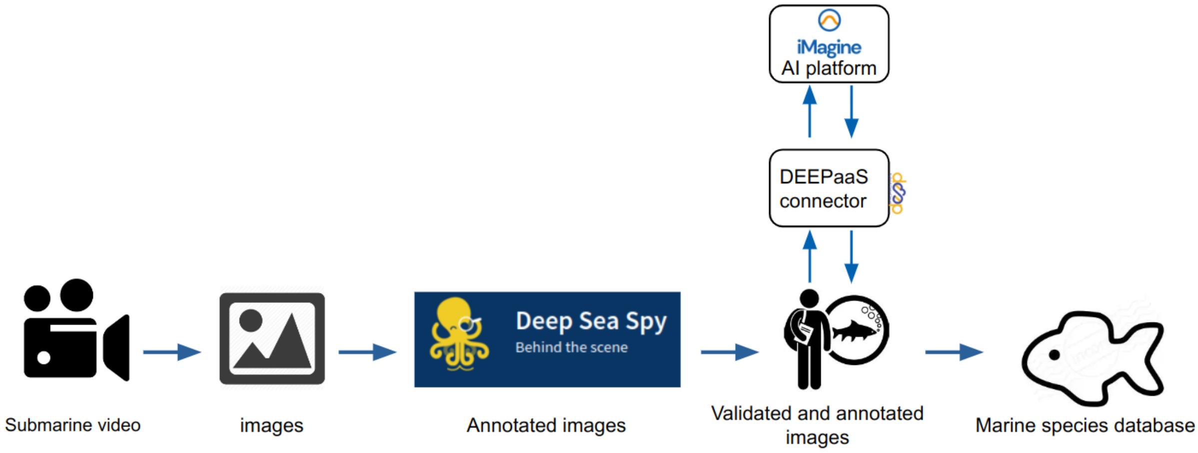

Figure UC3a.1 - High-level architecture of the UC3-Azores image service

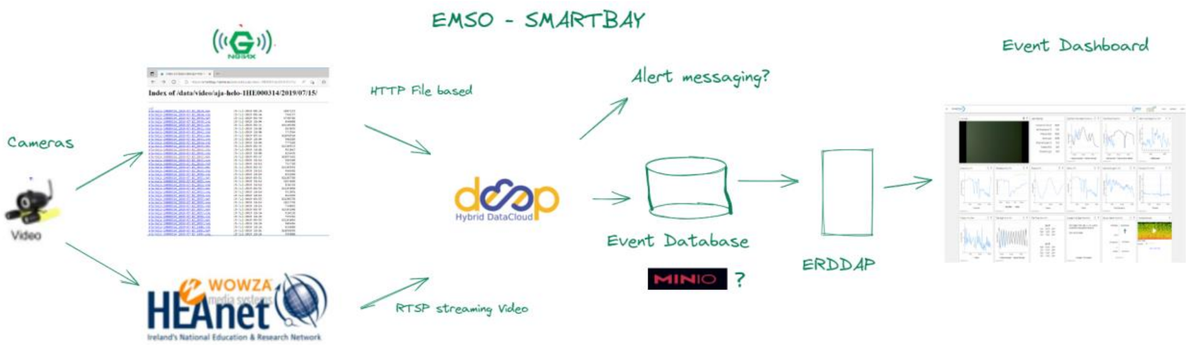

Figure UC3s.1 - High-level architecture of the UC3-SmartBay image service

Epics and User stories

In the process of capturing user requirements, the Epics and User stories listed in the Tables UC3o.1, UC3a.1, UC3s.1 were spotted for the three EMSO sites.

Epic UC3o.E1 "Increase usability of underwater pictures"	
Personas	1. A data manager with a lot of underwater images 2. A marine scientist studying fish communities in shallow waters
UC3o.E1.US1	As a data manager, I want to develop and deploy a Deep Learning based service for underwater images, so that to increase the outreach of underwater pictures that are produced in real-time
UC3o.E1.US2	As a marine scientist, I want to have a service for automated analysis of underwater images, so that I can study the fish community abundance and behavioural patterns in a certain location

Table UC3o.1 - Identified Epics and User stories for UC3-Obsea

Epic UC3a.E1 "Improve dataset of annotated and validated submarine images"	
Personas	1. A citizen using https://www.deepseaspy.com 2. A biological imaging engineer validating annotations done by citizens 3. A biology researcher, who is analysing the deep sea with the data collected
UC3a.E1.US1	As a citizen I want to help scientists to identify marine species on images on the participative science project https://www.deepseaspy.com. This is important because he helps to explore the deep sea. I want an AI-based service which helps the annotation, so that annotation of images can be faster and more images can be processed. For now images that are randomly presented to citizens in deepseespy.com, could be grouped by species to accelerate annotation.
UC3a.E1.US2	As a biological imaging engineer, I want to use AI-based service to validate images annotated by citizens in the participative science project (https://www.deepseaspy.com), so that the validation process is less time-consuming and more images can be processed. The expert would be able to easily compare the manual annotations and the predictions, and can decide to re-train the model to obtain more accurate predictions.
UC3a.E1.US3	As a biology researcher, I want to analyse marine species on a large dataset of annotated and validated submarine images to improve knowledge in marine science

Table UC3a.1 - Identified Epics and User stories for UC3-Azores

Epic UC3s.E1 "Identify events of interest in the underwater videos"	
Personas	1. A data manager at EMSO - Smartbay 2. A Scientific Technical officer at EMSO - Smartbay
UC3s.E1.US1	As a data manager, I want to identify time periods, poor quality video footage (Dark, poor visibility, technical Faults, Glass Fouling), so that I can remove unusable footage and preserve storage for more useful footage
UC3s.E1.US2	As a data manager, I want to monitor video footage in real-time to identify poor video quality events, so that a real time alert on visual quality issues is produced and footage issues can be addressed
UC3s.E1.US3	As a scientific technical officer, I want to identify video footage in the Archive or in real-time with unusual species detection events, so that valuable news items are reported on e.g. SmartBay website promoting the Underwater Observatory
UC3s.E1.US4	As a scientific technical officer, I want to identify and count Prawns(Nephrops) and prawn Burrows, so that to assist research projects

Table UC3s.1 - Identified Epics and User stories for UC3-SmartBay

Gap and Bottlenecks analysis

All three sites collected a large amount of images and videos, which can be more efficiently exploited using AI methods and the iMagine AI platform.

EMSO-OBSEA site

There is a lot of image data from OBSEA that is not exploited. This data is gathered from an underwater camera, where different species of fish are observed. It would be nice to exploit this data to increase the scientific outreach of OBSEA's data. Applying AI tools to existing images would make it possible to extract important biological content from the pictures, generating derived datasets that could be easily used by marine scientists to extract ecological conclusions. There are thousands (eight years) of pictures that have not been analysed. Analysing these images manually is very time-consuming and is a major drawback for large image datasets. However, analysing only a subset of the dataset represents a loss of important information. We are going to exploit the iMagine platform to train and deploy a Deep Learning service to get species' abundance data from existing (and future) images.

These derived data will be important for marine scientists to create and analyse time-series of species presence/absence and changes in abundance along different years. Relating these time-series to the environmental parameters, collected by the OBSEA environmental sensors, will help to get conclusions on the effect of climate change on the local fish community. Moreover, performing time series analysis on long time-series of fish counting will be important to better describe the biological rhythms (at seasonal and diel level) of the different species present at the OBSEA.

EMSO-Azores site

The imagery data collected with the EMSO-Azores observatory should be analysed. The annotation of images can be carried out by citizens through the Deep Sea Spy platform. Data produced thanks to that participative science project needs to be validated by experts. Currently this is done manually and is time-consuming. Expanding the dataset of annotated and validated submarine images is important for biology researchers to improve knowledge in marine science. The iMagine AI platform is going to be used to develop and deploy the AI models which will help to annotate images automatically and to validate annotated images.

EMSO-SmartBay site

It is important to flag poor quality video footage in the Observatory Archive and in real-time because, e.g. Complete Darkness, Algal growth, suspended particulate matter reduction, Equipment failure affect the utility of observatory footage. Manual inspection of the video archive would be overly time consuming. Similarly, to inspect footage for interesting observations or "Novelty" occurrences, or to detect and count Prawn burrows in the field of vision of 2 observatory cameras, is time consuming. This is where the iMagine AI platform may help to develop and deploy the service to allow quick detection of issues and quick responses or for flagging and referencing of interesting "Novelty" footage, or to detect and enumerate Prawns and prawn burrows.

Development roadmaps

Each of the EMSO sites participating in the project established their planning for the developments within the iMagine project, which are presented in Figures UC3o.2, UC3a.2, UC3s.2 below. The sites will look to share data and experiences using labelling and training tools etc. The utility of both image and video Classifiers will be investigated as part of the development road map, as all 3 sites record video as well as imagery.

EMSO-OBSEA site

The development roadmap (Fig. UC3o.2) for EMSO-OBSEA will consist in creating a workflow to automatically process underwater pictures in real-time, extracting fish abundance and taxa. This workflow consists of two different steps: segmentation and classification. The segmentation focuses on selecting the region of interest where a fish specimen is present. After segmentation, the extracted regions of interest will be passed to a classification algorithm that will determine the taxa. The abundance and taxa

information will be compiled into time series datasets, which will be much easier to analyse by scientists than the raw images.
Due to the relatively large dataset already available, from month 6 to 20 several state-of-the-art segmentation / classification algorithms will be benchmarked (model development / training). Although there is already a 2-year long labelled dataset, it is expected that some adjustments on the dataset will be required by the models (data preparation). Due to the ambient variability the concept of dataset shift will be investigated in a later stage to improve the accuracy of the predictions.
Once a final model is developed and deployed, it is expected to ingest the legacy data into the system from month 18 to 24. Afterwards, from month 22 to 36 the workflow will be put into production to analyse underwater images in real-time. In parallel, the predictions will be scientifically exploited to extract information on the long-term biological rhythms of the fish community.

EMSO-Azores site

The development roadmap (Fig. UC3a.2) consists of creating the pipeline for images annotated in deepseaspy.com. This pipeline includes development of software to transform annotations in a suitable format for image segmentation by AI model; existing labelling tools can be tested and used. The development and/or implementation of software tools for the analysis and validation of training and test datasets. The training of several existing segmentation models with several training dataset augmentation techniques (like increasing image contrast, vertical and horizontal flipping, rotating …). The video analysis will be investigated for motion detection, and video segmentation of animals species.

EMSO-SmartBay site

Fig. UC3s.2 shows the development road map for the EMSO-SmartBay Observatory. For the implementation of the user stories described in the Table UC3s.1, we start with the data integration into the platform and updating of the corresponding workflows of the SmartBay site. In parallel, we work on the video data labelling for enlargement of the training dataset. We will investigate various segmentation, object detection and classification algorithms for video data to identify poor quality video footage. The Concept of "Dataset shift" where video image quality deteriorates over time affecting predictions due to algae, dirt algae etc growing on the camera glass in the case of the imagery/video will also be investigated. Once a good AI solution is found, it will be integrated in the SmartBay service. It will be made available as a part of the service around month 20-22. After that we start collecting the feedback and exploiting the new functionality.

	2022		2023						2024						2025			
Task \ Project Month	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**	**36**
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
Data Preparation (UC3o.E1.US1)																		
Model development (UC3o.E1.US1)																		
Model Training (UC3o.E1.US1)																		
Data Integration (UC3o.E1.US1)																		
Model Serving (UC3o.E1.US1-2)																		
Data exploitation (UC3o.E1.US2)																		

Figure UC3o.2 - Development timeline for UC3-Obsea

	2022		2023						2024						2025			
Task \ Project Month	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**	**36**
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
Data Integration (UC3a.E1.US1-3)																		
SQA implementation (UC3a.E1.US1-3)																		
Model development, training, validation (UC3a.E1.US1-3)																		
Model Serving (UC3a.E1.US1-3)																		

Figure UC3a.2 - Development timeline for UC3-Azores

	2022		2023						2024						2025			
Task \ Project Month	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**	**36**
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
Data integration (UC3s.E1.US1-4)																		
SQA implementation (UC3s.E1.US1-4)																		
Model development, training, validation (UC3s.E1.US1-4)																		
Model Serving (UC3s.E1.US1-4)																		
Model Output exploitation																		

Figure UC3s.2 - Development timeline for UC3-SmartBay

UC4 Oil spill detection

Use case overview

The oil spill monitoring and forecasting system OKEANOS is currently in place and fully operational, supporting public institutions and the private sector. Within the iMagine project, we aim to establish an operational service at the iMagine platform for automatic processing of satellite images for detecting oil spills as an extra component with higher accuracy and spatially refined oil spill forecasts. The technology used for the OKEANOS forecasting component relies on open and quality-controlled inputs (meteo-oceanographic fields, bathymetry and coastline geometry). The monitoring component equally relies on open and quality controlled satellite imagery (Sentinel 1, 2 and 3 constellations). The high-level architecture of the service with the iMagine platform extension is shown in the Fig. UC4.1.

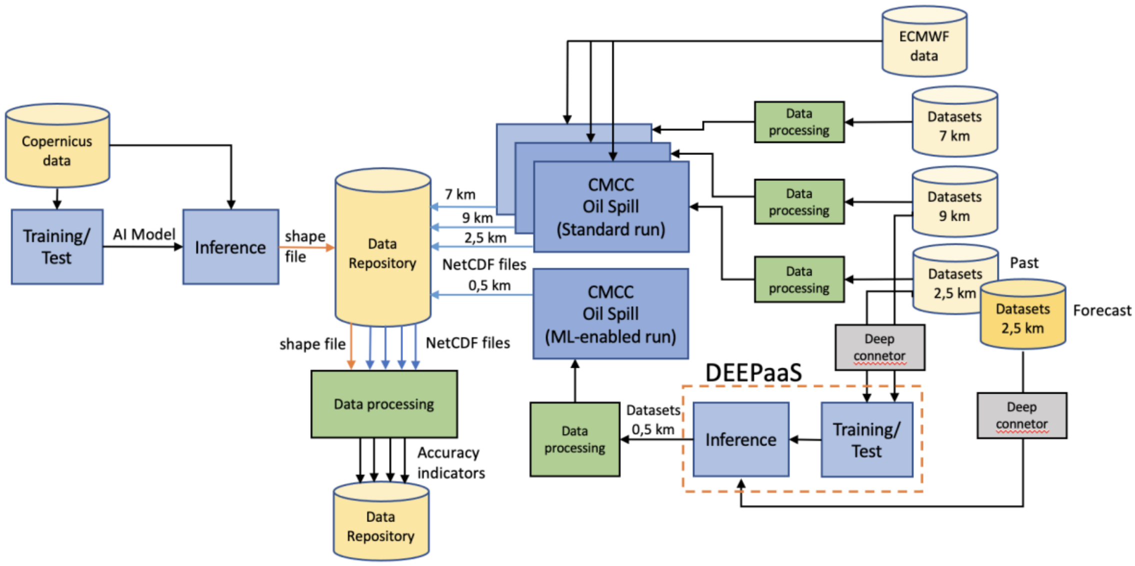

Figure UC4.1 - High-level architecture of the UC4 image service

Epics and User stories

In the course of the project we are going to work on the following user stories, listed in the Table UC4.1

Epic UC4.E1 "Improve overall oil spill monitoring & forecasting system accuracy"	
Personas	1. A provider of oil spill monitoring services 2. An expert in oil spill modelling
UC4.E1.US1	As a provider of oil spill monitoring services and/or expert in oil spill modelling, I would like to improve my ML algorithm in place increasing the number of hits and decreasing false alarms and misses so as to set a higher standard in the market
UC4.E1.US2	As a provider of oil spill forecasting services and/or expert in oil spill modelling, I would like to quantify the accuracy of my forecasts so as to (1) give a clear idea of their uncertainties to users and (2) set a higher standard in the market
UC4.E1.US3	As a provider of oil spill forecasting services, I would like to deliver high resolution and accurate forecasts to fulfil my user requirements without significantly impacting costs
Epic UC4.E2 "Seamless end-to-end workflow"	
Personas	1. A manager of an operational oil spill forecasting chain
UC4.E2.US1	As manager of an operational oil spill forecasting chain, I would like to implement a "smooth" workflow reducing downtime and average production time.
Epic UC4.E3 "FAIR-enabled marine products catalogue"	
Personas	1. A marine scientist
UC4.E3.US1	As a scientist I would like to perform data browsing and access to browse collections and download data easily
UC4.E3.US2	As a scientist I would like to have a FAIR-enabled data repository for my products, so that our repo could be better exposed to the iMagine marine scientists and beyond
UC4.E3.US3	As a scientist I would like to perform search & discovery on the catalogue of products from a nice User Interface, in order to easily find products

Table UC4.1 - Identified Epics and User stories for UC4

Gap and Bottlenecks analysis

OKEANOS oil spill forecasts still lack an appropriate quantification of uncertainties. The identified issue is a general problem in the oil spill forecasting field due to the lack of quality-controlled observations and well-established validation methods. The limited capability of ocean and atmospheric models to reproduce small scale features of a few hundreds of meters has played a major role in the oil spill forecast accuracy. Although possible, the implementation and operation of very high resolution meteo-oceanographic models to supply the oil spill model with equally resolved inputs was found to be expensive and time consuming. We are going to leverage the iMagine AI platform for:

- Improving the accuracy of algorithms for automatic oil spill detection and classification using Sentinel 1 & 2 and Landsat 8 imagery;
- Improving the accuracy of numerical oil spill forecasts, paramount in predicting the impacts of detected slicks and identifying polluters.

Development roadmap

Fig. UC4.2 shows our development timeline for the use cases described above. We start with the improvement of AI/ML algorithms, planning of the "smooth" workflows, and organisation of already existing data into downloadable collections. Then, we continue working on the high resolution inputs and accurate forecasts, also quantifying the accuracy of the forecasts. The data repository of marine products will be prepared in accordance with FAIR principles with easy search and discovery. It is planned to have two releases of services before M24.

	2022		2023						2024						2025			
Task \ Project Month	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**	**36**
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
UC4.E1.US1																		
UC4.E1.US2																		
UC4.E1.US3																		
UC4.E2.US1																		
UC4.E3.US1																		
UC4.E3.US2																		
UC4.E3.US3																		
Model serving																		

Figure UC4.2 - Development timeline[12] for UC4

UC5 Flowcam plankton identification

Use case overview

Phytoplankton has a key function in the aquatic food web and produces energy for other marine life. The use case aims to establish an operational service at the iMagine platform for ingestion, storage, analysis and processing of FlowCAM images for determining

[12] *marked in blue: first release, grey colour indicates second release*

taxonomic composition of phytoplankton samples. The technology to be used is based on a deep learning image recognition algorithm based on a Convolutional Neural Network (CNN) in combination with a NoSQL MongoDB database. Adaptation of model Parameters and determining classes to stain on is done through Python scripts. The existing workflow is going to be improved leveraging the iMagine AI platform. For high-level architecture, see Fig. UC5.1.

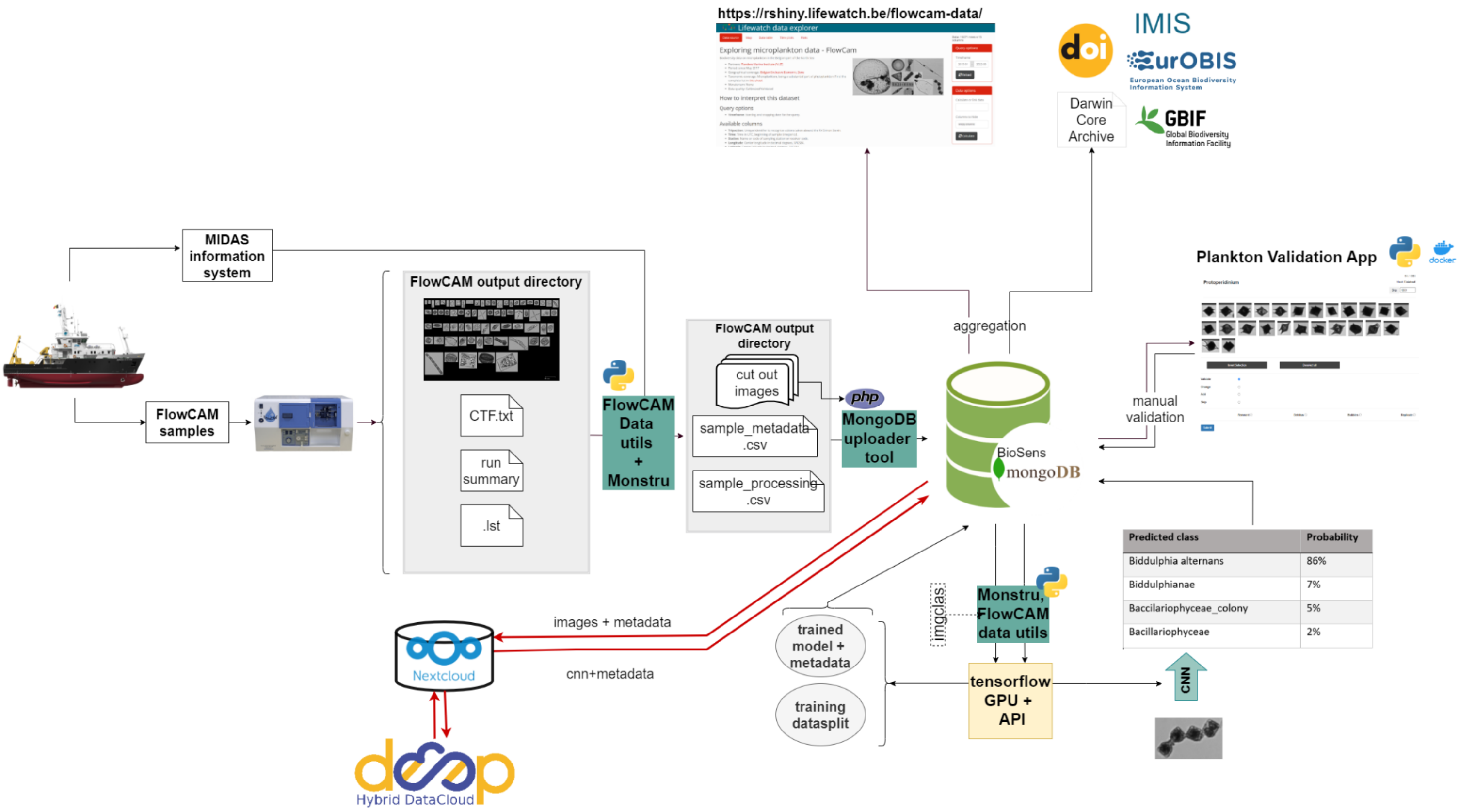

Figure UC5.1 - High-level architecture of the UC5 image service

Epics and User stories

During the iMagine project, we envisage the User stories identified in the analysis phase and described in the Table UC5.1.

Epic UC5.E1 "Improve biodiversity monitoring using pre-trained AI models"	
Personas	1. A researcher studying sediment samples 2. A taxonomist who needs to validate images 3. A scientist from a Research Infrastructure (RI) monitoring plankton through imaging techniques
UC5.E1.US1	As a sediment researcher, I want to easily finetune model training on my own training set so that the model is better optimised for my

	needs.
UC5.E1.US2	As a taxonomist, I want to use existing well-performing models in order to speed up my work on validating FlowCAM images.
UC5.E1.US3	As a scientist at RI, I want to label my plankton images using pre-trained AI models, so that the process is less time-consuming and not labour-intensive.

Table UC5.1 - Identified Epics and User stories for UC5

Gap and Bottlenecks analysis

The following challenges were identified and will be tackled within the project using the iMagine AI platform:

- Optimise existing data ingestion pipeline from sensor to database
- Improve current metadata & data output formats towards compliance with community-based standards and vocabularies
- Improve the service to incorporate the context input and increase the classification accuracy
- Extend the training dataset by identification of additional particles currently grouped under a rest class
- Prepare the data and processing components for connection, synchronisation and migration to enable access from the iMagine platform
- Ecotaxa comparison: need to make same training set available and train similar models

Development roadmap

We are going to work on all three user stories in the course of the whole project (Fig. UC5.2). In order to properly address the user stories, we start with data integration and model development. Once the model is well-trained and satisfies the minimum defined performance, we start providing it to the end users utilising the iMagine AI platform. We monitor their feedback in order to steer next iterations of the service and model development.

	2022		2023						2024						2025			
Task \ Project Month	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
UC5.E1.US1													—>>	—>>	—>>	—>>	—>>	->>
UC5.E1.US2													User feedback monitoring					
UC5.E1.US3													—>>	—>>	—>>	—>>	—>>	->>
Data Integration																		
SQA implementation																		
Model Development																		
Model Serving																		

Figure UC5.2 - Development timeline for UC5

UC6 Underwater noise identification

Use case overview

Underwater sound is essential for most aquatic life and an important tool to survive. It is a complex mixture of biotic, abiotic and man-made sound sources. Underwater noise is recognised as a pollutant by EU MSFD. The use case is going 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). The general high-level architecture of the prototype service is depicted in [Fig. UC6.1](#).

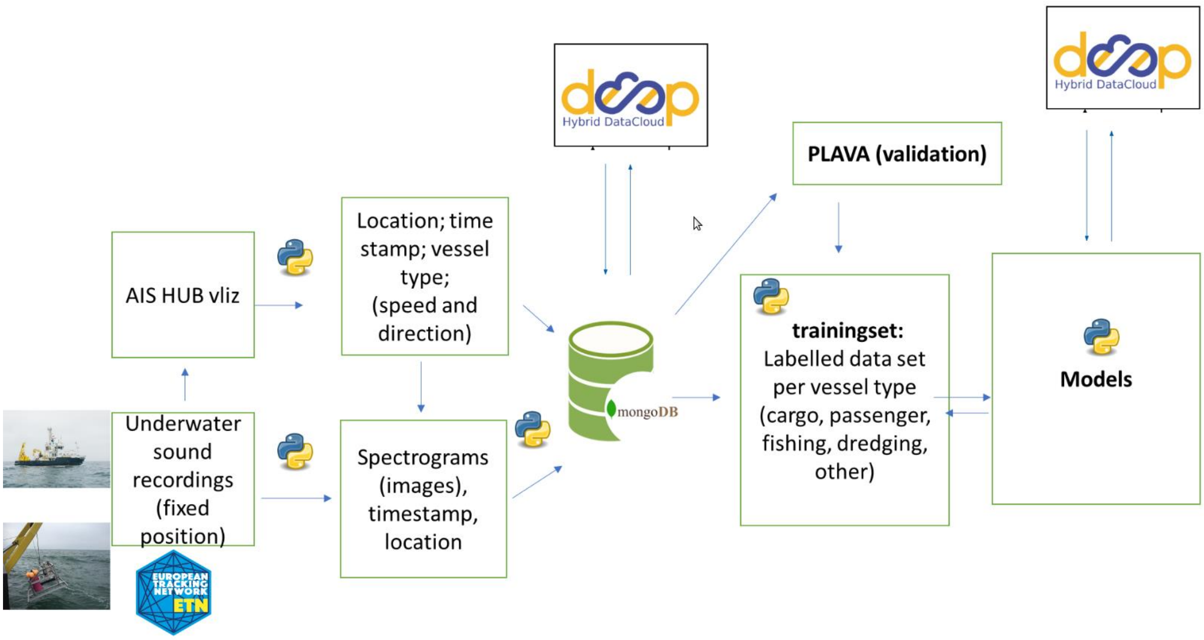

Figure UC6.1 - Planned high-level architecture of the UC6 image service

Epics and User stories

Epic UC6.E1 “Identify sound sources from audio recordings”	
Personas	1. A data scientist training AI for sound recognition 2. A domain scientist interested in identifying underwater sounds
UC6.E1.US1	As a data scientist, I want to label underwater sound data, so that I can develop an efficient AI model for sound identification
UC6.E1.US2	As a domain scientist, I want to have a tool which will allows me to process acoustic underwater recordings for identification and recognition of marine species and other types to improve our knowledge on the ocean health

Table UC6.1 - Identified Epics and User stories for UC6

Gap and Bottlenecks analysis

There is currently 1.5 years of underwater sound data and the data collection is ongoing. However, the processing of these data and identification of sound sources is very time consuming and individual effort to derive the sources is needed. The process also lacks automation. To fill the gap, we will use the iMagine AI platform and know-how available in the project to improve the labelling of data and try different AI approaches for sound recognition and identification.

Development roadmap

For the development of the solution, we start with the data ingestion of raw sound data into our database (MongoDB) (see Fig. UC6.1). We improve the labelling and validation interface for more efficient data labelling in order to prepare the training dataset. Once there is enough training data, we develop, train, and validate various AI models. With the well performing AI model in-place, we can work on the automation of the sound identification process.

		2022		2023						2024						2025			
Task \	Project Month	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36
	Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
Develop data ingestion pipeline from sensor to DB																			
Improve current data validation interface for more efficient validation and labelling																			
Develop & validate models																			
Prepares the used data and processing components for connection, synchronisation and migration																			
Prototype validation with external users																			

Figure UC6.2 - Development timeline for UC6

UC7 Beach monitoring

Use case overview

From 2011, SOCIB has set up a systematic and continuous monitoring of beaches, using cameras, generating long-term time-series of data, available to the scientific community, coastal management authorities, and citizens. The data is already used for shoreline tracking. In the iMagine project, we are going to develop a prototype service for processing video images from beach cameras for monitoring formation and dismantling events of seagrass beach berms (Posidonia oceanica) and detecting rip-currents.

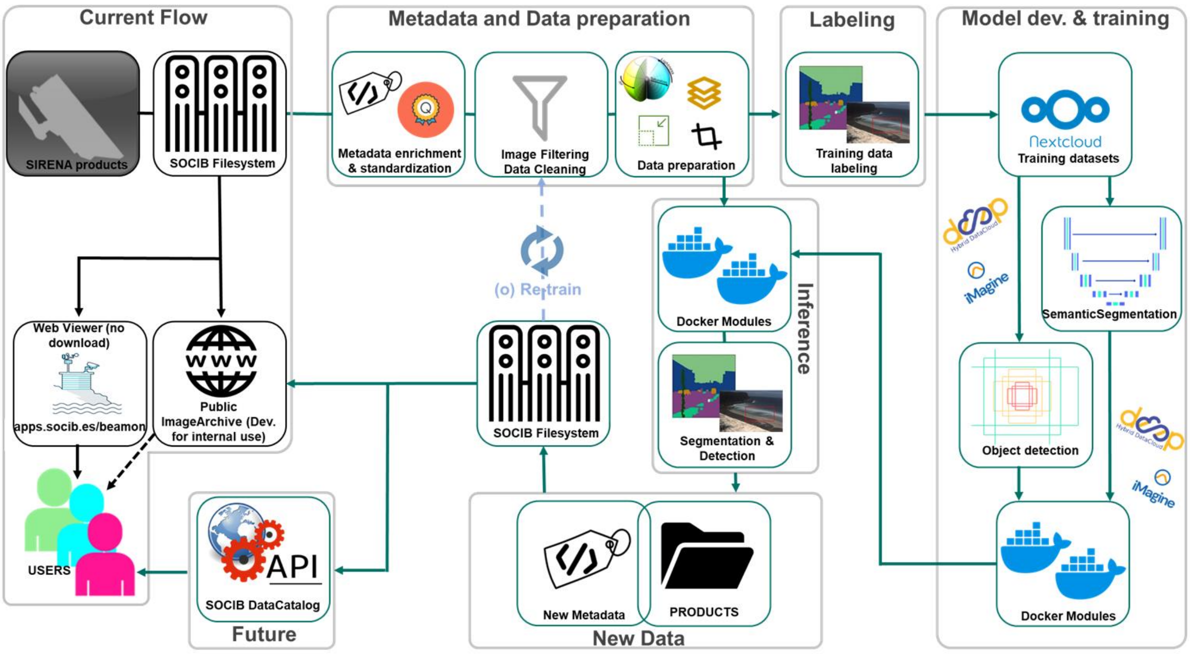

Figure UC7.1 - Planned high-level architecture of the UC7 image service

Epics and User stories

Epic UC7.E1 "Create a training dataset"	
Personas	1. A marine science researcher with a background in remote sensing for the monitoring and management of coastal areas 2. A software engineer that provides support to researchers
UC7.E1.US1	As a marine science researcher, I want to create a training dataset for image segmentation and a training dataset for object detection based on the image metadata and labels so that I can test various AI models and approaches.
UC7.E1.US2	As a software engineer, I want that all images have standard and adequate metadata so that a marine science researcher can select the appropriate images for the preparation of an AI training dataset
Epic UC7.E2 "Implement image segmentation and object detection based on deep learning"	
Personas	1. A marine science researcher with a background in remote sensing for the monitoring and management of coastal areas
UC7.E2.US1	As a marine science researcher, I want to apply AI-based approach (e.g. image segmentation and object detection) on the beach monitoring images so that I get information on presence and distribution of key coastal features with importance in beach monitoring and management or for emergency services and forecasting models (e.g rip-currents).

Table UC7.1 - Identified Epics and User stories for UC7

Gap and Bottlenecks analysis

The main coastal feature which is currently being extracted from the video-monitoring system (SIRENA) is the shoreline position. Shoreline is extracted manually, at one timestamp every ~15 days, but SIRENA images are taken several times per day. No other feature is extracted, while images offer the possibility to get different information of beach features related to biogeophysical and socioeconomic processes such as *Posidonia* berms and rip currents identification, determination of beach width, swash zone and run-up. Using DL architectures intended for image segmentation will allow to get information on presence and distribution of important features such as sand, water, white scum, *Posidonia* berms, 'humans', and vessels, which would allow to automate the process of shoreline extraction (at almost all available timestamps), *Posidonia* berms characterization, and others with importance in beach monitoring and management. DL applied to object identification could be useful for identification of rip currents (importance for emergency services, forecasting models and early-warning systems).

Development roadmap

We start with the metadata enhancement and selecting deep learning and labelling methods (Fig. UC7.2) in order to proceed with the data preparation and labelling. We then transfer the training dataset to the iMagine AI platform so that we can start with the model development, training, and validation. Once the model is ready, we implement it into the prototype service.

	2022		2023						2024						2025			
Task \ Project Month	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36
Actual month	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*	*10*	*12*	*2*	*4*	*6*	*8*
koM, Tutorials, Webinar, 'Learning'																		
SirenaPi - python, Workshop																		
Metadata enhancement																		
Selecting DL & Labelling methods																		
Data preparation and Labelling																		
First training dataset to Nextcloud																		
Exp. training dataset to Nextcloud																		
Model development and training																		
Testing: Posidonia berms area																		
Testing: RipC. NRT warning system																		
SQA implementation																		
Prototype validation with external users																		

Figure UC7.2 - Development timeline[13] for UC7

UC8 Freshwater diatoms identification

Use case overview

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

[13] *Blue: Principal timeline (must); Orange: Additional time for extending training dataset (if possible)*

bioindication service able to identify diatom species but also to quantify key morphological features (Fig. UC8.1), leveraging the iMagine AI platform.

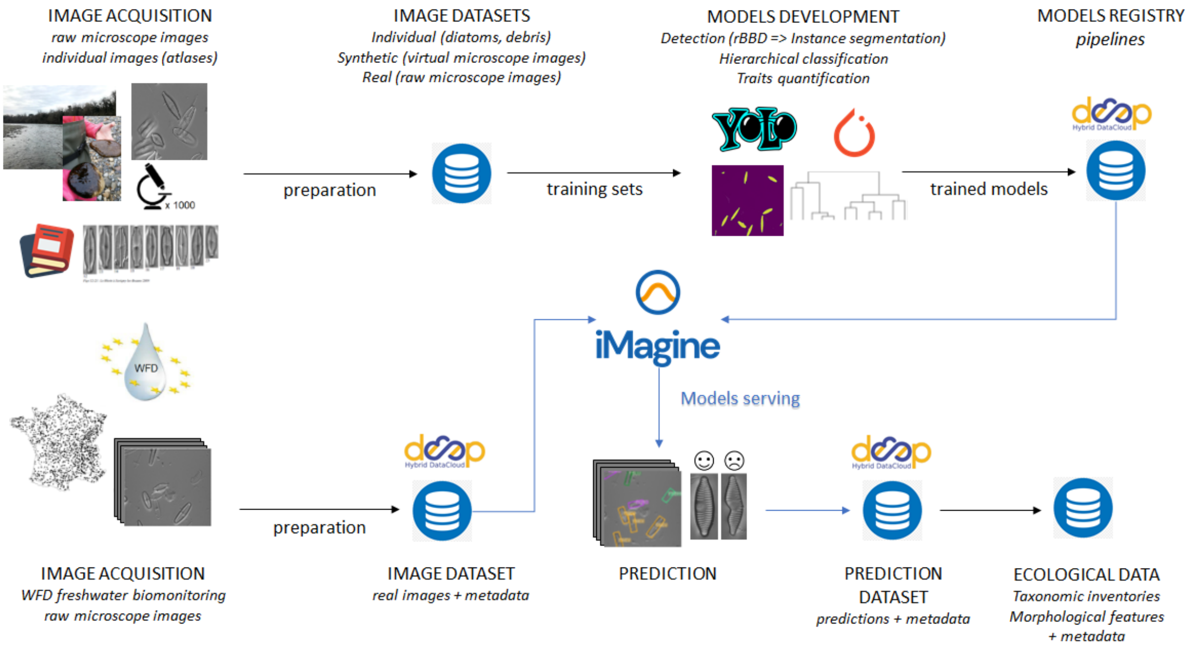

Figure UC8.1 - Planned high-level architecture of the UC8 image service

Epics and User stories

For the development of the prototype service we plan to focus on the user stories listed in the Table UC8.1 .

Epic UC8.E1 "Improve diatom identification"	
Personas	1. A domain scientist using diatoms for monitoring the ecological quality of rivers 2. Taxonomist(s) validating diatom inventories
UC8.E1.US1	As a domain scientist, I want to perform diatom taxonomic and morphometric analysis in a standardised way using an automated approach so that it is less time-consuming and less prone to multiple biases (e.g. operator experience, image quality), for both research and teaching process.
UC8.E1.US2	As a taxonomist involved in the diatom-based biomonitoring, I want to have a pre-screening approach based on AI so that it can help for getting in a much faster way diatom taxonomy and morphometry, in order to focus only on the most difficult cases.
Epic UC8.E2 "Create high-quality training dataset(s) of diatoms"	
Personas	1. A data scientist training AI

UC8.E2.US1	As a data scientist, I want to have high-quality training datasets of diatoms so that I can develop AI models with good prediction

Table UC8.1 - Identified Epics and User stories for UC8

Gap and Bottlenecks analysis

Diatom-based freshwater quality indices are calculated from the inventory of indicator diatom species present in a natural sample. These species are identified under a microscope on the basis of morphological characteristics, which is currently a time-consuming step often subject to multiple biases (operator experience, image quality). This can be improved by standardising the process using an AI-based automated approach.

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

Development roadmap

The development roadmap (Fig. UC8.2) consists of setting up the annotation workflow for labelling real microscope images which will be acquired during the first part of the project. Annotation tasks will allow to expand training sets for diatom classification (currently 150 to as much as several hundreds of species) but also to create training sets for segmentation which will be needed for traits quantification (size, deformations). In parallel, model developments will consist in fine tuning the already available end-to-end pipeline for diatom classification (probabilistic approach) but also exploring different AI approaches for diatom morphological traits quantification. Once the models are validated, we transfer them on the iMagine platform and implement the prototype service.

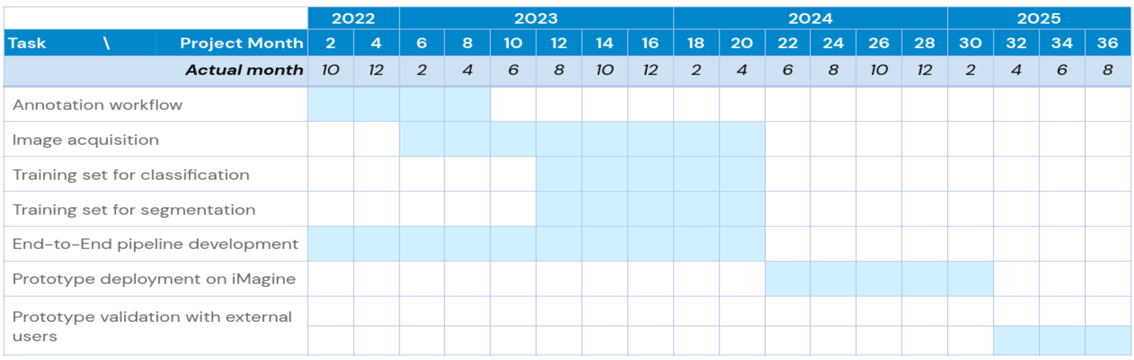

Figure UC8.2 - Development timeline for UC8

(initial) Requirements for the platform

The attempt to define full and detailed requirements early in the project is often proven to be counterproductive and restrictive[14]. As the project progresses, the use cases and AI Platform providers understand each other's needs better and can refine the details. Therefore in the performed analysis, the high-level use case requirements for the AI Platform were collected and the requirements tracking was set (see "Requirements tracking" section). Since the AI Platform consists of two installations: T4.1 - AI Application Development and T4.2 - AI Applications as a Service, we collected requirements separately for both installations. They are represented in the following subsections and Tables R1-R6. Currently they mainly concern the Storage, Computing, and Network requirements. The document[15] allows us to monitor them and collect more user requests in the course of the project (see "Requirements tracking" section below). The review of the requirements yields the average and median values for a typical use case, which facilitates forecasting resources for new external use cases. As many use cases have to work on the AI development first, requirements for serving the AI-based services are often not fully defined at the current stage (TBD status) but will be complete in accordance with the project timeline (Fig. G1).

Requirements for the AI Development

Storage Requirements

Use Case	Storage space required for development[16]	Access bandwidth required[17]	Privacy concerns
UC1	1054 MB (PLD = 709 MB ; PLQ = 345 MB)	>25 Mbps	Regulated in WP6, "Ethics management"
UC2	< 1 TB	Upload once and then use	None

[14] See footnote 1.
[15] WP3 - AI Platform Requirements Tracking: https://docs.google.com/spreadsheets/d/1P1NBEzdQGlmOxcb0gN7fd6D6eEaNeUSL-BK9OAuPtwM/
[16] This includes all data which have to be stored on the platform for the successful model training e.g. raw data, pre-processed data (if relevant), and training data.
[17] If an operational use case did not put any bandwidth constraints, we put low limit of 25 Mbps as a pretty much guaranteed median internet speed, see https://www.speedtest.net/global-index.

Use Case	Storage space required for development[16]	Access bandwidth required[17]	Privacy concerns
UC3o	< 100 GB	>25 Mbps	None
UC3a	1 TB	>25 Mbps	None
UC3s	< 1 TB	<300 Mbps	None
UC4	O(10) GiB	1000 Mbps	None
UC5	100 GB	>25 Mbps	None
UC6	TBD	TBD	None
UC7	<200 GB	>25 Mbps	None
UC8	TBD	TBD	None
Total	**3483 GB**	**1000 Mbps[18]**	—
Average	**435 GB**	**181 Mbps**	—
Median	**150 GB**	**25 Mbps**	—

Table R1: Storage requirements of use cases for the AI model development

Computing requirements

The Table R2 presents numbers estimated for the *average* load (e.g. CPU and GPU usage by time). It is understood that *peak* values, especially in the case of the AI Development installation, may significantly differ from the listed averages. Nevertheless, those numbers provide good quantitative understanding of the user's needs: for example, if all use cases request GPUs, we would need to allocate at least 13 GPUs for 24/7 during active development periods and aim to double the number.

Use Case	Estimated CPU usage by time (h/week)	RAM required	Estimated GPU usage by time (h/week)	Estimated number of GPUs per training job	GPU memory per card
UC1	18 h/week (Detector 12h, Quantifier: 6h)	8 GB	-	-	-
UC2	<20 h/week	>16 GB	<10 h/week	1	>24 GB

[18] For the bandwidth, the requested maximum is taken

Use Case	Estimated CPU usage by time (h/week)	RAM required	Estimated GPU usage by time (h/week)	Estimated number of GPUs per training job	GPU memory per card
UC3o	<20 h/week	TBD	<10 h/week	1	TBD
UC3a	<15 h/week	32GB	<10 h/week	1	8 GB
UC3s	<168 h/week	16 GB	<24 h/week	1	8 GB
UC4	TBD	TBD	TBD	4	16
UC5	<40 h/week	>8 GB	<20 h/week	2	>4 GB
UC6	<20 h/week	>8 GB	<10 h/week	1	>8 GB
UC7	<20 h/week	>8 GB	<10 h/week	1	>8 GB
UC8	<20 h/week	16 GB	<10 h/week	1	>8 GB
Total	**341 h/week**	**112 GB**	**104 h/week**	**13**	**76 GB**
Average	**38 h/week**	**14 GB**	**13 h/week**	**1.5**	**11 GB**
Median	**20 h/week**	**12 GB**	**10 h/week**	**1**	**8 GB**

Table R2: Computing requirements of use cases for the AI model development

Further requirements

Use Case	Requirement title	Description
UC1-8	Information about available resources	General information about available resources, e.g. a r the GPU memory size; amount of available storage etc

Table R3: Further requirements of use cases for the AI model development

Requirements for the AI Application Serving

Storage requirements

Use Case	Permanent storage space required	Access bandwidth required[19]	Privacy concerns
UC1	Depends how much of	>25 Mbps	Regulated in WP6,

[19] If an operational use case did not put any bandwidth constraints, we put low limit of 25 Mbps as a pretty much guaranteed median internet speed, see https://www.speedtest.net/global-index

Use Case	Permanent storage space required	Access bandwidth required[19]	Privacy concerns
	the generated data is stored permanently, 100 GB must be enough		"Ethics management"
UC2	200 GB	2-5GB / week	None
UC3o	<100 GB	>25 Mbps	None
UC3a	< 1 TB	>25 Mbps	None
UC3s	< 1 TB	300 Mbps	None
UC4	O(100)GiB	1024 Mbps	None
UC5	50 GB	>25 Mbps	None
UC6	<50GB	>25 Mbps	None
UC7	TBD	TBD	None
UC8	TBD	TBD	None
Total	**2598 GB**	**1024 Mbps[20]**	
Average	**371 GB**	**207 Mbps**	
Median	**100 GB**	**25 Mbps**	

Table R4: Storage requirements of use cases for the AI model serving

Computing requirements

Use Case	Estimated CPU usage by time (h/week)	RAM required	Estimated GPU usage by time (h/week)	If service scalability is required
UC1	1214 h/week	8 GB	-	TBD
UC2	TBD	TBD	TBD	TBD
UC3o	TBD	TBD	TBD	TBD
UC3a	1h/week	8GB	1h/week	No
UC3s	<168 h/week	16 GB	<24 h/week	TBD

[20] For the bandwidth, the requested maximum is taken

Use Case	Estimated CPU usage by time (h/week)	RAM required	Estimated GPU usage by time (h/week)	If service scalability is required
UC4	15 h/week	32 GB	TBD	Yes
UC5	TBD	TBD	1 h/week	TBD
UC6	TBD	TBD	TBD	TBD
UC7	TBD	TBD		
UC8	TBD	TBD	TBD	TBD
Total	**1398 h/week**	**64 GB**	**26 h/week**	
Average	**350 h/week**	**16 GB**	**9 h/week**	
Median	**92 h/week**	**12 GB**	**1 h/week**	

Table R5: Computing requirements of use cases for the AI model serving

Further requirements

Use Case	Requirement title	Description
UC3s	Data store for Labelled Data	Need to consider where to store time series of event detections

Table R6: Further requirements of use cases for the AI model serving

Requirements tracking

In order to keep track of the technical requirements for the AI platform, a set of processes based on the collaborative tools used within the project has been defined:

1. User requirements are tracked in the dedicated document[21]. This includes Requirement ID, Title, Corresponding AI Platform installation, Value, Priority, corresponding Epic and other fields (Fig. R1). The Status of each requirement (TBD, Defined, In Progress, In Review, Done) can be updated and new requirements can be added. The document is also accessible by other work packages, including WP4 - AI and Infrastructure Services.
2. The set teleconference channel[22] allows regular direct discussions with users and monitoring of the feedback, as well as know-how exchange.

[21] See footnote 15

[22] WP3 regular communication channel: https://confluence.egi.eu/display/IMP/WP3+Meetings

3. The available email list imagine-wp3@mailman.egi.eu lets offline discussions and collection of user requests.
4. There are three planned Competence Centre Workshops at M5 (already happened[23]), M16, and M22 in the project. They provide users with the training on the platform and the opportunity to discuss in-person AI and IT related questions with corresponding experts.

A	B	C	D	E	F	G	H	I	J	K	L
		UC5 - Flowcam plankton identification									
CID	ID	Title	AI Platform Installation	Value	Priority	EPIC	Category	Requester	Status	Rev. Version	Revision date
CD.Req001	UC5.Req001	Storage space (dev)	Development	< 100 GB	Must have	UC5.E1	Storage	UC5	Defined	1.0	15.2.2023
CD.Req002	UC5.Req002	Access bandwidth (dev)	Development	> 25 Mbps	Must have	UC5.E1	Network	UC5	Defined	1.0	15.2.2023
CD.Req003	UC5.Req003	CPU usage (dev)	Development	< 40 h/week	Must have	UC5.E1	Computing	UC5	Defined	1.0	15.2.2023
CD.Req004	UC5.Req004	RAM required (dev)	Development	< 8 GB	Must have	UC5.E1	Computing	UC5	Defined	1.0	15.2.2023
CD.Req005	UC5.Req005	GPU usage (dev)	Development	< 20 h/week	Must have	UC5.E1	Computing	UC5	Defined	1.0	15.2.2023
CD.Req006	UC5.Req006	Nr. GPUs per training task (dev)	Development	2	Must have	UC5.E1	Computing	UC5	Defined	1.0	15.2.2023
CD.Req007	UC5.Req007	GPU memory per card (dev)	Development	> 4 GB	Must have	UC5.E1	Computing	UC5	Defined	1.0	15.2.2023
CS.Req001	UC5.Req008	Permanent storage (srv)	Serving	50 GB	Moderate	UC5.E1	Storage	UC5	Defined	1.1	17.2.2023
CS.Req002	UC5.Req009	Access bandwidth (srv)	Serving	> 25 Mbps	Must have	UC5.E1	Network	UC5	Defined	1.0	15.2.2023
CS.Req003	UC5.Req010	CPU usage (srv)	Serving	TBD	Must have	UC5.E1	Computing	UC5	TBD	1.0	15.2.2023

Figure R1 - AI Platform Requirement Tracking, example for one of the use cases

Conclusion

The document presents an analysis of eight use cases of iMagine, which represent different domains in aquatic science. The diversity and complementarity of the use cases, their developments and experiences within the project will formulate Best practices for future adopters in the field.

First, the deliverable describes the methodology to understand the needs and analyse the use cases. The user stories and epics, existing gaps and bottlenecks are spotted. The use cases established the development roadmaps with corresponding timelines. Those timelines are in accordance with the general timeline of the project, where mature use cases are expected to release the AI-based components of services at full scale on M24. Additionally, we collected technical requirements for the AI platform, individually for the AI Development and AI Serving installations. The means to monitor and expand those requirements in the course of the project are provided. Analysis of the requirements allows us to understand an average and median case for an image service in aquatic science for future adopters.

[23] iMagine Competence Centre Workshop in Villefranche, 30-31.01.2023, https://indico.egi.eu/event/5999/

Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
BODC	British Oceanographic Data Centre
CI/CD	Continuous Integration / Continuous Delivery
CPU	Central Processing Unit
CNN	Convolutional Neural Network
DevOps	Development and Operations
DL	Deep Learning
DSDM	(stands for) Dynamic System Development Method
DwCA (DwC-A)	Darwin Core Archive
EOSC	European Open Science Cloud
FAIR	principles of Findability, Accessibility, Interoperability, and Reusability
GPU	Graphics Processing Unit
KER	Key Exploitable Result
ML	Machine Learning
(EU) MSFD	EU Marine Strategy Framework Directive
NGO	Non-governmental organisation
UC	Use Case
US	User Story
UAV	Unmanned Aerial Vehicle