EU MISSIONS
RESTORE OUR OCEAN AND WATERS

Scalable full-cycle marine litter remediation in the Mediterranean: Robotic and participatory solutions

SeaClear 2.0

https://www.seaclear2.eu

D2.4
Marine Litter Dataset
WP2 — Concept Design & Technical Specification

Grant Agreement no. 101093822

Lead beneficiary: Fraunhofer

Date: 28/06/2024
Type: OTHER
Dissemination level: PU

Document Information

Grant agreement no.	101093822
Acronym:	SeaClear2.0
Full title:	Scalable full-cycle marine litter remediation in the Mediterranean: Robotic and participatory solutions
Start date of the project	01/01/2023
Duration of the project	48 months
Deliverable	D2.4: Marine Litter Dataset
Work package	WP2: Concept Design & Technical Specification
Deliverable leader	Fraunhofer
Delivery date	Contractual: 30/06/2024 Actual: 28/06/2024
Status	Draft ☐ Final ✓
Type[1]	R ☐ DEM ☐ OTHER ✓ DMP ☐
Dissemination level[2]	PU ✓ C-UE/EU-C ☐ SEN ☐
Author(s)	Cosmin Delea (Fraunhofer) Ibtehaj Khan (Fraunhofer) Kaya ter Burg (TU Delft)
Responsible author	Cosmin Delea, email: cosmin.delea@cml.fraunhofer.de Fraunhofer
Deliverable description	Creating a database of marine litter by gathering information from marine scientific reports and field studies, categorising waste fractions by properties such as size, material, weight, solidity, location in the designated demonstration areas.

[1] R = Document, report, DEM = demonstrator, OTHER: Software, technical diagram, etc. DMP = Data Management Plan
[2] PU = Public, C-UE/EU-C = EU Confidential under Decision 2015/444, SEN = Sensitive

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Document History

Name	Date	Version	Description
C. Delea, I. Khan, K. ter Burg	29/05/2024	V0.1	Initial Draft
C. Au, J. Oeffner, C. Delea	04/06/2024	V0.2	revised fundamentals, cloud service selection and enhanced figures
C. Au, J. Oeffner	28/06/2024	V1.0	adjusted footnotes, references, and figures, included reviewer comments

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Disclaimer of Warranties

This document has been prepared by SeaClear2.0 project partners as an account of work carried out within the framework of Grant Agreement no. 101093822. Neither the Project Coordinator, nor any signatory party of the SeaClear2.0 Project Consortium Agreement, nor any person acting on behalf of any of them:

- makes any warranty or representation whatsoever, express or implied, with respect to the use of any information, apparatus, method, process, or similar item disclosed in this document, including merchantability and fitness for a particular purpose, that such use does not infringe on or interfere with privately owned rights, including any party's intellectual property; or
- makes any warranty or representation whatsoever, express or implied, that this document is suitable to any particular user's circumstance; or
- assumes responsibility for any damages or other liability whatsoever (including any consequential damages, even if the Project Coordinator or any representative of a signatory party of the Project Consortium Agreement, has been advised of the possibility of such damages) resulting from your selection or use of this document or any information, apparatus, method, process, or similar item disclosed in this document.

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

Table of Contents

Definitions

- **Beneficiary:** A legal entity that is signatory of the EC Grant Agreement no. 101093822.
- **Consortium:** The SeaClear2.0 Consortium, comprising the below-mentioned list of beneficiaries.
- **Consortium Agreement:** Agreement concluded amongst SeaClear2.0 Beneficiaries for the implementation of the Grant Agreement.
- **Grant Agreement:** The agreement signed between the beneficiaries and the EC for the undertaking of the SeaClear 2.0 project (Grant Agreement no. 101093822).

Beneficiaries of the SeaClear2.0 Consortium are referred to herein according to the following abbreviations:

- **TU Delft:** TECHNISCHE UNIVERSITEIT DELFT
- **DUNEA:** REGIONALNA AGENCIJA DUNEA
- **Fraunhofer:** FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV
- **HPA:** HAMBURG PORT AUTHORITY
- **ISOTECH**: ISOTECH LTD
- **MDanchor**: M. DANCHOR LTD
- **Subsea Tech:** SUBSEA TECH SAS
- **TECNOSUB**: TÉCNICAS Y OBRAS SUBACUÁTICAS, SLU
- **TUM**: TECHNISCHE UNIVERSITAET MUENCHEN
- **UNIDU:** SVEUCILISTE U DUBROVNIKU
- **UTC**: UNIVERSITATEA TEHNICA CLUJ-NAPOCA
- **VEO**: VEOLIA PROPRETE
- **VLPF**: VENICE LAGOON PLASTIC FREE

Abbreviations

- **EC:** European Commission
- **GA:** Grant Agreement
- **WP**: Work Package

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Acronyms

ADLS Azure Data Lake Storage. 21, 22, 25, 26

AML Azure Machine Learning. 21, 22, 25, 26

API Application Programming Interface. 20

CSV Comma Separated Values. 18

EC European Comission. 28

EMODnet European Marine Observation and Data Network. 26, 28

GES Good Environmental Status. 10

IDE Integrated development environment. 18

ML Machine Learning. 8–10, 12, 17, 18, 20–28

MSFD Marine Strategy Framework Directive. 10

RBAC Role-based Access Control. 20–22

ROV Remotely Operated Vehicle. 26

UNEP United Nations Environment Programme. 12

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Executive Summary

This report exploits the SeaClear2.0 project technical requirements in reference to the creation of a database for storing, curating and organising data needed for training the Machine Learning (ML) models used within the project's robotic solution for detection and classification of marine litter. In addition to the technical requirements, the marine litter database enforces the key findings from marine scientific reports and field studies, categorising waste by size, material, weight, solidity, and location (water surface, column, floor, beach) within the specified demo and pilot test areas.

Based on the aforementioned requirements, the software solution, consisting in a Microsoft Azure cloud service, hosting multiple tools and services, is described and the management of data by different types of users is presented. The process of quality assurance involves enforcing a directory structure and naming convention. A step-by-step walkthrough in the cloud service solution is presented in an annex.

Cloud-specific data collection tools are used for enforcing the data compliance. Specific survey parameters for collecting data from the demo and pilots sites from the seabed, water column, and water surface are, further on, presented in this report. These parameters are in-line with the already established technical specifications of the SeaClear2.0 robotic system, that are derived from the boundary conditions established in previous deliverables.

Finally, the alignment of the newly created database with the SeaClear2.0 project objectives are summarised, alongside further recommendations on handling and up-scaling the database. These include input from existing data sets from related projects and the expert knowledge of partners and other stakeholders.

An overview of the user interface for the pertinent Microsoft Azure cloud services is included at the conclusion of the deliverable.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

1 Introduction

This section gives the introductory concepts for understanding the scope and requirements on the dataset. Throughout the document the words *database*, *data lake* and *dataset* will be intensively used, as these are the core elements of the current work. In this context, the *dataset* represents the collection of data points, which are used in the scope of training or performance validation of the ML algorithms used for detecting and classifying marine litter. In the context of marine litter detection and classification, the *dataset* would include images of litter items along with their labels (e.g., types of litter, locations). The usage of the word "database" emerged since late 1970s [3], and it may refer to several different but interrelated concepts such as the information, the software, or their combination [1]. The word "data lake" emerged into popularity since mid-2010s [4]. Here we redefine these words For the sake of clarity.

Definition 1.1. A *database* is a conceptual entity that allows storage, retrieval, and management of a collection of digital information. In the context of this report, the *database* stores the raw data collected from various sources (e.g., images) and potentially metadata (e.g., timestamps, locations). This data could then be used to create datasets for training and evaluating ML models.

Definition 1.2. A *data lake* is a scalable storage and analysis system for data of any type, retained in their native format and used mainly by data specialists (statisticians, data scientists or analysts) for knowledge extraction [2]. Its characteristics include:

1. a metadata catalogue that enforces data quality;
2. data governance policies and tools;
3. accessibility to various kinds of users;
4. integration of any type of data;
5. a logical and physical organisation; and
6. scalability in terms of storage and processing.

A *data lake* is used to store and manage all the data collected during the marine litter detection process. This includes raw sensor data, images, videos, metadata, logs, and other relevant information. Apart from storage, the data lake offers a centralised data management interface and the ML models can access the data stored in the data lake for training and evaluation purposes.

In the context of marine litter detection and classification, it can be summarised that a *data lake* is utilised as the storage system for saving and analysing all marine litter *datasets*. This includes raw sensor data, images, and videos, which are ingested into the *data lake* in their native formats. The proposed data lake cloud-based architecture ensures simple and efficient storage and retrieval of data, supporting the development and evaluation of ML models, whilst also ensuring high horizontal scalability via the public service provider.

[3] https://books.google.com/ngrams/graph?content=database&year_start=1950&year_end=2019&corpus=en-2019&smoothing=1&case_insensitive=true

[4] https://books.google.com/ngrams/graph?content=data+lake&year_start=2000&year_end=2019&corpus=en-2019&smoothing=1&case_insensitive=true

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

1.1 SeaClear2.0 Project Objectives

The SeaClear2.0 project will develop an integrated approach to address the entire cycle of marine litter ML training. The project focuses on reducing marine pollution, specifically from marine litter, in the Mediterranean. This will be achieved by using teams of autonomous, intelligent robots to monitor and collect marine seafloor and surface litter, and through participatory practices to identify site-specific measures for marine litter prevention and reduction. By identifying ways to valorise litter and extend policy-making, SeaClear2.0 will provide innovative solutions for effective marine litter management, further promoting the health of oceans, seas, and water bodies.

The main objective of this report is to document the selection of the medium used for storing the datasets required for the training of neural network models for identifying and classifying marine litter. Furthermore, the process of creation and then the usage of this data store is explained in detail. The choices are based on the research done by the SeaClear2.0 partners in documenting the general situation of marine debris in the Mediterranean, the most recent data and acknowledgements , as well as the relevant environmental challenges addressed by the project.

SeaClear2.0 has as objective contribution to all expected outcomes of the call "*Mediterranean sea basin lighthouse – Prevent and eliminate pollution of our ocean, seas and waters*". In order to follow up this objective and to qualify the SeaClear2.0 system benefits, but also to identify the challenges, a correlation of the project use cases with Marine Strategy Framework Directive (MSFD) qualitative descriptors has been synthesised in [3]. This analysis also showcases how Good Environmental Status (GES) can be achieved along with the direct contribution to MSFD.

Sea debris poses environmental, economic, health, aesthetic, and cultural threats, with huge degradation consequences for marine and coastal habitats and ecosystems that incur socioeconomic losses in marine-based sectors. It can be observed everywhere in the oceans, with the Mediterranean Sea being drastically impacted because of its specific geographical and oceanographic setting. Debris enters oceans from both land and water-based sources and can travel long distances before being deposited on shores and the seafloor. SeaClear2.0 system will be challenged with various marine litter types and sizes and site conditions. Continuous removal of marine debris is a key factor in fighting this enormous environmental problem, where innovative solutions such as the SeaClear2.0 robotic system, depicted in Figure 1, plays the crucial role of solving the problem and mitigating the consequences of further accumulation of garbage in our seas, specifically on depths that are difficult to be reached by human divers.

1.2 Marine Litter Terminology

The scope of the SeaClear2.0 project is set to underwater litter, which encompassed a plethora of materials and consumer products. Globally, there are inconsistencies in the analysis methods and categorisation of marine litter fractions. Within [3], the authors incorporated the official EU terminology in describing the types of marine litter of each demo or pilot site within SeaClear2.0. For these, the analysis yielded that approximately 50% of the litter are either D- or E-sized, meaning dimensions in between $20\,\mathrm{cm}$ to $100\,\mathrm{cm}$ (per each length) in size, as can be depicted in Figure 2.

The aforementioned sizes of litter need to be scaled for the actual test environments. Given the oper-

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

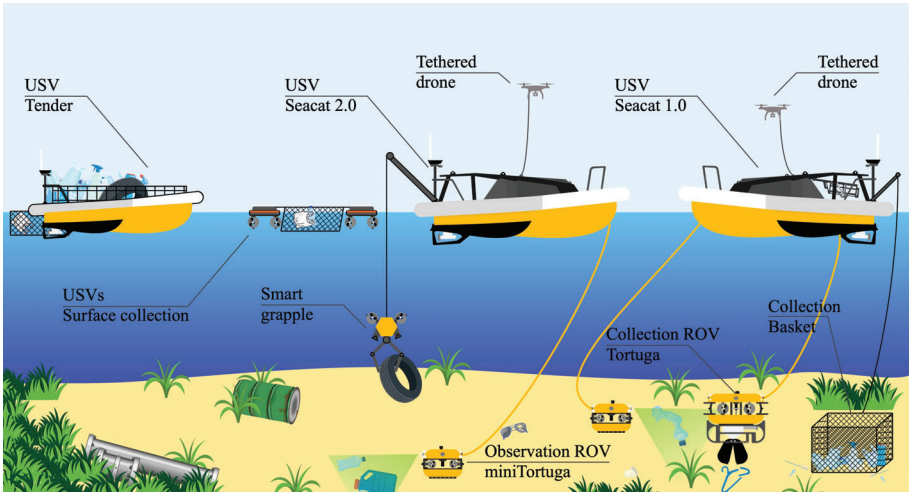

Figure 1: The SeaClear2.0 system concept

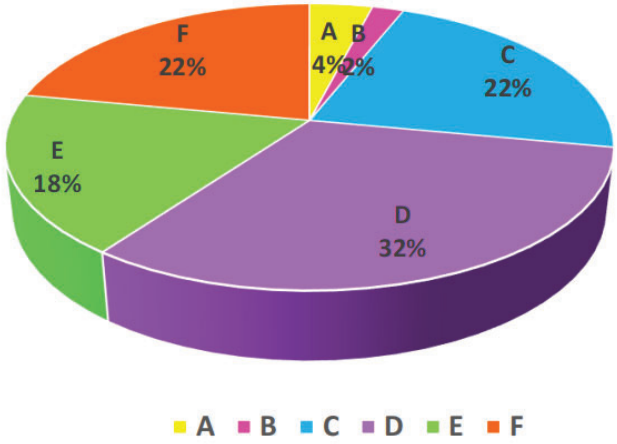

Figure 2: Percentages for each size grade for the list of 100 most frequent marine litter fractions within SeaClear2.0 demo and pilot sites: A. $<$ 5cmx5cm = 25cm^2; B. $<$ 10cmx10cm = 100cm^2; C. $<$ 20cmx20cm = 400cm^2; D. $<$ 50cmx50cm = 2500cm^2; E. $<$ 100cm-100cm = 10000cm^2 = 1m^2; F. $>$ 100cm-100cm = 10000cm^2 = 1m^2.(source [3])

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

ational range of $1000\,\mathrm{m}$ (in open sea) and depths of up to $300\,\mathrm{m}$[5] established in [4] for the SeaClear2.0 and the densities ranging from $50\,\mathrm{items/km^2}$ to $1161\,\mathrm{items/km^2}$, this implies that, on average, hundreds of items are expected to be located within the operational area of the SeaClear2.0 system. The marine litter densities for each demo or pilot site are depicted in Figure 3.

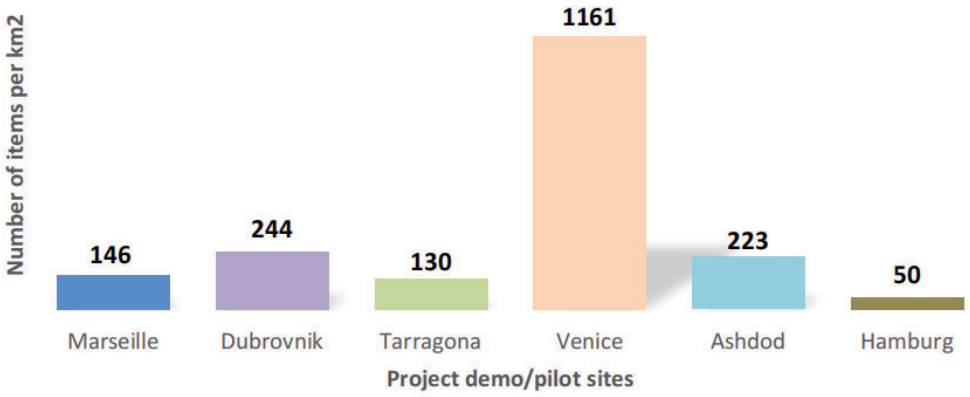

Figure 3: Marine litter densities for depths up to $100\,\mathrm{m}$ within the demo and pilot areas (source [3])

The study conducted in [3] follows on the standardisation efforts conducted in [5], which represent the most recent EU standards for monitoring litter, applicable on a global scale. The classification system contained in [5] is intended for use across various marine settings, including coastlines, the surface layer of the water column, the seabed, and within *biota*. Given the critical nature of marine litter, which has been addressed by international bodies like United Nations Environment Programme (UNEP) and the G7/G20 summits, this methodology and its classifications are designed to foster global consistency and standardisation. The complete list of categorised marine litter resulting from the aforementioned study is publicly available within the EU's Online Photo Catalogue of the Joint List of Litter Categories[6].

In pursuit for establishing standard categories for the marine litter, the same categories will be expected for the database. Taking into account the marine litter occurrence analysis from [3], done in the demo and pilot sites, a non-exhaustive list of marine litter elements, distributed either at seafloor or on surface, has been created and summarised in Table 1. In order to train the ML algorithms onto detecting and classifying these, a sufficient amount of data from each category has to be available in the database.

[5] https://www.subsea-tech.com/
[6] https://mcc.jrc.ec.europa.eu/main/photocatalogue.py

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Table 1: Distribution of litter types per demo and pilot sites classified by the material type and name used for labelling, categorised by the size (see Figure 2) and occurrence (source [3] and [4])

MATERIAL	#	Item General Name	Category	Location	
				Surface	Seafloor
PLASTIC	1	Shopping plastic bags	D	x	x
	2	Small plastic bags	C	x	x
	3	Plastic bottles $\leq$ 0.5 L	C	x	x
	4	Plastic bottles $>$ 0.5 L	D	x	x
	5	Plastic buckets	D	x	x
	6	Cleaner bottles	D	x	x
	7	Food containers	D	x	x
	8	Cosmetic bottles $\leq$20 cm	C	x	x
	9	Cosmetic bottles $<$50 cm	D	x	x
	10	Engine oil bottles $<$50 cm	D	x	x
	11	Engine oil bottles $>$50 cm	E	x	x
	12	Various crates and containers	D	x	x
	13	Plastic cups and lids	A	x	x
	14	Various plastic items and fractions $\leq$20cm	C	x	x
	15	Cutlery and trays	C	x	x
	16	Straws	C	x	x
	17	Cover/packaging	D	x	x
	18	Mussel/oyster nets - pergolars	F		x
	19	Synthetic ropes	E		x
	20	Fishing net $<$ 1m2	E		x
	21	Fishing net $>$ 1m2	F		x
	22	Fishing line	F	x	x
	23	Tangled nets	F		x
	24	Fish boxes	D	x	x
	25	Floats for fishing nets	C	x	x

Continued on next page

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Table 1 – continued from previous page

MATERIAL	#	Item General Name	Category	Location	
				Surface	Seafloor
	26	Buoys ≥50cm	D	x	x
	27	Sheets, industrial packaging, plastic sheeting	D	x	x
	28	Car parts	E	x	x
	29	Hard hats/Helmets	D	x	x
	30	Shoes/sandals	D	x	x
	31	Traffic cones	D	x	x
	32	Foam sponge	C	x	x
	33	Telephone (incl. parts)	C	x	x
	34	Plastic construction waste	F	x	x
	35	Cable ties	F	x	x
	36	Sanitary towels/panty liners/backing strips	C	x	
	37	Toilet fresheners	C	x	x
	38	Diapers/nappies	D		x
	39	Medical/Pharmaceuticals containers/tubes	C	x	x
	40	Flip-flops	D	x	x
	41	Sunbeds	F		x
	42	Styrofoam packaging (boxes, etc.) >50cm	D	x	x
RUBBER	43	Cigarette butts and filters	A	x	x
	44	Various plastic fragments <20cm	C	x	x
	45	Rubber gloves	C	x	
	46	Balls	D	x	x
	47	Rubber boots (scuba diving bots)	D	x	x
	48	Bicycle tyres	E	x	
	49	Car tyres	F	x	
	50	Large industrial tyres	F		x

Continued on next page

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Table 1 – continued from previous page

MATERIAL	#	Item General Name	Category	Location	
				Surface	Seafloor
CLOTH	51	Tractor tyres	F		x
	52	Rubber belts	F	x	
	53	Various wheels	F	x	
	54	Other rubber pieces $\leq$50 cm	D	x	
	55	Other rubber pieces $>$50 cm	F	x	
	56	Clothing (various clothes)	E	x	
	57	Shoes	D	x	
	58	Clothing rugs and towels	F	x	
	59	Backpacks and bags	E	x	
	60	Carpets and furnishing	F	x	
	61	Ropes	F	x	
	62	Other textile pieces $>$50cm	D	x	
METAL	63	Aerosol/Spray cans industry	D	x	x
	64	Cans (beverages and food)	C	x	
	65	Foil wrappers	C	x	
	66	Disposable BBQ's	E	x	
	67	Appliances (refrigerators, washers, etc.)	F	x	
	68	Tableware (plates, cups, cutlery)	D	x	
	69	Fishing related (weights, sinkers, lures, hooks)	C	x	x
	70	Middle size containers $<$50 cm	D	x	
	71	Gas bottles, drums and buckets ($>$ 4 L)	D	x	
	72	Wire, wire mesh, barbed wire	E	x	
	73	Barrels	F	x	
	74	Car parts / batteries	F	x	
	75	Cables	F	x	
	76	Household Batteries	B	x	
	77	Large metallic objects (various)	F	x	

Continued on next page

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Table 1 – continued from previous page

MATERIAL	#	Item General Name	Category	Location	
				Surface	Seafloor
GLASS	78	Other metal pieces <50 cm	D	x	
	79	Other metal pieces > 50 cm	E	x	
	80	Bottles incl. pieces	C	x	
	81	Jars incl. pieces	D	x	
	82	Light bulbs	C	x	
	83	Tableware (plates, cups)	C	x	
	84	Construction material (brick, cement, pipes)	F	x	
	85	Glass buoys	E	x	x
	86	Glass or ceramic fragments >2.5cm	B	x	
	87	Large glass objects >50 cm	E	x	
PAPER	88	Paper/Cardboard	E	x	x
	89	Cardboard (boxes & fragments)	E	x	x
	90	Cups, food trays, food wrappers, drink containers	D	x	x
	91	Newspapers & magazines	D	x	x
	92	Cartons/Tetra pack Milk	C	x	x
	93	Other paper items ≤20 cm	C	x	x
	94	Other paper items ≤50 cm	D	x	x
WOOD	95	Corks	A	x	x
	96	Processed timber	E	x	
	97	Crates	E	x	
	98	Various boxes (fish boxes)	E	x	x
	99	Ice-cream sticks, chip forks, chopsticks, toothpicks	A	x	
	100	Various processed wooden items >50cm	E	x	

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

1.3 Marine Litter Database Requirements

For establishing the requirements of the marine litter database, the general SeaClear2.0 system requirements, available in [4] have been analysed. An overview of the SeaClear2.0 system requirements that have an impact on the the marine litter database, can be found in Table 1.3. These have to be treated as requirements for the overall performance of the ML-based approach, which in turns is highly dependent on the quality of the data onto which it was developed. Hence, such requirements can be applied on the database itself. The dataset format and categorisation shall also complement the SeaClear1.0 dataset developed in [6], which provides 40 object categories, encompassing not only litter but also observed animals, plants, and robot parts.

Table 2: SeaClear2.0 system requirements related to marine litter (source [4])

Category	Description	Design Objectives
Technical Constraints		
Litter sizes	Identified litter in the test areas range from 5x5cm to more than 100x100cm in cross section.	Enabled grasping capacity to address up to 78% of the identified litter sizes by enabling grasping of objects up to 100cm by 100cm in size.
Object recognition accuracy	Inaccurate identification of underwater litter objects.	Improve object recognition algorithms to achieve 90% accuracy in identifying and classifying litter items.
Adaptability to different types of litter	Inability to effectively handle various types of underwater litter.	Enhance the system's adaptability to different shapes and sizes of underwater litter objects.
User Interface and Human-Robot Interactions	Limited user-friendliness in the control interface.	Develop an intuitive and user-friendly control interface for operators overseeing the cleaning process.
Regulatory and Compliance Constraints		
Data Privacy and Security	Stakeholders/End-users impose regulations regarding the handling and storage of data collected during underwater cleaning operations.	Develop protocols and data management plans for secure data transmission and storage.
Safety and Security Constraints		
Underwater Obstacles	Safety concerns related to potential collisions with underwater obstacles, especially in low-visibility/high-turbidity test sites.	Integrate sonar imaging technologies for obstacle detection and avoidance.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Based on the SeaClear2.0 system requirements described in Table 1.3, the database-specific requirements can be formulated. The database needs to be able to offer a scalable solution, where vast amounts of data, including raw, unstructured or structured data can be uploaded and referenced. Various data types should be allowed, that complement the information in the raw images. It should allow storing the data in its original form, as long as the established structure can be imposed. Furthermore, the database should support alongside storing, also the possibility to analyse the use of the data for training and evaluation purposes on typical ML models. The users of the database should benefit of facile integration with commonly used tools and Integrated development environments (IDEs) for ML developments. The access should be allowed to multiple partners over internet, each having their rights align with their role in the development of the SeaClear2.0 ML algorithms. The requirements on the database

Table 3: Database-specific user requirements, derived from Table 1.3. SeaClear2.0 general requirements

Category	Description
Technical Constraints	
Litter sizes	Database should include images and data on litter ranging from 5x5cm to more than 100x100cm in cross section. Database should include various document types, including images, text documents, Comma Separated Values (CSV) files, Jupyter Notebook and Python scripts, XML and JSON files
Object recognition accuracy	Database needs to contain accurately labeled images of various underwater litter objects. The training environment should be tightly integrated with cloud compute services for running analyses using specific machine learning algorithms
Adaptability to different types of litter	Include diverse types of underwater litter in the database, such as e-scooters, bikes, wheels, and ghost nets.
User Interface and Human-Robot Interactions	Should support resource browsing directly from the IDE Database should be accessible to all project partners, allowing granular control for role-specific access control Images and data should support resource browsing directly from the IDE
Regulatory and Compliance Constraints	
Data Privacy and Security	Ensure compliance with regulations regarding the handling and storage of data collected.
Safety and Security Constraints	

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Underwater Obstacles	Include data on underwater obstacles to train AI in detection and avoidance.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

2 Data lake creation and management

The current section presents the reasoning for selecting the software solution as well the Role-based Access Control (RBAC) enforced on the users and the methods for assuring quality of the data.

2.1 Data storage solution

Within the internal infrastructure available at Fraunhofer, there are two main cloud services available: **Private Cloud** and **Public Cloud** services. Both make use of various cloud tools and web applications, depicted in Figure 4, that deliver a scalable and secure infrastructure for using and developing cloud applications.

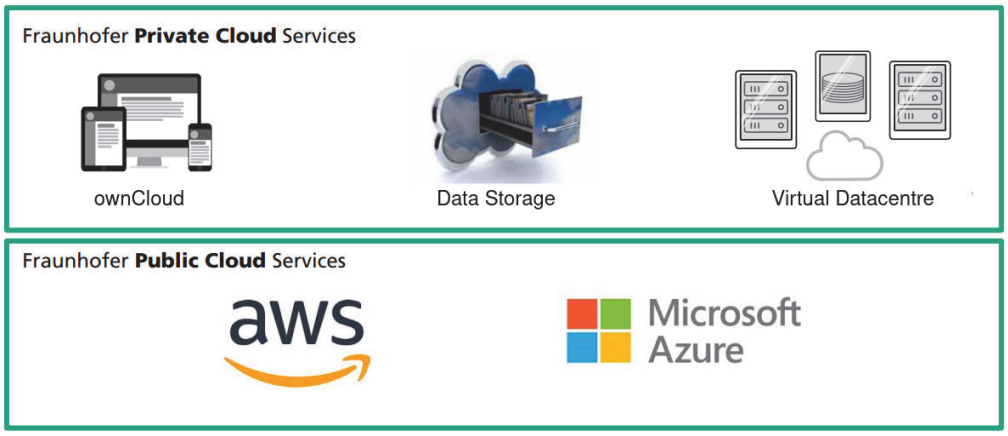

Figure 4: Overview of the available cloud architecture at Fraunhofer (source Fraunhofer)

The two services are complementary to each other and serve slightly different purposes. The **Private Cloud** services are aimed for use-cases where data flexibility is of major importance and where the data doesn't need to be well structured. Moreover, RBAC is very simplistic and offers simple access rights to the teams. However, complex hierarchical roles, with various access rights are not their strength.

On the other hand, the **Public Cloud** services present a scalable solution to use-cases where data automation (through means of Application Programming Interface (API)) is important. These services enforce a standardised data structure, providess computing resources on-demand, and have the benefit of externalising the computation on well-established data centres, that are compliant with the EU laws on data protection.

As depicted in Figure 5, cloud services can be used together within the same cloud architecture. However, for the application purpose, it is advisable to choose the one that fits the purpose the best.

In the context of setting up a database for managing marine litter, adhering to the specifications outlined in Section 1.3, the best option is to utilise a **Public Cloud** service. This choice ensures secure data management and provides the necessary scalability. Additionally, standardising the data structure is beneficial without restricting the data types. Furthermore, with multiple roles involved in dataset creation, ML model training, and data curation, a well-defined RBAC is ideal. Lastly, computing resources

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

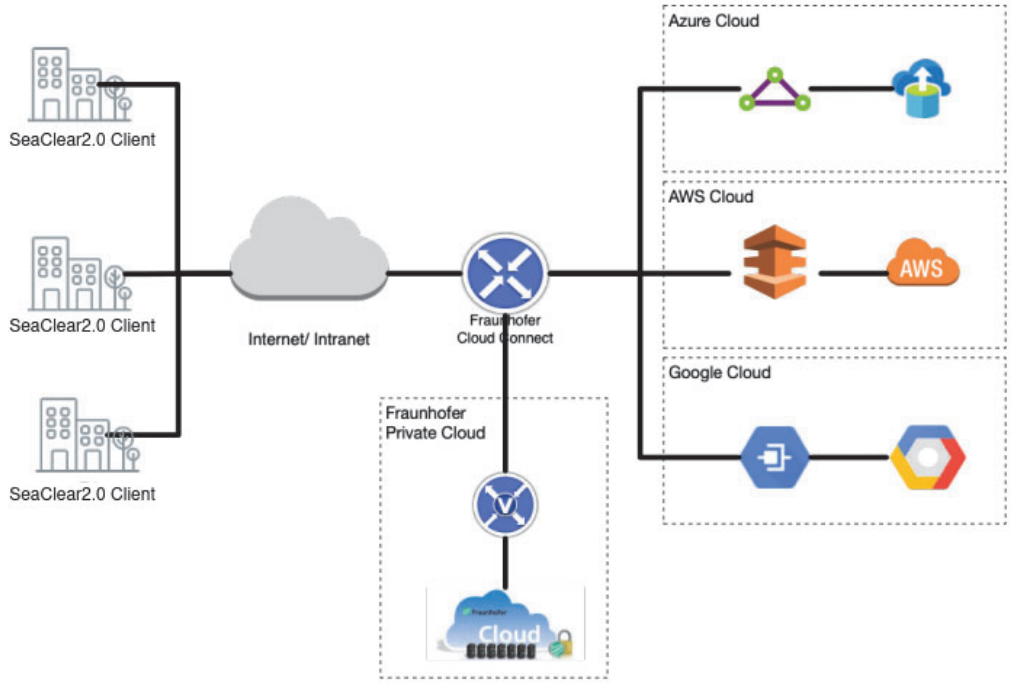

Figure 5: Possibility of using both cloud services within the same cloud architecture (source Fraunhofer)

are needed only during the training and validation phases of ML models, not during dataset creation, so the demand for computational resources on request is not problematic. From the two **Public Cloud** services offered, the Microsoft Azure solution was selected. Azure's Azure Data Lake Storage (ADLS) provides a direct and scalable solution for handling a wide spectrum of data and it is optimised for big data analytics, [7]. As service providers, Amazon AWS and Microsoft Azure have similar features available, but the latter was chosen due to the usage of other cloud-based services by the majority of the partners in SeaClear2.0, hence being more accustomed to the user interfaces and nomenclature.

Within the Microsoft Azure cloud services, the specific database type for supporting the dataset creation, maintenance and usage was chosen. The ADLS Gen2 was selected as the base for the data lake providing file system semantics, file-level security and scale alongside low-cost, tiered storage, with high availability/disaster recovery capabilities, as mentioned in the requirements in 1.3. For more information about the ADLS Gen2, Microsoft's learning platform[7], can be referred. The bulk of the data, image files, as well as their XML or annotations will be stored on a separate ADLS instance, while the code files (e.g. Python and Jupyter Notebook scripts) can be stored on a separate store created while creating an Azure Machine Learning (AML) instance. Code files will be expected to be separately stored on a shared online Git repository. ADLS' RBAC allows for granular access control, while also supporting a hierarchical folder structure. ADLS, being part of Azure's ecosystem, is also complemented by its analysis tools, such as AML. More specifically, the SeaClear2.0 data lake will consist in a series of `Data Storage Containers` located within the `Storage account` record, situated in the

[7] https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction#data-lake-storage-gen2

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Azure `Resource group` service.

2.2 Role Management

Azure cloud services allow granular access to data, called RBAC. This is accomplished through the Access control (IAM) sidebar option, see Figure 6. The list of individuals and their roles can be monitored in the Role assignments tab and modified accordingly. Depending on the role chosen, different access rights to the services (allowed by the subscription) are granted. For example, users having the role of *Azure AI Developer* are granted access to the data lake itself, contained within the `Storage Account`, which can be found under the path: Home > Resource Group > Storage Account). In addition, they can also work with the ML models in `Azure Machine Learning Studio` contained within `Azure Machine Learning workspace`, by navigating to the following path:
Home > Resource Group > Azure Machine Learning workspace. So far the administrative right is managed by Fraunhofer ADS (Agricultural Data Space), who is also responsible for providing access rights to all subsequent users.

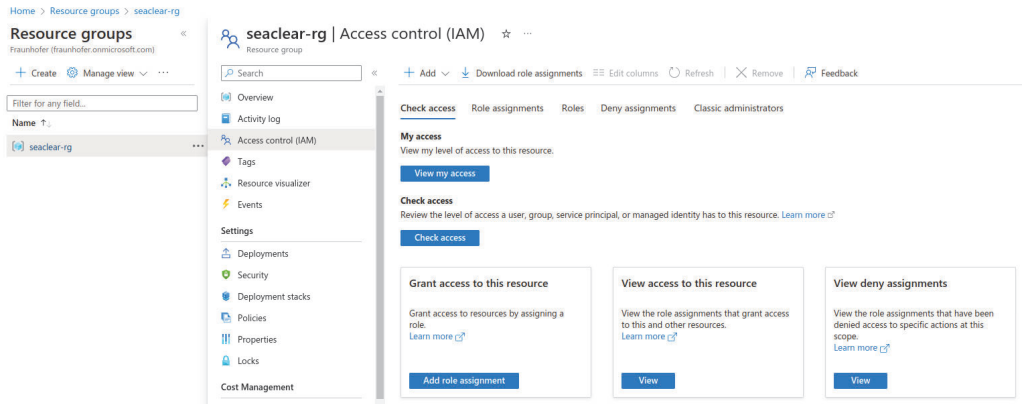

Figure 6: View of the member management menu Access control (IAM)

However, the SeaClear2.0 data lake is contained within a specific resource group and each member must be added to the resource group in order to use the aforementioned tools within the SeaClear2.0 data lake. The new members can be added simply through ADLS account or through an AML workspace. This action can be undertaken only by privileged users having administrative rights, such as *Owners*, *Management Group Contributors*, or *User Access Administrators*. Members can be added to either one resource group or to a specific resource/service within the resource group. Privileged users can use the following path for adding new members Access control (IAM) > Add role assignment > (Assign specific role) > Next > Select members > Select > Review + assign. The added member will see a new subscription in their subscriptions as well as the resource group they have been added to now, if the addition goes through.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

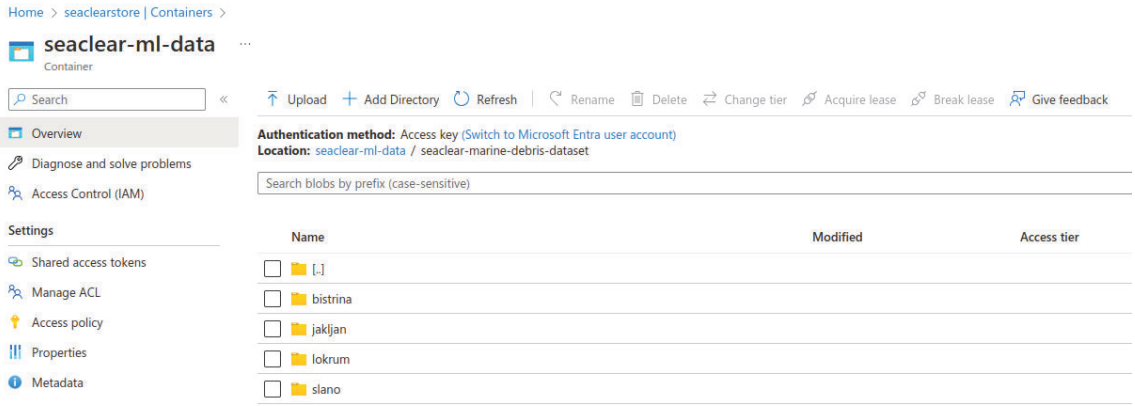

Figure 7: Example of the site-oriented folder structure within the data lake

2.3 Quality Assurance

To guarantee the efficient execution of ML algorithms, it is essential to have the data properly formatted and structured. For this purpose, specific conventions should be adhered to when uploading data or creating scripts that generate annotation files.

Firstly, the directory structure organisation must be respected. The SeaClear2.0 data lake established earlier is a `seaclear-ml-data` container within `resource group` service. For the SeaClear2.0 data lake, a dedicated directory for each demo and pilot site has been created within the `seaclear-ml-data` directory located in the `seaclear-ml-data` container, as depicted in Figure 7. Inside each site folder, there are additional directories for different hardware used to capture images. For instance, `oculus-1200` is used for sonar log and images from the imaging sonar `BluePrint Oculus 1200d`. Images from a particular piece of hardware at a specific site should be placed in their respective directories, and new folders should be created if the hardware changes.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

Within the collected data itself, due to the fact that the specific use-case for ML is object detection and classification, within each hardware folder, there needs to be a folder dedicated for images and one for labels. Images can be taken from different sources that are available on the same robot (e.g. a set from a camera and one from a sonar), but, as mentioned earlier, these must be placed in the folder of the corresponding sensor. Pairs should be identifiable by usage of a common naming method or an unique ID.

Labels consist of either segmentation maps or bounding boxes (or both) for every (relevant) instance that occurs in the image. Each bounding box or segmentation map should be accompanied by a corresponding class label. The complete set of classes used for labelling can be found in Table 4. These are grouped based on the expected litter fragments described in [3] and summarised in Table 1.

Table 4: List of all class labels sorted by super-category.

Litter	Litter	Litter	Bio	Robot
tarp_plastic	can_metal	brick_clay	plant	rov_cable
container_plastic	container_middle_size_metal	cup_ceramic	animal_etc	rov_tortuga
bottle_plastic	wreckage_metal	branch_wood	animal_sponge	rov_vehicle_leg
pipe_plastic	cable_metal	furniture_wood	animal_shells	rov_bluerov
net_plastic	boot_rubber	cardboard_paper	animal_urchin	
cup_plastic	tire_rubber	snack_wrapper_paper	animal_fish	
bag_plastic	bottle_glass	unknown_instance	animal_starfish	
sanitaries_plastic	jar_glass			
snack_wrapper_plastic	rope_fiber			
lid_plastic	clothing_fiber			
rope_plastic	tube_cement			

In addition to the labelled images, their metadata should be uploaded alongside, in a different directory, at the same level with the images and label directories, but under the parent directory having the name of the hardware used in gathering the data.

Specific codes for identification of the raw data, the labelled images or the labels themselves will be agreed within the consortium in a separate document.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

3 Data Collection

This section provides an overview of the envisioned methods and parameters for collecting data and the tools used in the cloud infrastructure.

3.1 Tools and Frameworks

The main cloud services used to collect, store and process data are ADLS and AML. There are, also, some ancillary services like *Cost Management*, *App Registrations* and *Subscriptions* that are necessary for the setup as well as monitoring of these services.

ADLS

ADLS is used to store the data. Data is organized and managed using the GUI as explained in A.2. ADLS data can also be accessed by other services for purposes of analysis.

AML

AML is an Azure service that allows us to consume data, pre-process it and run analyses using ML algorithms on it. These ML tasks can also be automated in pipelines to allow for real-time data to be processed. A small example is given in A.3.

Cost Management

To monitor the costs of each resource, or a resource group, the "Cost Management" service is used, accessible through the General > All services (accessed by clicking on the *More Services* right-arrow icon on the home screen). See Figure 8. Inside Cost analysis > Accumulated Costs and Cost analysis > Resource the historical, predicted and resource-wise breakdown of costs is shown.

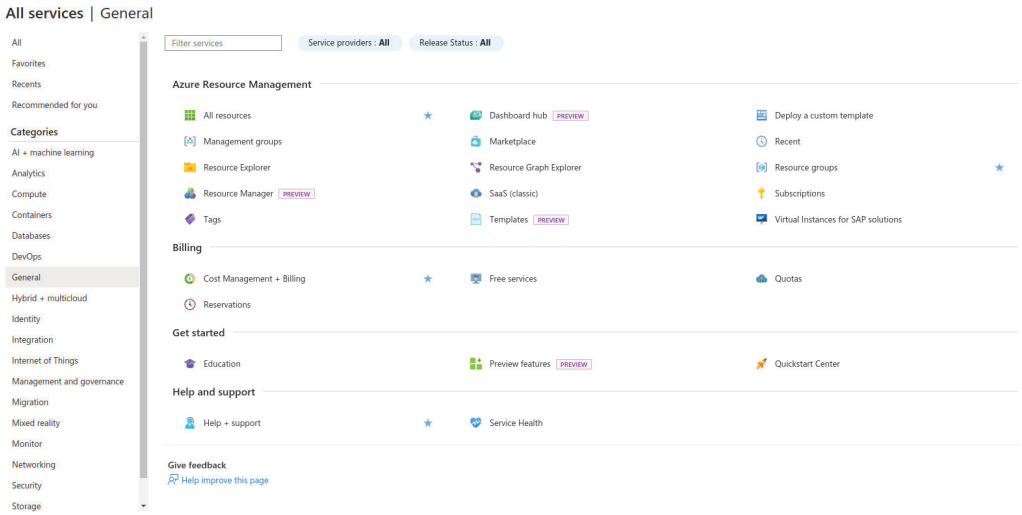

Figure 8: Cost Management

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

App Registrations

App Registrations is an administrator-level service that is used to register Azure services like AML as applications in order to create *Service Principals* required to access data in a data store like ADLS in AML.

3.2 Methods and Parameters

1. Selection and description of designated demonstration areas. Rationale for choosing specific locations.
2. Consideration of boundary conditions for technical developments.
3. Methods for collecting and processing data on marine litter.
4. Use of photo and video recording for AI training.
5. Integration of existing data from related projects and expert knowledge.

The main methods for ingesting new data foreseen in the SeaClear2.0 project is by field surveys done by the partners from the demo and pilot sites using similar sensory equipment. In addition to the field surveys, integration of external data is desirable, as long as the collection parameters are comparable. A possible public source for external data integration is European Marine Observation and Data Network (EMODnet), which aims to create the most comprehensive data sharing platform across marine research and monitoring initiatives at the European level.

The main parameters for data collection are summarised as following:

- **Equipment**: the used camera or imaging sonar must have similar working parameters. In addition, the Remotely Operated Vehicle (ROV) used for data collection is important for knowing the exact position of the sensor equipment in reference to the other sensors. Lastly, lights onboard the ROV are expected.
- **Site conditions**: the site environmental conditions have to be thoroughly described and in-line with the boundary conditions presented in [4].
- **Survey design**: operational parameters have to be include details such as altitude, speed or the route followed by the operator with the ROV in aquiring data.
- **Data management**: the output information has to be uploaded to the data lake respecting the file formats and the structure established in Section 2.3.

By aligning the field survey to the aforementioned parameters accurate, reliable, and suitable data for training the ML models will be stored in the SeaClear2.0 data lake. This will allow better performance in the detection and classification tasks and, ultimately, achieving the goals of the SeaClear2.0 project.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

4 Conclusion and Future Directions

4.1 Alignment with Project Objectives

The integration of a data lake into the project's architecture is a strategic decision that aligns closely with our overarching project objectives. The primary aim of the project is to enhance the detection and classification of marine litter through the use of advanced ML technologies. This requires a robust and scalable data management solution that can handle diverse and voluminous datasets, which is where the data lake proves to be invaluable. The annotation format shall be compatible and complementary to its predecessor dataset from SeaClear 1.0 [6].

1. **Centralised Data Repository**: The data lake offers a unified storage solution for diverse data types such as structured, semi-structured, and unstructured data. This centralisation simplifies data access and management, guaranteeing that all pertinent data gathered from various sources (e.g., underwater sensors, cameras, and external datasets) are easily accessible for analysis and ML models training.

2. **Scalability and Flexibility**: One of the core objectives of the SeaClear2.0 system is to develop a system that can adapt to the evolving nature of marine litter detection. The data lake's inherent scalability allows for the seamless addition of new data sources and expansion of storage capacity as needed. This flexibility supports the continuous improvement and training of ML models by incorporating new and diverse data, leading to more accurate and comprehensive detection capabilities.

3. **Enhanced Data Processing Capabilities**: Efficient data processing is crucial for real-time and near-real-time applications of marine litter detection. The data lake supports advanced data processing frameworks that enable quick ingestion, transformation, and analysis of large datasets. This capability aligns with the project's goal of developing a responsive and efficient system that can operate effectively in various marine environments.

4. **Support for Advanced Analytics**: The project aims to leverage advanced analytics and machine learning to improve the accuracy of litter detection and classification. The data lake's ability to store raw data in its native format allows partners to perform more detailed and complex analyses. This supports the development of sophisticated ML models and algorithms that are essential for achieving high levels of accuracy and reliability in marine litter detection.

5. **Interoperability and Integration**: The data lake architecture is designed to be highly interoperable, allowing easy integration with various data processing and analytical tools. This interoperability ensures that the project can utilise the best available technologies for data analytics, machine learning, and visualisation, thus enhancing the overall effectiveness and efficiency of the marine litter detection system.

By aligning the implementation of the data lake with these project objectives, it is ensured that the data management infrastructure not only supports but also enhances the SeaClear2.0 project's capabilities in achieving its mission of improving marine litter detection and contributing to environmental conservation efforts.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

4.2 Summary and Recommendations

In summary, the deployment of a data lake for storing and managing datasets pertinent to marine litter detection represents a critical advancement in the SeaClear2.0 project. The data lake offers a centralised, scalable, and flexible repository that supports the diverse data requirements inherent to this initiative. By facilitating efficient data processing, advanced analytics, and seamless integration with various tools, the data lake aligns with and enhances the project's primary objectives of improving the accuracy and efficiency of underwater litter detection and classification and, also, addresses the technical requirements formulated in [4].

Recommendations:

1. **Continuous Data Ingestion**: Establish ongoing processes for the regular ingestion of new data from various sources. This will ensure that the ML models are continually trained on the most current and comprehensive datasets. An example at the European Comission (EC) level is EMODnet, briefly described in Section 4.2.1.
2. **Data Quality Management**: Enforce data quality management protocols to ensure that the data stored in the data lake is accurate, reliable, and suitable for training the ML models. This includes regular validation and cleaning of the data.
3. **Performance Optimisation**: Continuously monitor and optimise the performance of the data lake to ensure that it meets the demands of large-scale data processing and analytics. This might involve scaling resources or refining data processing workflows.
4. **Stakeholder Training and Engagement**: Provide comprehensive training for stakeholders and project partners on how to effectively utilise the data lake for their respective roles. This will maximise the utility of the data lake and ensure that all users are capable of leveraging its full potential.

4.2.1 Integration of Existing Data (EMODNet)

Integrating the data lake with the EMODnet would enhance data sharing and collaboration across marine research and monitoring initiatives. EMODnet provides access to a wealth of marine data, products, and services that are crucial for understanding and managing Europe's seas and oceans. By aligning with EMODnet, the data lake can facilitate more efficient data ingestion, standardised data formats, and improved accessibility for researchers, policymakers, and stakeholders involved in marine conservation and environmental protection. This integration would support the creation of comprehensive datasets, fostering innovation and informed decision-making in marine management and policy development.

4.3 Potential Applications

The implementation of a data lake within the project framework opens up a wide range of potential applications that can significantly enhance our understanding and management of marine litter. The comprehensive and centralised nature of a data lake allows for versatile use cases across various domains, including environmental research, policy-making, and technological innovation.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

References

[1] Wiktionary, "database — wiktionary, the free dictionary," 2024, [Online; accessed 19-June-2024]. [Online]. Available: https://en.wiktionary.org/w/index.php?title=database&oldid=80240375

[2] P. Sawadogo and J. Darmont, "On data lake architectures and metadata management," *Journal of Intelligent Information Systems*, vol. 56, no. 1, pp. 97–120, 2021.

[3] I. Pozniak, "D2.1: Marine litter occurrence domains report," Regional agency DUNEA, Croatia, Tech. Rep., October 2023. [Online]. Available: https://www.seaclear2.eu/results/

[4] C. H. ten Eikelder, "D2.2: Public demonstrations, pilot & showcases plan," Hamburg Port Authority, Tech. Rep., January 2024. [Online]. Available: https://www.seaclear2.eu/results/

[5] D. Fleet, T. Vlachogianni, G. Hanke *et al.*, "A joint list of litter categories for marine macrolitter monitoring," *EUR*, vol. 30348, p. 52, 2021.

[6]

[7] R. Ramakrishnan, B. Sridharan, J. R. Douceur, P. Kasturi, B. Krishnamachari-Sampath, K. Krishnamoorthy, P. Li, M. Manu, S. Michaylov, R. Ramos *et al.*, "Azure data lake store: a hyperscale distributed file service for big data analytics," in *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017, pp. 51–63.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

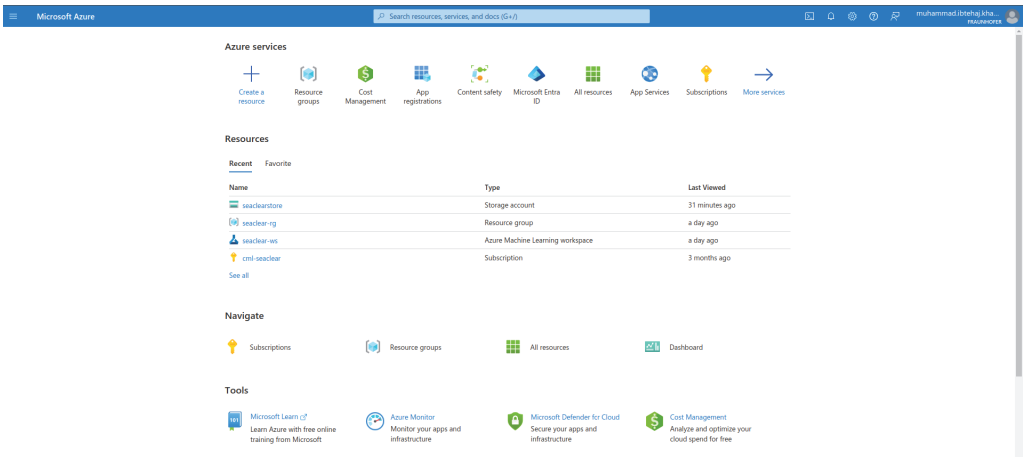

Figure 9: Azure homepage

A Annex A: Azure Cloud Walkthrough

This project makes extensive use of the Azure cloud, for storing litter images as well as running analyses on them.

This section intends be a simple guide to accessing the data assets, creation of the data store, as well as running analyses.

A.1 Overview

To access the portal, go to https://portal.azure.com. Here, you'll have to sign in using your institutional account.

After being logged in, you'll land on the homepage, where you'll see, among other things, your "Subscriptions" (yellow key icon) and "Resource Groups" (light blue box in grey brackets), see Figure 9 for reference.

In "Subscriptions", you should see "cml-seaclear". This is like an access key, allowing you to access Azure resources.

In "Resource Groups", you'll see "seaclear-rg". This is a collection of all the Azure cloud resources of SeaClear.

In order to use the Azure cloud to store data or run analyses on it, we need to "Create a resource" by clicking on its icon with a plus sign. There are several kinds of resources to be created, but for a data store, we can go to "Storage" and to "Analytics" for "Azure Machine Learning". Both of these categories can be seen on the left of the screen in a scroll-able menu. See Figure 10 for reference.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

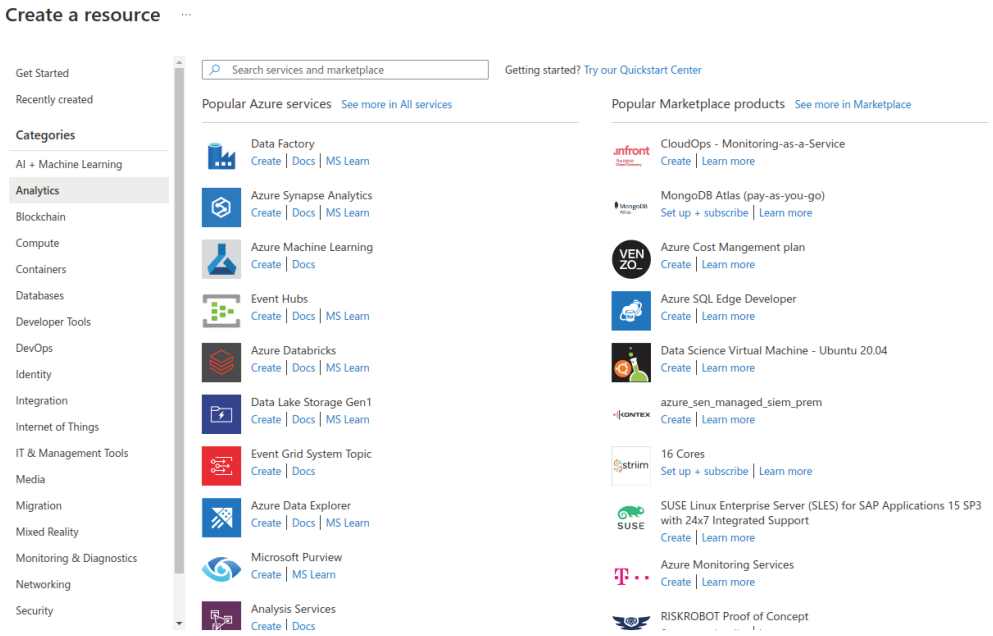

Figure 10: Creating a resource

A.2 Data Store Creation

For creating a data store, such as Azure Data Lake Storage Gen2, we can navigate to “Create a resource”, then “Storage”, then click create under “Storage account”.

In the basics tab, we’ll pick the “cml-seaclear” subscription. Under “Resource group”, we will create a new resource group by giving it an appropriate name. Next, we’ll give an appropriate name to the storage account and in “Redundancy”, pick “Zone Redundant Storage (ZRS)”.

Note: The reason for choosing ZRS is to save our data in case of some catastrophe

Click on “Next” to enter the “Advanced” tab. Here, we just check the “Enable hierarchical namespace” checkbox.

Note: Hierarchical namespace allows us to have a complex folder structure

Click on “Review + create” and then “Create”.

The storage resources will now be provisioned. It may take a short while.

When done, click on “Go to resource” on the resulting screen. Alternatively, go to “Resource groups” from the homepage, in your newly created resource group, in “Overview” and then click on your storage account, bearing the name that you gave it earlier. You will be directed to a page similar to the one in Figure 11.

To ensure that folders can be created, please make sure that Figure 11 shows that “Hierarchical namespace” is “Enabled”

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

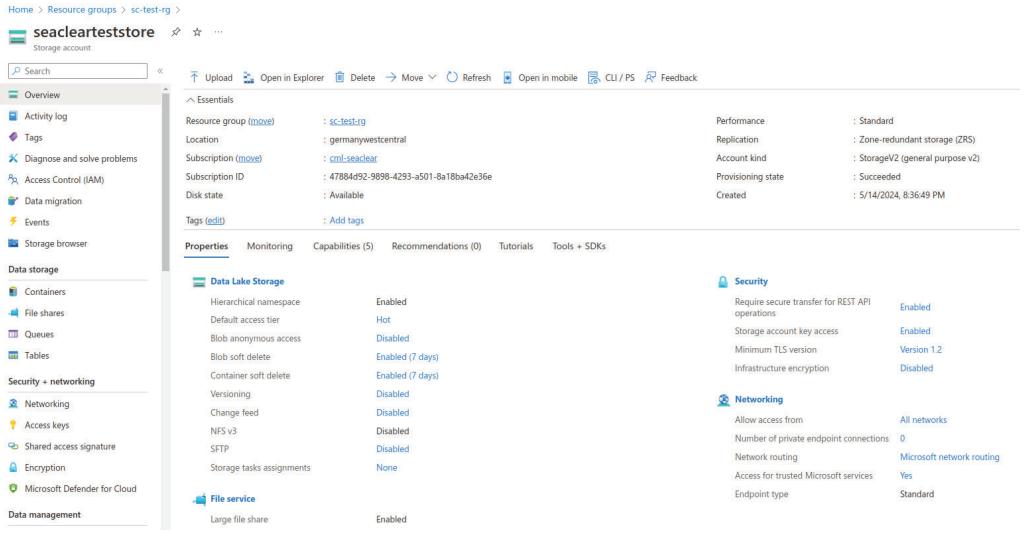

Figure 11: Storage account view

Next, we need to give ourselves permission to access the contents of the containers that we create. To do that, open the resource group that you created, navigate to “Access Control (IAM)”, click on “Add role assignment”, click on “Storage Blob Data Owner”, then “next”, then, “Select members”, search your email address, click on it and then “select”, then “Review + assign”.

In order to be able to save data to this storage account, we need to create at least one container. Containers are meant to keep unrelated data separate and can also be used for access control, meaning, only authorised users can access the contents of these containers.

To create a container, simply go to “Containers” under “Data storage” on the left and then click on the button with a plus sign next to “Container”. Next, name the container and click on “Create”. An entry should now appear with the name that you assigned to the container.

You may enter this container and then use “Add Directory” to create a folder and “Upload” to simply upload files. After uploading files, they too appear as list items in the center of the screen. An example is shown in Figure 12.

A.3 Accessing Data Store for Analyses

To be able to run analyses like training a Machine Learning algorithm on our data, it is necessary to create a Machine Learning workspace and connect our storage account created in the A.2 to the workspace to provide it access to the stored data.

To create a Machine Learning instance, we go to “Create a resource”, in “Analytics”, “Create” Azure Machine Learning. In the page that opens, we use the same “resource group” as the one we created in A.2, give it an appropriate name and click on “Review + create” and then finally “Create”. Please wait while the resource is deployed.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

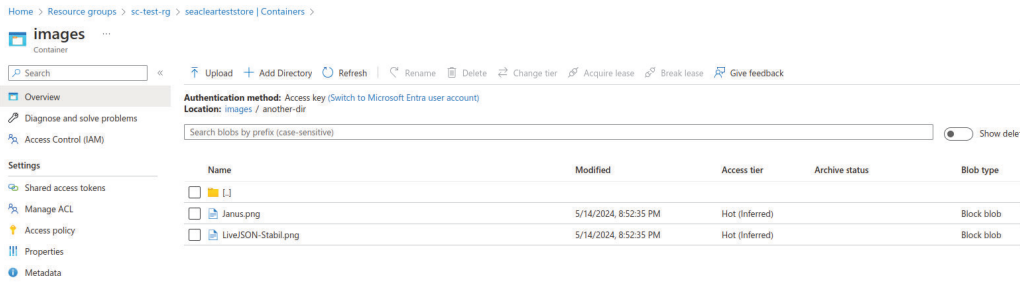

Figure 12: Files and directories in the storage account

Once the resource is deployed, click on “Go to resource”, or open the resource through the “Resource groups” where you’ll now see see several new items including the “Azure Machine Learning workspace” with a navy-blue beaker outline icon. See Figure 13. To go to the workspace, click on the icon and then “Launch studio” on the page that opens. A new tab will open. This is our Machine Learning workspace.

We still need to connect our storage account to this workspace. To do that, we go to “Data” under “Assets” (on the left pane), then “Create”, give it an appropriate name, leave type as “Folder”, “Next”, in “Data source”, “From Azure storage”, “Next”, “Create new datastore” in “Source storage type”, give an appropriate name, choose “Azure Data Lake Storage Gen2” in “Datastore type”, leave fields until “Store name” as they are where you’ll select your storage account and then the container that you want to attach. In “Authentication type”, we’ll have “Service principal” and then fill in the “Client ID” and “Client secret” (Tenant ID is pre-filled). The Client ID and Client secret are given to us by the administrators of Azure from our organisation when requested. After filling in these, we also enable (switch on the blue toggle button) beside “Use workspace managed identity for data preview and profiling in Azure Machine Learning studio”, then we click on “Create” and see that the newly created datastore shows up in the list at “Source storage type” step when the “Datastore type” is “Azure Data Lake Storage Gen2”. We select it and then “Next”. In the next step, we can’t move forward without giving a “Storage path”. A workaround here is to do “Back” and then “Next”, and then again the newly enabled “Next”. Then “Create”.

Now, when we click on “Data” under “Assets”, we see the new datastore(s) that we created. Clicking on it reveals a tab (among others) called “Explore”. Clicking on it reveals the files that we have in the container that we have attached, meaning that data has successfully been made accessible to the workspace and can be used in analyses.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

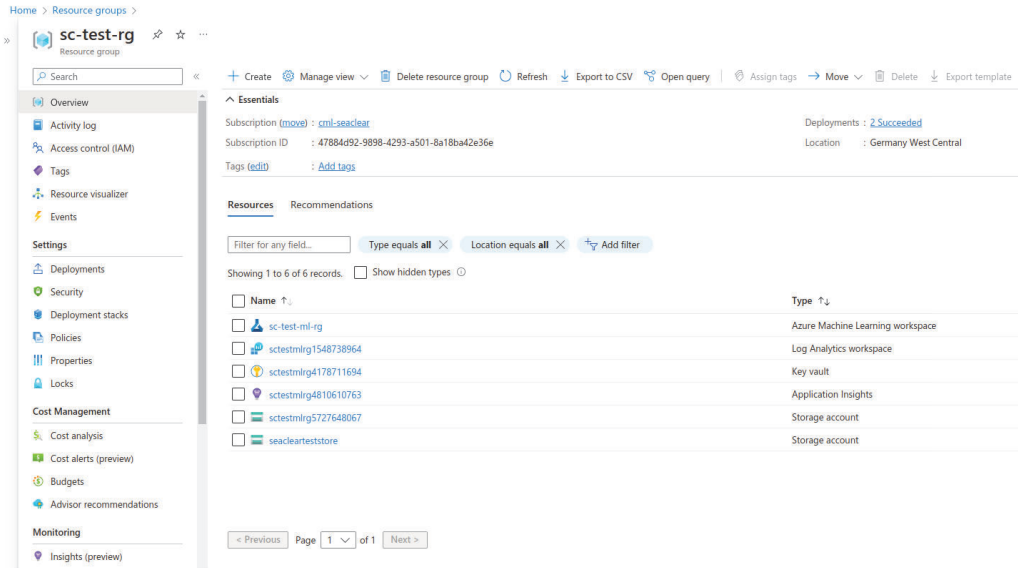

Figure 13: Azure ML workspace in the resource group

A.4 Running Machine Learning Analysis

In order to run arbitrary code in an Azure ML workspace to analyse some data, we need to create an environment, a Jupyter Notebook script and import data to be analysed.

We start with importing some CSV data. In the workspace, we go to “Data”, then “Create”, give it an appropriate name, “Next”, “From local files”, let the “Datastore type” be “Azure Blob Storage”, then “workspaceblobstore”, “Next”, “Upload files or folder”, “Upload files”, upload a CSV file, then “Next”, then “Create”. Now, in the workspace, in “Data” under “Assets”, your new data store should appear, bearing the name that you gave it. Clicking on it and then “Explore” will show it as a list item as well as give a preview of the CSV data. See Figure 14

Next, we will create the environment as well as the script necessary to access and analyse this data.

We start by creating a script. To do that, please click on “Notebooks” under “Authoring” on the left pane of your workspace. Click on the blue “Files” button in the middle of the screen with a plus sign and then “Create new file”. Name appropriately and then “Create”.

Paste the following code in the first cell, after inserting your own subscription_id, resource_group_name and workspace_name:

```
from azure.ai.ml import MLClient
from azure.identity import DefaultAzureCredential
from azure.ai.ml.entities import Data
from azure.ai.ml.constants import AssetTypes

# authenticate
credential = DefaultAzureCredential()

# Get a handle to the workspace
```


Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

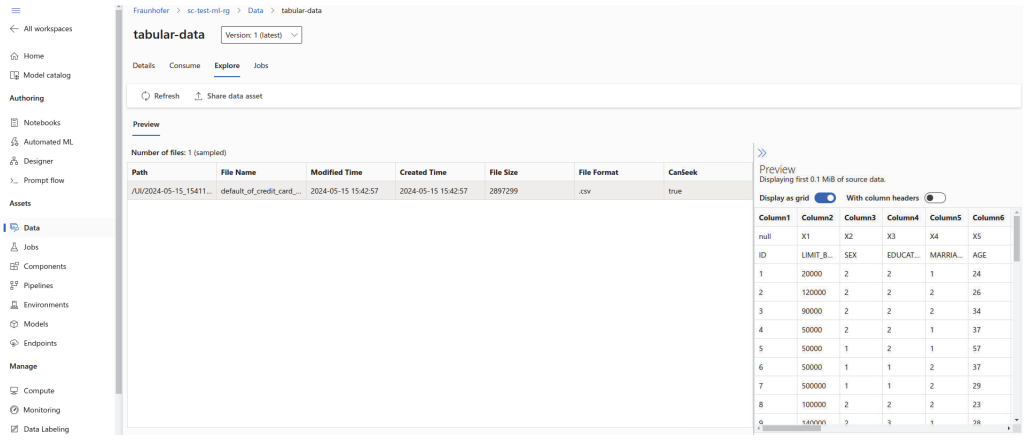

Figure 14: CSV data with the preview

```
ml_client = MLClient(
    credential=credential,
    subscription_id="your-id-here",
    resource_group_name="your-resource-group-name-here",
    workspace_name="your-workspace-name-here",
)

# click on the subscription name at the top right to find the subscription id,
    resource group and workspace name
```

Create a new cell (press Shift + Enter after clicking inside the first cell or hover under the first cell and the option will show up, use “Code” in that case) and then paste the following code after replacing the “your-storage-acc-name” by the name of your CSV data storage account. (If unsure, go to “Data” under “Assets” and investigate which of the displayed list items have your CSV data):

```
import pandas as pd

# get a handle of the data asset and print the URI
data_asset = ml_client.data.get(name="your-storage-acc-name", version=1)
print(f"Data asset URI: {data_asset.path}")

# read into pandas - note that you will see 2 headers in your data frame - that is
    ok, for now
df = pd.read_csv(data_asset.path)
df.head()
```

Save the script (Ctrl + S, or the save icon in the toolbar).

Next, we need to create a compute instance (also called creating an environment) to run our script. Click on the plus icon next to the “Compute” text box in the toolbar and pick a virtual machine, giving it an appropriate name in the process. “Review + Create”. Wait for the instance to be created, the black circle beside “Compute” will turn blue and then finally green when the process is complete. See Figure 15.

Finally, run the two cells, one after the other, if all goes well, you should see the first five rows of the

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them

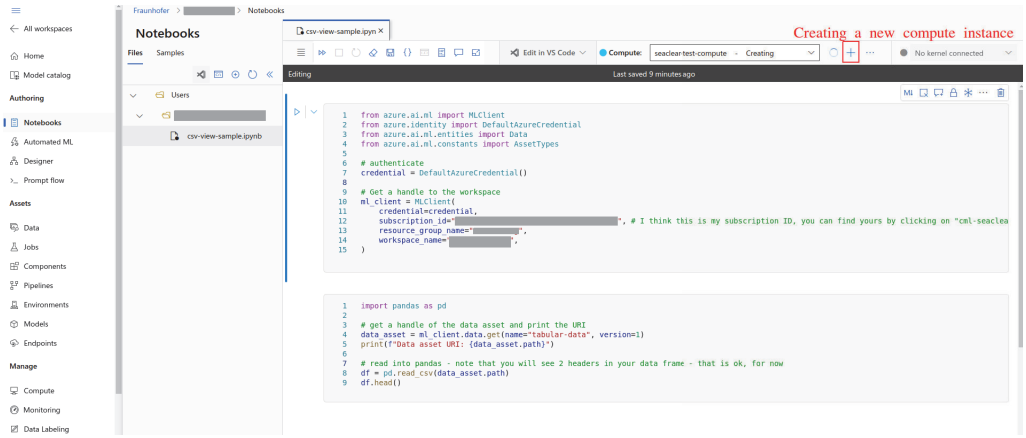

Figure 15: Creating a compute instance in Azure

CSV data that you uploaded, showing that data is now available to be ingested by a Machine Learning algorithm.

Co-funded by the European Union

SeaClear2.0 is co-funded by the European Union under the Horizon Europe Programme (Grant Agreement 101093822). Views and opinions expressed herein are those of the author(s) and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them