Politecnico di Milano

Department of Civil and Environmental Engineering

Master of Science in Geoinformatics Engineering

Towards a Digital Twin of the Italian Coast

Juan Pablo Duque Ordoñez

Supervisor: Prof. Maria Antonia Brovelli

Student code 10730745

Academic year 2021/2022

Abstract

In the framework of the Destination Earth initiative, a long-term project to take advantage of the great amount of data produced by European agencies and scientific organizations around the world, the goal of building a Digital Twin of the Earth was born. The Digital Twin Earth is a high-precision digital model of the Earth that integrates various aspects of the Earth's system in order to monitor and simulate natural phenomena and related human activities, and that is able to explore the past and present and predict the future.

To build a Digital Twin Earth it is required the scientific cooperation of European institutions, alongside with a set of technological tools such as High-Performance Computing (HPC), Cloud Computing and connectivity, Big Data, interoperable data and data standards, and Artificial Intelligence (AI) to simulate and model the Earth's systems processes.

On this work it is presented a web platform that uses open-source technologies to integrate a wide set of more than 60 geospatial layers of environmental data, provided openly and for free by Copernicus Marine Service, and Copernicus Land Monitoring Service, in addition to demographics data provided also open and free by WorldPop. With the data integrated on this platform a user is able to explore and analyze many land and sea layers. This platform is focused solely on the coastal areas of Italy, but its modular and extensible design is suitable for extending it and replicating it to other parts of the world.

The platform addresses the Digital Twin Earth's Big Data and interoperability component by integrating several geospatial data sources using a mediator-wrapper integration architecture that leverages the Open Geospatial Consortium (OGC) standards for geospatial data, the Cloud Computing and connectivity component by providing a web-based interface to explore and analyze the integrated data, and the scientific cooperation component by enabling the possibility to save and share the analysis and discoveries made through the platform.

This work constitutes a proof of concept and an approach of what a Digital Twin of the Earth is capable of. To show it, a case study is presented analyzing ocean acidification on Genova and is whereabouts.

The development of the platform is a work in progress, which means that many more features and functionalities are to be included in following versions, having in mind a tool that is open, data-centric, and a good example of a Digital Twin Earth.

Abstract in Italian

Nell'ambito dell'iniziativa Destination Earth, un progetto a lungo termine per sfruttare la grande quantità di dati prodotti dalle agenzie europee e dalle organizzazioni scientifiche di tutto il mondo, è nato l'obiettivo di costruire un Digital Twin of the Earth. Il Digital Twin Earth è un modello digitale della Terra ad alta precisione che integra vari aspetti del sistema terrestre al fine di monitorare e simulare i fenomeni naturali e le attività umane correlate, ed che è in grado di esplorare il passato e il presente, e prevedere il futuro .

Per costruire un Digital Twin Earth è necessaria la cooperazione scientifica delle istituzioni europee, insieme a strumenti tecnologici come High-Performance Computing (HPC), Cloud Computing e connettività, Big Data, dati interoperabili e standard dei dati, e Intelligenza artificiale (AI) per simulare e modellare i processi dei sistemi terrestri.

In questo lavoro viene presentata una piattaforma web che utilizza tecnologie open-source per integrare un'ampia serie di oltre 60 strati geospaziali di dati ambientali, forniti apertamente e gratuitamente da Copernicus Marine Service e Copernicus Land Monitoring Service, oltre ai dati demografici forniti anche aperti e gratuiti da WorldPop. Con i dati integrati su questa piattaforma un utente è in grado di esplorare e analizzare molti strati di terra e mare. Questa piattaforma è focalizzata esclusivamente sulle aree costiere italiane, ma il suo design modulare ed estensibile è adatta per estenderla e replicarla in altre parti del mondo.

La piattaforma affronta la componente Big Data e interoperabilità di Digital Twin Earth integrando diverse fonti di dati geospaziali utilizzando un'architettura di integrazione mediator-wrapper che sfrutta gli standard Open Geospatial Consortium (OGC) per i dati geospaziali, la componente Cloud Computing e connettività fornendo una interfaccia web per esplorare e analizzare i dati integrati, e la componente di cooperazione scientifica consentendo la possibilità di salvare e condividere le analisi e le scoperte fatte attraverso la piattaforma.

Questo lavoro costituisce un proof of concept e un approccio di ciò di cui è capace un Digital Twin of the Earth. Per dimostrarlo, vengono presentati due casi studio, che analizzano l'acidificazione degli oceani a Genova e un'analisi della variazione biogeochimica sul delta del Po, nella Regione Veneto.

Lo sviluppo della piattaforma è un work in progress, il che significa che molte più caratteristiche e funzionalità devono essere incluse nelle versioni successive, avendo in mente uno strumento aperto, incentrato sui dati e un buon esempio di Digital Twin Earth.

Acknowledgements

This work is the conclusion of a 2-year wild journey called Laurea Magistrale. I would like to give special thanks to my advisor, Maria Brovelli, that guided me through the process of developing this thesis with perseverance and patience. Also, to EBWorld for providing me this topic of research and for providing me their immense support.

I would like to thank all the people that has been part of the journey, both in Italy and in Colombia: to my parents Juan Carlos and Patricia, to my girlfriend Angelly, to my friends, and to my teachers. Going through these two weird and amazing years would have not been possible without your support and company.

Last, but not least, to me. *mic drop*

Contents

List of figures

List of tables

Introduction

The European Commission initiative called "Destination Earth" is a long-term project that encompasses the creation of a Digital Twin Earth (DTE), which is a high-precision digital model of the Earth that integrates various aspects of the Earth's system in order to monitor and simulate natural phenomena and related human activities. The main aspect of a Digital Twin Earth is that it offers the possibility to explore the past, understand the present, and build predictive models of the future, adding value to the unprecedented amount of data that is being generated each day. A tool like it will make it possible for decision makers and the scientific community to have a deeper understanding of all processes that interact on the Earth's system to tackle major problems such as global warming.

A Digital Twin Earth is a technological tool that needs many elements to work, such as strong computation capabilities to process high amounts of data and models, connectivity, cloud computing, Artificial Intelligence (AI), models that can describe physical, chemical, and social phenomena, scientific collaboration to further advance the models, high volumes of good quality data –big data –, and interoperability to glue everything together.

A full-scope Digital Twin Earth is an enormous task that may require years to be built and be fully functional. This work presents an initial approach to address some of the elements that constitute the Digital Twin Earth, with a special emphasis on the interoperability component and focused solely on the Italian coastal zones, which are of high importance for social and economic development, ecology, and tourism. Since ancient times, coasts have played a fundamental part in human civilization, being a critical element for development, economy, transportation, and tourism. Biologically, coasts host an important portion of global biodiversity and ecosystem richness, which is endangered by global warming and pollution.

Through this thesis, the Digital Twin Earth elements of big data, interoperability, cloud computing and connectivity, and scientific collaboration will be addressed through the development of an innovative platform that combines all those elements. This platform main emphasis is providing a tool that integrates multiple data sources that provide geospatial information regarding the Italian coasts by using the well-known virtual integration architecture of mediator-wrapper (or mediator-based), with the goal of adding value to the high amount of high-quality data that is available on the internet by providing data analysis and sharing capabilities. The final product presented is a work in progress and a proof of concept of the things that could be done with a Digital Twin Earth, and by integrating many datasets that are not easily put together.

Although this thesis is only focused on the Italian coasts, the whole development and planning process of the platform has been made having modularity and reproducibility in mind to be able to replicate, reuse, and extend it for other coasts of the world or any type of area.

Chapter 1 includes a state-of-the-art review where the concept of Digital Twin Earth is defined on the framework of the Destination Earth project, an analysis of similar platforms that integrate and use several data sources is presented, along with a review of the use of data integration methods and the mediator-wrapper architecture for geospatial data.

Chapter 2 contains the theoretical framework of the thesis, which is an introduction to important concepts and definitions necessary to fully understand its content. In this chapter the definitions and concepts of maritime boundaries, georeferenced data standards, and data integration are treated.

Chapter 3 contains the information on all the data that is used and integrated on the developed platform, including a short description of the data sources and each one of the layers that are integrated on the web platform. It includes at the end a table summarizing all the integrated layers with its technical information about spatial resolution, data source, temporal resolution, etc.

Chapter 4 is the technical description of the developed platform, where the architectural design, used technologies, implemented features, application workflow, and screenshots of the application are included. This chapter contains diagrams and specific descriptions of the implementation of the thesis.

Chapter 5 includes a case study that was performed using the platform as a way to show its potential use. On it, the ocean acidification of the Genova port and the Portofino marine protected area is analyzed by using the available maritime, land, and demographic layers available on the platform.

Finally, the conclusion of this thesis is presented, along with a set of ideas to be added to the platform as future work.

Chapter 1: State of the art

1.1 Digital Twin Earth – DTE

The concept of Digital Twin Earth (DTE) is very recent. In fact, Digital Twin Earth was only extensively described by the European Space Agency (ESA) in the Phi-Week 2020, where a discussion was held on how Earth Observation can contribute to the creation of a digital twin of the Earth. ESA defines the Digital Twin Earth as a "dynamic, digital replica of our planet which accurately mimics Earth's behavior" [1]. The topic of the 2020 Phi-Week was selected following the proposal of the Destination Earth initiative of the European Commission, as a response to the emergent problems due to climate change.

Destination Earth, first mentioned in 2020 on the framework of the European Strategy for Data [2], and officially launched in March 2022, is a European Commission initiative introduced by the European Strategy for Data which aims to use the big amount of generated data in actions on climate change, circular economy, zero pollution, biodiversity, deforestation and compliance assurance, supporting the European Green Deal [3]. The final goal of Destination Earth is to assemble the European industrial and scientific excellence to develop a very-high-precision, interactive, dynamic, multi-dimensional, and data intensive replica of the Earth as a digital platform that offers visualization, monitoring, and forecast of natural and human activity on the planet, as well as interaction with the immense amount of available socio-economic and environmental data. This initiative supports the sustainable development goals and Europe's efforts for improving the environment, as stated in the Europe's Green Deal [2][4].

Destination Earth aims also to contribute to the European Commission digital strategy based on principles such as ethics, democracy, fairness, and open autonomy. Copernicus satellites and other constellations that are constantly producing big data streams contribute to shape the core of the project [5].

The core concept behind the Digital Twin Earth is the Digital Twin (DT). This concept was firstly introduced in the manufacturing sector but then emerged in industries such as aerospace, automation, shipbuilding, healthcare, and energy [6]. With the advance of technology, this concept moved also to the likes of smart cities and finally to the concept of Digital Twin Earth. However, the concept of Digital Twin is not yet well defined, and a consensus has not been yet reached, but certain elements are common on the definitions such as physical entities, virtual models, data, connections and services [7]. On the 2020 Phi-Week featuring the Digital Twin Earth, a set of building blocks for the Digital Twin Earth were proposed, such as inclusion of Earth Observation data, AI models for Earth system's, High-Performance Computing (HPC), Cloud Computing and connectivity, scientific collaboration, data interoperability and big data integration [12].

The implementation of Digital Twins is also very recent because in order to implement such a model two main things are necessary: processing capacity and data. Computers nowadays provide huge amounts of processing power, and there is high data availability, which creates a perfect storm for the creation of Digital Twins. Data for Digital Twins refers to both having several datasets that are relevant to describe different aspects of the area of interest, and the proper

integration of such datasets. The concept of Digital Twin can be applied to local, national, and global level [6].

The Digital Twin Earth proposed by the ESA is a global DT, which will be constantly fed with earth observation data in combination with in-situ measurements and artificial intelligence to be able to visualize and forecast natural and human activity on Earth. The ESA's DTE will be able to perform simulations on the Earth's interconnected system considering natural phenomena and human behavior, in order to monitor the Earth's health and to support environmental policies [1].

The actual implementation of the Digital Twin Earth is planned to be modular, by creating separate Digital Twin of different portions of the Earth's system, called Digital Twin Precursors. Within the precursors are included the Digital Twins of Forests, Hydrology, Antarctic, Food Systems, Ocean and Climate Hot Spots. Each precursor is a separate project on their own and is assigned to different organizations and universities across Europe [1]. Data integration and interoperability is fundamental for this initiative, as different DTs should be able to communicate between them to produce a single, complete Digital Twin Earth.

The first Digital Twin precursors defined by the Destination Earth program are:

- Digital Twin Antarctica (DTA): created by the University of Edinburgh and supported by ESA, DTA uses satellite observations, numerical simulations and artificial intelligence to build a replica of the Antarctic ice sheet system, that includes its hydrology, the surrounding ocean, the atmosphere and the biosphere. It tracks the locations of melting events occurring on and under the ice sheet and explores how the edge of ice shelves melt under different hydrological scenarios [1][8].

- Digital Twin Food Systems (DTFS): developed by the ESA's British contractor CGI, the DTFS integrates modular simulations of agricultural activities and the interactions within agricultural ecosystems with the aim to support advanced science-based decision making, allowing enhanced predictive and simulation capacity to respond to environmental challenges [1][9].

- Digital Twin Hydrology (DTH): developed by a consortium headed by the National Research Council (Italy), and five collaborators: Vienna University of Technology (TU Wien, Austria), CIMA Research Foundation (Italy), Ghent University (Belgium), Earth Observation Data Centre (EODC, Austria) and the University of Bologna (Italy). The DTH is a 4D reconstruction of dynamic hydrology at very-high-resolution based on the integration of Earth Observation data, and an advanced modeling system. The DTH prototype has been used for water resources management, and to identify the time and locations that are at risk of landslides and flooding events in the Po River basin in Italy [1][10].

- Digital Twin Climate Impacts (DTCI): developed by a collaboration between the British contractor Telespazio, the University of Leicester, and other members. The DTCI was initially focused on African drought, implementing an innovative combination of earth observation, environmental modeling, and machine learning to provide enhanced decision support capabilities. DTCI emulates complex models and deploys them as fast and simple cloud tools [1][11].

- Digital Twin Forest (DTF): developed by VTT Technical Research Centre of Finland, DTF provides a reconstruction of the forest system at a high level of detail. DTF relies on satellite-based Earth Observation, principally the high-quality Copernicus Sentinel data, to get unique and uniform information from the Earth's forests [1][13].

- Digital Twin Ocean (DTO): created by the French Institute of Research for the Exploitation of the Sea (IFREMER), DTO will focus on exploring the potential of artificial intelligence to learn directly from past data, and the behavior of the Earth system, in order to forecast oceanic events. DTO addresses distinct phenomena on different ocean basins, and sea ice dynamics to help understanding the Arctic amplification. To develop the DTO, two approaches are being followed, a data-driven approach using numerical models and data assimilation to analyze and simulate dynamic systems, and a model-driven approach using numerical simulations to assess large scale and long-term effects from small scale events [1][14][15].

Moving away from the Destination Earth initiative, other studies and works tries to define what is a Digital Twin. In [16], according to the authors, Digital Twins are defined as real-time realistic digital representations of physical entities. Digital Twins in the technological sector are seen as a promising tool for efficiency and optimization, meanwhile in the governmental agencies, it is seen as a means for improving decision making to meet sustainability goals. The last point is supported by the fact that the European Commission is willing to delegate a significant role on the creation of Digital Twins to address climate change and the Green Deal policies.

A similar project to Destination Earth is the Earth Archive project [17]. This project is a scientific effort towards the generation of an open-source, 3D Digital Twin of the entire Earth's surface and everything on top of it. The project aims at scanning the land's surface with very high-resolution LiDAR and posteriorly generating a true 3D Digital Twin for the next generations, reflecting the landscape exactly as it was at the time of scanning.

The Earth Archive Digital Twin will serve as a baseline to understand and visually explore the Earth. It will also support conservation studies and provide necessary information to advance towards sustainable development. Therefore, the Amazon Basin was chosen as the first region to be scanned due to its key role in the Earth's climate and ecological diversity, but also due to its danger of disappearing due to deforestation [17].

A study closer to the one treated in this thesis is the one presented in [18]. The Digital Twin Earth – Coasts is a 2021 study that featured the development of a flood and ocean model which is able to accurately predict sea surface height in most regions. This study also features a web application that is no longer available. The work is framed as a Digital Twin Earth, and on it they tried to develop an end-to-end modular platform, which allows it to easily extend to other simulations and machine learning surrogate models. This plays a key role in an open-source platform to allow extensiveness.

1.2 Similar Platforms

There are not many platforms that integrate several layers for visualization and analysis. Most of the platforms on the internet are focused on displaying proprietary layers, providing means to use or download subsets of data or merely visualize data, but there are not many platforms that integrate and use several data sources in a single platform and allow processing and analysis.

Several web platforms were analyzed and studied to gather inspiration and to understand how they manage data, provide analysis, and offer services in general.

Copernicus Marine Service (CMEMS) has a platform called MyOcean Viewer [19], in which they provide all their layers in an interactive environment. It is a web application developed with the aim of explore and subset all the CMEMS catalog of data. The platform is available on https://myocean.marine.copernicus.eu/.
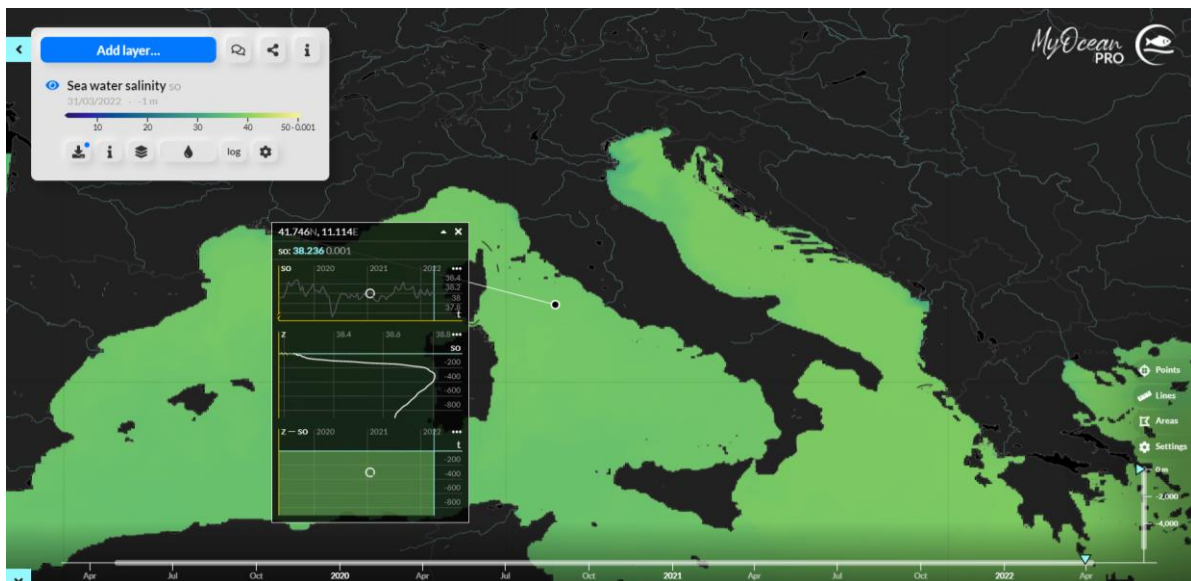
Figure 1. Copernicus MyOcean Viewer

On the platform, it is possible to generate the time series and the depth profiles of CMEMS layers, visualize them, generate area analysis using sampling and share the status of the platform by copying the URL, which contains the applications state.

Although the platform is an excellent example of the use of CMEMS geospatial layers and the capabilities provided by them, it does not integrate any other data sources. Also, their method of generating the time series is very slow since they do the queries day by day.

Another platform that was analyzed for the development of the thesis was the ARIES for SEEA Explorer [20], which features a platform that uses the power of AI models to study ecosystem accounts. It uses the k.LAB Engine [21], an integrated modelling environment to produce AI models. This allows the platform to be extended, not only by the developers, but also by the community with the addition of new AI models developed on k.LAB.

15

On their platform, they integrate several data sources and implement a "context" selection that allows the user to select a range of years, data resolution, and a spatial extent to perform an analysis. However, they do not provide direct visualization of layers, but they provide layers after processing that is done through the k.LAB engine. It is a very good scientific tool since they provide all the documentation of the layers produced, model schema, allow to download data as raster, and provide useful analysis using global data. The tool is available on https://klab.officialstatistics.org/modeler/?app=aries.seea.en#/login but it requires k.LAB Engine credentials in order to access.
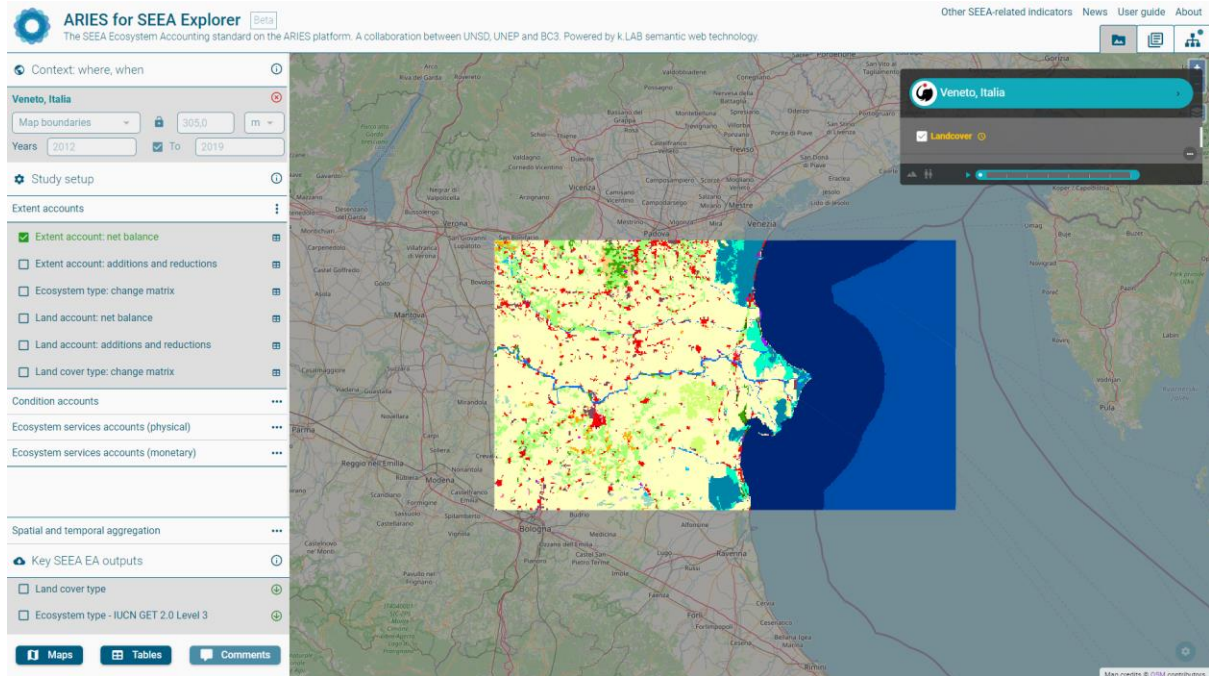
Figure 2. ARIES for SEEA

The USGS Earth Explorer [22] is a tool that allows to explore several datasets, principally from Earth Observation and satellite imagery. It is an interesting tool since it has an intuitive workflow to visualize data and connect with the providers. They provide an intuitive and step-wise workflow that allows to set a spatio-temporal context in which to search for data, and then select the desired available layers, filtered by the context. The tool is available for free at https://earthexplorer.usgs.gov/. The drawback is that this platform allows only visualization and do not allow further analysis or subsetting of the data.
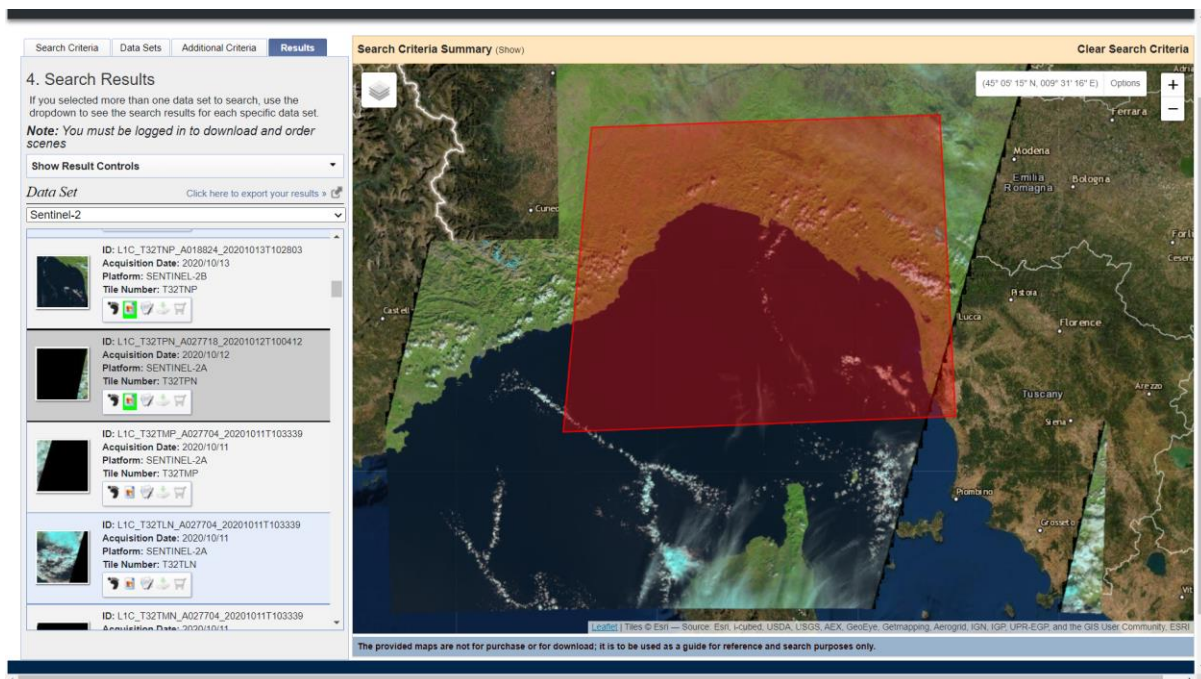
Figure 3. USGS Explorer

The Italian Ministero dell'Ambiente has a platform that allows to visualize all the data available by the ministry and the national geoportal (Geoportale Nazionale) [23]. The platform is available on http://www.pcn.minambiente.it/viewer/ and is a viewer that allows users to import and observe very-high resolution Italian data such as the Digital Elevation Model (DEM) of the entire Italian territory up to 20 meters, orthophotos, coastal data, and in general, all geospatial data provided by the national geoportal. Nevertheless, their platform does not provide means of data analysis and download, just visualization capabilities, and although there is a service to add external Web Map Services (WMS - see section 2.2.1) and ArcGIS Server Services, it works only for their own data services.
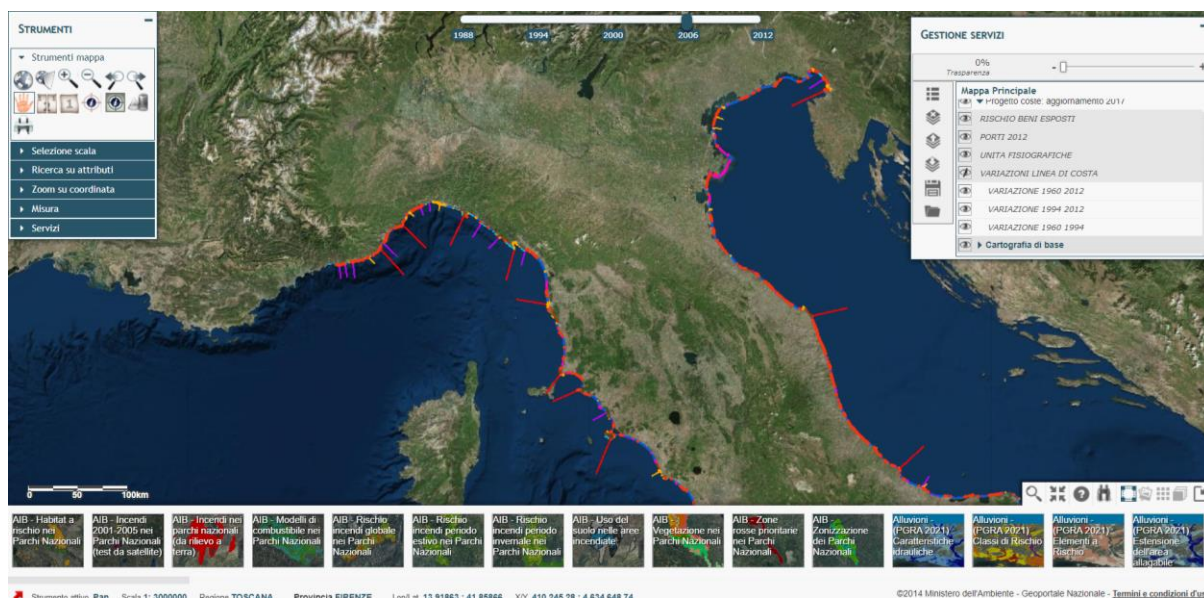
Figure 4. Ministero dell'Ambiente - Geoportale Nazionale Viewer

1.3 Spatial Data integration for geospatial applications

Data integration is a computer science discipline whose goal is to "offer uniform access to a set of autonomous and heterogeneous data sources" [25]. This means to provide a way of accessing and consuming data from different data sources in a single, generic way. The best way to tackle data integration is to have standards on how the data is presented on different data sources. However, this is not always possible because different organizations may have their own data (or the interfaces to provide it) in different formats, different units of measurement, different processing levels for the same data, or do not follow standards at all. This problem is visible even in Copernicus, the main Earth Observation data provider for Europe, where data from different services is not fully interoperable and there is no way to connect them without performing previous processing.

Geospatial data is data with a geographic reference and is very important for the exploration and understanding of our planet and their systems. Geospatial data must follow certain standards that are compiled by the Open Geospatial Consortium (OGC) and ISO (International Organization for Standardization). Section 2.2 explains in depth those standards and how they are used on the platform.

Geospatial data integration is important, because in order to understand and model different Earth's systems it is necessary to integrate multiple kinds of data that may not be provided by the same service. In the literature, many works have been found that use data integration strategies to put together geospatial data. In [26] a mediator-based architecture is used to integrate sensor data provided in a standardized way using the OGC Sensor Observation Service (SOS). The authors in [27] do a similar work by providing a mediator/wrapper architecture to perform virtual integration of sensor data. This virtual integration schemes are popular on the sensor's environment because of the high amount of raw and unprocessed data, as well as the weak standards and heterogeneous data formats that are present.

In [28] it is exposed how geospatial data integration can be benefited by High-Performance Computing (HPC) since processing of geospatial data is often very power-consuming and expensive. The use of HPC is stated on the Destination Earth initiative and a very important part of the Digital Twin Earth.

One of the most used strategies include the use of semantic technology and ontologies for automatic data integration. Studies such as [29][30][31][32] show how semantic technology can aid in data integration schemes. When combined with virtual integration architectures as the mediator-based, they can help to develop faster, and in an automatic fashion, modules to integrate new data. Although semantic technology and ontologies were not implemented on the platform, it is useful for automatic data integration and is intended to be implemented in further versions of the platform.

Chapter 2: Theoretical framework

To fully understand the contents of this thesis, it is necessary to expand and explain some concepts as a contextualization for the data exploration exposed in Chapter 3 and the description of the developed platform that is presented in Chapter 4.

This chapter introduces the concept of data integration and shows the mediator-wrapper architecture that was used. Also, deals with the definition, historical background, and core of the georeferenced data standards, which are fundamental to understand how data is presented and retrieved from data sources. Finally, the concept of maritime boundaries will be addressed, along with a piece of its historical background, and important definitions.

2.1 Data Integration

The practice of integrating several data sources is a matter of the subject of data integration. In the era of big data there is wide availability of information all over the internet, and for creating a Digital Twin, applications should be able to take advantage of that and build knowledge from the existing data. Although standards are in place to help data integration, not all data sources deliver information in the same format and with the same structure, making data integration a complex topic.

As a fact, WMS standards are consistent in the way they present information visually but lack standardization on how to retrieve data from data sources. Some data sources may present the result of the queries as a JSON object, others as XML, others may not deliver the information at all or do it through an API. This lack of consistency between responses supposes an additional step to integrate multiple data sources in a single platform. The way to tackle this problem is through the implementation of the mediator-wrapper architecture as the virtual data integration strategy, which is the main element, and the most important contribution, of the platform developed for this thesis.

2.1.1 Mediator-wrapper architecture

Mediator-wrapper architecture, also called mediator-based, is an architectural pattern for virtual data integration in which multiple data sources are integrated using a centralized mediator who receives generic requests from the client and directs them to a specific wrapper, which manages the complexity of communicating with the data sources, and then translates the response of the wrapper to a generic response. This pattern is well known and widely used in data integration for its flexibility and ease of use. The architecture is visualized explained by *Figure 5*.

Virtual integration means to integrate data without the actual physical data available, integrating the interfaces that are available by means of an integration strategy. The mediator-wrapper architecture is a virtual integration strategy since it uses the mediator to centralize the request and response schemes and communication with the client, and the wrappers as the modular elements that communicate with the services' interfaces and provide integration.

The specific implementation of the architecture on the developed platform is further described in Chapter 6.
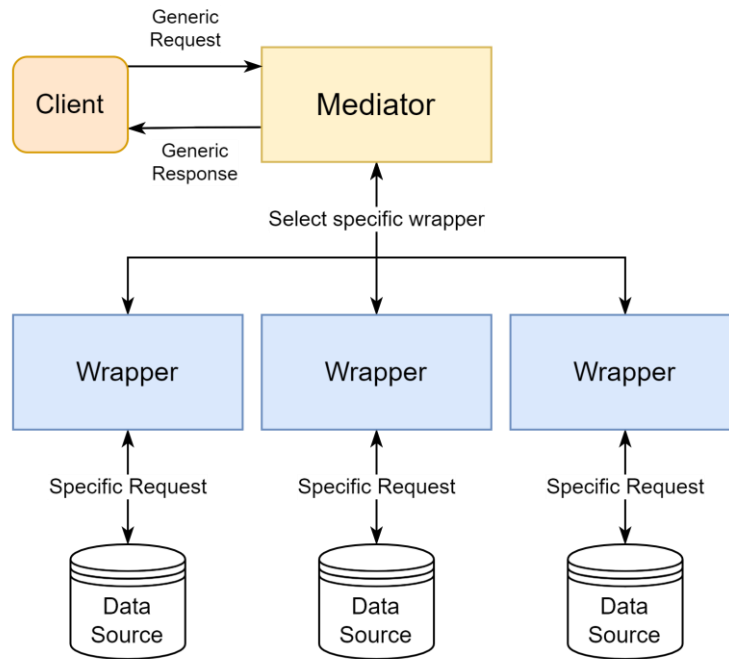
Figure 5. Mediator-wrapper architecture

2.2 Georeferenced data standards

The need to spatially locate ourselves and our surroundings has been a concern for humanity since the dawn of civilization. The earliest global map ever found is a stone tablet called "The Babylonian World Map" dated down to the year 600 B.C. Through the years, maps have improved and evolved from stone, paper, and pencil to digital maps and Geographic Information System (GIS) software.

Advances in technology and the need of governmental institutions of territorial defense and natural resources management helped grow the GIS software industry. Many public and private GIS software solutions started to appear, as vendors created their own solutions to process the newly generated georeferenced data. The increased GIS usage and availability created a new problem, sharing data between different GIS systems was not possible because each company used their own standards and formats, making interoperability and collaboration a very complex task.

The Open GIS Consortium, nowadays called Open Geospatial Consortium (OGC), was founded in 1994 to tackle the problem of interoperability between GIS systems and the growing need of standards for geoprocessing and georeferenced data. Both OGC and ISO (International Organization for Standardization) oversee the standards on georeferenced data. ISO standards are specified in the document ISO 19115-1:2014 [33] by the technical committee ISO/TC 211, in charge of Geographic information/Geomatics standards. OGC standards are published as documents called OGC Specifications and are sent to ISO for review and posterior publication as international standards. OGC collaborates with government, academic, and private sector

organizations to deliver their standards and has a rigorous review process for changes on the specifications.

Among the OGC specifications they mainly cover data delivery, cataloging, and data processing. All the modern mapping software libraries have adopted these specifications and can display and consume services that follow them, making interoperability and data sharing very easy. Through this section are covered diverse OGC Web Services (OWS) specifications that include Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), and Web Map Tile Service (WMTS), as well as data standards for files and objects as GeoTiff and GeoJSON, and finally an introduction to the ArcGIS Server technology and its API that is used by the Copernicus Land Monitoring Service (CLMS) – see section 3.4.2 – to share the geospatial data that is used in the prototype platform developed for this thesis.

2.2.1 Web Map Service – WMS

The Web Map Service Interface Standard (WMS) provides a simple HTTP interface for requesting georeferenced map images from one or more geospatial databases. A WMS request requires the definition the geographic layer (or layers) and area of interest to be processed. The response to the request is one or more georeferenced map images (returned as JPEG, PNG, etc.). Those images can be displayed by an application using any modern web mapping tool or library. The interface also supports the ability to specify whether the returned images should be transparent so that layers from multiple servers can be combined or not [34].

The WMS specification has 4 published versions, 1.0, 1.1, 1.1.1, and 1.3. The version must be specified on the query parameters of any request made to a WMS. This is done for version reconciliation purposes, since most recent versions of the standards may have functionalities that older versions do not have.

According to the specification, a WMS can be basic or queryable. A basic WMS shall support two basic operations: GetCapabilities and GetMap, while a queryable WMS shall also implement the GetFeatureInfo operation. Modern web mapping software libraries know how to manage the retrieval of information from a WMS, both for visualization and querying purposes.

-   **GetCapabilities:**
    Operation which returns in XML format, and a specific structure defined by the standard, all the metadata regarding the service such as names and titles, general metadata, capabilities metadata, layers properties and styles, coordinate reference systems, scale denominators, sampling dimensions and attribution.

    The following table shows the request parameters available for a GetCapabilities request.

Request parameter	Description	Mandatory
VERSION=version	Request version. Can be 1.0, 1.1, 1.1.1 or 1.3	No
SERVICE=WMS	Service type. Must be WMS.	Yes
REQUEST=GetCapabilities	Request name. Must be GetCapabilities.	Yes
FORMAT=MIME_type	Output format of service metadata. By default, is "text/XML"	No
UPDATESEQUENCE=string	Sequence number or string for cache control	No

Table 1. WMS GetCapabilities request

- **GetMap:** Operation which returns the requested map image in the format specified by the request and with the specified parameters. The specific parameters, which are used to locate and build the image to return, are defined in the following table.

Request parameter	Description	Mandatory
VERSION=version	Request version. Can be 1.0, 1.1, 1.1.1 or 1.3	Yes
REQUEST=GetMap	Request name. Must be GetMap.	Yes
LAYERS=layer_list	Comma-separated list of one or more map layers, as defined in the GetCapabilities.	Yes
STYLES=style_list	Comma-separated list of one rendering style per each requested layer. Although it is mandatory it can be empty.	Yes
CRS=namespace:identifier	Coordinate reference system. The namespaces and identifiers available for each layer are specified in the GetCapabilities response.	Yes
BBOX=minx,miny,maxx,maxy	Bounding box corners (lower left, upper right) in CRS units.	Yes

WIDTH=output_width	Width in pixels of map picture.	Yes
HEIGHT=output_height	Height in pixels of map picture.	Yes
FORMAT=output_format	Output format of map.	Yes
TRANSPARENT=boolean	Background transparency of map (the default is FALSE).	No
BGCOLOR=color_value	Hexadecimal red-green-blue color value for the background color (default=0xFFFFFF).	No
EXCEPTIONS=exception format	The format in which exceptions are to be reported by the WMS (default=XML).	No
TIME=time	Time value of layer desired. Only if the layer has "time" as one of its sampling dimensions.	No
ELEVATION=elevation	Elevation of layer desired. Only if the layer has "elevation" as one of its sampling dimensions.	No
Other sampling dimension(s)	Value of other dimensions as appropriate.	No

Table 2. WMS GetMap request

- **GetFeatureInfo:** Operation that returns information about a specific layer on a specific location defined in the request parameters. This operation is optional and, if implemented, the WMS becomes a queryable WMS. In the GetCapabilities metadata it is possible to identify which layers are queryable, because they have the attribute queryable="1" (true).

  The final response is at the discretion of the service provider, but the data shall be related to the feature or features nearest to the (i, j) pixels specified in the request [34].

  The parameters that the GetFeatureInfo request can receive are defined in the following table.

Request parameter	Description	Mandatory
VERSION=version	Request version. Can be 1.0, 1.1, 1.1.1 or 1.3	Yes
REQUEST=GetFeatureInfo	Request name. Must be GetFeatureInfo.	Yes
map request part	Partial copy of the GetMap request parameters that generated the map for which information is desired.	Yes
QUERY_LAYERS=layer_list	Comma-separated list of the layers to query. The layer names are as specified in the GetCapabilities.	Yes
INFO_FORMAT=output_format	Return format of feature information (MIME type). Usually text/xml or application/json.	Yes
FEATURE_COUNT=number	Number of features about which to return information (default=1).	No
I=pixel_column (version < 1.3) or X=pixel_column (version 1.3)	i (or x) coordinate in pixels in the map Coordinate System. Corresponds to the column.	Yes
J=pixel_row (version < 1.3) or Y=pixel_row (version 1.3)	j (or y) coordinate in pixels in the map Coordinate System. Corresponds to the row.	Yes
EXCEPTIONS=exception format	The format in which exceptions are to be reported by the WMS (default= XML)	No

Table 3. WMS GetFeatureInfo request

2.2.2 Web Map Tile Service – WMTS

The Web Map Tile Service (WMTS) [35] standard provides a standard-based solution to serve digital maps using predefined image tiles. This standard complements the previously explained WMS.

WMTS differs from WMS because it serves the map of the source data as individual tiles, instead of the whole image as WMS does. This offers additional flexibility as it is possible to request smaller portions of a map and load it quicker than loading a complete image, but it also requires more power from the server side since caching is more difficult.

It also differs from WMS in the way it calls the data. WMS uses GetMap while WMTS uses the GetTile operation. WMTS uses a local space that is defined by the so-called tile matrix set and tile matrix. Then a tile is located within this space by the tile's column and row. The tile matrix set, and tile matrix represent the tile space, which is defined by the server according to the division of the tiles on the space according to the scale of the map. This allows for different tiles to be fetched depending on the scale in what is called an image pyramid, depicted in *Figure 6*.
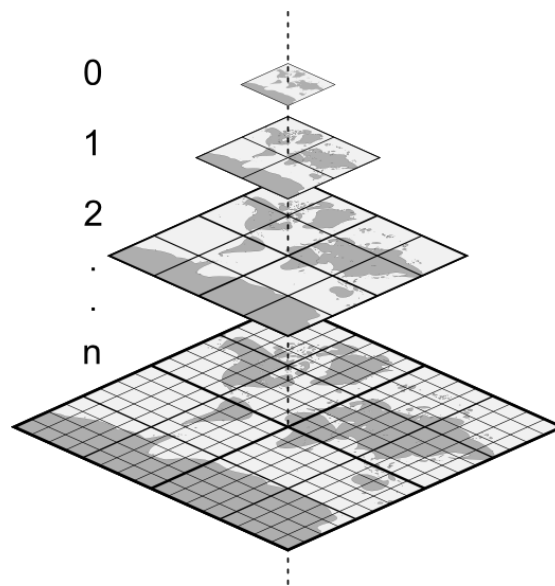
Figure 6. Example of Image Pyramid

2.2.3 Web Feature Service – WFS

The Web Feature Service (WFS) [36] represents the way in which geographic information is created, modified and exchanged on the Internet by offering direct, fine-grained access to geographic information at the feature and feature property level. This service exchanges GIS vectorial data (points, lines, and polygons).

This standard has 3 released versions, 1.0.0, 1.1.0, and 2.0.0. Through the versions, syntax remains the same but newer versions have access to more operations.

This standard specifies operations of discovery, querying, locking, transactions and management of stored and parameterized query expressions. The latest version, 2.0.0, defines the following eleven operations:

-   **GetCapabilities**: Metadata discovery operation. It also shows the operations that the WFS supports.
-   **DescribeFeatureType**: Schema description of feature types. Defines how a WFS expects feature instances to be encoded on input, and how feature instances shall be encoded on output.

- **GetPropertyValue:** Retrieves from the data store the value of a feature property or part of the value or a complex feature property, for a set of features using a query expression.
- **GetFeature:** Returns a selection of features from a data store. Data is returned as a document to the client that contains zero or more feature instances that satisfy the query expressions specified in the request in the format specified on the request.
- **GetFeatureWithLock:** Like the GetFeature operation, except that in a GetFeatureWithLock operation the WFS returns the document with the features and locks the features in the result set. Locks prevent the resources to be updated, thus avoiding corruption of information while a transaction is being executed.
- **LockFeature:** Performs a lock on a set of features.
- **Transaction:** Used to describe data transformation operations to be applied to feature instances. Using the Transaction operation clients can create, modify, replace and delete features in the WFS data store.
- **CreateStoredQuery**: Used to create a stored query. Stored queries are procedures that are stored on the WFS and can be executed to query the data store. A stored query contains metadata that describes it, a list of zero or more arguments that the stored query receives and one or more query expressions that are executed when the stored query is invoked. Stored queries can be created using this operation or be predefined in the WFS.
- **DropStoredQuery**: Allows previously created stored queries to be removed from the system
- **ListStoredQueries**: Retrieves the list of stored queries on the WFS.
- **DescribeStoredQueries**: Provides detailed metadata about each stored query expression that a WFS offers.

This service is mainly a feature access service but can also include elements of a feature type service, coordinate conversion/transformation service, and geographic format conversion service.

Although on the prototype platform developed for this thesis no WFS services were used, WFS was considered on the modularization of the platform and is included on the layer types.

2.2.4 Web Coverage Service – WCS

A Web Coverage Service (WCS) [37] offers multi-dimensional coverage data for access over the Internet. WCS supports electronic retrieval of geospatial data as "coverages." Coverages are digital geospatial data representing space/time varying phenomena, specifically spatio-temporal regular and irregular grids, point clouds, and general meshes. A raster is an example of a regular coverage representing multidimensional data as its bands. The most updated version of the WCS standard is 2.1.

WCS and WMS are very similar and can return the same formats, but WCS allows more formats, raw data, extract specific portions of the data, and more precise queries, while WMS is usually only for displaying images. On the other hand, WCS visualization is complex and is usually not supported natively by web mapping libraries, so in practice it relies on WMS for visualization. The operations that a WCS provides are GetCapabilities, DescribeCoverage, and GetCoverage.

- **GetCapabilities:** Operation which returns in XML format all metadata related to the service. It includes a list of all operations, services, and coverages that are offered by the service. The parameters accepted by the request are shown on the following table.

Request parameter	Description	Mandatory
VERSION=version	Request version.	Yes
SERVICE=WCS	Service type. Must be WCS.	Yes
REQUEST=GetCapabilities	Request name. Must be GetCapabilities.	Yes

Table 4. WCS GetCapabilities request

- **DescribeCoverage:** This operation returns all information about a specific coverage offered by the WCS. The parameters are similar to the GetCapabilities parameters, with the addition of the COVERAGEID parameter, which specifies one or more coverages from which the service will extract the information. The coverage ID is a short, unique indicator for each one of the coverages of the service and it can be found on the GetCapabilities response.

Request parameter	Description	Mandatory
COVERAGEID	One or more coverage Identifiers. Available coverage identifiers are specified in the GetCapabilities response.	Yes

Table 5. WCS DescribeCoverage request

- **GetCoverage:** Operation that returns the data from the requested coverage. It allows the domain subsetting operation, meaning that the user can get only a portion of the coverage inside a specified bounding box and a set of specific parameters.
It requires the same parameters of the GetCapabilities request with the addition of extension, coverageId, format, mediaType, and DimensionSubset, which are further explained on the following table.

Request parameter	Description	Mandatory
EXTENSION	Any ancillary information to be sent from client to server. Optional	No

COVERAGEID	Identifier of coverage evaluated. Must be only one.	Yes
FORMAT	MIME type identifier of the format in which the coverage returned is encoded.	Yes
MEDIATYPE	If present, enforces a multipart encoding. Optional.	No
DIMENSIONSUBSET	Subsetting specifications, one per subsetting dimension.	No

Table 6. WCS GetCoverage request

2.2.5 File format standards

The file format dictates the way in which the data of a file is structured, so specialized programs know exactly how to execute or use the data contained in them. Geospatial data is not the exception and there exist a wide variety of file formats.

OGC contains several standard formats for files which are widely used such as GeoJSON, GeoTIFF, HDF, Geopackage of netCDF [41], while other formats, although not standardized, such as ESRI's shapefile are also widely used within GIS applications.

A very comprehensive and extensive list of GIS formats is compiled on [40], covering file formats for vector data, raster data, multidimensional gridded data, CAD, LiDAR, databases, among others.

The developed platform is mainly based on web technologies and does not heavily rely on static files, but it is important to understand the individualities of some data formats that are mentioned in the thesis and that are used. Those formats are GeoJSON and GeoTIFF [38] [39].

- GeoJSON: this is not a file format itself, but an object notation. GeoJSON is an extension of the JSON (JavaScript Object Notation) standard that is a way to share data across the internet. A GeoJSON object is a JSON object with a certain standard structure that allows it to hold geospatial information. When put in a file, it can be read and rendered by any modern web mapping library. GeoJSON objects contain vector data, such as points, lines, and polygons.

- GeoTIFF: this OGC Standard defines the Geographic Tagged Image File Format (GeoTIFF) by specifying requirements and encoding rules for using the Tagged Image File Format (TIFF) for the exchange of georeferenced or geocoded imagery [42]. It has become the industry standard for sharing image data with a georeference and is widely used in GIS and remote sensing applications.

2.2.6 ArcGIS Server

ESRI is one of the principal vendors for GIS software in the world. ArcGIS Server is an ESRI product used by multiple organizations around the world to publish their georeferenced data following OGC standards in a fast and easy way. ArcGIS server provides a vast set of options and functionalities for hosting and sharing geospatial data and is compliant with OGC standards.

Two specific types of services provided by ArcGIS Server are important for this thesis because some data sources rely on this technology to deliver information: ArcGIS Map Service and ArcGIS Image Service.

- ArcGIS Map Service: provides access to the contents of a map hosted on a server. It is mainly used for visualization, which is provided through a WMS or through the ArcGIS REST API, but it allows querying. It can perform the operations of Export Map, Find, Generate KML, Identify, Generate Renderer, Query, Query Attachments, Query Related Records, Estimate Export Tile Size, and Export Tiles. Using the Identify operation it is possible to extract data from the layer.

- ArcGIS Image Service: provides access to raster data through a web service. This service provides more functionalities than the Map Service since it uses the raster data source directly. It provides visualization of the map using WMS and also provides several raster operation capabilities. Among the most important operations it supports are Add Rasters, Calculate Volume, Compute Class Statistics, Compute Histograms, Compute Multidimensional Info, Compute Pixel Location, Compute Statistics and Histograms, Compute Tie Points, Download Rasters, Export Image, Export Tiles, Get Samples, Identify, Measure, and Query.

Copernicus Land Monitoring Service (see section 3.4.2) provides its data using the Image Service until 2015 and uses Map Service for the layers provided from 2015 on.

2.3 Maritime boundaries

Every country has a set of limits to determine the administrative borders and territory. Modern maritime boundaries and law is a complex and rather recent. Maritime law became a topic of interest around the XV century with European colonization on waters outside Europe where seas were still free and outlawed for all countries (*mare libero*). The first definition of territorial sea was the 3 nautical miles (n.m) boundary, stating that the portion of the sea outside the coastline up to 3 n.m belonged to the country. This concept of territorial sea evolved to the modern 12 n.m territorial sea boundary and, in addition, the further expansion of boundaries with the exclusive economic zone (EEZ), that sets the maritime boundary of 200 n.m from the baseline as the space where countries can perform exploitation activities such as fishing or mining. In 1982 most countries agreed on a comprehensive maritime regime to rule the usage and sovereignty of the sea: the United Nations Convention on the Law of the Sea (UNCLOS) [45].

UNCLOS is an international agreement that establishes the legal framework for all maritime related activities. This agreement is composed of a set of customs, treaties, and international agreements between governments to maintain order, productivity, and peaceful relations on the

sea. One of the concerns of UNCLOS is the definition of boundaries between coastal states, which are set by treaties between countries. However, maritime territorial disputes are not the concern of UNCLOS and are managed directly by the states in bilateral negotiation or, in few cases, by international courts [43].

2.3.1 Definitions

In the UNCLOS definitions there are some important concepts to understand the legal maritime boundaries. The nautical mile measurement unit, the territorial sea baseline, territorial sea, contiguous zone, exclusive economic zone, and continental shelf.

2.3.1.1 Nautical Mile

A nautical mile (n.m) is a unit of distance equal to 1852 meters. This value was adopted by the International Hydrographic Conference in 1929 and has also been adopted by the International Bureau of Weights and Measures [44].

2.3.1.2 Territorial Sea Baseline

It Is the line from which all the outward maritime limits are calculated. It is different from the coastline because it does not correspond to the boundary between land and sea, but to a line that marks the start of the territorial sea with respect to internal waters. There are 3 things to have in mind to calculate the baseline of a coastal state [44]:

- Normal Baseline corresponds to the low water line along the coast, including the coasts of islands. It is drawn around low tide elevations which are defined as naturally formed areas of land surrounded by water at low tide but submerged at high tide, provided they are completely or partly within 12 n.m of the coast, meaning that no baseline can be outside the 12 n.m limit.

- Straight baselines is a system of straight lines that join specific or discrete points on the low-water line, usually known as straight baseline end points. These may be used in localities where the coastline is very sharp, or where there are many small islands along the coast in its immediate vicinity.

- Bay or river closing lines are straight lines drawn between the respective low-water marks of the natural mouths of bays or rivers.

2.3.1.3 Territorial Sea

The Territorial Sea is a belt of water not exceeding 12 n.m in width measured from the baseline. The coastal state has the sovereignty of the seabed and subsoil below, and air, and space above [44].

2.3.1.4 Contiguous Zone

The Contiguous Zone is a belt of water contiguous to the territorial sea of another 12 n.m. In total the contiguous zone is measured as the 24 n.m limit from the territorial sea baseline. In this zone, the coastal state can take measures to prevent and punish infringement of its law [44].

2.3.1.5 Exclusive Economic Zone (EEZ)

The Exclusive Economic Zone or EEZ of a coastal state is the 200 n.m limit with respect to the baseline. On the EEZ, coastal countries have rights over exploration, conservation, and management of all natural resources of the waters of the seabed and its subsoil together with other activities such as the production of energy. Jurisdiction is also extended to the establishment and use of artificial islands, installations and structures, marine scientific research, the protection and preservation of the marine environment, and other rights and duties [44].
In case of conflicts between states on where to delimit the EEZ, the resolution often leads to a equidistant solution of a median line between both countries. Another popular resolution is to assign the zone based on the equitative principle, where the state with more coastal area or dimension gains access to a bigger portion of sea [43].

2.3.1.6 Continental Shelf

The continental shelf of a coastal state comprises the seabed and subsoil of the submarine areas that extend beyond its territorial sea throughout the natural prolongation of its land territory to the outer edge of the continental margin, or to a distance of 200 nautical miles from the territorial sea baseline where the outer edge of the continental margin does not extend up to that distance [45].

Chapter 3: Data exploration

To create a DTE it is necessary to have plenty of data. There are many data sources available, both public and private, from which it is possible to obtain the necessary data to give insights, create models and observe the past. This thesis is centered on using only public or open datasets, especially Copernicus data, but other datasets were considered such as WorldPop for population density. In this chapter, each data source used in this thesis and its corresponding explanation are presented. There is a definition of the boundaries used for both the sea and the land and their reasoning, a section presenting the base maps used, the different data frequencies available and the ones used, an extensive explanation of the Copernicus data sources, in particular the Copernicus Marine Service and the Copernicus Land Monitoring Service, an overview of the WorldPop datasets, and finally a summary of all the categories and variables available through the datasets in the form of a table.

3.1 Defining the coastal limit

To set a scope on where to spatially delimit the retrieval of information from data sources it is necessary to understand where are the official limits that Italy has with respect to the sea and also to understand what portion of land is considered as coastal zone.

Two maritime boundaries were considered for the developed platform: the Italian official sea limits that considers the EEZ and continental shelf boundaries of Italy, and a more coastal-focused boundary that considers only the territorial waters. As for the land boundaries, a 10km in-land buffer from the coastline is considered. This limit corresponds to the limit used by Copernicus for the Coastal Zones (CZ) dataset presented in section 3.4.2.5.

The way in which the maritime limits are added to the developed platform is as static files encoded as GeoJSON objects, containing a single polygon.

3.1.1 Italian official maritime boundaries

Italy holds maritime frontiers with 9 states from which boundaries has been set through a set of series of treaties: Albania (1982), Croatia (1968), France (1989), Greece (1982), Slovenia (1968), Spain (1978), Tunisia (1979), Serbia (1968) and Montenegro (1968).

Each treaty contains not only a legal core and the agreements between states, but also the set of geographical coordinates that define the limits. From these limits it is possible to build georeference objects and create maps using those limits.

Marine Regions is an organization managed by the Flanders Marine Institute that publishes and maintains a geodatabase containing all the maritime limits as georeferenced files. From their platform it is possible to download the baselines, limits of territorial waters, contiguous zones, exclusive economic zones, and all the available limits for any coastal states. Unfortunately, this geodatabase does not provide the limits as a web-based service, but only as static, downloadable files [47].

Figure 7 depicts the maritime boundaries of Italy on a map as downloaded from the geodatabase of Marine Regions, including the baseline, territorial sea
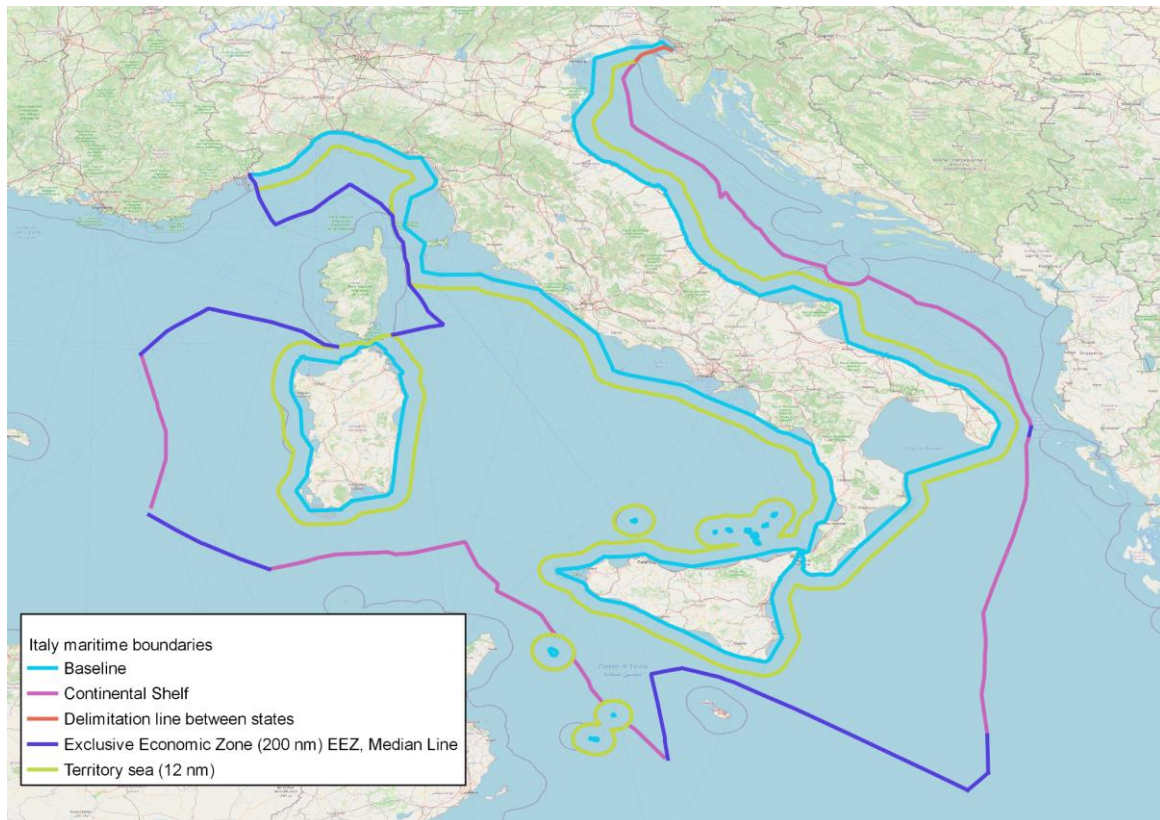
Figure 7. Italian maritime boundaries

From the boundaries, two layers were produced, serving as masks for the data retrieved from data sources. One layer contains a limit of 12 n.m from the coastal baseline and a 10km inland buffer and is visible in *Figure 8*, while the other contains the maximum limit of the Italian sea with the same 10km inland buffer. The second one is visible in *Figure 9*.

Figure 8. Complete Italian maritime limits with a 10km inland buffer

Figure 9. 12 n.m sea limit with a 10km inland buffer

3.2 Base maps

A base map is a graphic representation at a specified scale of selected fundamental map information; It is used as a framework upon which additional data of a specialized nature may be compiled (American Society of Photogrammetry, 1980).

Base maps are a way to enrich a map, and, in a Web GIS, they are very important because they provide the background setting to it, add aesthetic appeal, help orient the user and provide it with a means to understand spatially what any additional data means and put it into a geographical context.

Two base maps were integrated on the platform developed for this thesis: OpenStreetMap as a base map for showing different features such as streets, buildings, administrative boundaries, etc., and the EOX Cloudless Sentinel-2 2020 map for showing satellite imagery.

3.2.1 OpenStreetMap (OSM)

OpenStreetMap is a free, open-source, editable map of the whole world built by volunteers largely from scratch. OSM is made by the community and contains information on roads, trails, buildings, railway stations, and many more all over the world.

OSM is open data and is licensed under the Open Data Commons Open Database License (ODbL) which means that it is possible to copy, distribute, transmit and adapt OSM data, as long as there is a proper attribution of OSM and its contributors, and that if there is any alteration of their data, of anything is built upon it, those changes must be distributed under the same license type as OSM.

3.2.2 Sentinel-2 Cloudless by EOX IT Services

For a base map that displays satellite imagery the chosen map was the Sentinel-2 Cloudless map produced by EOX. This layer is produced by combining trillions of pixels collected during different weather conditions during the whole year from Sentinel-2 satellite images and merging them into a homogeneous mosaic that is almost free from satellite and atmospheric effects.

This base map is available as a WMS and WMTS and is free to use, having a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

They provide several layers through the WMS and WMTS services, including a complete Sentinel-2 cloudless layer for each year from 2016 to 2020. The layer that was chosen as base map is the Sentinel-2 cloudless layer for 2020, which has a spatial resolution of 10 meters in concordance to the Sentinel-2 resolution.

3.3 Data temporal resolution

Across the different layers available in the many datasets considered there are some similarities in the temporal resolution in which they present data. Datasets that represent physical phenomena tend to have mostly monthly and daily frequencies but also, less commonly, hourly and 15-minutes temporal resolutions. For the scope of the thesis only monthly and daily data is considered, since most of the datasets have these frequencies and other, more granular frequencies, are too specific, making data retrieval slow even for short periods of time because of the amount of data to process.

Other datasets are mono temporal or represent a large period of time in a single reference date. E.g., the CORINE Land Cover datasets (see section 3.4.2.1) are produced and distributed on a specific reference year, but every map represents a range of a few years.

Daily average datasets present the values as the calculation of the average of at least two appropriately spaced measurements taken over a period of one day.

Monthly average datasets, like daily average datasets, present the values as a calculation of the average of the measurements of the entire month.

3.4 Copernicus Data Sources

Copernicus is the European Union's Earth Observation program, looking at our planet and its environment for the benefit of humanity. It offers information services that draw from satellite Earth Observation and in-situ (non-space) data.

Copernicus is composed of several services. Those services cover not only the physical systems of the Earth but also offer a gamma of services that help in decision-making and territorial sovereignty.

The services are:

-   **Copernicus Atmosphere Monitoring Service (CAMS)**: provides data and information on the atmospheric composition worldwide, including forecasts. It focuses on five areas: air quality and atmospheric composition, ozone layer and ultraviolet radiation, emissions and surface fluxes, solar radiation, and climate forcing.

-   **Copernicus Marine Environment Monitoring Service (CMEMS):** provides data on the physical and biogeochemical state, variability, and dynamics of the ocean and marine ecosystems for the entire globe, but with higher accuracy for European regional seas.

-   **Copernicus Land Monitoring Service (CLMS):** provides information on land cover and its changes, as well as land use, vegetation state, and Earth's surface energy variables of both Europe and the entire world. It supports applications in urban planning, forest and water management, agriculture and food security, nature conservation and restoration, amongst others.

-   **Copernicus Climate Change Service (C3S):** provides authoritative data about the past, present, and future climate in Europe and the rest of the World. Also, it provides a set of tools to support policy makers and businesses with climate change mitigation and adaptation strategies.

-   **Copernicus service for Security:** aims to support European Union policies by means of providing information in response to Europe's security challenges. It is focused on providing crisis prevention, preparedness, and response in three key areas: border surveillance, maritime surveillance, and support to EU External Action.

-   **Copernicus Emergency Management Service (Copernicus EMS):** provides all actors involved in the management of natural disasters, emergency situations, and humanitarian crises with timely and accurate geospatial information derived from satellite remote sensing and completed by available in situ or open data sources.
    It consists of two main components: a mapping component that provides maps based on satellite imagery, and an early warning component.

In this thesis, it was decided to integrate in the developed platform only the Marine Service and the Land Monitoring Service data, which can provide an accurate representation of the interface between the land and the sea and describe this interaction in the coastal zones of Italy. In the following sections, both services will be detailed in-depth, and the description of the used datasets will be included.

3.4.1 Copernicus Marine Environment Monitoring Service - CMEMS:

Copernicus Marine Service covers the global ocean; however, its focus is on European Oceans. In fact, it offers more localized datasets of regional domains for the Global Arctic Ocean, Antarctic Ocean, Baltic Sea, European North-West Shelf Seas, Iberian-Biscay-Ireland Regional Seas, Mediterranean Sea, and the Black Sea. For this thesis, it was decided to use only datasets belonging to the Mediterranean Sea, since all the Italian Coasts are engulfed on the Mediterranean and having a more localized dataset is positive in terms of accuracy, resolution, and faster processing and querying, due to the smaller size with respect to the entire world.

The filtering of datasets that cover the Mediterranean Sea, spanning between 1992 (oldest available date) to 2022 shows 39 datasets available, called products.

Although there is a high quantity of products available, many products offer the same variables or deliver them with different levels of processing. There are five levels of processing (from L0 to L4) which represent the amount of processing that the data has been subjected to. Each level represents different processing depending on the dataset, but generally, L0 corresponds to unprocessed instrument data, while L4 corresponds to the greatest amount of processing applied, even including data corresponding to computational models and processed measurements from several satellites and several days. It is important to mention that all input data in L4 products is validated. For this thesis, the only datasets considered were the ones corresponding to the L4 processing level.

In total from CMEMS there are a total of 45 layers, that cover the categories of Sea Physics, Ocean Biogeochemistry, and Sea Winds

3.4.1.1 Mediterranean Sea Physics Analysis and Forecast

This product, officially named MEDSEA_ANALYSISFORECAST_PHY_006_013 covers 7 physical variables for the Mediterranean Sea in a monthly, daily, hourly, and 15-minutes frequency. It has a spatial resolution of 0.042° × 0.042° (approximately 3.5 km at the 42N parallel) defined on a regular grid and has 141 vertical levels unevenly spaced from -5500 meters to 0 meters. The geographical coverage spans from 17.29°W to 36.29°E and 30.19°N to 45.98°N, as seen in the example on *Figure 10*. This product is of processing level 4 (L4) and provides data from 1987 to the present.

Figure 10. Geographical coverage for MEDSEA_ANALYSISFORECAST_PHY_006_013

Although this product is provided by Copernicus, it requires the calculation of many variables and processing of several measurements from a wide range of satellites. As stated on the technical documentation of the product, measurements to calculate in-situ vertical profiles of Temperature and Salinity are provided by ARGO and XBT (eXpendable BathyThermograph). To calculate the Sea Level Anomaly (SLA), Jason 2 & 3 (NASA and CNES), Saral-Altika (ISRO), CryoSat (ESA), and Sentinel-3A/3B (ESA) are used. Additionally, objective Analyses-Sea Surface Temperature (SST) fields are used to correct surface heat fluxes. The initial conditions of the model are provided by World Ocean Atlas (WOA) 2013 V2, winter climatology at 01/01/2015. [50].

Data for this product is delivered in several formats, but for the concern of this thesis the interest is on the WMS provided by CMEMS.

The product is organized into 21 datasets, from which 10 were considered based on the frequency of the datasets —daily and monthly. Those 10 datasets correspond to 5 monthly-mean and 5 daily-mean datasets that provide 3D (from top to bottom) information on potential temperature, salinity and currents, and 2D information on sea surface height, bottom potential temperature, and mixed-layer thickness. From those 10 datasets, 16 layers are used.

The layers correspond to 8 different variables on a monthly-mean and daily-mean average and with a spatial resolution of 0.042° × 0.042°:

Physical Variable	Layer name	Units
Eastward ocean current velocity	uo	m/s
Northward Sea Water Velocity	vo	m/s
Sea Water Velocity – Automatically-generated vector field, composed of the northward and eastward components.	sea_water_velocity	m/s
Ocean Mixed Layer Thickness	mlotst	m
Sea Water Potential Temperature	thetao	°C
Sea Water Potential Temperature At Sea Floor	bottomT	°C

Sea Surface Height Above Geoid	zos	m
Sea Water Salinity	so	psu (Practical Salinity Unit)

Table 7. Mediterranean Sea Physics Analysis and Forecast layers

3.4.1.2 Mediterranean Sea Biogeochemistry Analysis and Forecast

This product, named MEDSEA_ANALYSISFORECAST_BGC_006_014 covers 17 biogeochemical variables for the Mediterranean Sea in a monthly and daily mean frequency. It has a spatial resolution of 0.042° × 0.042° (approximately 3.5 km at the 42N parallel) defined on a regular grid of 1005x380 and has 125 vertical levels unevenly spaced from -4153 meters to 0 meters. The geographical coverage spans from 5.54°W to 36.30°E and 30.18°N to 45.98°N, as seen in the example on *Figure 11*. This product is of processing level 4 (L4), and provides data from 1999 to the present.
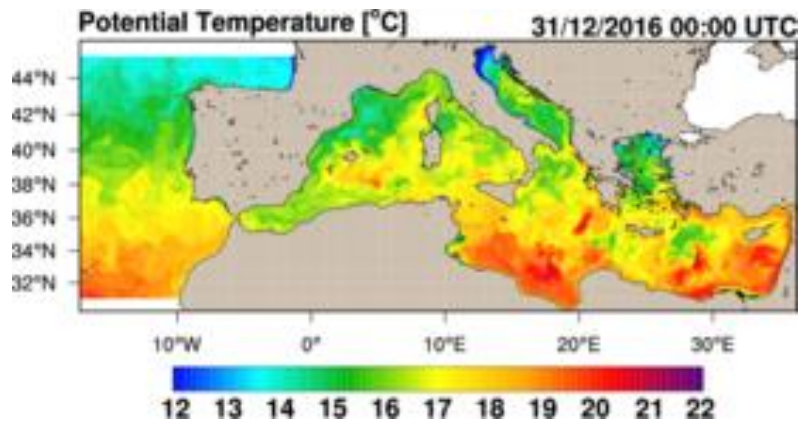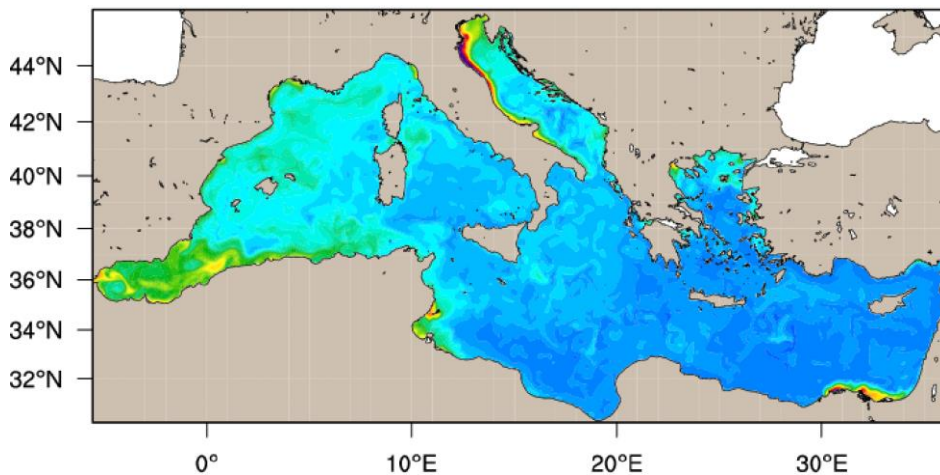

Figure 11. Geographical coverage for MEDSEA_ANALYSISFORECAST_BGC_006_014

As stated on the technical documentation of the product, to calculate its variables this product uses measurements provided by many satellites and instruments. For calculating surface chlorophyll, it uses multi-sensor satellite data provided by CMEMS Ocean Colour Thematic Center — OCTAC, from the MODIS-AQUA, NOAA20-VIIRS, NPP-VIIRS and Sentinel 3A-OLCI sensors. In-situ chlorophyll and nitrate profiles are provided from Biogeochemical Argo floats, which are provided by CORIOLIS data repository. Initial conditions for the model are extracted from the Sub-basin climatological profiles from EMODnet2018_int dataset.

The biogeochemical analysis and forecasts are produced by means of the MedBFM [52] model system, which is the physical-biogeochemical OGSTM-BFM model coupled with a 3DVarBio assimilation scheme.

Data for this product is delivered in several formats, but for the concern of this thesis the interest is on the WMS provided by CMEMS.

This product is organized in 11 datasets; 10 of those datasets contain the values for the biogeochemical variables and were the ones considered. The other dataset contains static information for the other datasets such as coordinates, mask and bathymetry. The 10 datasets provide 3D daily and monthly mean data of the concentration of chlorophyll, nitrate, phosphate, ammonium, silicate, primary production, oxygen, phytoplankton carbon biomass, zooplankton carbon biomass, pH, dissolved inorganic carbon, and alkalinity, and 2D daily and monthly mean data of surface partial pressure of CO2 and surface flux of CO2.

From those 10 datasets 28 layers are used, which correspond to daily and monthly data of 14 biogeochemical variables with a spatial resolution of 0.042° × 0.042°. Some datasets provide information on more than one variable. Information on the used layers is contained in the following table.

Biogeochemical Variable	Layer name	Units
Ammonium - Mole Concentration of Ammonium in Sea Water	nh4	$mmol/m^3$
Nitrate - Mole Concentration of Nitrate in Sea Water	no3	$mmol/m^3$
Phosphate - Mole Concentration of Phosphate in Sea Water	po4	$mmol/m^3$
Silicate - Mole Concentration of Silicate in Sea Water	si	$mmol/m^3$
Mass Concentration of Chlorophyll-A In Sea Water	chl	mg/m^3
Mole Concentration of Dissolved Inorganic Carbon in Sea Water	dissic	mol/m^3
Mole Concentration of Dissolved Molecular Oxygen in Sea Water	o2	$mmol/m^3$
Phytoplankton Carbon Biomass	phyc	$mmol/m^3$
Zooplankton Carbon Biomass	zooc	$mmol/m^3$
Net Primary Production of Biomass	nppv	mg/m^3 day
Sea Water Alkalinity	talk	mol/ m^3
Sea Water pH	ph	Adimensional
Surface Carbon Dioxide Flux	fpco2	kg/m^2 s
Surface Partial Pressure of Carbon Dioxide in Sea Water	spco2	Pa

Table 8. Mediterranean Sea Biogeochemistry Analysis and Forecast layers

3.4.1.3 Global Ocean Wind L4 Reprocessed Monthly Mean Observations

This service identified as WIND_GLO_PHY_CLIMATE_L4_REP_012_003 offers a time series of monthly averaged wind variables calculated over the global oceans. The 8 variables are wind speed, wind zonal component, wind meridional component, wind stress amplitude, wind stress zonal component, wind stress meridional component, root mean square (rms), and sampling length. The estimation of the monthly wind products and their calibration are derived from the scatterometers ASCAT-A and ASCAT-B with a 0.25° x 0.25° spatial resolution, considering at least 25 daily values at each grid point. The ASCAT scatterometers are onboard the METOP-A and METOP-B satellites, respectively. The product is available from April 2007 to December 2020.
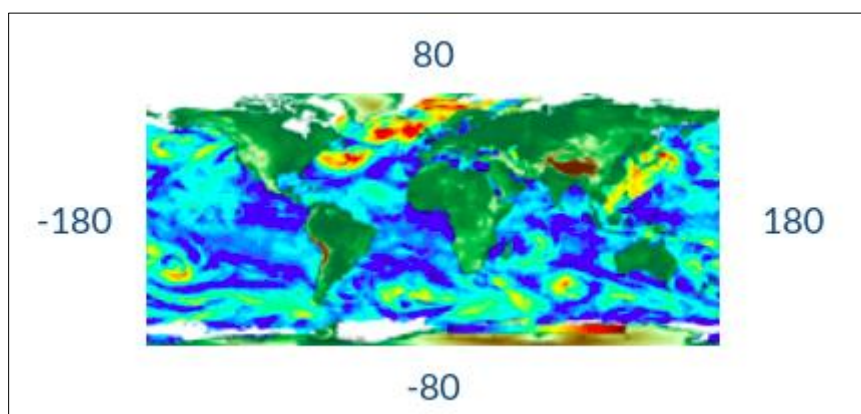
Figure 12. Geographical coverage for WIND_GLO_PHY_CLIMATE_L4_REP_012_003

The layer Wind Speed Root Mean Square (RMS) is the only one used from this product, and its basic information is reported on the following table.

Wind variable	Layer name	Units
Root mean square	wind_speed_rms	m/s

Table 9. Global Ocean Wind L4 Reprocessed Monthly Mean Observations layers

3.4.2 Copernicus Land Monitoring Service:

The Copernicus Land Monitoring Service (CLMS) is divided into four main components: Global, Pan-European, Local, and Imagery and reference data. For this thesis, datasets from the Pan-European and Local components were selected because they all provide a WMS to serve the maps, and have a better spatial resolution due to being focused solely on Europe. In particular, the CORINE Land Cover (CLC), Imperviousness, Forests, and Water and Wetness products were used from the Pan-European component, while the Coastal Zones product, containing the land cover for European coastal zones, was selected from the Local component. A total of 19 layers from CLMS are used.

3.4.2.1 CORINE Land Cover (CLC)

CORINE stands for Coordination of Information on the Environment; the CORINE Land Cover products provide information on the biophysical characteristics of the Earth's surface for Europe. It is produced by most countries by photo interpretation of high-resolution satellite imagery. Additionally, some countries apply semi-automatic solutions, using national in-situ data, satellite image processing, GIS integration and generalization.
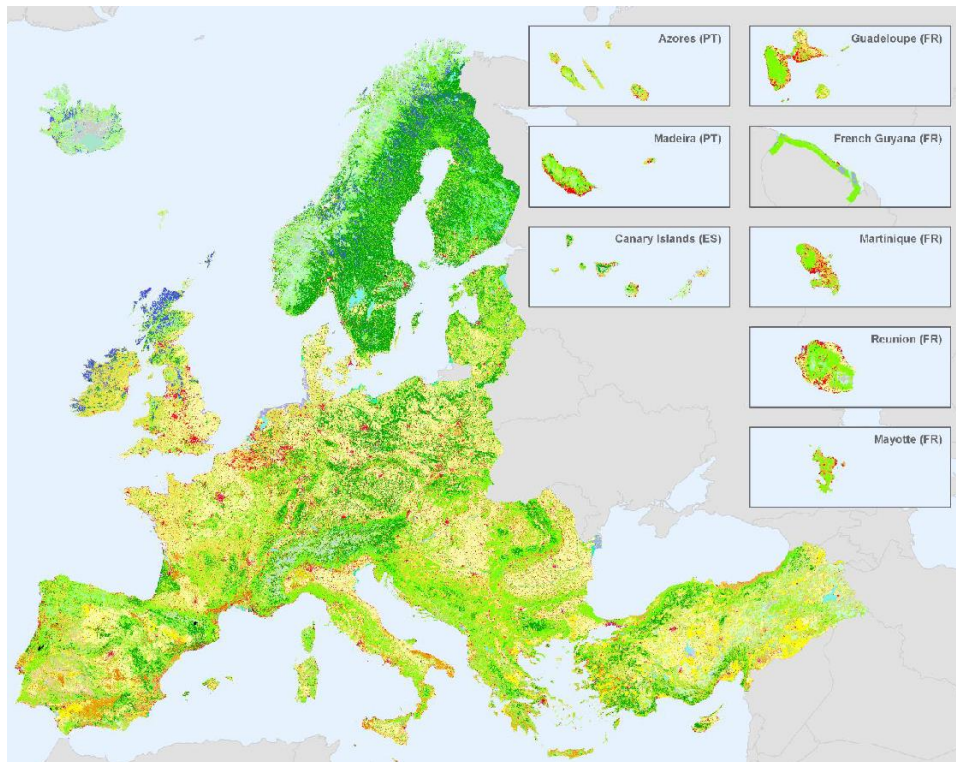
Figure 13. CLC 2018 coverage

The CLC inventory started in 1985 (reference year 1990), and updates have been produced in 2000, 2006, 2012, and 2018. In order to simplify the developed platform, only the layers for the reference years of 2018 and 2012 are considered.

The CLC products in use differentiate the land cover on 44 categories over 5 major groups: artificial surfaces, agricultural areas, forests and semi-natural areas, wetlands, and water bodies. They have free access for all users and are available through a WMS provided as an ArcGIS REST Service MapServer.

Both products contain 2 layers, a vector and a raster layer, which are available for large and small scales, respectively. The CLC itself is calculated as polygons with a minimum mapping unit of 25 hectares (approximately 500m x 500m) and the raster layer is generated from those polygons and has a 100-meter spatial resolution. Both the raster and the vector layer were considered for the thesis for the 2 reference years, meaning that a total of 4 layers are used from the CLC products.

Details on the 2012 and 2018 products are described in the following table:

	CORINE Land Cover 2012	CORINE Land Cover 2018
Thematic Classes	44 Classes	44 Classes
Satellites used to produce it	IRS Resourcesat-1/2, SPOT-4/5, RapidEye constellation	Sentinel-2 and optionally Landsat-8
Min. mapping unit/width	25 ha/ 100m	25 ha/ 100m
Time Consistency	2011-2012	2017, 2018 for gap filling
Geometric accuracy of satellite imagery	≤ 25 m	≤ 10 m
Production Time	2 years	1.5 years

Table 10. CORINE Land Cover products detail

3.4.2.2 Imperviousness

The Imperviousness products from CLMS capture the percentage and change of soil sealing. Sealed or impervious areas are the ones that have seen the substitution of the original natural or semi-natural land cover or water surface with an artificial, usually impervious cover. The artificial surfaces are usually maintained over long periods of time. The imperviousness High Resolution Layers (HRL) represent the spatial distribution of artificially sealed areas, that includes the level of sealing of the soil per unit of area. The level of sealed soil, or imperviousness degree between 1% and 100%, is produced using semi-automated classification, based on calibrated NDVI [56].

The product presents 2 status layers, the Imperviousness Density (IMD) layer that maps the percentage of sealed area, and the Impervious Built-up (IBU) layer that shows built-up areas, which are the part of the sealed surfaces where buildings can be found. It is worth mentioning that built-up areas are a subgroup of the sealed areas, and they refer to the areas where above-ground constructions and buildings can be found.

This product has the IMD layer available for the reference years of 2006, 2009, 2012, 2015, and 2018, and the IBU layer for 2018. The six available layers are considered for this thesis. All layers are considered High Resolution Layers (HRL).

Although they represent the same variable, there are differences between the different layers, especially on spatial resolution, as shown on the following table extracted from [54].

Name of product	Available reference years or periods	Spatial Resolution (Pixel size)
IMD – Imperviousness Density	2006, 2009, 2012, 2015, 2018	10m (2018) 20m (2006, 2009, 2012, 2015)
IBU - Impervious Built-up	2018	10m

Table 11. Imperviousness layers details

All the layers are available through a WMS provided by Copernicus using an ArcGIS MapServer as images and are queryable using the ArcGIS REST API.

The 2006, 2009, and 2012 imperviousness layers were calculated using re-processing and re-analysis of historical layers and the existing IMD 2006, 2009, and 2012 layers. For the 2015 layer the calculation was performed using high-resolution satellite imagery from IRS-P6/Resourcesat-2 LISS-III (ISRO – India), SPOT 5 (CNES – France) and Landsat 8 (NASA – USA). The final product uses supervised classification of built-up/non-built-up areas with posterior visual improvement of classification results, and derivation of degree of imperviousness based on continuous multi-temporal image composites (3 per year). [55].

The 2018 imperviousness Density layer was produced using high resolution optical (Sentinel-2) and SAR (Sentinel-1) satellite images. It uses a similar methodology of the 2015 layer of supervised classification plus visual improvement of classification and derivation of degree of imperviousness using image composites. The Impervious Built-up 2018 dataset was produced based on optical Sentinel-2 satellite images and using the MASADA Toolbox (Massive Spatial Automatic Data Analytics), developed by the Joint Research Centre (JRC) for the derivation of the European Settlement Map of 2012 [56].

Figure 14. Imperviousness Density 2018 layer over Italy

3.4.2.3 Forests

CLMS pan-European products include a series of High-Resolution Layers for the status of forests. There are 3 products, corresponding to the reference years of 2012, 2015, and 2018, that containing information on the variables of Tree Cover Density – TCD, which is the level of tree cover density in a range from 0% to 100%, Dominant Leaf Type – DLT, pointing the broadleaved or coniferous leaf type majority on a pixel, and Forest Type – FTY, which tells if a pixel belongs to a coniferous forest or a broadleaved forest. [57]
For this thesis only the forest Tree Cover Density layers were used.

Copernicus defines Tree Cover Density as the "vertical projection of tree crowns to a horizontal earth's surface", that means, the proportional crown coverage per pixel.

In total 3 layers are used from the Forests products, which are the TCD status layers of 2012, 2015, and 2018. All of them are provided through a WMS that uses ArcGIS REST Services MapServer. The layers tell from 0 to 100 the forest cover density.

The 2012 and 2015 layers have a spatial resolution (pixel size) of 20 meters and were derived from multispectral High-Resolution satellite imagery using Very High Resolution (VHR) satellite

imagery and/or aerial ortho-imagery as reference data. Sentinel 2-A and Landsat 8 are the main input data source for the 2015 products. [58]

The 2018 layer has a 10-meter spatial resolution and was derived using multi-temporal imagery of the Sentinel-2 satellite.
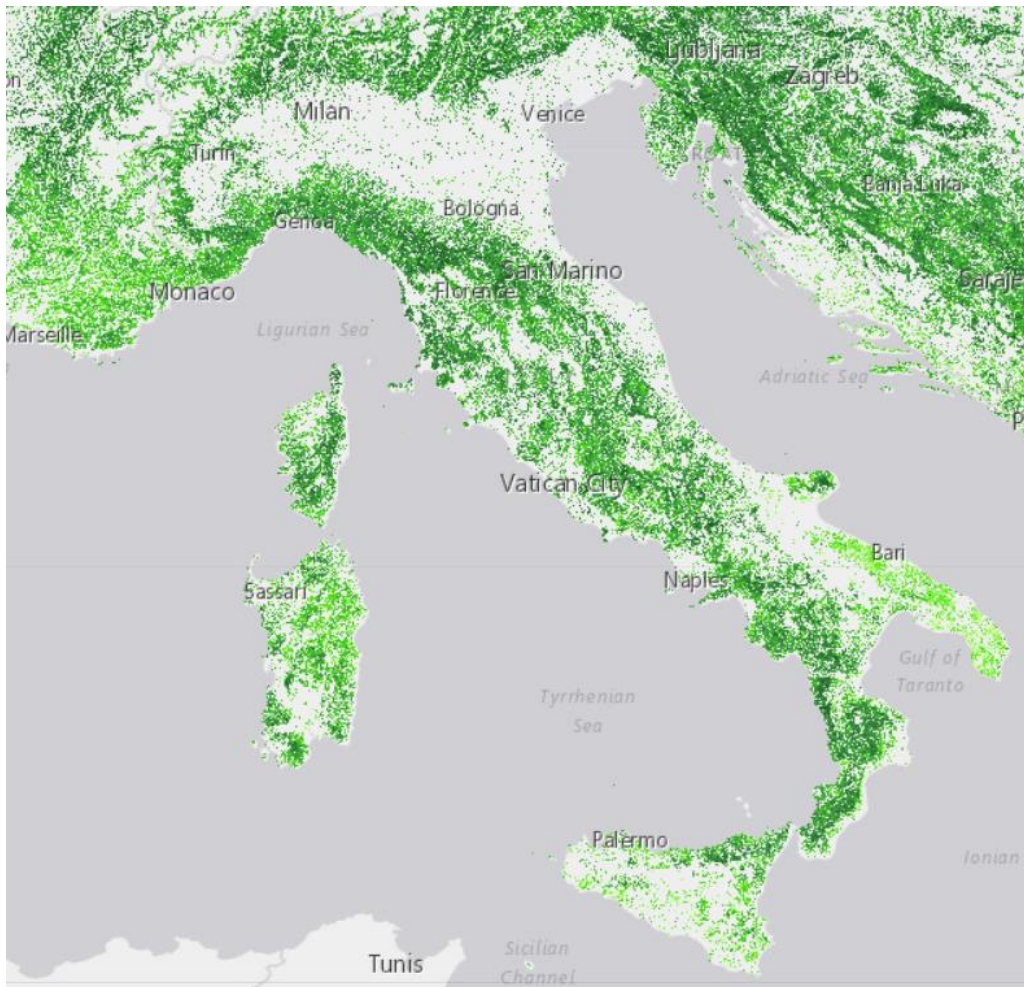
Figure 15. Tree Cover Density 2018 layer over Italy

3.4.2.4 Water and Wetness

The Water and Wetness (WAW) products belong to the pan-European High-Resolution layers. The combined Water and Wetness product is a thematic product showing the occurrence of water and wet surfaces over the reference years of from 2009 to 2018, on the reference years of 2015 to 2018. These layers are based on multi-temporal and multi-seasonal optical high-resolution satellite imagery. [60]

Two layers are provided in the WAW product corresponding to the reference years mentioned before. The 2015 layer shows the occurrence of water and wet surfaces over the period from 2009 to 2015 on 5 categories: no water, permanent water (e.g. rivers, lakes), temporary water (e.g. temporarily inundated areas), permanently wet areas (e.g. areas of permanently high soil moisture), and temporary wet areas (e.g. areas of changing soil moisture). The spatial resolution

of this layer is 20 meters and is produced using Landsat seasonal composites from Landsat 5, 6, 7 and 8, Sentinel-1A, Copernicus DWH HR_IMAGE_2012 datasets, and ENVISAT-ASAR Wide Swath [61].

The 2018 layer is based on imagery from 2012 to 2018 and, in addition to the 5 categories previously mentioned, the category Sea Water was added. This layer has a spatial resolution of 10 meters and was produced using Landsat 5, 7, and 8 for the period from 2012 to 2015 and subsequently on the fully pre-processed Sentinel-2 and Sentinel-1 time series from 2016 to 2018 [62].

The reason why the WAW layers are calculated using such time intervals is to capture the intra-annual dynamics as much as possible, using one image composite per season, each season covered by 3 months, and a composite for each year during the observation period.
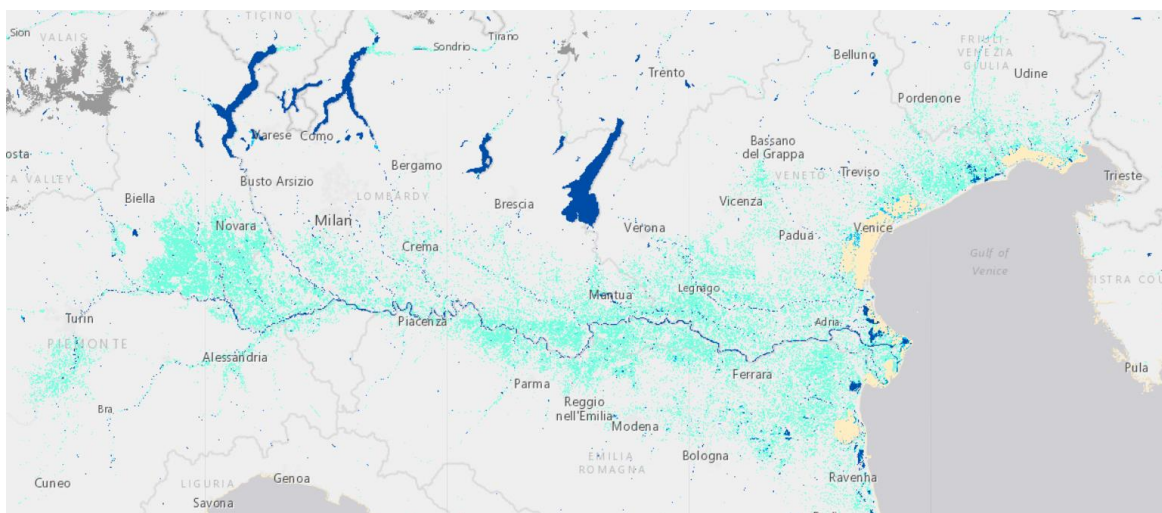
Figure 16. Water and Wetness dataset on the north of Italy and the Po River basin

3.4.2.5 Coastal Zones

The Coastal Zones Land Cover/Land Use (CZ LC/LU) product offer detailed Land Cover and Land Use information for areas along the marine coastline of the European Economic Area (EEA-39) countries. The area of interest of the CZ layer is a 10 km inland buffer zone and a seawards buffer of 12 nautical miles along the coastline of the CZ mapping. It also includes other relevant areas that might be outside of the boundaries such as estuaries, coastal lowlands, and nature reserves. The total covered area on land along all European coastline is approximately 715.000 km^2.

This product has 2 status layers, which are the LC/LU status maps for the reference years of 2012 and 2018. CZ layers differentiates 71 thematic classes divided in 8 categories: Urban, Cropland, Woodland and forest, Grassland, Heathland and scrub, Open spaces with little or no vegetation, Wetland, and Water.

Similarly to the CLC products, the CZ products come in 2 layers, a raster layer and a vector layer that are visible for large and small scales, respectively. A large set of satellite imagery, very high resolution (VHR) images, and additional products are used on the production of the CZ layers,

which is performed using semi-automatic LC/LU classification of VHR satellite data and computer assisted visual refinement. Additionally, visual interpretation of LC/LU classes is performed for very specific classification [63].

Additional details on the CZ products are shown on the following table:

Minimum Mapping Unit	0.5 hectares
Minimum Mapping Length	10 meters
Temporal Reference	Reference year 2012: 2010 – 2014 Reference year 2018: 2017 – 2019
Nomenclature	71 thematic classes

Table 12. Coastal Zones layers details

Figure 17. Coastal Zones 2018 layer over Italy

48

3.5 Population Grid

According to Eurostat [64] a population grid is a lattice usually composed of square grid cells overlaying a particular territory, for which information is collected relating to the number of inhabitants. Population grids are a powerful tool to describe the spatial distribution of population and can be used to analyze the relationship between human activities and the environment. In general, the population grid is stable over time and is not dependent on changes in administrative boundaries.

There are many population grids available with different modeled levels. Highly modeled datasets include WorldPop, World Population Estimate (WPE), History Database of the Global Environment (HYDE), High Resolution Settlement Layer (HRSL), and European GHS Population Grid (GHS-POP-EUROSTAT).

From those datasets, WorldPop was selected to be used as the population grid for this thesis, because it provides a good resolution (100 meters) and is available for many reference years (2000 to 2020). Additionally, it is available as a WMS layer, making integration easier. Although other datasets may offer a better accuracy, they offered only downloadable products and offered only one or very few reference years.

3.5.1 WorldPop

WorldPop offers open spatial demographic data and research, using transparent, fully documented, and peer-reviewed methods to produce easily updatable maps and their corresponding uncertainties. WorldPop is produced by the University of Southampton, and, on their website, there are more than 44.000 datasets available.

The dataset that will be used in this study is the WorldPop gridded population counts, which includes in their WMS, layers from the 2000 to the 2020. The layers from 2015 to 2020 were the ones considered and integrated. This dataset is provided as a single WCS and as WMS which contains the one layer for each year separately.

WorldPop population counts dataset has a spatial resolution of 3 arcseconds, which correspond to approximately 70 meters at the latitude of Italy. The values on each pixel of the layer indicate the estimated number of people that are in that pixel, with country totals adjusted to match the corresponding official population estimates for each country. [65]

The way in which the dataset is produced is using a Random Forest-based dasymetric redistribution with remotely sensed ancillary data, explained in [66]. The datasets for WorldPop are produced on a country level, that means that there is a dataset for each country and for each year. There are 4890 population counts datasets in total, but the WMS service provides a singular dataset with all of them already integrated for the entire world available as a single layer for a reference year.
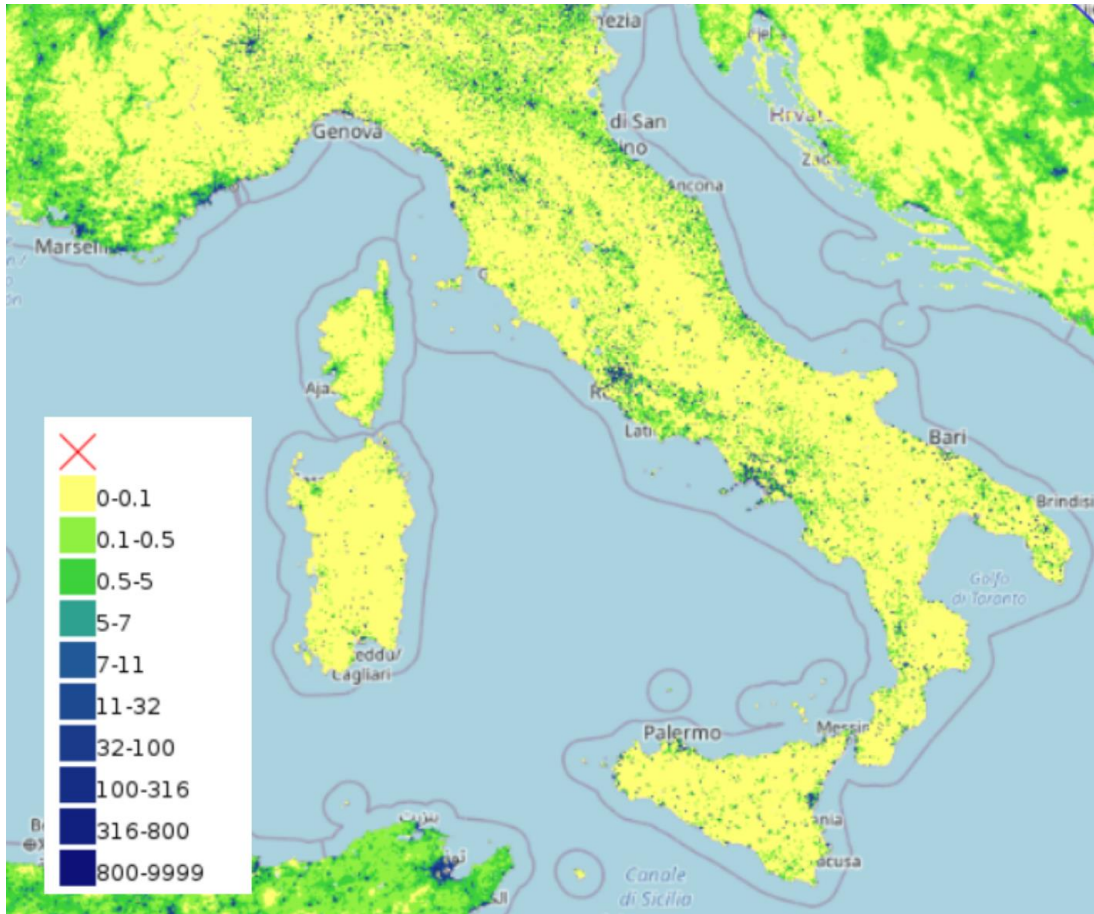
Figure 18. WorldPop 2020 layer over Italy

3.6 Summary

Table 13 presents the summary of all the layers that are included in the prototype of the thesis. The table includes 44 rows that display the most relevant information of the 68 layers integrated on the platform

Layer Name	Time Period	Spatial Resolution	Temporal Resolution	Measurement Unit	Service Typology	Source	Category	Source URL
Eastward Sea Water Velocity	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	m/s	WMS	CMEMS	Sea Physics	[Link]
Northward Sea Water Velocity	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	m/s	WMS	CMEMS	Sea Physics	[Link]
Sea Water Velocity (Vector field)	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	m/s	WMS	CMEMS	Sea Physics	[Link]
Ocean Mixed Layer Thicknes	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	meters	WMS	CMEMS	Sea Physics	[Link]
Sea Water Potential Temperature	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	°C	WMS	CMEMS	Sea Physics	[Link]
Sea Water Potential Temperature At Sea Floor	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	°C	WMS	CMEMS	Sea Physics	[Link]
Sea Surface Height Above Geoid	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	meters	WMS	CMEMS	Sea Physics	[Link]
Sea Water Salinity	16-01-1987 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	psu (Practical Salinity Unit)	WMS	CMEMS	Sea Physics	[Link]

Ammonium - Mole Concentration in Sea Water	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mmol/m3	WMS	CMEMS	Sea Biogeochemistry	Link
Nitrate - Mole Concentration in Sea Water	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mmol/m3	WMS	CMEMS	Sea Biogeochemistry	Link
Phosphate - Mole Concentration in Sea Water	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mmol/m3	WMS	CMEMS	Sea Biogeochemistry	Link
Silicate - Mole Concentration in Sea Water	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mmol/m3	WMS	CMEMS	Sea Biogeochemistry	Link
Mass Concentration of Chlorophyll-A	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mg/m3	WMS	CMEMS	Sea Biogeochemistry	Link
Mole Concentration of Dissolved Inorganic Carbon	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mol/m^3	WMS	CMEMS	Sea Biogeochemistry	Link
Mole Concentration of Dissolved Molecular Oxygen	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	$mmol/m^3$	WMS	CMEMS	Sea Biogeochemistry	Link
Phytoplankton Carbon Biomass	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	$mmol/m^3$	WMS	CMEMS	Sea Biogeochemistry	Link

Zooplankton Carbon Biomass	1999-01-01 to present	1/24 (Approx. 3.5km)	- Monthly - Daily	mmol/m^3	WMS	CMEMS	Sea Biogeochemistry	[Link](#)
Net Primary Production of Biomass	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mg/m^3 per day	WMS	CMEMS	Sea Biogeochemistry	[Link](#)
Sea Water Alkalinity	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	mol/ m^3	WMS	CMEMS	Sea Biogeochemistry	[Link](#)
Sea Water pH	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	Adimensional	WMS	CMEMS	Sea Biogeochemistry	[Link](#)
Surface Carbon Dioxide Flux	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	kg/m^2 s	WMS	CMEMS	Sea Biogeochemistry	[Link](#)
Surface Partial Pressure of Carbon Dioxide	1999-01-01 to present	1/24 ° (Approx. 3.5km)	- Monthly - Daily	Pa	WMS	CMEMS	Sea Biogeochemistry	[Link](#)
Wind Speed RMS (Root Mean Square)	2007-05-16 to 2020-12-31	1/4° (Approx. 20 km)	- Monthly - Daily	m/s	WMS	CMEMS	Sea Wind	[Link](#)
CORINE Land Cover 2012	01-01-2010 to 31-12-2014	100m	Monotemporal	44 thematic classes	ArcGIS MapServer	CLMS	Land Cover	[Link](#)
CORINE Land Cover 2018	01-01-2017 to present	100m	Monotemporal	44 thematic classes	ArcGIS MapServer	CLMS	Land Cover	[Link](#)
Imperviousness Density (IMD) 2006	01-01-2006 to 31-12-2008	20m	Monotemporal	Imperviousness value percentage	ArcGIS MapServer	CLMS	Imperviousness	[Link](#)

Imperviousness Density (IMD) 2009	01-01-2009 to 31-12-2011	20m	Monotemporal	Imperviousness value percentage	ArcGIS MapServer	CLMS	Imperviousness	Link
Imperviousness Density (IMD) 2012	01-01-2012 to 31-12-2014	20m	Monotemporal	Imperviousness value percentage	ArcGIS MapServer	CLMS	Imperviousness	Link
Imperviousness Density (IMD) 2015	01-01-2015 to 31-12-2017	20m	Monotemporal	Imperviousness value percentage	ArcGIS MapServer	CLMS	Imperviousness	Link
Imperviousness Density (IMD) 2018	01-01-2018 to present	10m	Monotemporal	Imperviousness value percentage	ArcGIS ImageServer	CLMS	Imperviousness	Link
Impervious Built-up 2018	01-01-2018 to present	10m	Monotemporal	Binary Built/Non-built value	ArcGIS ImageServer	CLMS	Imperviousness	Link
Tree Cover Density (TCD) 2012	01-01-2012 to 31-12-2014	20m	Monotemporal	Tree cover percentage	ArcGIS MapServer	CLMS	Forests	Link
Tree Cover Density (TCD) 2015	01-01-2015 to 31-12-2017	20m	Monotemporal	Tree cover percentage	ArcGIS MapServer	CLMS	Forests	Link
Tree Cover Density (TCD) 2018	01-01-2018 to present	10m	Monotemporal	Tree cover percentage	ArcGIS ImageServer	CLMS	Forests	Link
Water & Wetness (WAW) 2015	01-01-2015 to 31-12-2017	20m	Monotemporal	5 thematic classes	ArcGIS MapServer	CLMS	Water and Wetness	Link
Water & Wetness (WAW) 2018	01-01-2018 to present	10m	Monotemporal	6 thematic classes	ArcGIS ImageServer	CLMS	Water and Wetness	Link

Coastal Zones 2012	01-01-2010 to 31-12-2017	10m	Monotemporal	71 thematic classes	ArcGIS MapServer	CLMS	Land Cover	[Link](#)
Coastal Zones 2018	01-01-2017 to present	10m	Monotemporal	71 thematic classes	ArcGIS MapServer	CLMS	Land Cover	[Link](#)
WorldPop People Count 2015	01-01-2015 to 31-12-2015	3 arcseconds (approx. 70m)	Monotemporal	People per pixel	WMS, WCS	WorldPop	Demographics	[Link](#)
WorldPop People Count 2016	01-01-2016 to 31-12-2016	3 arcseconds (approx. 70m)	Monotemporal	People per pixel	WMS, WCS	WorldPop	Demographics	[Link](#)
WorldPop People Count 2017	01-01-2017 to 31-12-2017	3 arcseconds (approx. 70m)	Monotemporal	People per pixel	WMS, WCS	WorldPop	Demographics	[Link](#)
WorldPop People Count 2018	01-01-2018 to 31-12-2018	3 arcseconds (approx. 70m)	Monotemporal	People per pixel	WMS, WCS	WorldPop	Demographics	[Link](#)
WorldPop People Count 2019	01-01-2019 to 31-12-2019	3 arcseconds (approx. 70m)	Monotemporal	People per pixel	WMS, WCS	WorldPop	Demographics	[Link](#)
WorldPop People Count 2020	01-01-2020 to 31-12-2020	3 arcseconds (approx. 70m)	Monotemporal	People per pixel	WMS, WCS	WorldPop	Demographics	[Link](#)

Table 13. Integrated layers summary

Chapter 4: Software description

This chapter includes the technical description of the developed platform that is presented in this thesis. On it, the purpose of the platform is stated, and its scope delimited, an explanation and presentation of the architectural design of the platform is provided with the general architecture of the platform and the specific mediator-wrapper implementation. The used technologies are presented with a short discussion on other alternatives and the reason behind the selection of that specific technology. Finally, a description of the most relevant features that were implemented is included alongside with the intended workflow of the application.

4.1 Purpose

As stated before, doing a complete digital twin earth is a long and complex task, and the European Commission is taking an incremental approach with the Destination Earth project and the creation of Digital Twin precursors as smaller parts of a complete digital twin.

The platform developed in this thesis is a prototype of a Digital Twin of the Italian coasts as a proof of concept, focusing on performing data integration of different data sources from different services (which is currently not done by other web platforms), modularization to easily expand the data sources in use and to replicate it in other parts of the world, and the use of data analysis to add value to the already available data on the internet.

Coastal areas are of great importance for countries; economically coasts are fundamental for commerce and tourism, and ecologically coastal areas hold a great portion of biodiversity and ecosystems. Understanding the role of the coasts and its processes is of great importance for governments and for science. Hence, a Digital Twin of the coasts is a considerable effort that is worth doing.

4.2 Scope

The platform developed for this thesis is a proof of concept of a data integration scheme for geospatial data provided by web services for a Digital Twin Earth. It is important to highlight what are the boundaries of the application, and to specify what functionalities and processes are in the scope of the project.

There are four fundamental parts that this platform has: the physical extent, the workflow, the data available, and the data analysis. In these four areas it is necessary to provide a scope to highlight the total reach of the platform developed for this thesis.

- **Physical extent:**
  The extent of the platform focuses on coastal areas, in particular Italian coasts, and on the use of land and marine geospatial data to understand and describe land and sea processes, the interaction between them, and the human impact on and from those processes. For this, two limits have been established for the portion of sea to analyze and a single limit for the land portion. One sea limit is bounded to the territorial sea –12 n.m.– while the

other includes the complete extent of the Italian sea –up to the exclusive economic zone. The land limit is composed of a buffer of 10km inland from the coastal baseline.

- **Workflow:**
  The workflow designed for the platform consists of five steps: context definition, layer selection, layer exploration, data analysis, and sharing.
  Section 4.6 will explain the workflow in depth.

- **Data available:**
  The data to be integrated and to be used correspond to the layers specified in Chapter 3. All the layers summarized on the summary table are the ones that got integrated for visualization and querying on the platform. They correspond to three data sources (CMEMS, CLMS, WorldPop) with a total of 68 layers. The integrated layers are, in practice, more than the ones specified in the summary table because layers that are both daily and monthly are, in fact, two different layers while in the table are specified as one.

- **Data analysis methods:**
  The data analysis for the developed platform is currently limited to point and area analysis. Point analysis methods include time series, depth profile, and value extraction for a maximum of five layers in a single analysis. Area analysis includes summary statistics (minimum, maximum, median, average, and standard deviation), area sampling of low (10x10 sampling space) and high (20x20 sampling space) resolution and histograms of the samples. For layers that allow it, area analysis is performed directly on the raw data instead of on sampled data.

The platform focuses on implementing a virtual data integration strategy for multiple data sources that allows visualization and querying, without duplicating data by consuming it directly from online data sources.

Sharing refers to taking a snapshot of the complete state of the platform at a specific moment by means of a file or a database record. This state consists of the context, the selected layers, the visualization parameters of each layer, and the analysis performed with its parameters. The sharing feature is designed to allow reproducibility and to build scientific collaboration.

4.3 Architectural Design

4.3.1 General architecture

The platform developed for this platform is a Client-Server application. The client (frontend) of the platform is an Angular application, using OpenLayers as its mapping library, that is in charge of receiving user input and displaying all the required geospatial information. The frontend also contains the mediator, which is a fundamental component of the mediator-wrapper architecture used for data integration.

The server application (backend) is a GeoDjango web application with python as the programming language. It performs all the heavy processing of geospatial information and communicating with data sources and the database. It contains the wrappers of the mediator-wrapper architecture which are in charge of communicating with external data sources, parsing the responses of the data sources, and delivering data to the mediator in a generic way. This interaction with data sources is a complex and computationally expensive task, especially because responses are in XML format and communication with data sources may take time.

The database is a PostgreSQL database with the PostGIS extension for managing geospatial data. It is connected to the backend, and it is located on the server component of the application. The database contains information about all the integrated layers, users' credentials and data, and the stored scenarios.

Figure 19 displays the general architecture of the platform showing the two main components of client and server, its internal components, and the connection between them.
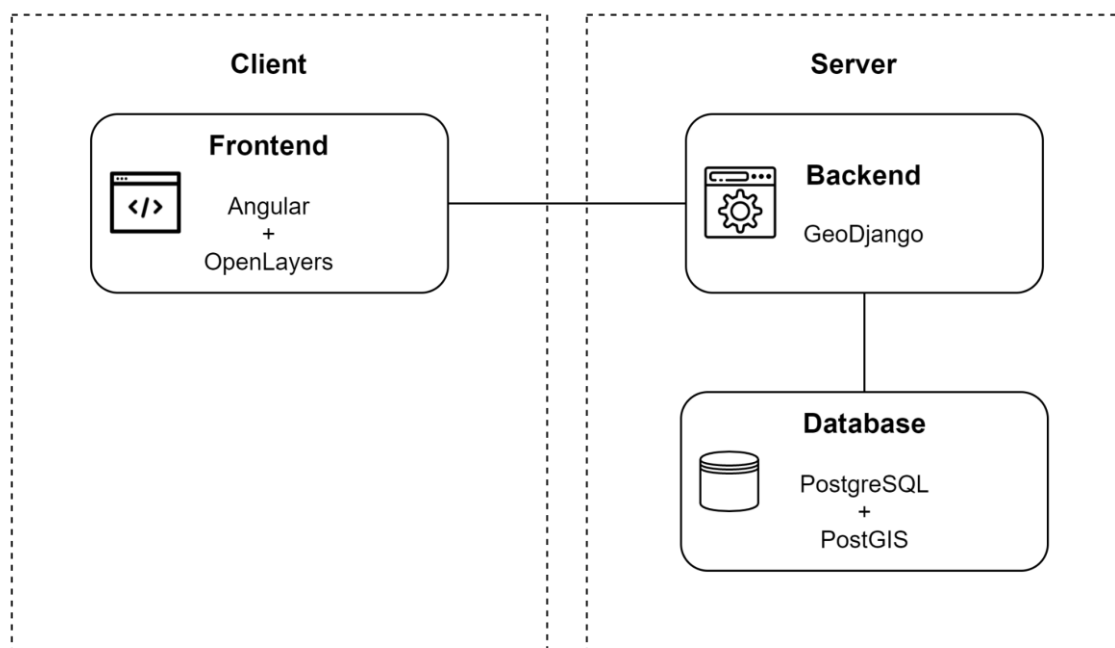

Figure 19. General architecture

4.3.2 Mediator-Wrapper architecture implementation

The data integration strategy for the developed platform of this thesis is based on the mediator-wrapper architecture. This architecture, also called mediator-based, is a very well-known and used architecture for virtual data integration and geospatial data.

The architecture consists of a mediator component, which is exposed to the client and is the one that receives generic requests for information from the client. Internally, the mediator knows how to deal with those requests, eliminating the burden of building specific requests to the user. When the mediator receives the generic request, it interprets it and sends it to one of the available wrappers, which manage the communication with the data source.

A wrapper is a modular component that contains the logic to translate from a generic request to a specific request to a particular data source. It is modular because multiple individual wrappers may be connected to the mediator and are interchangeable as long as they provide the interfaces needed by the mediator, meaning that in order to expand the data integration capabilities of the architecture it is sufficient to add new wrappers that connect with new data sources instead of doing major modifications to the mediator each time a new data source is integrated.

Once the wrapper receives the generic request and translates it to a request for a specific data source, it communicates with the data source and receives the desired information. Internally, the wrapper knows how to process the incoming data and translate it back as a generic response that is then passed to the mediator. Finally, the mediator receives the generic response from the wrapper and delivers it to the client.

Figure 20 displaying the relation of the two macro-components of the application with respect to the data integration architecture.
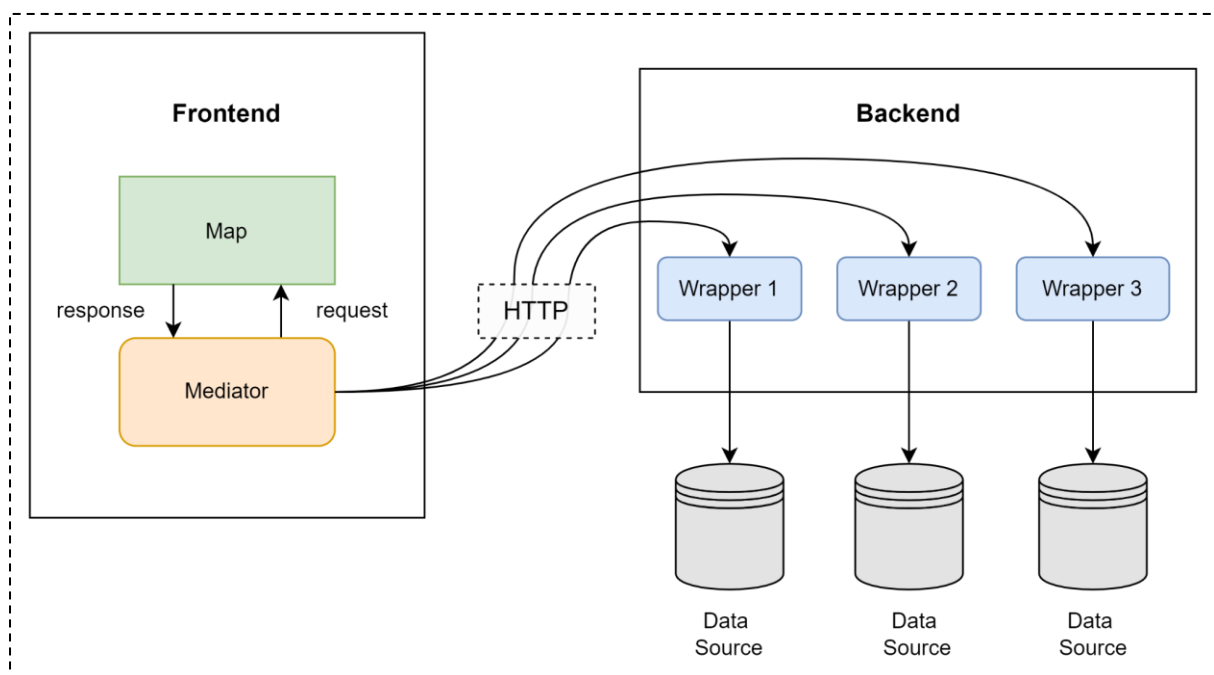

Figure 20. Mediator-Wrapper architecture

In particular, the implementation of the mediator-wrapper architecture in the platform uses both the backend and the frontend to perform the communication with the data source. On the platform, the mediator is located on the frontend and the wrappers on the backend. The use of the frontend on the implementation of the mediator is to leverage the power of OpenLayers for building the parametrized GetFeatureInfo URLs for WMS layers.

The implementation adds an additional layer for the mediator component, called drivers. The drivers handle the selection and communication with the specific wrapper. This communication between drivers and wrappers is done through HTTP, since drivers are located on the frontend with the mediator and wrappers on the backend. Another reason for having the wrappers on the backend is to bypass CORS (Cross-Origin Resource Sharing), since servers usually block requests

that come from browsers and do not come from the same host for security reasons, but requests coming from a non-client web application are usually allowed.

Wrappers are located on the backend and are accessible through HTTP requests. The drivers perform a request to the specific wrapper by using a particular URL and receive the response through HTTP. On the backend, the wrappers perform requests to the data sources also through HTTP and parse the response accordingly.

Figure 21 shows the components of the actual implementation of the mediator-wrapper architecture on the platform, and the connections between them.
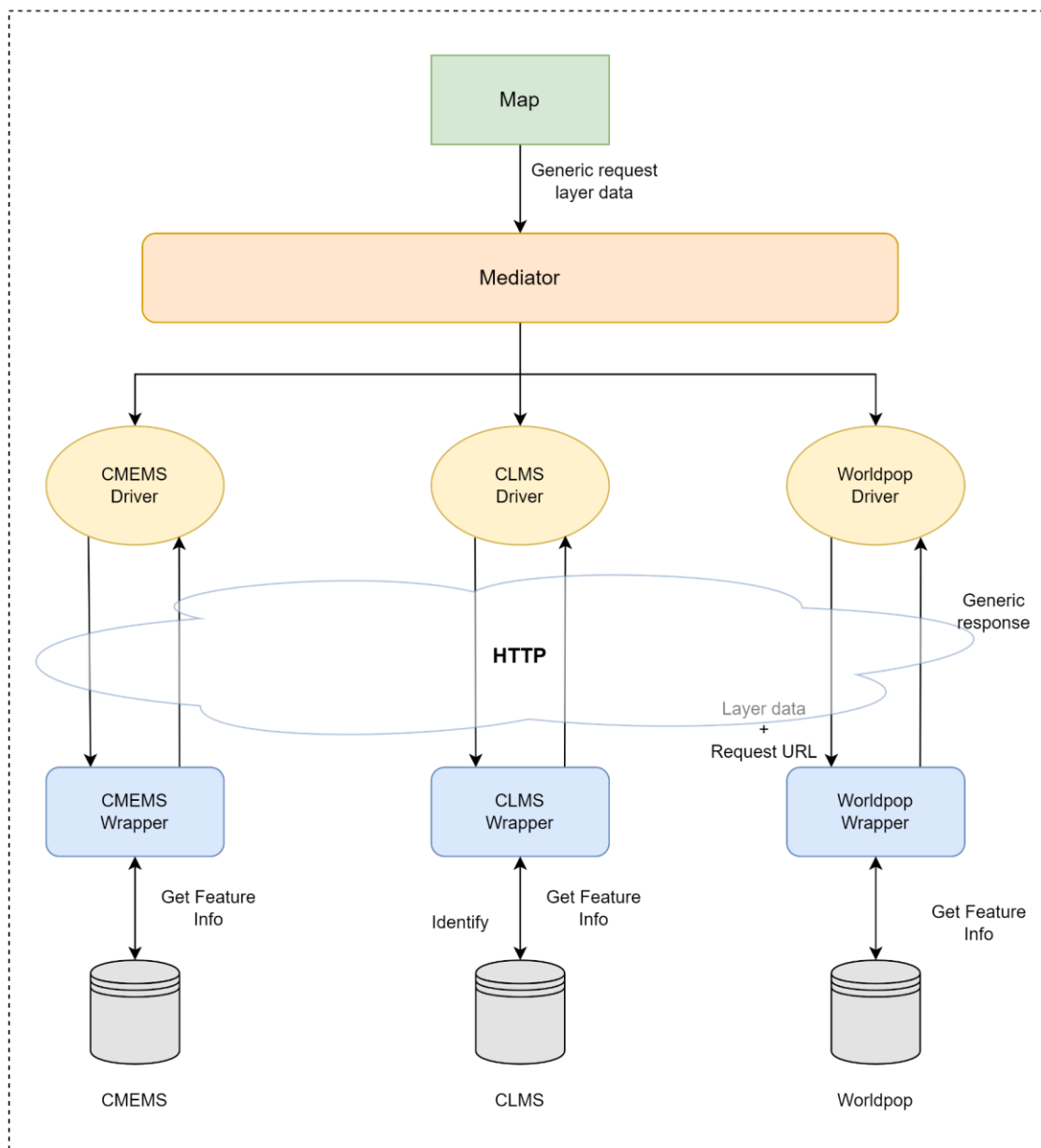

Figure 21. Mediator-Wrapper implementation components

4.3.3 Data Model

The data model is the way in which the tables are arranged in the database, what information its being stored, and what connections are made between the tables. Data is stored on the database on four tables, shown on *Figure 22*.
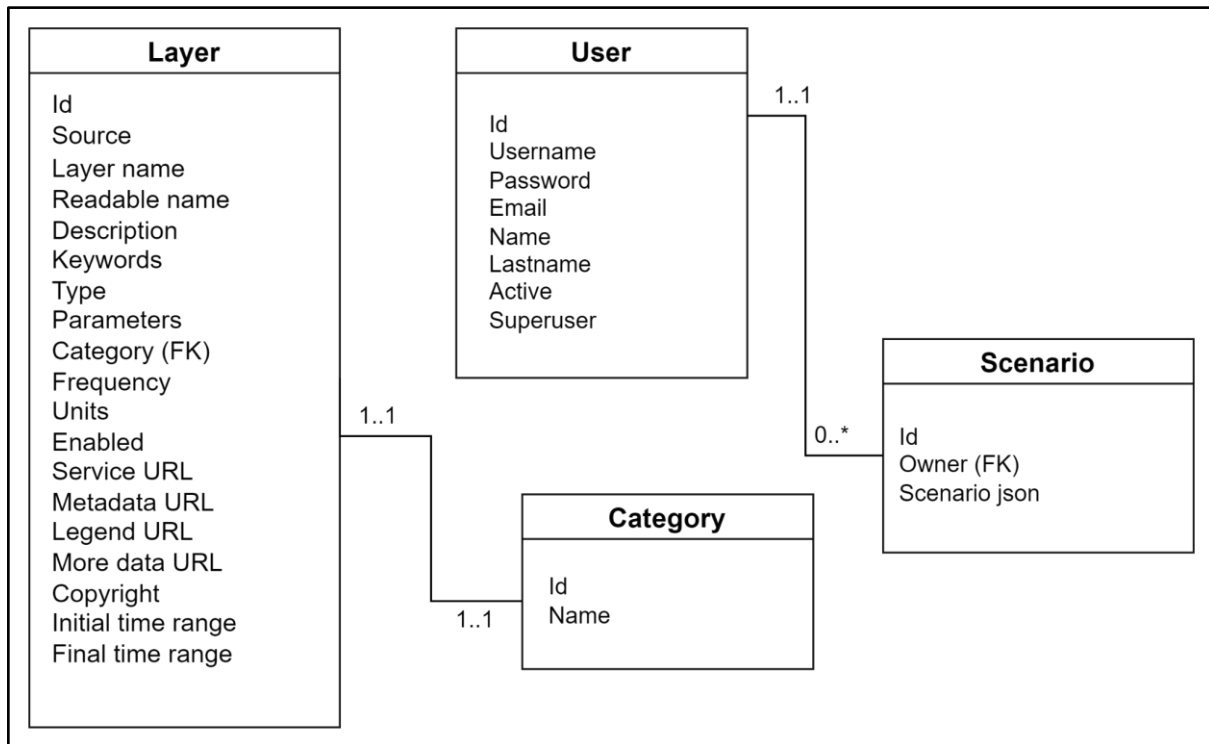

Figure 22. Database data model

On the following list, each of the tables will be described, along with each of the fields of the table:

- **Layer**: contains all the necessary metadata for each of the integrated layers. This data is presented to the user when the list of available layers is requested. The list of available layers is defined by the context selected by the user.

  - Id: layer unique identifier.
  - Source: data source of the layer. e.g. CMEMS, CLMS, WorldPop
  - Layer name: layer name as specified by the service.
  - Readable name: comprehensive and human-readable name of the layer.
  - Description: short description of the layer. Usually corresponds to the layer abstract.
  - Keywords: words that describe the layer
  - Type: layer typology. e.g WMS, WCS, ArcGIS Map Service.
  - Parameters: parameters that the layer accepts on its queries. e.g., time, elevation.
  - Category: category of the layer. It is related to the Category table with a one-to-many relationship, meaning that one layer shall have one and only one category, but a single category might have several layers associated.
  - Frequency: temporal resolution of the layer. e.g., Monotemporal, daily, monthly.
  - Units: units of measurement of the layer. e.g., m/s, mol/m^3.

- Enabled: defines if the layer is enabled for appearing on the platform.
- Service URL: the URL where the service is located.
- Metadata URL: layer with metadata of the service. Usually corresponds to the GetCapabilities request of the service.
- Legend URL: path to locate the legend of the layer. It corresponds to the legend specified in the GetCapabilities response.
- More data URL: URL that points to the description of the layer or to the source website.
- Copyright: any additional copyright information is stored on this field.
- Initial time range: initial data from which data is available on the service.
- Final time range: last day from which data is available on the service.

- **Category**: table to store the category names available on the service. By having categories, the layers can be organized by the type of data they represent.
    - Id: category unique identifier.
    - Name: name of the category.

- **User**: The user table contains all the users and their credentials. The user table also controls the authentication of users on the platform for scenario saving and sharing.
    - Id: user unique identifier.
    - Username: selected username used for authenticating on the system. This field is unique.
    - Password: user password used for authentication on the system. Passwords are hashed on the database.
    - Email: email of the user.
    - Name: name of the user. Optional.
    - Lastname: last name of the user. Optional.
    - Active: this field defines if the user is active and can login into the system.
    - Superuser: this field defines if the user can access the administration application provided by Django. Superusers can modify the data on the database.

- **Scenario**: The scenario table stores all the data of the saved scenarios of the user.
    - Id: scenario unique identified.
    - Owner: the user that owns the scenario. It is related to the user table with a one-to-many relationship. This means that a scenario shall only belong to a single user, but a user shall have multiple scenarios associated.
    - Scenario json: JSON object of the scenario that is used to restore the platform state to the state when the scenario was saved. This JSON can also be downloaded and stored as text.

4.4 Used Technologies

The platform developed for this thesis relies on web technologies. The selection of technologies is important since it defines the limitations of the application, the coding style, underlying technologies for hosting and serving data, and speed of the application. For the developed platform there are four main choices of technologies, corresponding to the two macro components explained before (frontend and backend), the web mapping library, and the database management system.

In this section the selected technology is presented along with a discussion on other potential technologies and the reason why the technology was selected.

4.4.1 Frontend framework

As the frontend framework, the used technology is Angular [67], in particular the version 12 since it was the latest stable version at the moment of development. Angular is an open-source web application framework developed by the Angular Team at Google and by the community. It offers a complete web development platform for building robust, complex, and enterprise-level applications on the web. It is written on TypeScript, which is a programming language built on top of JavaScript that offers typing capabilities. This means that variables can have a data type declared from the beginning instead of being dynamic, which improves robustness and performance by reducing errors and making it easier to detect and correct them.

Other options for frontend frameworks include two popular technologies: React and Vue. Both are JavaScript frameworks that focus on User Interface (UI) creation and offer a wide range of functionalities. The main difference between Angular and other frameworks relies on the robustness and the use of TypeScript, but also that Angular is more complex and has a higher learning curve.

In the end Angular was chosen for the robustness as a development platform, which provides many functionalities, and because it uses a more reliable programming language than base JavaScript.

4.4.2 Web mapping library

To display geospatial information on a web platform, it is necessary to use a web mapping library. OpenLayers was selected as the web mapping library since it is open source, and it is the most complete alternative. OpenLayers [68] is an open-source, client side (meaning that it is executed on the user's browser) JavaScript library for creating interactive web maps. By interactive it means that a user can visualize, move, and even modify the elements on the map. This library offers a wide variety of functionalities such as visualization of layers from multiple standards such as WMS, WMTS, WCS, WFS, ArcGIS Server, GeoJSON, GeoTIFF, among others, as well as querying capabilities. Additionally, OpenLayers has an Angular library which provides most of the OpenLayers features as components as well as all the data types for the OpenLayers library, which makes development easier and faster.

Another popular open-source web mapping library is Leaflet. It is a simpler and less powerful alternative to OpenLayers for creating simpler mapping applications but provides a good number

of features and support for all OGC standards. It was discarded because the platform required a more robust library and a wider set of functionalities.

4.4.3 Backend framework

The backend technology is very important since it is the one that manages all the retrieval of information from data sources, builds the data analysis results, and manages high traffic of requests and responses. Also, it is in charge of communicating with the database.

For the backend technology, the technology selected was Python with Django as the web framework. This decision is based on the power of Python as a processing tool for data analysis as well as the good support that it has for geospatial data. In addition, Django has an extension called GeoDjango that specializes on building web applications that interact with geospatial data.

Python is an interpreted, high-level, object-oriented programming language. It is mainly used for web applications and data science projects since it provides very good processing capabilities and libraries specialized in data analysis and visualization [69]. Django is a web framework for Python that allows the creation of web applications written in Python. It is particularly known for being very easy to configure and learn yet being a powerful and complete tool for creating web applications [Django ref]. Another advantage of Django is that it provides a geospatial module called GeoDjango, that focuses on geospatial data [70].

Another backend technology to consider is NodeJS, a JavaScript runtime library. Although it also has support for geospatial data, it was ruled out in favor of Python's data management capabilities.

4.4.4 Database Management System

For the developed platform the use of the database management system (DBMS) is bound to the storage of the layers information, e.g., name, service URL, parameters, etc., the categories and the scenarios. For the geospatial nature of the platform, the DBMS must have geospatial capabilities in addition to the usual reliability and speed.

The selected DBMS is PostgreSQL, a very popular and powerful open-source relational database which is especially useful for geospatial applications thanks to the PostGIS extension, which adds support for geographic objects and geospatial queries.

Other open-source alternatives include MySQL and SQL Server, which also has support for geospatial data types, but fall short in terms of functionalities with respect to PostgreSQL. The PostGIS extension is a strong reason to pick PostgreSQL over other database technologies, being the more robust and reliable GIS database extension.

4.5 Developed Features

The most important features that are included into the developed platform are listed and explained in this section. Feature refers to a specific functionality that is implemented into the platform and that is important for the application.

4.5.1 Area of interest selection by polygon drawing tool and feature import

A fundamental part of the context is the area of interest, since it is where the processing is focused, and it is a limit for the entire analysis session. The area of interest dictates the limits in which queries can be performed, layers are visualized, and area analysis is executed. Therefore, it is important that users are able to select their area of interest in any way they need.

The two ways in which users are able to select their area of interest is by drawing one using the drawing tool implemented for this end or importing a polygon feature in GeoJSON format.

The drawing tool allows the user to create a polygon and save it on the map. This tool is particularly useful for fast prototyping and loose selection of an area of interest, since it is not necessary to process any geospatial data and create new files for data visualization and analysis. Feature import is good for dealing with complex areas of interest or with areas of interest for more specific analysis, such as city areas, parks, etc.

4.5.2 Layer metadata storage using a database

The actual layer's data is not managed by the platform. Layers are virtually integrated by accessing the data source service URL with the necessary parameters. The way in which the platform connects to the data sources, and it is aware of the available layers is by storing all the necessary metadata of each of the layers available on a database.

By doing this it is possible to manage the layers in a centralized, and code-independent way, as well as adding more layers by simply adding the necessary metadata, making the platform easily extensible.

On the database the layers are displayed in a categorical manner. This means that each layer belongs to a category. On the platform, the way in which layers are obtained from the backend is in the format of a tree, with the layers being the leaves and the categories the branches. The structure of the database is presented in section 4.3.3, and the Layer model is the one that stores the metadata of each one of the layers integrated.

The communication with the database is managed by the backend, and it is this component that is responsible for delivering the metadata of the available layers to the front end depending on the parameters set on the context.

4.5.3 Mediator-Wrapper implementation

Data integration is implemented by following the mediator-wrapper architecture. This design pattern is presented on section 2.1 and the implementation architecture is presented on section

4.3.2. To recap, the data integration component of the application works by taking a generic request done by the frontend and redirecting it to a backend wrapper that is in charge of doing the request to a specific data source and of translating the response to a generic response, to send it back to the mediator so it can deliver a generic response to the client.

Due to this feature, it is possible to have many data sources integrated in a single platform and to make specific queries and deliver the information all with the same format. Additional layers and data sources are integrated by developing new wrappers that can respond to the generic requests from the mediator.

4.5.4 Layer parametrized visualization

OGC standards are very complete and allow a great flexibility in terms of parametrization. Most of the layers that are integrated allow parameterized view, in particular the CMEMS layers. This parametric visualization and querying of the layers are implemented on the platform.

Layer metadata stored on the database include the available parameters for each of the layers integrated. Parameter's names are stored as a list of comma-separated values, and the possible values for each of the parameters are extracted from the GetCapabilities request of each of the layer's service URL. The only parameters that are available on many of the layers are time and elevation, which are also standard parameters of WMS. Nevertheless, the platform is designed to be dynamic and include any number and any kind of parameters.

OpenLayers is able to manage WMS requests with parameters, correctly managing the update of the layer when a parameter is changed and generating a valid URL for the querying of the data. The capability of OpenLayers to generate the GetFeatureInfo URL for a WMS layer with parameters is used for the data analysis and for extracting the values on each layer when a click is made on top of the map.

4.5.5 Point querying for layers

It is possible to get the value of a layer on a coordinate by querying the data sources for information of its available layers. This is done using the capabilities offered by OGC web services and ArcGIS server, in particular the GetFeatureInfo operation of WMS and the identify operation of ArcGIS Server, respectively.

When a layer is queried, using the mediator, the generic response contains the value of the layer on the specified coordinates and the units of the layer. The units can also be the category description for a categorical layer such as CORINE Land Cover layers (see section 3.4.2.1).

Queries are performed when a user clicks on the area of interest while at least one layer is active. This event triggers the querying for all the active layers, on the coordinates clicked, and with the selected parameters that each layer has in that moment. When the data is received, it is shown on top of the map on a popup, one record for each layer active.

Figure 23 shows this feature in action. There are 3 layers active on the map at the moment of the click, which correspond to Mass Concentration of Chlorophyll-A, Wind Speed, and Coastal Zones,

all from different categories and data sources. The popup shows a generic response of the value of each one of the layers, showing the value and units for the layers with continuous data, and showing the category code and name for categorical layers.
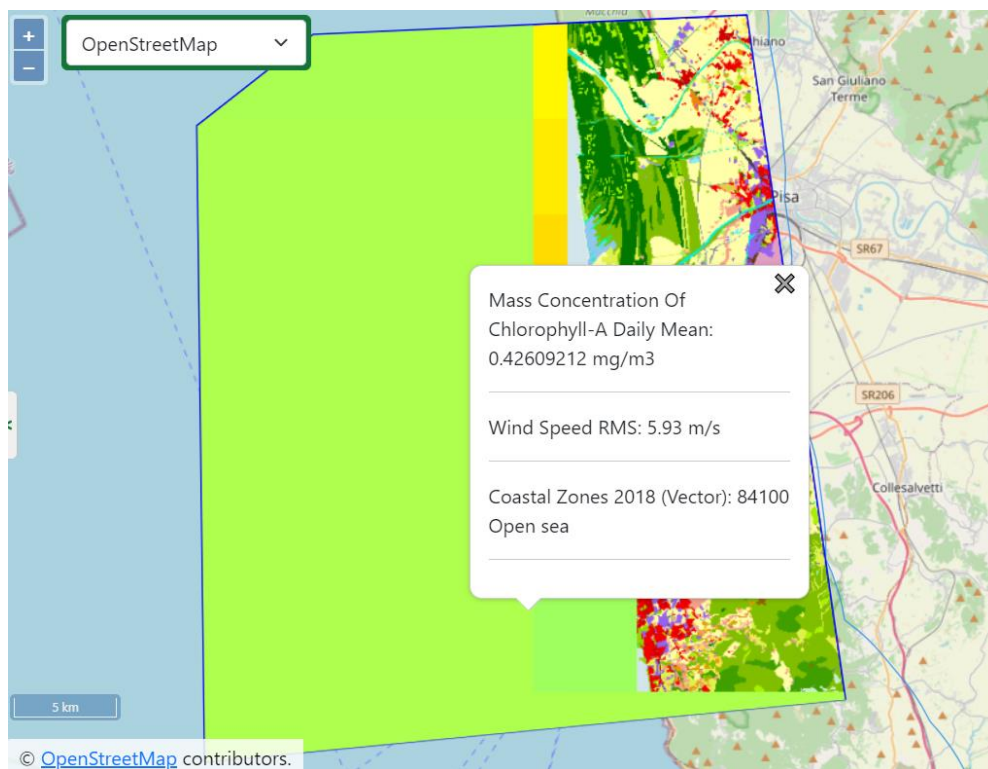

Figure 23. Point querying for layers

4.5.6 Depth profile and time series

Point analysis allows the user to perform a depth profile and a time series analysis of one or more layers. Depth profile and time series are explained more in depth on section 4.7. Layers that contain the "time" parameter are capable of requesting a time series, while layers with the "elevation" parameter can request a depth profile

A depth profile is an analysis of the different levels of a variable at different elevation levels. Layers that have "elevation" as a parameter can do this kind of analysis since it is possible to request the values of the layer at different depths, thus building the data for this type of analysis.

A time series is a visualization of the layer on a single point over time. Layers with "time" as a parameter can request a value multiple on different times and build a plot of the value over time.

These two analysis types make use of the data integration strategy to generate reports and to build knowledge around the layers. Although both analyses are composed of many requests to the data source, they are performed by the wrappers, instead of the mediator. This means that the mediator has a functionality exposed to obtain the time series and depth profiles instead of building the several requests and then putting all the data together.

This approach is faster for the frontend since it lowers the load of generating several requests, receiving the data, and building the data structures necessary for displaying the plots, but puts more work on the backend that is in charge of performing the requests and building the data. For some layers that rely on WMS for querying it is possible to speed up the time series analysis since GetFeatureInfo allows queries with time ranges. This means that a single request returns all the values at once for all the dates in a range.

4.5.7 Area Sampling

Another capability that the platform offers is area analysis. Although WMS layers do not possess any capability to perform analysis over a defined area, they are the more prominent layer type integrated on the platform, so a way to perform it is a must-have feature.

For WMS layers the way to perform data analysis is by performing a sampling of the layer values over the area of interest defined by the user. The process used to sample the layers is further explained on section 4.7.4.

With the sampling data it is possible to calculate statistics on the analyzed layer. The analysis performed for the sampling are the minimum, maximum, average, median, and standard deviation values, as well as a histogram of the samples over the whole area. Although it is not as accurate as performing an analysis over the original data, it is the only approach that can be used to extract any kind of area analysis for the layers provided as WMS.

On the other hand, layers that have built-in area analysis capabilities, i.e., ArcGIS Image Server, can directly perform sampling and the analyses mentioned above directly on the original data.

4.5.8 Sharing of the platform state - Scenarios

To comply with the scientific collaboration part of the Digital Twin Earth, and to allow reproducibility of the results and visualization of the layers on the platform, a means of sharing the state of the platform is offered.

The sharing of the application state can be done by two ways, generating a file with all necessary information such as context definition, active layers, visualization parameters, and analysis parameters that can be loaded again to rebuild the application state, or through scenario saving.

The application state file is a JSON file that contains all the information necessary to rebuild the application status as it was at the moment of generation. With the file it is possible to save and share many scenarios in an easy way.

Scenarios stored on the database are linked to users, meaning that in order to create a scenario it is necessary to register and log in to the application using the authentication provided by the platform. Saved scenarios are stored on the database. After logging in, the "Scenarios" tab will appear on the platform allowing the user to manage and share their scenarios. On the scenario tab it is possible also to download the scenario as a JSON file, which is the same application state file mentioned before. Using scenarios, it is possible to share the application state to other users, which duplicates the scenario at the moment of sharing as it was on the database.

4.6 Application Workflow

The main goal of the application is to allow users to visualize coastal-related data from several data sources in a single place, as well as doing analysis on the data. As the platform integrates an important amount of data from multiple sources, it is important that the user is able to restrict the amount of data for visualization and analysis, since a lot of data can be confusing and overwhelming for a user that is not familiar with the platform and with the data integrated in general. For this end, a workflow was designed to let users utilize the application in an intuitive manner by following certain steps.

The application workflow is divided into five steps, context definition, layer selection, layer exploration, data analysis, and sharing. *Figure 24* in the next page shows a diagram to explain the movement between steps and the processes that take place in each one of them.

This workflow is mainly inspired by the following platforms:
- USGS earth explorer (https://earthexplorer.usgs.gov/)
- ARIES for SEEA platform (https://seea.un.org/content/aries-for-seea)
- MyOcean Viewer (https://myocean.marine.copernicus.eu/data)

Through this section, the workflow of the platform will be explained by following a "walkthrough" approach. Screenshots from the platform will be available as a visual reference of the processes that take place in each of the steps of the workflow.
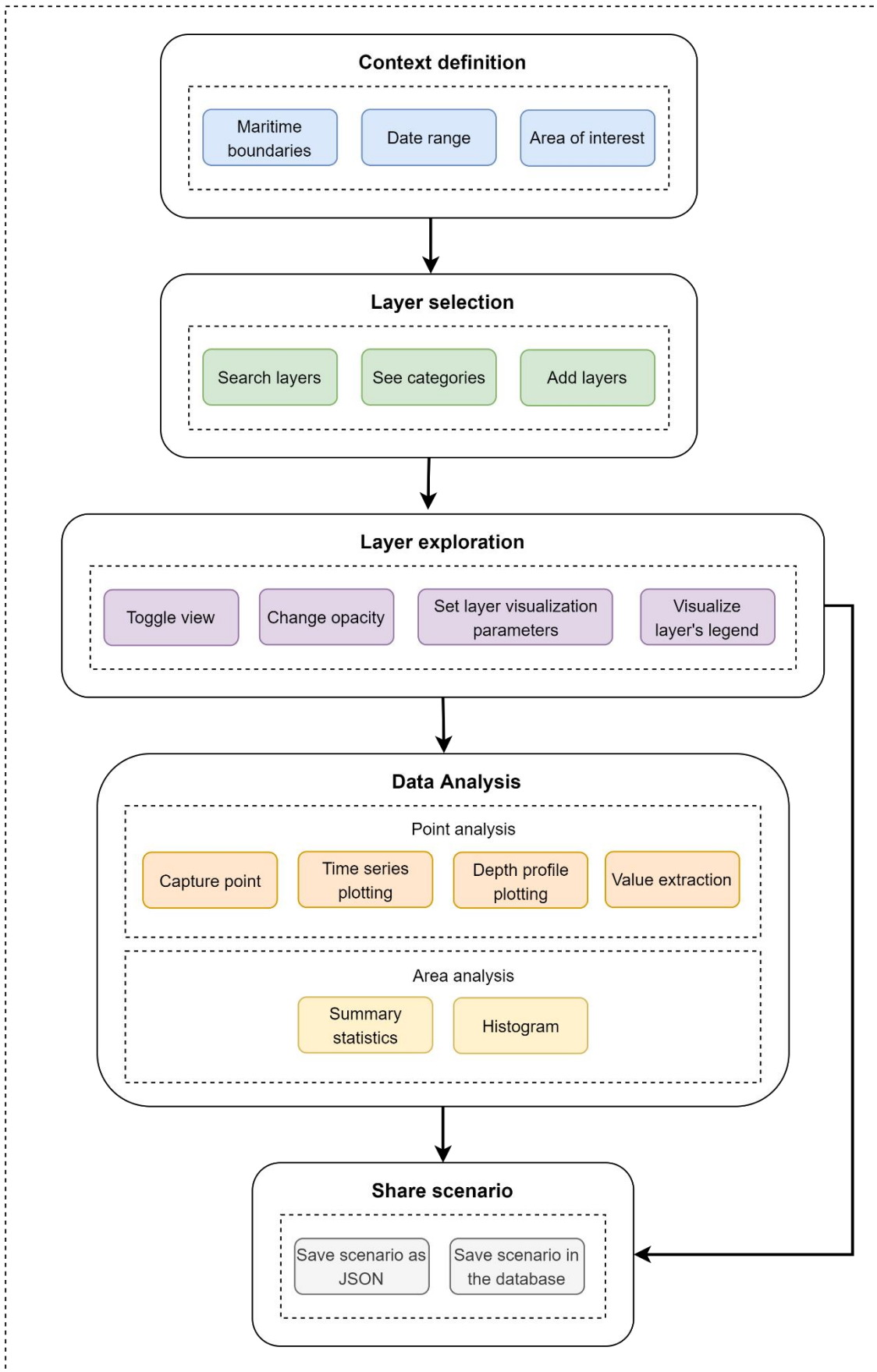
Figure 24. Application Workflow diagram

4.6.1 Context Definition

This first step consists of setting the spatiotemporal context in which the data visualization and analysis is to be performed. This is done by setting an area of interest (AOI) and the time range in which the data is going to be requested. Additionally, the maritime boundaries are set in this step between the available boundaries defined in section 3.2.

The reason behind setting a context is to filter the available layers, so only layers with data on a given time range are available for selecting and visualizing. Another reason to set a context before layer selection and exploration is to limit the data requested to the sources to speed up and lower the load of the retrieval of images from WMS services, and instead of fetching complete images they will retrieve an image bound to a specific, and smaller, region.

The "Context" tab is the first one of the left menus. *Figure 25* shows a snapshot of the platform in which a context has been set. The context corresponds to the months between January and April of 2020 (01/01/2020 to 01/04/2020), an area of interest that includes the Arno River outlet, a portion of Pisa, Livorno (an important Italian port), a natural park and a maritime protected area, and the official sea limits of Italy which correspond to the portion of the sea belonging to Italy up to the exclusive economic zone. The area of interest was selected using the draw feature.
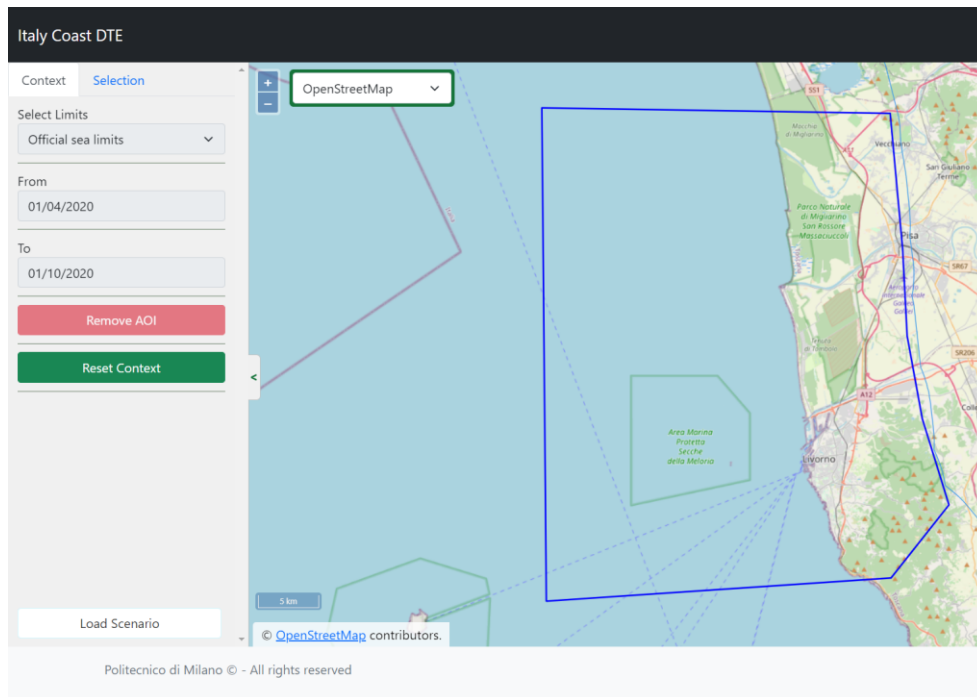

Figure 25. Context Area of Interest example

Figure 26. Context definition layout

This is a zone of high interest both ecologically and economically, because it comprises a port with high activity and zones that are of biological importance where the understanding of processes that relate water and land are important, such as the outlet of the Arno River near Pisa, and the protected maritime area.

After the context is defined, a request is made to the backend to obtain the available layers for the specific context from the database. In *Figure 27* it is possible to see the layers that are available for the period selected inside their corresponding category.
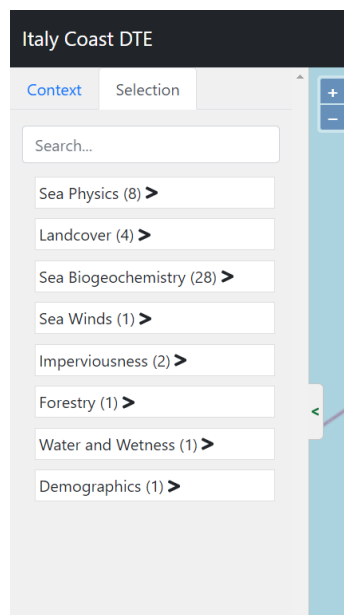
Figure 27. List of available categories example

4.6.2 Layer Selection

After the context is selected, a list of layers is loaded into the "Selection" tab coming from the database in concordance with the context parameters. This list of layers comes organized by categories and in alphabetical order. It is possible to filter the layers by name using the search field located on top of the layer's list.

Figure 28 shows the layer list with the category "Sea Biogeochemistry" expanded. The name that is displayed on the list for each layer corresponds to the readable name stored on the database. From the figure it is also possible to see that each layer has a "switch" button on the left that allows the user to activate or deactivate the layer for visualization.

When a layer is activated, it is automatically added to the list of active layers that is shown on the "Layers" tab on the right menu and gets displayed. Deselecting a layer removes all the data about the layer from the active layers.

Figure 29 shows the layer "Ammonium - Mole Concentration Daily Mean" layer being displayed after being selected from the list of available layers.

In case a context provides no available layers, the list will be empty. Also, in case a category has no available layers it is not displayed on the list.
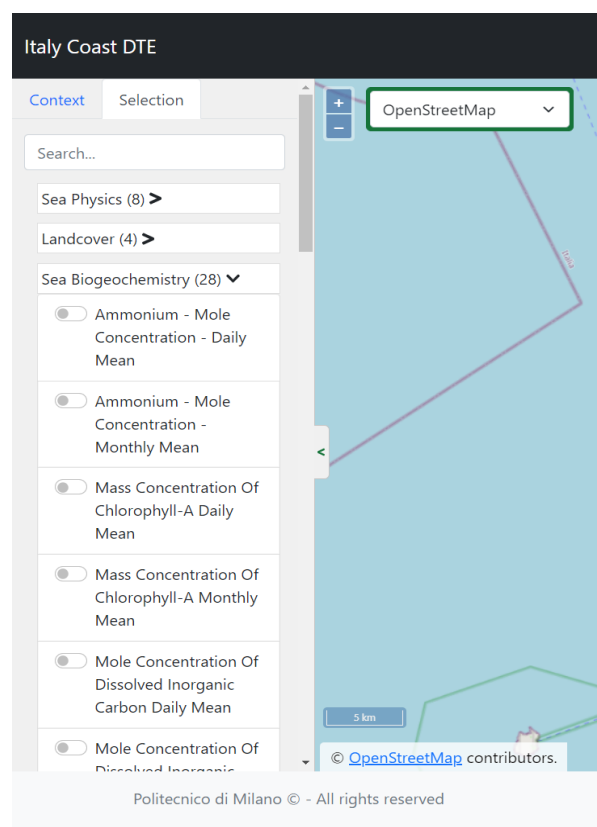

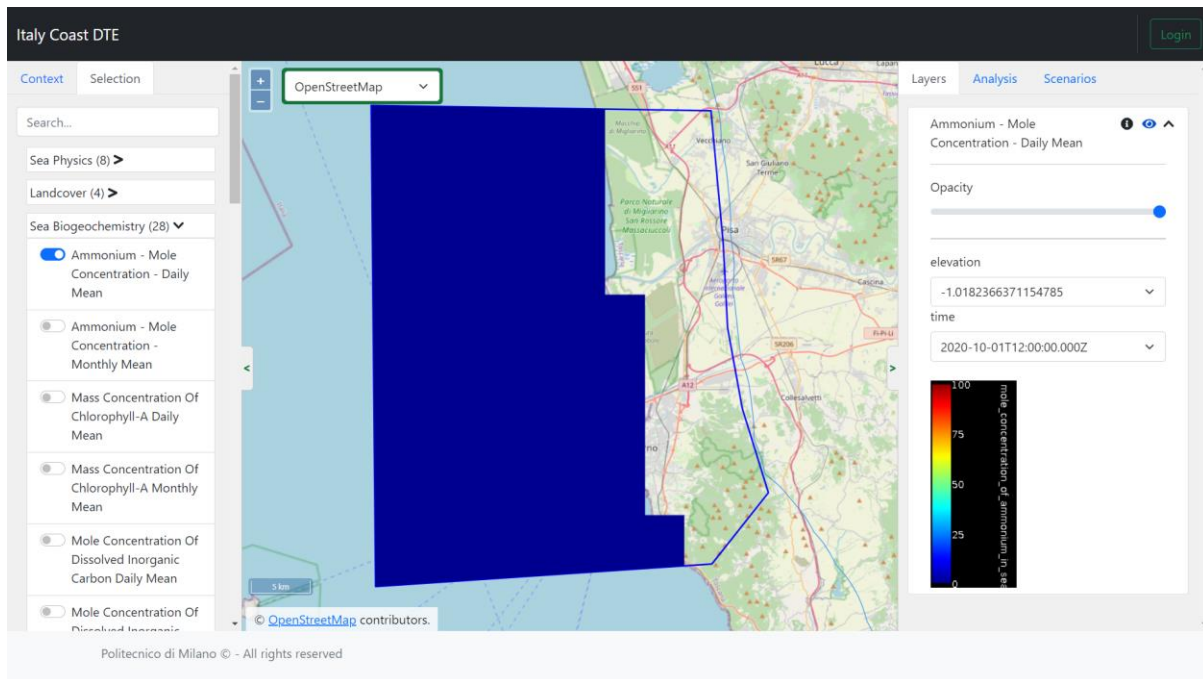*Figure 28. Layer selection with category expanded*

73

Figure 29. Layer selected and displayed

4.6.3 Layer Exploration

After selecting the layers to be visualized it is now possible to interact with those layers and explore the data available. In the "Layers" tab located in the right menu all the layers that are selected from the layer selection step are listed, ready to be manipulated. As each layer has its particularities, depending on the nature of each layer, the user can set different parameters for layer visualization. The parameters are defined for each layer on the database and its values are extracted from the GetCapabilities of each layer, if available.

For layers that have any form of temporal frequency it is possible to visualize its data on different timestamps. Some layers also allow the user to get data at different elevations. Dynamically, each layer has a set of dimensions that can be modified to visualize its data in the way the user wants. This behavior is shown on the CMEMS layers, for which all layers have "time" available as parameter and some of them have "elevation". Layers from other data sources do not present temporal frequency nor any other dimension.

It is also possible to manipulate the layer visibility and opacity, so it is possible to visualize many layers at once and see their data by playing with the opacity and hiding or showing the layer. By clicking on the map when one or more layers are active it is also possible to see the value of the layer on a specific point following the parameters of the dimensions set for each layer on the "Layers" tab.

Figure 30 shows the previous feature in action and shows the visualization of the same layer on different timestamps. Image on the left shows the Mass Concentration of Chlorophyll-A for the day 01-10-2020 in a point near the outlet of river Arno having a value of 0.25288 mg/m^3, while the image on the right shows the same layer but one month before on 01-09-2020 with a value of 0.11269 mg/m^3.
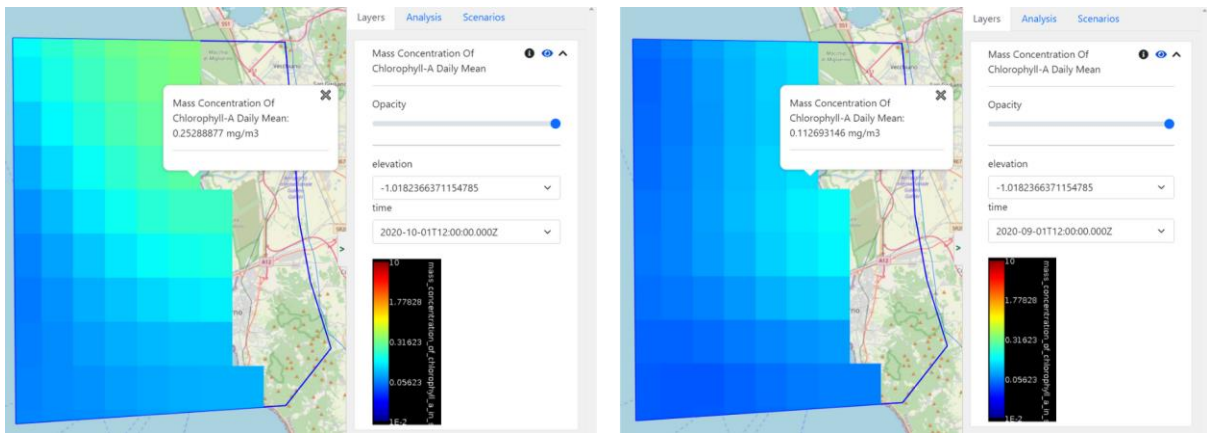
74

Figure 30. Layer visualization at different timestamps

Figure 31 shows the same layer as before at 01-10-2020 but for two different depths (elevations) corresponding to -1 meters on the left, and -10 meters on the right. It is worth noting that at different depths the number of pixels that are visible for the layer may be less, depending on the depth of the sea floor.
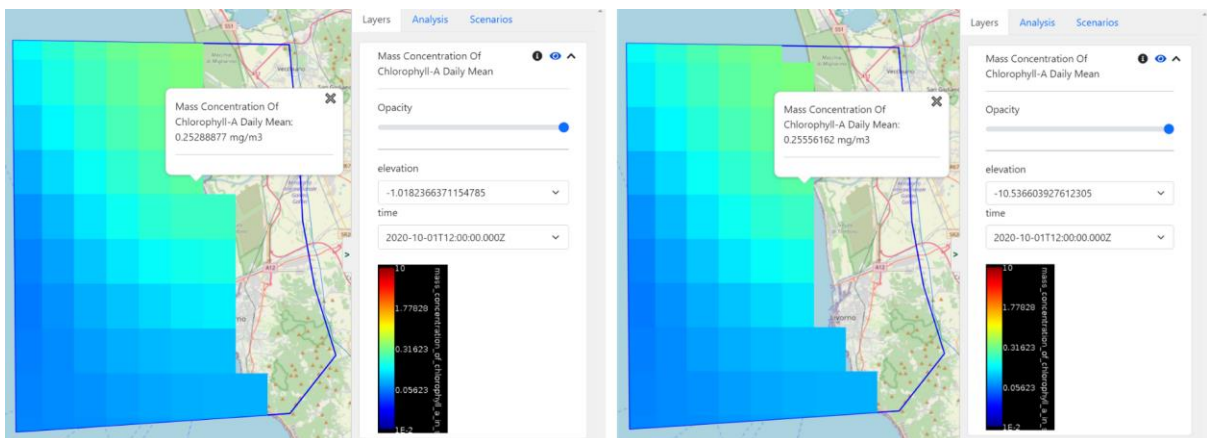

Figure 31. Layer visualization at different depths

The layer legend is displayed as an image on the "Layers" tab, if available. Legends are retrieved online from the URL specified in the metadata of the layer that is stored on the database with the layer information.

An example of the visualization of multiple layers is shown on *Figure 32*. On the figure, three layers are simultaneously displayed at different opacities: Coastal Zones 2018, Mass Concentration of Chlorophyll-A Daily Mean, and Imperviousness Density 2018. It also shows the legend of the Coastal Zones layer and the value for all 3 layers in a point located on the city of Livorno.
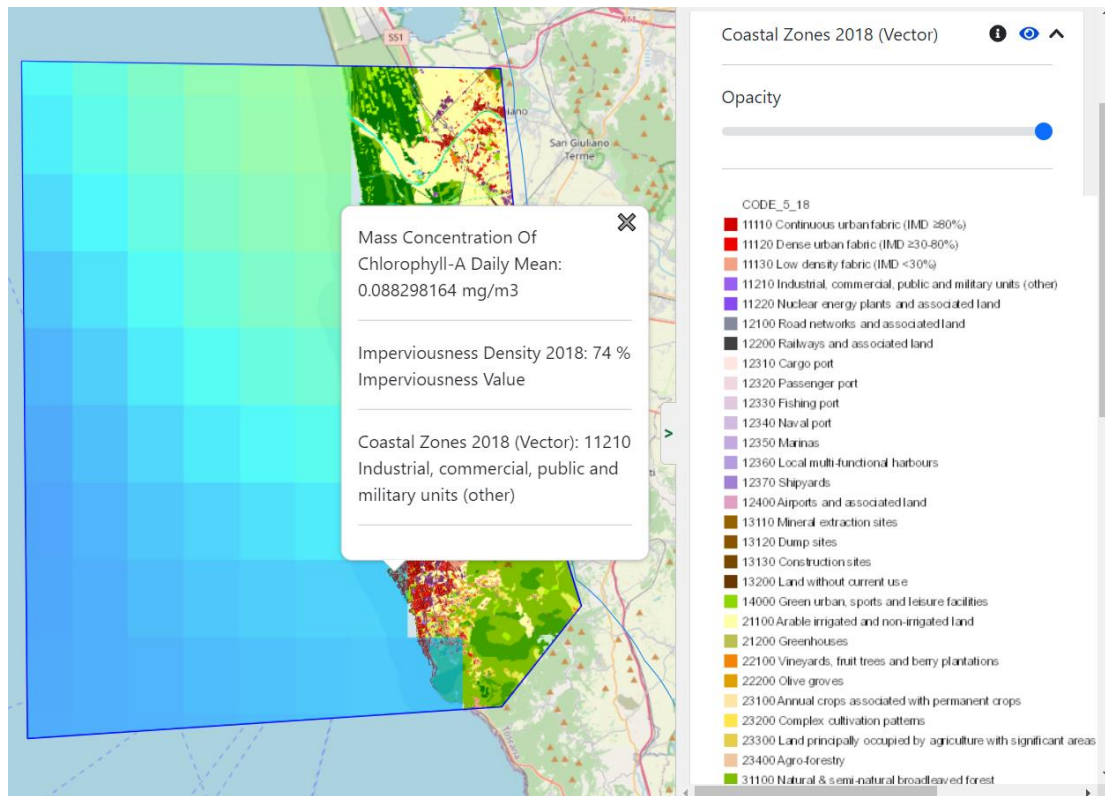
Figure 32. Visualization of multiple layers simultaneously

4.6.4 Data Analysis

After selecting the layers and observing the data it is possible to perform data analysis over the active layers. The data analysis capabilities are currently limited to performing point and area analysis for a maximum of 5 layers at a time. This limitation is due to the number of requests done simultaneously to the data sources, in an attempt to lower the load and avoid errors due to excess requests. The data analysis step is designed to add further utility to the integrated layers over simple visualization capabilities, adding value to the existing data and rendering it useful. Two types of analysis are included in the developed platform that correspond to point analysis and area analysis.

Point analysis consists of performing a layer's available analysis for a single location, specified as latitude and longitude by manually input the coordinates or capturing a point directly on the map by using the "capture point" feature. The available analyses for a layer can be time series (when the layer has the time parameter), depth profile (when the layer has the elevation parameter), and value extraction. When the point analysis is requested, the platform automatically detects the available analysis for the layer and performs them using the context information. It is also possible to download the data as CSV.

The time series and depth profiles are presented as plots and are downloadable as JPG, while the value extraction is presented as a table. The table with the values is also downloadable as a CSV, as well as the data for each of the time series and depth profile plots.
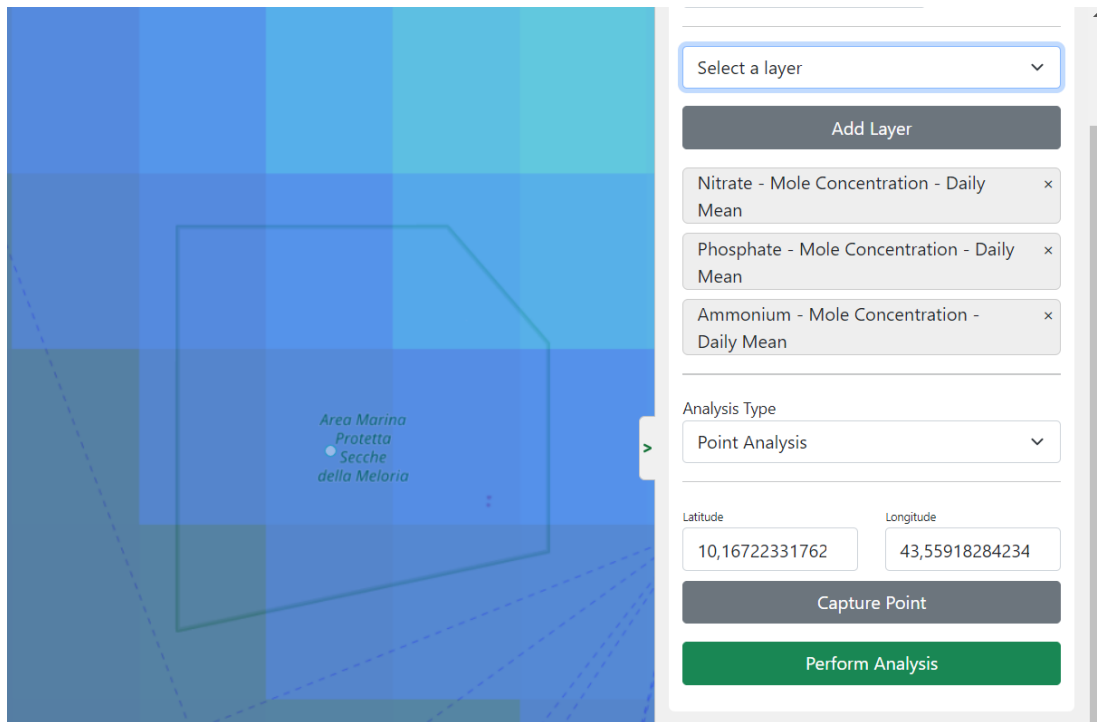
76

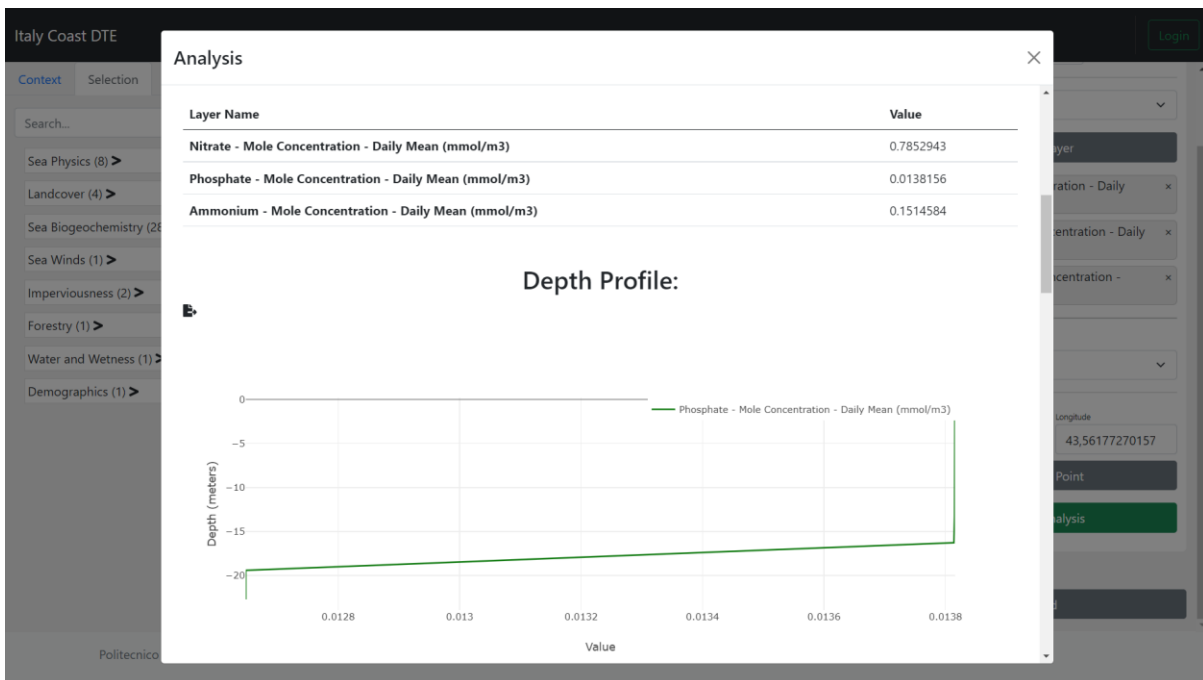*Figure 33. Point Analysis parameters*

Figure 34. Point Analysis results

Area analysis consists of an analysis of a set of layers over the area of interest defined in the context. It includes basic summary statistics of the layer such as minimum, maximum, median, average and standard deviation, and includes other analysis such as histogram to understand the value distribution of the layer over the area.

While some layers have built-in area analysis capabilities on the service, the layers that are integrated as plain WMS do not possess any analysis capabilities. To deal with the area analysis of WMS layers it is necessary to perform a sampling of the layer using the available data retrieval method from the service. This sampling is further explained in section 4.7.
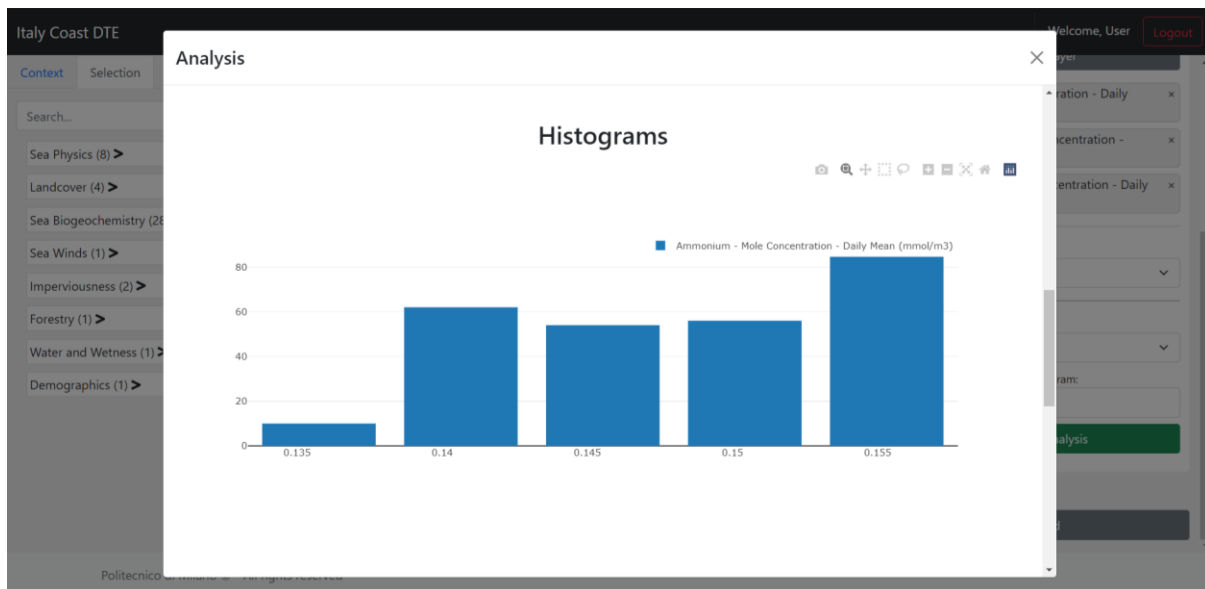

Figure 35. Area Analysis results

Figure 36. Histogram Result example

4.6.5 Sharing

After finalizing the analysis of the layers, a user may need to save their work to resume it later or to share its discoveries with their colleagues. To do that, sharing capabilities are added to the

platform in two ways, previously described on section 4.5.8, that are saving the state of the platform as a JSON file or saving a scenario to the database.

The object that is shared is a JSON object that contains all the necessary information to rebuild the application as it was at the time of generation. When generating the file, the platform will generate the JSON object and write it to a file, but when sharing an scenario that same object is stored into the database on the Scenario table. Scenarios are linked to users and can be further shared between them.
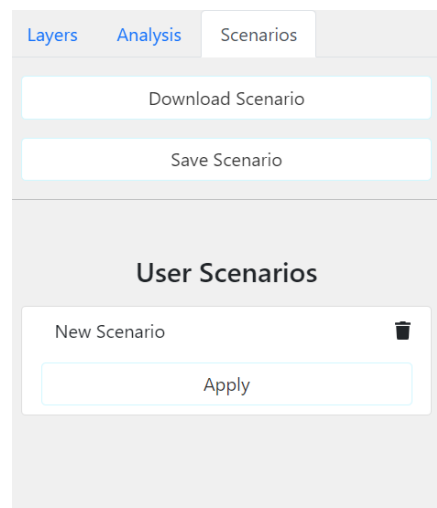

Figure 37. Sharing tab with a saved scenario

4.7 Data Analysis Methods

The available data analysis methods implemented on the platform are divided into two categories: point analysis and area analysis. Point analysis includes time series, depth profiles, and value extraction, while area analysis includes basic summary statistics, and area histogram, which are based on a previous area sampling performed over the layer that is being analyzed.

In this section all the data analysis methods are described and explained.

4.7.1 Time series

Time series is a point analysis method that allows visualization of the layer values on a single point over a period of time. This time range depends on the context. Time series analysis is performed by using the GetFeatureInfo operation available on WMS and is an operation supported by the mediator, meaning that the mediator can request a time series analysis by specifying the parameters to use. In practice, the time series is performed by a single GetFeatureInfo that receives a date range instead of a single timestamp, which is an incredibly useful feature that WMS provides.

The result of the time series is a dataset with dates and the corresponding value for each date. The results are plotted as a line plot using the library Plotly.js. In addition, time series data can be exported raw as a CSV file or as an image of the plot using Plotly.js built-in feature of image download.

Figure 38 shows an example of a time series plot produced by the platform. It shows the daily average phosphate concentration of sea water on the mouth of Arno River from April 2020 to October 2020. This plot was generated using Plotly.js and downloaded directly from the platform.
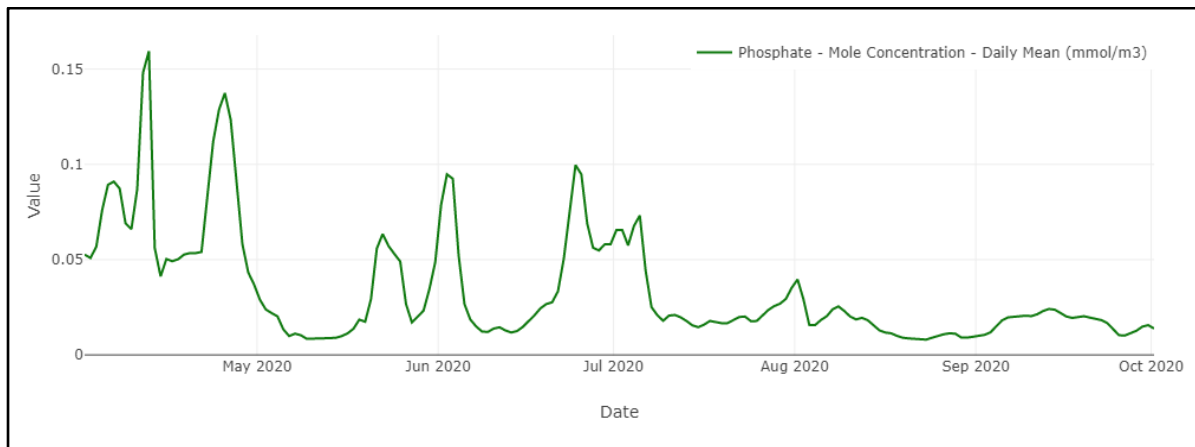

Figure 38. Time series of phosphate concentration on the outlet of the Arno River

4.7.2 Depth profile

Depth profile is a point analysis that consists of measuring the value of a layer on a single point, at a single timestamp, but at different depth levels. This analysis is made available when the layer has "elevation" as one of its parameters.

The depth profile implementation relies on the GetFeatureInfo operation of WMS to extract values at a single point, timestamp, and elevation. Ergo, it is necessary to request the value of the layer a number of times that is consistent with the number of available elevation levels. Most of the layers have up to 141 depth levels, but in practice, the coastal portion of the layers do not have a particularly deep sea floor, and do not have data from all the available depth levels. A way to improve the speed and to use resources accordingly is to request data at different depths starting from the one closer to the surface and start going deeper and deeper until a request does not return data. When this point is reached there are no more values under the invalid value, and it is safe to stop the requests and present the results.

The results of the depth profile are presented as a series of depth levels on the Y axis with their corresponding value on the X axis. As with the time series analysis, raw data of the analysis is available for download as CSV and the plot generated is available for download as an image.

Figure 39 shows an example of a depth profile plot, showing the temperature of seawater in a point located near the Livorno port at different depths for the day of April 9th, 2020.
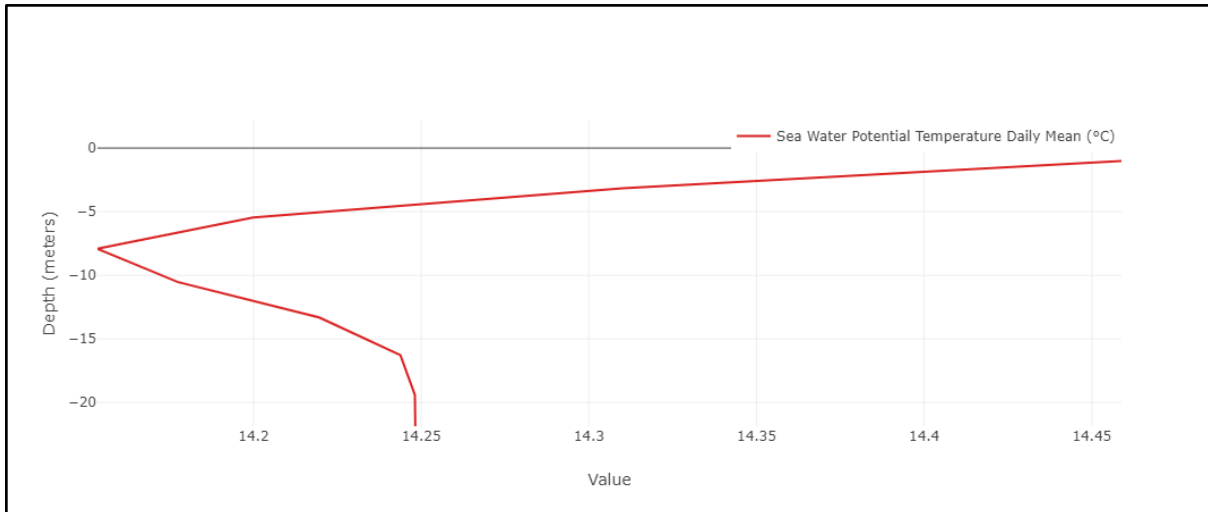
Figure 39. Depth profile of the sea temperature near Livorno

4.7.3 Value Extraction

This analysis consists of taking the values of the selected values at specific parameters and getting an organized table with the data. The extracted values can be downloaded as a CSV.

This analysis is performed using the GetFeatureInfo operation for WMS or the identify operation for ArcGIS layers for the layers that are selected for the analysis. In principle, it is the same feature as clicking on the map to obtain the values of the layers, but it is presented in a cleaner way, and it is exportable.

Layer Name	Value
Ocean Mixed Layer Thickness Daily Mean (m)	17.574026
Sea Water Salinity Daily Mean (0.001)	37.8872
Sea Water Potential Temperature Daily Mean (°C)	14.199914

Figure 40. Value Extraction table example

4.7.4 Area Sampling

To perform area analysis of a layer it is necessary to know the values of the data of the underlying layer. For WMS and ArcGIS Map Server it is not possible to process the original data, since the data that is delivered by the server is an image without any capabilities other than obtaining the value at a point. WCS and ArcGIS Image Server does provide some sort of area analysis such as histograms and area statistics and provide built-in sampling capabilities.

To perform area analysis, it is then necessary to sample the layer to be analyzed. Although this built-in sampling capability does not exist, it is possible to perform a sampling over an area by sending GetFeatureInfo requests for each of the points of the sampling region.

The sampling is done by creating a regular grid over the bounding box of the area of interest. The bounding box is the smaller rectangle that is able to fit the entire polygon of the area of interest.

Using the regular grid, the center of each one of the cells is calculated and, if the cell intersects the area of interest, a request to the underlying data source is performed. The grid obtained can be used as the data to build area analysis. This gridding process is automatically handled by WMS by using the width and height parameters of the GetFeatureInfo request as the resolution and the (X,Y) parameters to query each one of the grid cells individually. For reference on the GetFeatureInfo parameters refer to section 2.2.1. The WMS specification states that the process for obtaining the value for a point uses the width, height, X, and Y parameters and uses a center-point approach to get the value [34].
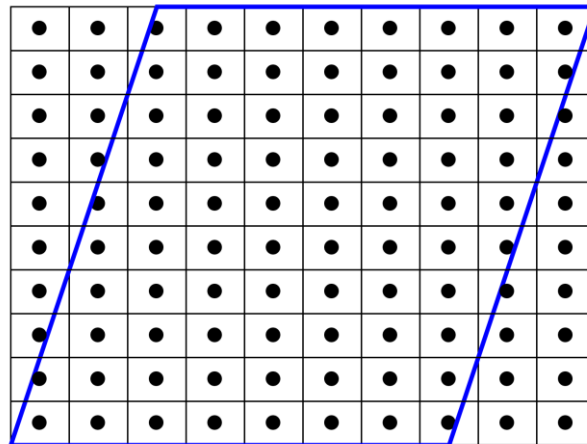

Figure 41. Sampling points over area of interest for a 10x10 sampling resolution

By querying only the grid cells that are located on the area of interest it is possible to lower the amount of queries performed, especially for irregular areas of interest. For a rectangular area of interest this does not present an improvement, since all the grid cells are within the area of interest. Null values are treated as NaN (Not a Number) by the backend and represent grid cells that do not have any value associated.

Sampling provides an approximation of the actual behavior of the layer data. Two sampling resolutions are offered on the platform: low resolution using a 10x10 sampling grid (100 data points), and high resolution using a 20x20 sampling grid (400 data points). Sampling may not reflect the real values of the layer if the original resolution of the layer is too high or if the area of interest is too big, but it provides an approximation of the values and it is useful for performing data analysis.

Sampling for CMEMS layers does provide good approximations, even for big areas, since the resolution of those layers is not very high (approximately 3km) but does not provide a very good approximation for high-resolution layers as the ones provided by CLMS, since those layers have very-high resolution.

4.7.5 Basic summary statistics

Using the data generated by the sampling or the built-in capabilities of each layer it is possible to perform basic summary statistics over an area of interest. Summary statistics are useful to understand the data over the area in a general way. The offered statistics are the minimum and maximum values, the average value, median value, and the standard deviation.

Minimum and maximum are useful for understanding the range in which values are over the area, the average is useful to understand the central tendency of the data, the median is also related to the central tendency of data but is less affected by outliers, and the standard deviation is used to understand the dispersion of the data and see which variables are probably outliers or outside average levels.

For WMS the basic statistics are calculated with the generated sampling explained in the previous section, while ArcGIS Image Server provides an operation to calculate the statistics directly on the data.

The statistics are presented as a table, with one row for each of the selected layers for analysis. The row contains the sample count, minimum, maximum, average, median, and standard deviation for each of the layers and it is downloadable as a CSV file.

Figure 42 shows the table produced by the area analysis summary statistics measuring the temperature of sea water on the previously used Area of Interest comprising the Arno River outlet, Livorno, Pisa, and a marine protected area. It was produced using a 20x20 sampling resolution, on the day April 9th, 2020.

Summary statistics

Layer Name	Min	Max	Average	Median	Standard Deviation	# of samples
Sea Water Potential Temperature Daily Mean (°C)	13.8928	15.1869	14.3412	14.2704	0.2618	272
Sea Water Salinity Daily Mean (0.001)	36.4116	38.0217	37.7731	37.8966	0.2749	272

Figure 42. Summary statistics table example

4.7.6 Area histogram

Another analysis that can be done over areas is a histogram of the layer values. A histogram shows the distribution of the values of the entire area, and it is calculated from the sampling over the area or directly on the data if the layer allows that operation.

A histogram is built by dividing the range of values into bins that are evenly spaced. Then the values of each of the samples are located into one of the bins and are plotted as a bar diagram. Using a histogram, it is possible to understand the distribution of the data, the skewness or bias, and to graphically see the tendency of the data.

The histogram for WMS layers is performed by calculating the ranges of N bins (N is selected by the user) and then classifying the sampled data on them. A histogram for each of the layers selected for analysis is calculated and shown as a plot. Plots are downloadable as images thanks to Plotly.js.

Figure 343 shows a histogram generated by the platform showing the temperature at the sea water surface on April 9th, 2020, for the same area of interest of the summary statistics.
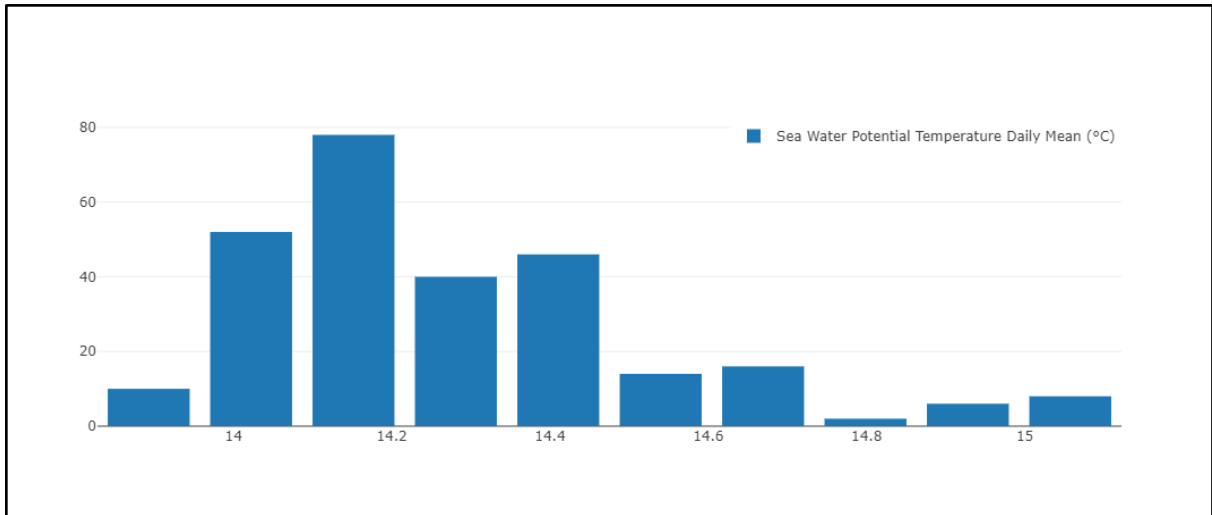
Figure 43. Histogram example of temperature over the area of interest

Chapter 5: Case Study

In this chapter, a case study will be presented as an example of the platform capabilities and to show the data integration of multiple sources in a real scenario. This study will address ocean acidification on Genova and its whereabouts, focusing on understanding the role of human activities and the land use on ocean acidification.

5.1 Genova Ocean acidification case study

5.1.1 Context

This case study is centered in the port of Genova and the Portofino marine protected area in the Ligurian Coast. It is a study of acidification of oceanic water by performing the analysis of pH, water temperature, and carbon dioxide surface partial pressure, which indicates the concentration of CO_2 in sea water, from the summer of 2020 to the end of the summer of 2021.

Ocean acidification is the reduction of pH on the sea water. This process is mainly driven by the increase of CO_2 dissolution in water that forms carbonic acid as shown on the following chemical reaction:

$$[CO_2] + [H_2O] <=> [H_2CO_3]$$
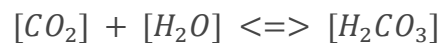
The pH is a measurement of acidity for aqueous solutions. It is a logarithmic scale from 0 to 14, where 7 is neutral, and anything lower than 7 is an acid and above 7 is basic.

Globally ocean pH is around 8.1, but it due to CO2 pollution oceans are slowly turning more acid and pH levels are getting lower. As the pH scale is logarithmic, even slight variations are very significant. A 0.1 drop in pH means that the solution is 1.3 times more acid, a 0.5 drop is 3.2 times more acid, and a drop of 1 point is 10 times more acid.

Ocean acidification greatly affects marine life, especially corals and organisms that depend on shells, since more acid waters slow down the formation of shells and skeletons, and even dissolve them, due to lower levels of carbonate ions which are the building blocks of shells.

To analyze ocean acidification, three fundamental variables, offered by the CMEMS, are used: pH, water temperature, and the surface partial pressure of carbon dioxide, that is an indicator of the atmospheric CO2 dissolution in seawater. Other layers that will help to analyze the context and environment of the specific places where measurements will be acquired are the Coastal Zones 2018 dataset, to better understand land use, WorldPop 2020 to understand how different population densities may affect ocean acidification levels, and the Imperviousness Density 2018 layer to track density of building area.

The purpose of this case study is to understand the yearly dynamics of the ocean acidification process in a very active area for tourism and commerce. The port of Genova is one of the busiest ports in Italy and the Mediterranean Sea, and the coasts surrounding the region are a major

tourist attraction during summer. Additionally, it will be possible to analyze the effect on the marine protected area and the relation between maritime activity and sea water acidification.

On Figures 45 and 46 it is possible to sea Genova to the west and the Portofino natural reserve to the east or the area of interest.
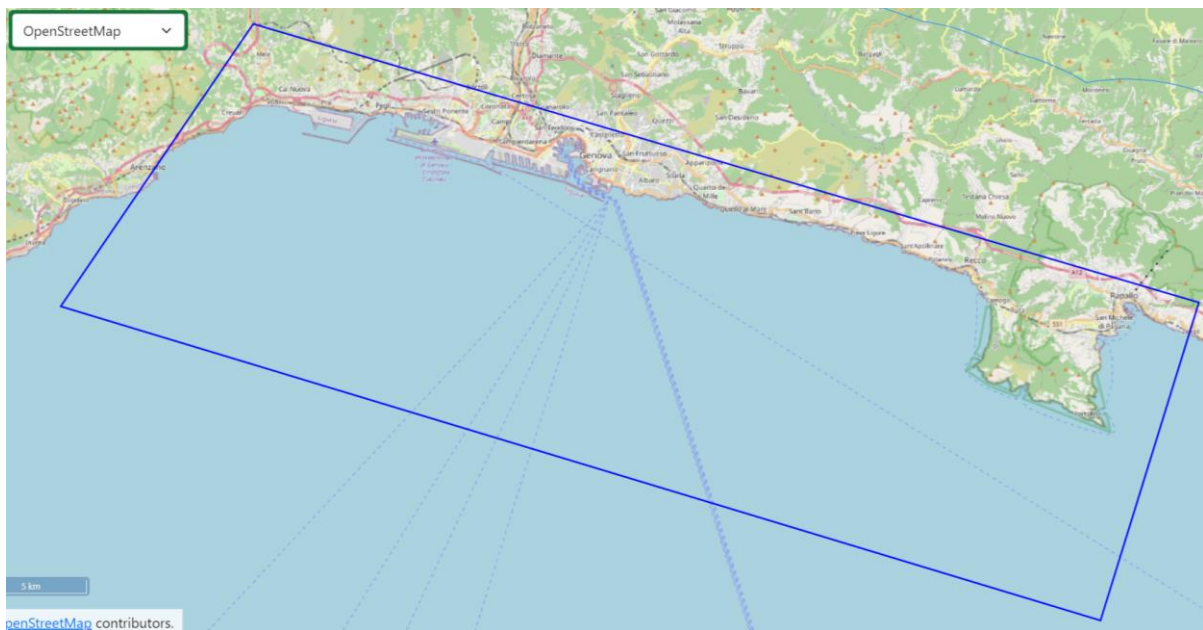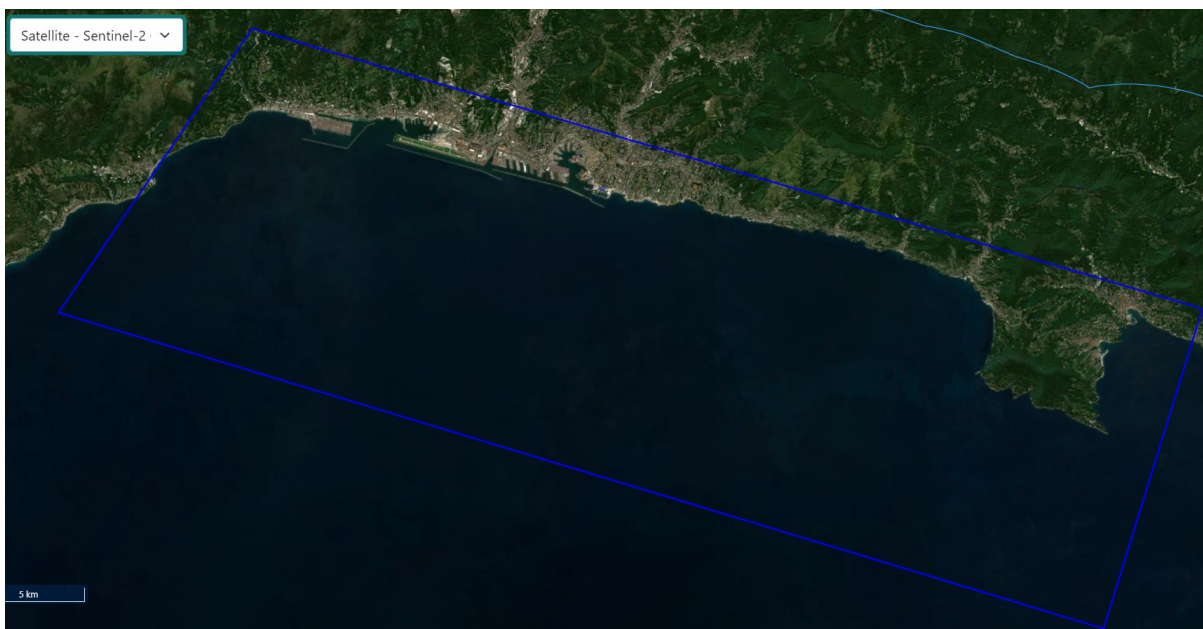

Figure 44. Genova case study area of interest

Figure 45. Genova case study area of interest satellite view

5.1.2 Layers

The layers used for this study come from all the data sources that are integrated on the platform. It will further show how data integration works in practice and how different layers can be used in a single, centralized way.

CMEMS layers are used to retrieve the biogeochemical and physical variables of the sea water on the area of interest: Sea Water pH Daily Mean, Surface Partial Pressure of Carbon Dioxide Daily Mean, and Sea Water Potential Temperature Daily Mean are the layers that are used. With CMEMS layers it is possible to perform time series and depth profiles using the variables mentioned on specific points of interest. The information returned by these layers is in XML format and is parsed and interpreted by the specific CMEMS wrapper before delivering a generic response to the mediator, that delivers a simple response to the application with the value and the units of the layer in the queried point.

CLMS layers are used as the land context of the zones being studied. The study includes Coastal Zones 2018 as the main reference of land use, and Imperviousness Density 2018 to show the human effect on the zones near the points of interest to be analyzed. With Coastal Zones 2018 it is possible to retrieve the category of land use for any specific point, and with Imperviousness Density 2018 it is possible to retrieve the percentage of change between the natural environment and synthetic structures. Some CLMS layers return their data in JSON format and others in XML format. Also, the way to query the data differs from the CLMS layers because they are provided through ArcGIS Server. The CLMS wrapper oversees parsing and interpreting the responses of the CLMS layers and delivers a generic response to the mediator and the web application.

From WorldPop, the 2020 layer is added to the study to track human population density in zones near the points of interest of the study. This will tell how different human densities may affect human acidification procedures. However, the temporal resolution of these datasets is not granular, meaning that the impact of people for shorter periods of time is not described by WorldPop. Other ways to track human impact on the points of interest is by checking for seasonal variability, e.g., summer vacations attract more tourists, winter months are less active for touristic activities.

5.1.3 Analysis

Figure 46 shows the city of Genova and its whereabouts under different views. On the top, it is displayed the satellite view of the city using the Sentinel-2 base map, the view in the middle corresponds to imperviousness, and the one in the bottom to the land use view using the Coastal Zones 2018 layer. A portion of the complete legend of the Coastal Zones 2018 layer is located on the right of the figure. From this figure it is possible to identify cargo ports, airport, marinas, shipyards, the road network between the port, and the commercial areas of the city. Close to the sea there are very few natural areas.
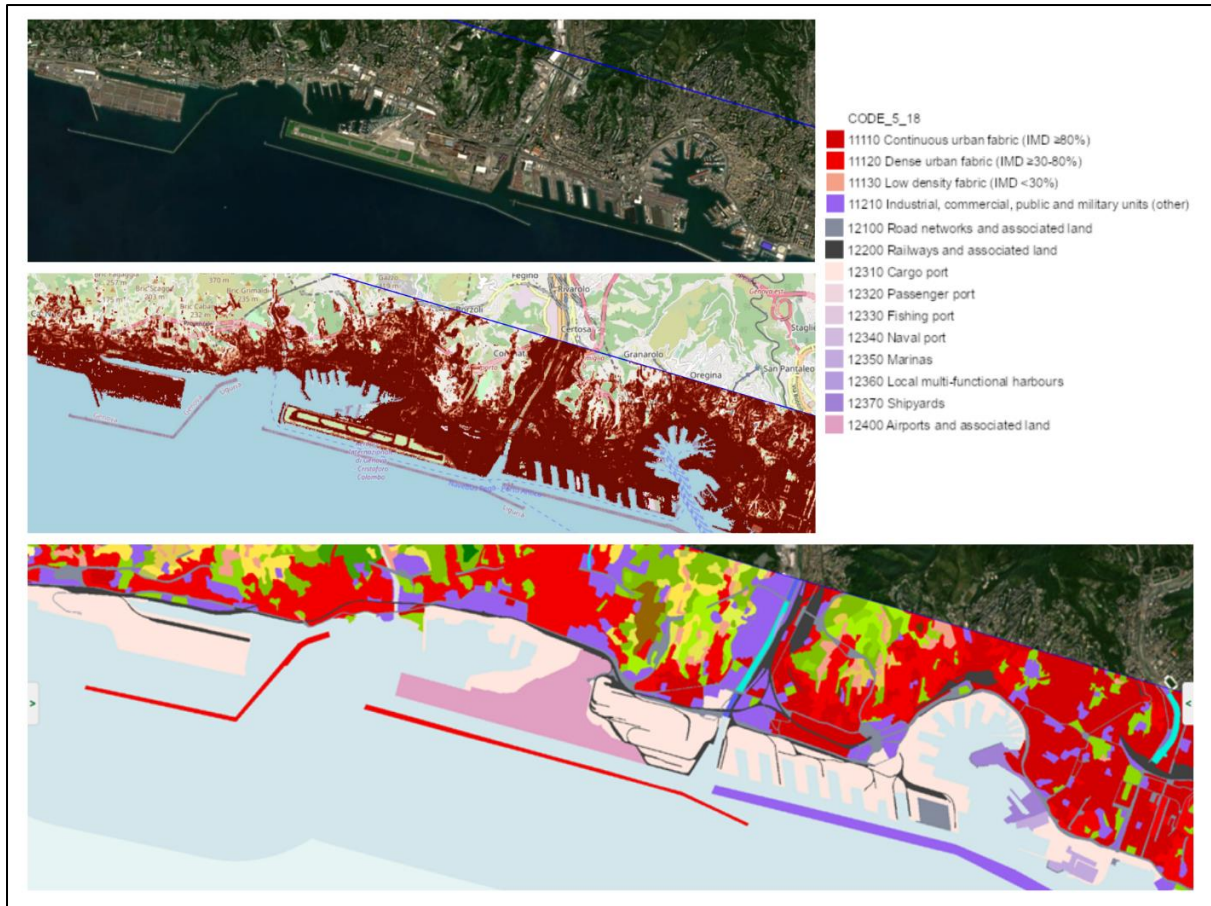
Figure 46. Genova area of interest

Figure 47 shows the second area of study: the Portofino natura reserve area. The image on the top-left corner shows the Sentinel-2 satellite view, on the bottom-left corner shows the imperviousness density, and on the top-right the Coastal Zones 2018 view. A partial portion of the legend of the Coastal Zones 2018 is displayed on the bottom-right. From these images it is possible to see the highly dense urban areas around the northeast that correspond to the towns of Rapallo and Santa Margherita Ligure, and at the south the small town of Portofino.

This zone is visibly more natural and contains several vegetated areas, natural & semi-natural forest, rocky environments, olive groves, and grasslands. On the adjacent land portion of the natural marine there is very little human settlement.
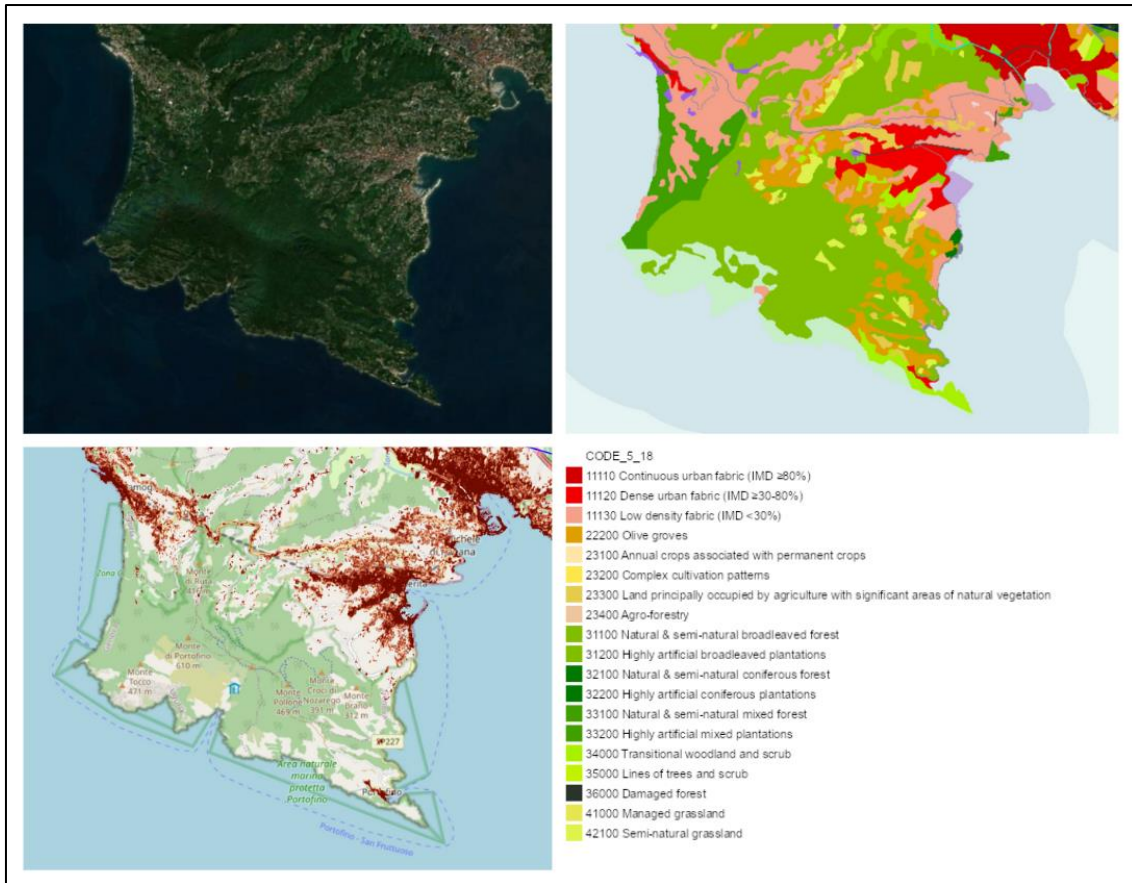
Figure 47. Portofino area of interest

The points of interest for the analysis are a point near the Genova port at the coordinates [8.921, 44.373], and a point on the western part of Portofino marine protected area at the coordinates [9.144, 44.336]. On each of the points, a point analysis was performed to see the time variation of the three variables of interest over the years of 2020 and 2021.

Figure 48 shows the time series plots corresponding to the pH (green), water potential temperature (blue), and CO_2 surface partial pressure (red) for the Genova port point, between the dates of July 1st, 2020, and September 1st, 2021, a total of 488 days.

From the plots it is easy to see that pH is inversely proportional to water potential temperature and CO_2 surface partial pressure, meaning that higher temperatures and higher CO_2 concentrations on sea water make the water more acid. There is also a visible seasonal variation of pH levels, reaching higher, and more normal levels in the winter months but reaching lower levels in the summer months, where human activity is higher over the coasts due to warmer temperatures.

pH reached its lower level on August 2nd, 2020, with a value of 7.927, and its higher level on January 18th, 2021, with a value of 8,124. This corresponds also with the higher and lower levels of sea water temperature and CO_2 surface pressure. These values are reported in *Table 14* and were calculated from the data downloaded from the platform as CSV.
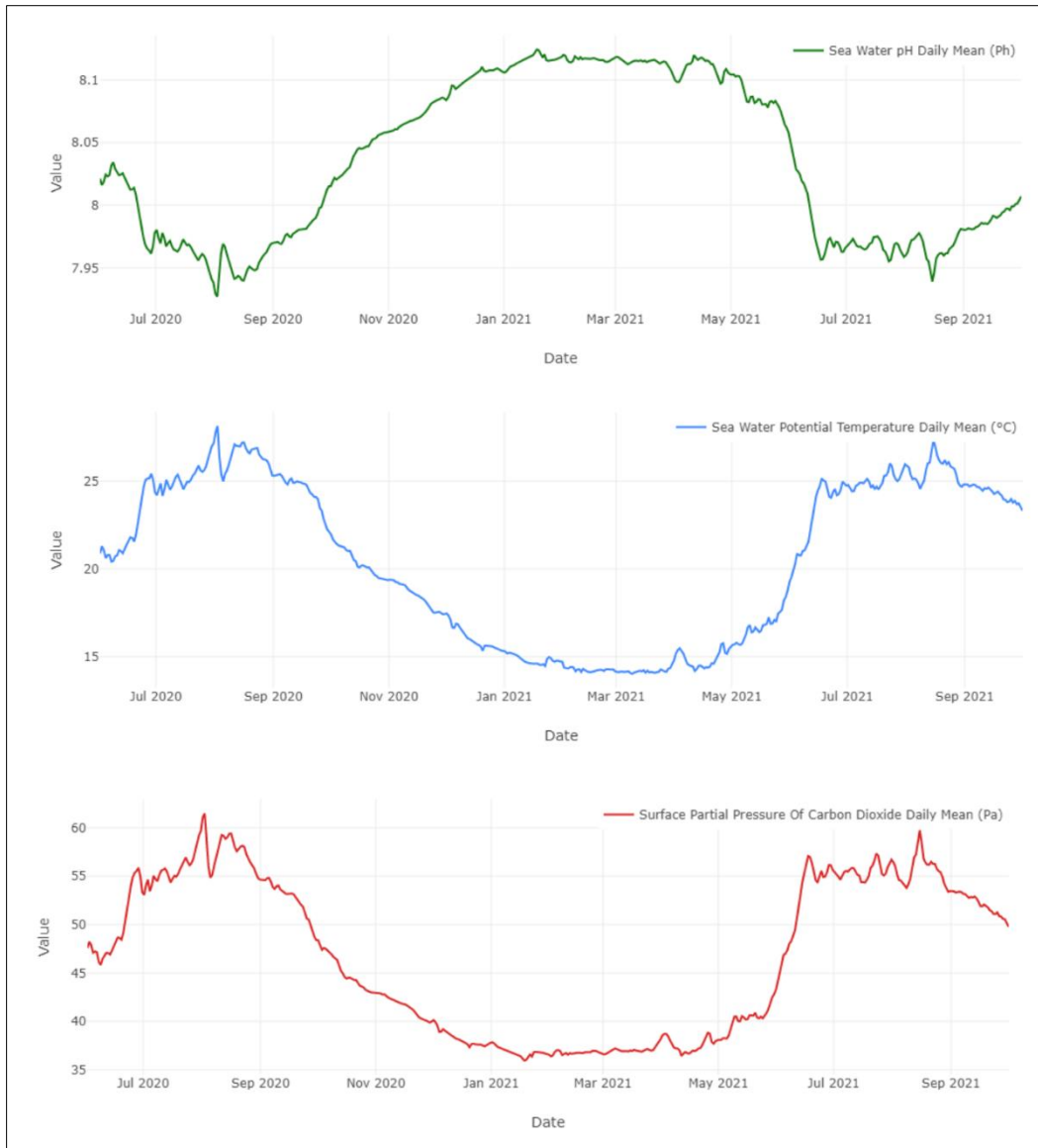
Figure 48. Ocean acidification time series plots at Genova port

	Minimum	Maximum	Average
CO$_2$ surface pressure (Pa)	35,95	61,4	46,59
Temperature (°C)	14,02	28,13	20,43
pH	7,927	8,124	8,035

Table 14. Ocean acidification variables at Genova port

Figure 49 shows the time series plots corresponding to the pH (green), water potential temperature (blue), and CO$_2$ surface partial pressure (red) for the Portofino marine protected area on the same time range as before.

The same peaks as the previous plots are also observed, with a minimum pH value of 7.924 (slightly lower than the port), and a maximum of 8.127 (slightly higher than the port). This shows

no major variations between the protected waters and the port waters in terms of pH, at least at the resolution provided by the Copernicus data. Values for the Portofino marine protected area are reported on *Table 15.*
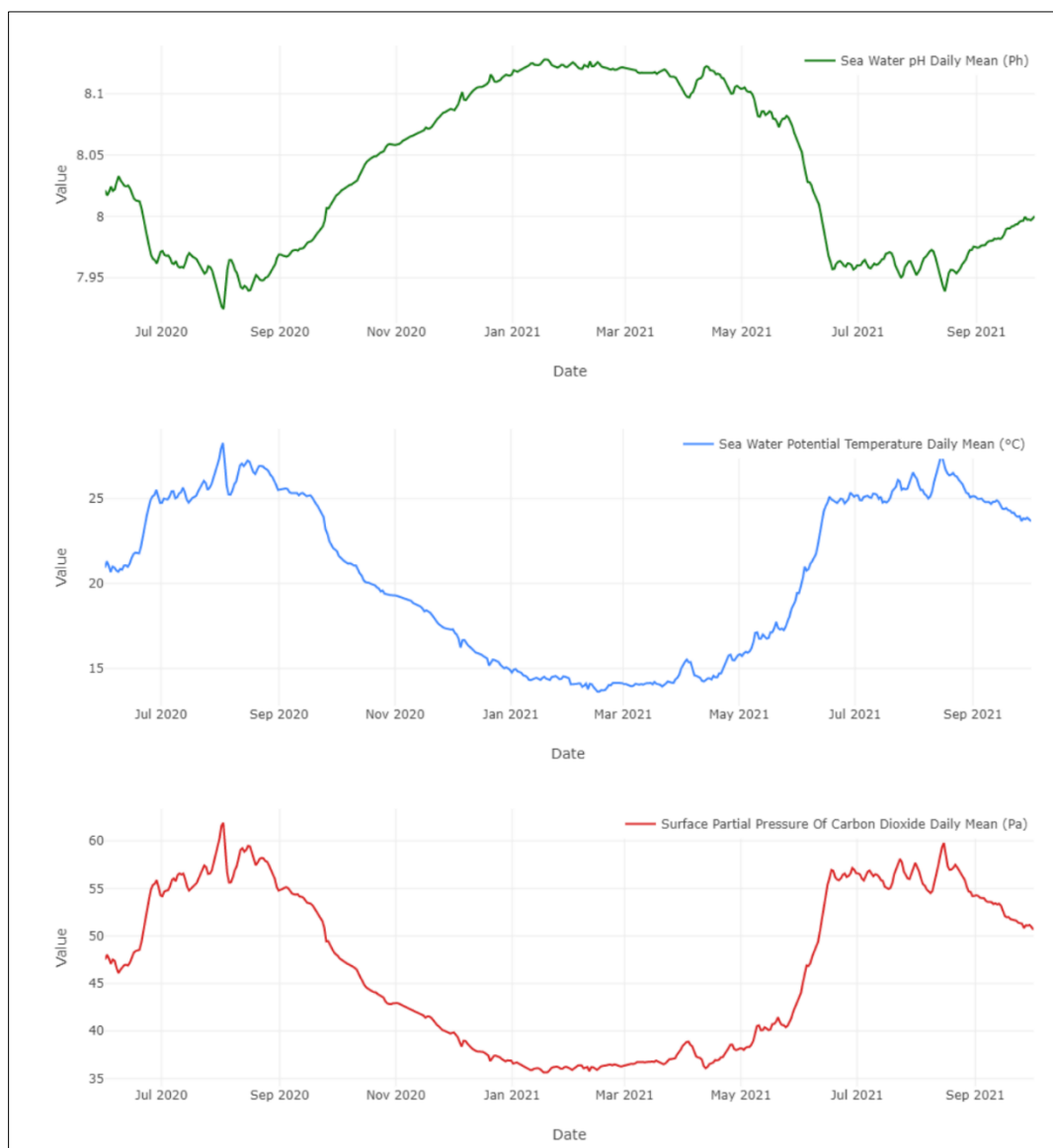

Figure 49. Ocean acidification time series plots at Portofino marine protected area

	Minimum	Maximum	Average
CO₂ surface pressure (Pa)	35,65	61,90	46,69
Temperature (°C)	13,64	28,26	20,49
pH	7,924	8,127	8,034

Table 15. Ocean acidification variables at Portofino marine protected area

As for the average values, there are also almost no variation between the Genova port and the Portofino marine protected area.

To show also the capabilities of depth profile and area analysis, the date of the lower pH level reported was taken to analyze the pH levels on different depths for that day for the two points of interest considered. Also, an area analysis of the entire area was performed on August 2nd, 2020, to see the distribution of values over the whole area and statistics.

Figures 50 and *51* show the depth profiles for the day August 2nd, 2020, for the Genova port area and the Portofino marine protected area, respectively. In the Genova port, pH level rapidly increases from 7,927 and reaches a normal level of 8.1 around 84 meters deep. Then it decreases again but does not go below 8 points as on the surface. Remembering that a 0.1 decrease in pH is equivalent to water 1.3 times more acid, means than water on the surface is approximately 1.3 times more acidic that water below 22 meters deep, where pH reaches 8.0266, and approximately 1.5 times more acid than water at 84 meters deep.
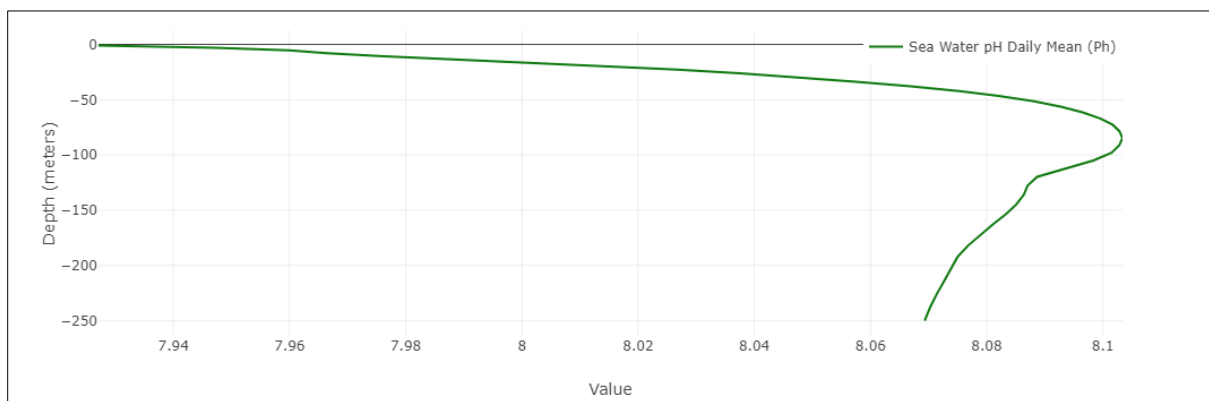

Figure 50. Depth profile August 2nd, 2020 - Genova port

As for the Portofino marine protected area, pH rapidly increases from 7,9245 to a maximum of 8.035 at 30 meters deep. The plot stops at 30 meters of depth because this area has more shallow waters, meaning that the seafloor is located around that depth and there are no more measurements of pH above that depth level.
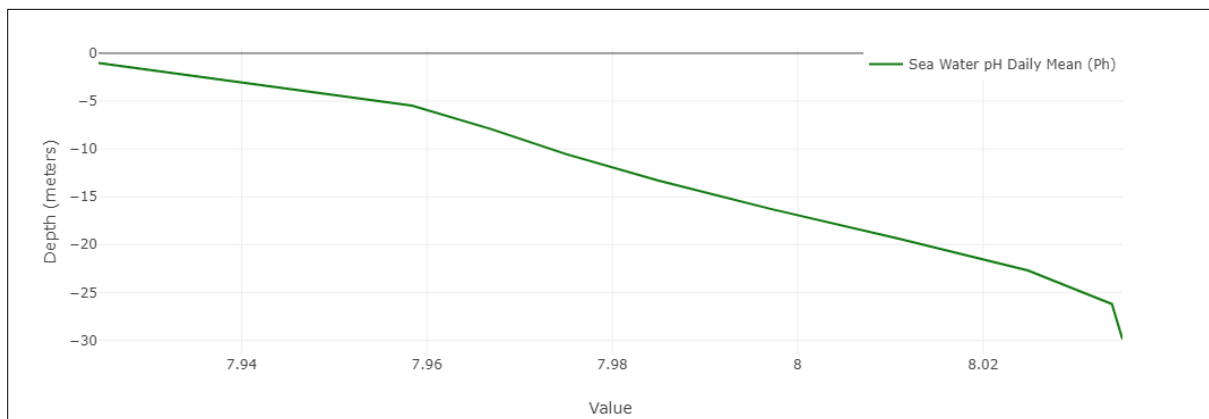

Figure 51. Depth profile August 2nd, 2020 - Portofino marine protected area

The area analysis for the area of interest corresponds to the day of 2nd August 2020, and was performed using a 20x20 sampling of the three marine layers that has been used for this study. From the total of 400 possible samples, 161 samples were taken from the area of interest. The summary statistics of the area analysis, featuring the minimum, maximum, average, median, and standard deviation are reported on *Table 16.*

	Minimum	Maximum	Average	Median	Std. Deviation
CO_2 surface pressure (Pa)	61.4469	62.1477	61.7199	61.7366	0.1367
Temperature (°C)	28.0371	28.3338	28.1803	28.1762	0.0728
pH	7.9231	7.9274	7.9255	7.9255	0.0009

Table 16. Case study area analysis variables statistics

From the summary statistics it is possible to see the little variation on pH on the entire area of interest, having minimum and maximum values that differ only in 0.0043 and a standard variation of only 0.0009. Temperature and CO_2 surface pressure also present very small variations between the highest and lower value, 0.3 °C and 0.7 Pa, respectively.

The histograms of each of the variables show the distribution of data. All histograms were produced with the sampling performed for the area analysis and with 5 classes.

The pH histogram on *Figure 52* presents a bias towards the right and a single value falling on the first, leftmost category. Although this tendency, all values remain very similar, and differences between categories only differ of a few millesimals.
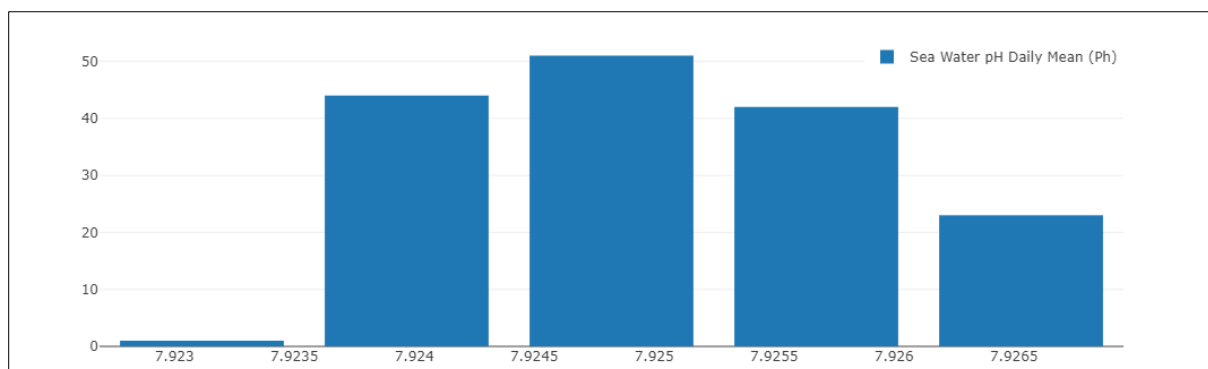
Figure 52. Histogram for pH over the area of interest

Surface partial pressure of CO_2 histogram on *Figure 53* presents a left bias. This means that only few samples indicated very high values, while most of the values indicated a pressure between 61.58 Pa and 61.86 Pa.

As for the sea water temperature histogram on *Figure 54,* it shows a central tendency but is very uniform and flat.

Figure 53. Histogram for CO$_2$ surface partial pressure over the area of interest

Figure 54. Histogram for water temperature over the area of interest

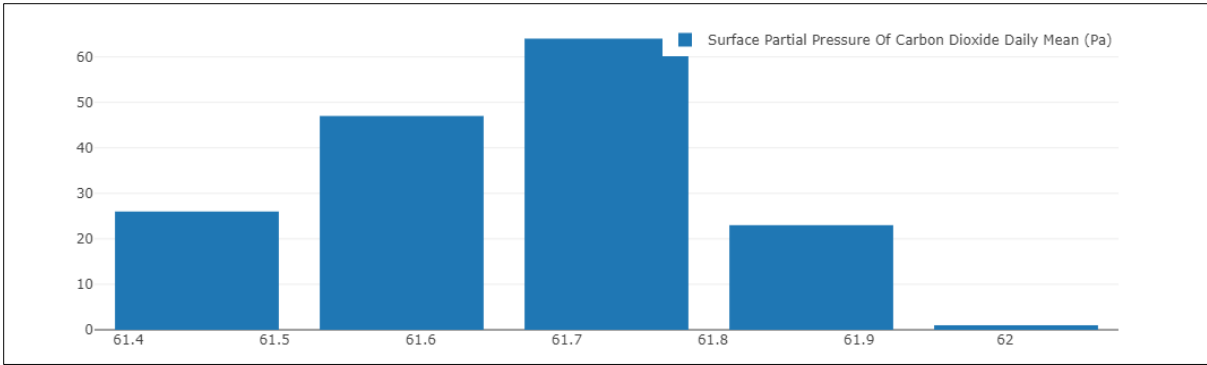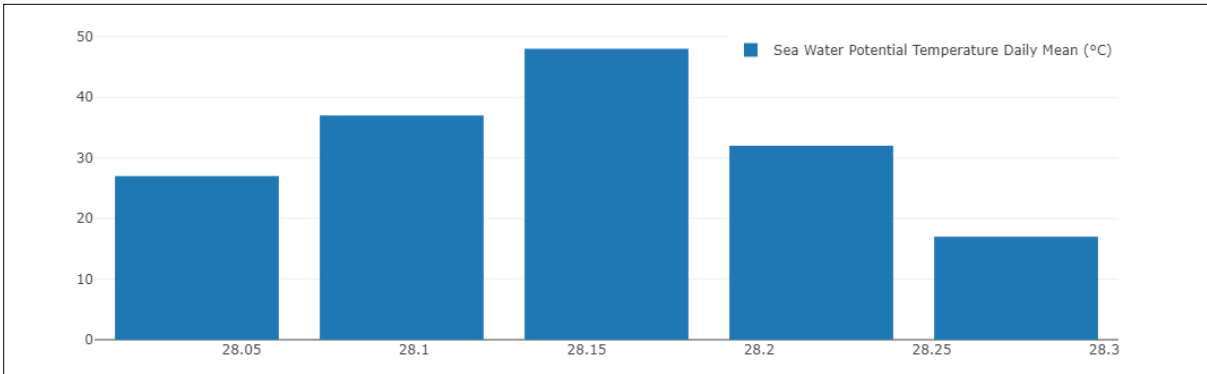# Conclusion

The work presented in this thesis constitutes a first approach to a Digital Twin Earth by addressing the problem of data interoperability through the implementation of an integration strategy using the mediator-wrapper architecture. This innovative platform allows us to integrate and use multiple data sources that are not easily put together, visualize them, explore them, and perform analysis over them.

Given the high amount of data available on the internet and the lack of platforms that integrate and use this data, the platform developed presents an innovative way of using the already existing data through a web application that is easily accessible and usable, without the user being an expert on the underlying data and without doing complicated processes to retrieve and combine it.

This thesis is done on the framework of the Destination Earth initiative to create a Digital Twin of the Earth, and it is scoped to a very important area for economy, tourism, and ecology, which is the coast.

A total of 68 layers were integrated on the platform coming from three data sources, Copernicus Land Monitoring Service (CLMS), Copernicus Marine and Monitoring Service (CMEMS), and WorldPop, that represent land, marine, and demographic data, respectively. To integrate them, the mediator-wrapper architecture was implemented as the virtual data integration strategy, providing a centralized way to utilize the data from those services and provide further utility by means of visualization and analysis. It is worth mentioning that these services are not easily integrated, highlighting the need of stronger interoperability between organizations and services, and the use of data integration strategies such as the one implemented on this thesis to provide further utility for the data.

The platform is a proof of concept of what could be achievable with a Digital Twin Earth and to show how data integration is an important task in its development. It is a work-in-progress and further development is planned.

The modular nature of the platform allows it to be replicated to multiple other contexts, including other coasts of the world or other types of environments. There are promising new features that can be added as future work, such as: addition of more analysis capabilities, integration of more data sources by the addition of new wrappers, automatic data integration by the inclusion of ontologies and semantic technologies, Artificial Intelligence (AI) models and forecasting, inclusion of external data sources on-the-fly, and addition of pre-defined indicators for the type of environment that is being studied.

The platform code is available on GitHub:
Frontend: https://github.com/Diuke/italy-coasts-dte-front
Backend: https://github.com/Diuke/italy-coasts-dte-back

References

[1] "ESA - Working towards a Digital Twin of Earth." https://www.esa.int/Applications/Observing_the_Earth/Working_towards_a_Digital_Twin_of_Earth (accessed 2022).

[2] European Commission, "COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS," 2020, [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0067&from=en

[3] S. Nativi, P. Mazzetti, and M. Craglia, "Digital Ecosystems for Developing Digital Twins of the Earth: The Destination Earth Case," Remote Sens. 2021, Vol. 13, Page 2119, vol. 13, no. 11, p. 2119, May 2021, doi: 10.3390/RS13112119.

[4] "A European Green Deal | European Commission." https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en (accessed 2022).

[5] "Policies | Shaping Europe's digital future." https://digital-strategy.ec.europa.eu/en/policies (accessed 2022).

[6] L. Deren, Y. Wenbo, and S. Zhenfeng, "Smart city based on digital twins," Comput. Urban Sci. 2021 11, vol. 1, no. 1, pp. 1–11, Mar. 2021, doi: 10.1007/S43762-021-00005-Y.

[7] A. Sharma, E. Kosasih, J. Zhang, A. Brintrup, and A. Calinescu, "Digital Twins: State of the Art Theory and Practice, Challenges, and Open Research Questions," Nov. 2020, Accessed: 2022. [Online]. Available: http://arxiv.org/abs/2011.02833

[8] "Scientists to create a Digital Twin of Antarctica | The University of Edinburgh." https://www.ed.ac.uk/informatics/news-events/stories/2020/scientists-to-create-a-digital-twin-of-antarctica (accessed 2022).

[9] "CGI to develop Digital Twin Precursor for the European Space Agency to advance understanding of global food systems | CGI UK." https://www.cgi.com/uk/en-gb/news/space/cgi-to-develop-digital-twin-precursor-for-the-european-space-agency-to-advance-understanding-of-global-food-systems (accessed 2022).

[10] "DTE Hydrology – hydrology." http://hydrology.irpi.cnr.it/projects/dte-hydrology/ (accessed 2022).

[11] "Telespazio UK Awarded Digital Twin Earth Precursor Contract to Advance Understanding of Local Impacts of Global Climate Change | Telespazio in the UK." https://telespazio.co.uk/en/news-and-stories-detail/-/detail/digital-twin-earth-precursor-contract (accessed 2022).

[12] "Scientific needs to realise Digital Twin Earth, Phi-Week 2020 summary", https://az659834.vo.msecnd.net/eventsairwesteuprod/production-nikal-public/289ac37bc1c44d92b1a0b6e6ac64b48f#:~:text=Advanced%20satellite%2Dbased%20Earth%20observation,the%20Digital%20Twin%20Earth%20vision (accessed 2022).

[13] M. Mõttus et al., "A Methodology for Implementing a Digital Twin of the Earth's Forests to Match the Requirements of Different User Groups", doi: 10.1553/giscience2021_01_s130.

[14]   "The Ocean and its Digital Twin: What's in for Copernicus Marine? | CMEMS." https://marine.copernicus.eu/news/ocean-and-its-digital-twin-whats-copernicus-marine (accessed 2022).

[15]   "Progressing toward a Digital Twin of the Ocean | Blue-Cloud." https://www.blue-cloud.org/news/digital-twin-ocean-marine-environmental-indicators (accessed 2022).

[16]   P. Korenhof, V. Blok, and S. Kloppenburg, "Steering Representations—Towards a Critical Understanding of Digital Twins," Philos. Technol., vol. 34, no. 4, pp. 1751–1773, Dec. 2021, doi: 10.1007/S13347-021-00484-1/TABLES/2.

[17]   "The Earth Archive." https://www.theeartharchive.com/ (accessed 2022).

[18]   P. Jiang et al., "Digital Twin Earth -- Coasts: Developing a fast and physics-informed surrogate model for coastal floods via neural operators," Oct. 2021, Accessed 2022. [Online]. Available: http://arxiv.org/abs/2110.07100

[19]   "CMEMS MyOcean Viewer." https://myocean.marine.copernicus.eu/ (accessed 2022).

[20]   "ARIES for SEEA | System of Environmental Economic Accounting." https://seea.un.org/content/aries-for-seea (accessed 2022).

[21]   "Getting started with k.LAB." http://integratedmodelling.org/statics/pages/gettingstarted.html (accessed 2022).

[22]   "USGS Earth Explorer" https://earthexplorer.usgs.gov/ (accessed 2022).

[23]   "Geoportale Nazionale" http://www.pcn.minambiente.it/viewer/ (accessed 2022).

[24]   S. Nativi, P. Mazzetti, and M. Craglia, "Digital Ecosystems for Developing Digital Twins of the Earth: The Destination Earth Case," Remote Sens. 2021, Vol. 13, Page 2119, vol. 13, no. 11, p. 2119, May 2021, doi: 10.3390/RS13112119.

[25]   D. AnHai, H. Alon, and I. Zachary, "Principles of Data Integration," Principles of Data Integration, 2012, doi: 10.1016/C2011-0-06130-6.

[26]   C. Y. Huang and S. Liang, "A sensor data mediator bridging the OGC Sensor Observation Service (SOS) and the OASIS Open Data Protocol (OData)," Ann. GIS, vol. 20, no. 4, pp. 279–293, Jan. 2014, doi: 10.1080/19475683.2014.942795.

[27]   M. A. Regueiro, J. R. R. Viqueira, J. A. Taboada, and J. M. Cotos, "Virtual integration of sensor observation data," Comput. Geosci., vol. 81, pp. 12–19, Aug. 2015, doi: 10.1016/J.CAGEO.2015.04.006.

[28]   S. Zhang et al., "Parallel spatial-data conversion engine: Enabling fast sharing of massive geospatial data," Symmetry (Basel)., vol. 12, no. 4, Apr. 2020, doi: 10.3390/SYM12040501.

[29]   W. Huang, K. Kazemzadeh, A. Mansourian, and L. Harrie, "Towards Knowledge-Based Geospatial Data Integration and Visualization: A Case of Visualizing Urban Bicycling Suitability," IEEE Access, vol. 8, pp. 85473–85489, 2020, doi: 10.1109/ACCESS.2020.2992023.

[30]   M. Butenuth, G. v. Gösseln, M. Tiedge, C. Heipke, U. Lipeck, and M. Sester, "Integration of heterogeneous geospatial data in a federated database," ISPRS J. Photogramm. Remote Sens., vol. 62, no. 5, pp. 328–346, Oct. 2007, doi: 10.1016/J.ISPRSJPRS.2007.04.003.

[31]   C. Prudhomme, T. Homburg, J. J. Ponciano, F. Boochs, C. Cruz, and A. M. Roxin, "Interpretation and automatic integration of geospatial data into the Semantic Web," Comput. 2019 1022, vol. 102, no. 2, pp. 365–391, Feb. 2019, doi: 10.1007/S00607-019-00701-Y.

[32]   L. Stoimenov and S. Djordjević-Kajan, "An architecture for interoperable GIS use in a local community environment," Comput. Geosci., vol. 31, no. 2, pp. 211–220, Mar. 2005, doi: 10.1016/J.CAGEO.2004.09.017.

[33]   "Geographic information — Metadata — Part 1: Fundamentals", ISO 19115-1:2014, ISO/TC 211 Geographic information/Geomatics Technical Comitee, 2014-04. [Online]. Available: https://www.iso.org/standard/53798.html

[34]   "Web Map Service.", 06-042, Open Geospatial Consortium, 2006 [Online]. Available: https://www.ogc.org/standards/wms

[35]   "Web Map Tile Service.", 07-057r7, Open Geospatial Consortium, 2010 [Online]. Available: https://www.ogc.org/standards/wmts

[36]   "Web Feature Service.", 09-025r2, Open Geospatial Consortium, 2014 [Online]. Available: https://www.ogc.org/standards/wfs

[37]   "Web Coverage Service.", 17-089r1, Open Geospatial Consortium, 2018[Online]. Available: https://www.ogc.org/standards/wcs

[38]   "OGC EO Dataset Metadata GeoJSON(-LD) Encoding Standard.", 17-003r1, Open Geospatial Consortium, 2019 [Online]. Available: https://docs.ogc.org/is/17-003r2/17-003r2.html

[39]   "OGC GeoTIFF Standard.", 19-008r4, Open Geospatial Consortium, 2019 [Online]. Available: http://docs.opengeospatial.org/is/19-008r4/19-008r4.html

[40]   "The Ultimate List of GIS Formats and Geospatial File Extensions - GIS Geography." https://gisgeography.com/gis-formats/ (accessed 2022).

[41]   "OGC Standards and Resources | OGC." https://www.ogc.org/standards (accessed 2022).

[42]   "ArcGIS Server documentation—ArcGIS Server | Documentation for ArcGIS Enterprise." https://enterprise.arcgis.com/en/server/latest/get-started/windows/about-arcgis-server-documentation.htm

[43]   A. Østhagen, "Maritime boundary disputes: What are they and why do they matter?," Mar. Policy, vol. 120, p. 104118, Oct. 2020, doi: 10.1016/J.MARPOL.2020.104118.

[44]   "Maritime Boundary Definitions | Geoscience Australia." https://www.ga.gov.au/scientific-topics/marine/jurisdiction/maritime-boundary-definitions (accessed 2022).

[45]   UNCLOS, United Nations convention on the law of the sea." 10 December 1982, doi: 10.1093/law/9780198715481.003.0002

[46]   "ITALY,"   www.un.org.
https://www.un.org/depts/los/LEGISLATIONANDTREATIES/STATEFILES/ITA.htm (accessed 2022).

[47]   "Maritime Boundaries Geodatabase, version 11.". Flanders Marine Institute, 2019. [Online].
Available: https://www.marineregions.org/. doi: https://doi.org/10.14284/382

[48]   "ArcGIS Server documentation—ArcGIS Server | Documentation for ArcGIS Enterprise."
https://enterprise.arcgis.com/en/server/latest/get-started/windows/about-arcgis-server-
documentation.htm (accessed 2022).

[49]   "Copernicus   Services | Copernicus."   https://www.copernicus.eu/en/copernicus-services
(accessed 2022).

[50]   R. Lecci, M. Drudi, A. Grandi, S. Cretì, and E. Clementi, "PRODUCT USER MANUAL For Mediterranean
Sea Physical Analysis and Forecasting Product MEDSEA_ANALYSISFORECAST_PHY_006_013 Issue: 2.1",
doi: 10.09.2021.

[51]   R. Lecci, S. Salon, G. Bolzon, and G. Cossarini, "PRODUCT USER MANUAL For Mediterranean Sea
Biogeochemical Analysis and Forecasting Product MEDSEA_ANALYSISFORECAST_BGC_006_014 Issue:
2.1," 2021, doi: 10.09.2021.

[52]   A. Teruzzi, P. Di Cerbo, G. Cossarini, E. Pascolo, and S. Salon, "Parallel implementation of a data
assimilation scheme for operational oceanography: The case of the MedBFM model system," Comput.
Geosci., vol. 124, pp. 103–114, Mar. 2019, doi: 10.1016/J.CAGEO.2019.01.003.

[53]   G. Büttner et al., "Copernicus Land Monitoring Service CORINE Land Cover User Manual CORINE
Land Cover Product User Manual (Version 1.0) Lead service provider for data production: NT (National
Teams) within EIONET National Reference Centres Land Cover (NRC) with support from EEA/ETC ULS
(European Environment Agency/European Topic Centre Urban Land and Soil) User Manual prepared by",
Accessed: 2022.

[54]   "Imperviousness — Copernicus Land Monitoring Service." https://land.copernicus.eu/pan-
european/high-resolution-layers/imperviousness (accessed 2022).

[55]   C. Tobias LANGANKE, "HRL Imperviousness Product Specifications Document Copernicus land
monitoring service-High Resolution Layer Imperviousness: Product Specifications Document Title
Copernicus Land Monitoring Service-High Resolution Layer Imperviousness: Product Specifications
Document".

[56]   "Copernicus Land Monitoring Service User Manual", Accessed: 2022. [Online]. Available:
https://land.copernicus.eu/user-corner/technical-library/imperviousness-2018-user-manual.pdf

[57]   "Forests   —   Copernicus   Land   Monitoring   Service."   https://land.copernicus.eu/pan-
european/high-resolution-layers/forests (accessed 2022).

[58]   C. Tobias LANGANKE, "Copernicus Land Monitoring Service-High Resolution Layer Forest: Product
Specifications Document Title Copernicus Land Monitoring Service-High Resolution Layer Forest: Product
Specifications Document".

[59]   "Copernicus Land Monitoring Service User Manual Consortium Partners - Tree-cover/forest and
change 2015-2018", Accessed: 2022. [Online]. Available: https://land.copernicus.eu/

[60]   "Water & Wetness — Copernicus Land Monitoring Service." https://land.copernicus.eu/pan-european/high-resolution-layers/water-wetness (accessed 2022).

[61]   P. Specifications and C. Tobias LANGANKE, "HRL Wetness & Water Product Specifications Document Copernicus Land Monitoring Service-High Resolution Layer Water and Wetness Copernicus land monitoring service-High Resolution Layer Water and Wetness: Product Specifications Document Title Copernicus Land Monitoring Service-High Resolution Layer Water and Wetness: Product Specifications Document".

[62]   "Copernicus Land Monitoring Service - Lot4: Water & Wetness 2018", Accessed: 2022. [Online]. Available: https://land.copernicus.eu/user-corner/technical-library/water-wetness-2018-user-manual.pdf

[63]   "Copernicus Local Land Monitoring Services – EEA/DIS/R0/18/008 Production of Very High Resolution Land Cover/Land Use dataset for Coastal Zones of the reference years 2012 and 2018", Accessed: 2022. [Online]. Available: https://land.copernicus.eu/user-corner/technical-library/coastal-zones-nomenclature-and-mapping-guideline.pdf

[64]   "Glossary:Population grid cell - Statistics Explained." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Population_grid_cell (accessed 2022).

[65]   "WorldPop :: Population Counts." https://www.worldpop.org/geodata/listing?id=69 (accessed 2022).

[66]   F. R. Stevens, A. E. Gaughan, C. Linard, and A. J. Tatem, "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data," PLoS One, vol. 10, no. 2, p. e0107042, Feb. 2015, doi: 10.1371/JOURNAL.PONE.0107042.

[67]   "Angular." https://angular.io/ (accessed 2022).

[68]   "OpenLayers - Welcome." https://openlayers.org/ (accessed 2022).

[69]   "Welcome to Python.org." https://www.python.org/ (accessed 2022).

[70]   "The web framework for perfectionists with deadlines | Django." https://www.djangoproject.com/ (accessed 2022).

[71]   V. J. Fabry, B. A. Seibel, R. A. Feely, and J. C. Orr, "Impacts of ocean acidification on marine fauna and ecosystem processes," ICES J. Mar. Sci., vol. 65, no. 3, pp. 414–432, Apr. 2008, doi: 10.1093/ICESJMS/FSN048.

[72]   "Partial pressure of carbon dioxide - OzCoasts." https://ozcoasts.org.au/indicators/biophysical-indicators/water_column_partial_pressure/#footnote_0_2942 (accessed 2022).

[73]   "Ocean Acidification | Smithsonian Ocean." https://ocean.si.edu/ocean-life/invertebrates/ocean-acidification#section_77 (accessed 2022).

[74]   "Ocean acidification | National Oceanic and Atmospheric Administration." https://www.noaa.gov/education/resource-collections/ocean-coasts/ocean-acidification (accessed 2022).