D3.1 Demonstrator general technical requirements

Work Package	WP3, Blue-Cloud Pilot Demonstrators
Lead Partner	IFREMER
Lead Author (Org)	MAUDIRE Gilbert MAUDIRE (IFREMER) & NYS Cécile (IFREMER)
Contributing Author(s)	BLONDEL Emmanuel (FAO), COCHRANE Guy (EMBL), DRUDI Massimiliano (CMCC), ELLENBROEK Anton (FAO), EVERAERT Gert (VLIZ), GENTILE Aureliano (FAO), LECCI Rita (CMCC), PESANT Stéphane (MARUM), SCHEPERS Lennert (VLIZ), TYBERGHEIN Lennert (VLIZ)
Reviewers	Dick Schaap (MARIS) & Pasquale Pagano (CNR)
Due Date	31.01.2020, M4
Submission Date	07.02.2020
Version	1.0

Dissemination Level

X	PU: Public
	PP: Restricted to other programm participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Blue-Cloud - Piloting Innovative services for Marine Research & the Blue Economy - has received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue-Cloud services, Grant Agreement n. 862409.

DISCLAIMER

"Blue-Cloud, Piloting Innovative services for Marine Research & the Blue Economy" has received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue-Cloud services, Grant Agreement n.862409.

This document contains information on Blue-Cloud core activities. Any reference to content in this document should clearly indicate the authors, source, organisation, and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the Blue-Cloud Consortium, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

COPYRIGHT NOTICE

VERSIONING AND CONTRIBUTION HISTORY

Version	Date	Authors	Notes
0.1	25.11.2019	Cécile NYS	Layout
0.2	16.12.2019	Cécile NYS	Drafting
0.3	13.01.2020	Cécile Nys	Storage and computing services & Data access and sources
0.4	15.01.2020	Gilbert Maudire	Introduction, WP3 summary, Summary of requirements & Software and services
0.5	15.01.2020	Cécile Nys	First revision & Appendix
0.6	24.01.2020	Guy Cochrane (EMBL)	Integration demonstrator #2 revisions
0.7	27.01.2020	Anton Ellenbroek (FAO)	Integration demonstrator #4 & #5 revisions
0.8	27.01.2020	Gilbert Maudire & Cécile Nys	Revision after Amsterdam Technical meeting
0.9	29.01.2020	Lennert Schepers	Integration of demonstrator #1 revisions
0.10	29.01.2020	Cécile Nys	Assembling of different revisions & Revision text and layout
0.11	03.02.2020	Pasquale Pagano	First reviewer
0.12	06.02.2020	Dick Schaap	Second reviewer
1.0	06.02.2020	Gilbert Maudire & Cécile Nys	Final version

Contents

Executive summary

The Blue-Cloud innovation potential will be explored and unlocked by developing five dedicated Demonstrators as Virtual Labs together with excellent marine researchers. For that purpose, Blue-Cloud selected five varied and domain-coverage rich scientific demonstrators. The objective is to develop and deploy a Virtual Lab for each demonstrator, which will become part of the Blue-Cloud Virtual Research Environment (VRE) and as such will be made accessible for users. The D4Science e-infrastructure will be the core platform for the VRE. This e-infrastructure already holds a generic VRE with many core services for building and running multiple Virtual Labs, dedicated to specific research targets. The D4Science e-infrastructure also has proven solutions for connecting to external computing platforms and means for orchestrating distributed services, which will be instrumental for smart connections to the other e-infrastructures in Blue-Cloud, namely EUDAT and DIAS (WekEO).

This deliverable D3.1 describes and summarizes the technical requirements towards the development and implementation of the Blue-Cloud demonstrators as Virtual Labs. These technical requirements have been derived from communications with the partners who are in charge of the Blue-Cloud Demonstrators. This implicates that the technical requirements are assessed and formulated from their perspective concerning the input, output, and workflow processes that are envisioned for the Demonstrators.

These requirements will be used in the further development of the Virtual Labs, which will take place in a close cooperation between WP3 and WP4, whereby WP4 will focus on the activities for making the existing VRE at D4Science more fit for purpose of the Blue-Cloud Demonstrators, which might involve expanding and/or upgrading functionalities. The technical requirements as summarized in this deliverable D3.1 will also contribute to the identification of the key challenges to address with the design of the overall technical architecture for the Blue-Cloud project.

As a next step in the cooperation and exchanges between WP3 and WP4, the Demonstrator groups will explore the existing VRE at D4Science, which will be supported by Webinars introducing and explaining VRE services. Moreover, the deliverable D3.2 ‘Demonstrator Implementation guidelines’ will be drafted in cooperation with WP4 and from the perspective of the existing D4Science e-infrastructure.

Abbreviations

Abbreviations	Signification
CPU(s)	Central Processing Unit(s)
EOSC	European Open Science Cloud
EOV	Essential Ocean Variable(s)
GIS	Geographical Information System
GPU(s)	Graphics Processing Unit(s)
OGC	Open GIS Consortium
VRE	Virtual Research Environment / Virtual Lab / Science Gateway
WP	Work-Package(s)

1 Introduction

This Blue-Cloud work-package #3 (WP3) deliverable D3.1, 'Demonstrator general technical requirements', summarizes the requirements from the demonstrators. These demonstrators will be developed, implemented and operated as Virtual Labs integrated on the Blue-Cloud Virtual Research Environment (VRE):

- Demonstrator #1 – Zoo- and Phytoplankton EOV products (led by VLIZ)
- Demonstrator #2 – Plankton Genomics (led by EMBL)
- Demonstrator #3 – Marine Environmental Indicators (led by CMCC)
- Demonstrator #4 – Fish, a matter of scales (led by FAO)
- Demonstrator #5 – Aquaculture Monitor (led by FAO)

This document compiles the **demonstrators' requirements** about:

- **Computer infrastructure capacity**: the necessary storage and computing capacity;
- **Existing software components**: the operating systems, software development languages and background API's which will be used and that already exist and are provided by computer vendors, or as Open Source components;
- **Data Sources and infrastructures**: The Data sources to be engaged by the demonstrators and the related access modes;
- **To-be implemented software and services**: The software and services - yet not fully provided - that have to be developed, upgraded or implemented for performing the demonstrators.

This document highlights, by order of priority:

1. The requirements requested by several demonstrators;
2. The high-priority requirements for a demonstrator, that are mandatory so that the demonstrator can be implemented;
3. The useful requirements but not mandatory.

These technical requirements are derived from the perspective of the partners who are in charge of developing the Demonstrators. The insights gained about the Demonstrators and their requirements will also contribute to the identification of the key challenges to address with the design of the overall technical architecture for the Blue-Cloud project.
The objective is to develop and deploy a Virtual Lab for each demonstrator, which will become part of the Blue-Cloud Virtual Research Environment (VRE) and as such will be made accessible for users. The D4Science e-infrastructure will be the core platform as this already provides a generic VRE with many core services for building and running multiple Virtual Labs.
The further development of the Blue-Cloud Virtual Labs will take place in a close cooperation between WP2, WP3 and WP4. WP2 will focus on building a common and smart Blue-Cloud data gateway and design the overall Blue-Cloud architecture, and WP4 will focus on the activities for making the existing VRE at D4Science more fit for purpose of the Blue-Cloud Demonstrators, which might involve expanding and/or upgrading functionalities.

2 Work Package 3 "Blue-Cloud Pilot Demonstrators" summary

The Blue-Cloud innovation potential will be explored and unlocked by developing five dedicated Demonstrators as Virtual Labs together with excellent marine researchers. For that purpose, Blue-Cloud selected five varied and domain-coverage rich scientific demonstrators:

- Demonstrator #1 – Zoo- and Phytoplankton EOV products.
  This demonstrator aims to produce phytoplankton, zooplankton and nutrients EOV products. They will contribute to improve knowledge and vastly reduce uncertainty regarding the present state of the marine plankton ecosystems and their response to ongoing and future climate change.
  The work plan includes:
    - Identifying and retrieving plankton and environmental data from various Blue-Cloud data resources;
    - Implementing a workflow to apply big data analysis and machine learning (e.g. neural networks).
- Demonstrator #2 – Plankton Genomics.
  This demonstrator aims to highlight a deep assessment of plankton distributions, dynamics and fine-grained diversity to molecular resolution by working across biomolecular, image and environmental data domains.
  It will focus on two areas:
  1. Discovery of undescribed biodiversity from genetic and morphological signals from the characterization of their geographical distributions, co-occurrences/exclusions and correlation with environmental contexts.
  2. Exploration of genetic and morphological markers of plankton diversity and abundance.

  The work plan includes
    - the construction of two notebooks making available and usable the intersection and link between biomolecular data, microscopy images and environmental datatypes;
    - the leveraging of workflows around functional genetic inference and image feature classification.

- Demonstrator #3 – Marine Environmental Indicators.
  This demonstrator aims to calculate and distribute online information and indicators on the environmental quality of the Mediterranean Sea, which will serve intermediate users such as environmental protection agencies. Tests will be conducted to extend the geographical area to the North-East Atlantic and demonstrators 1, 4 and 5, will use the results.
  The work plan includes the definition and the production of a set of marine environmental indicators (data products) such as:
    - Environmental indicators: Chl-a trends, TRIX, Chl-a statistical analysis, etc.;
    - Biological variables: e.g. plankton biomass and diversity;
    - Ocean products: wave, currents, temperature, salinity, Chl-a, primary production, etc.;
    - Wind and atmospheric variables;
    - Inorganic carbon indicators.
- Demonstrator #4 – Fish, a matter of scales.
  This demonstrator aims to expand data management and analytic capabilities for fisheries by:
    - Expanding the existing Virtual Lab for the FAO Tuna Atlas (tuna and billfish catch data) into the Fisheries Atlas. The Atlas should have features for data analysis (using indicators, interactive maps or dashboards, state-of-the-art analytical models);
    - Expanding the existing Global Record of Stocks and Fisheries (GRSF) Virtual Lab with new stocks and include and/or link to the results of approved status assessments of fisheries, including those from the Fisheries Atlas and other demonstrators.

  The Tuna Atlas work plan relies on a complete Spatial Data Infrastructure that already adopts OGC/INSPIRE standards, and includes widely used software components (PostgreSQL base, Geonetwork, Geoserver, Web mapping). It is thus capable to ingest and merge fisheries metadata and data with compliant data from other databases or Blue-Cloud data resources. The Tuna Atlas VRE can also expose data to the Global Record of Stocks and Fisheries VRE (GRSF) using UUIDs. The semantic Knowledge Base (KB) behind the GRSF allows organizing and merging information and endorse the content in a true publishing workflow.
- Demonstrator #5 – Aquaculture Monitor.
  This demonstrator aims at developing the remote sensing data capacity for the monitoring of aquaculture in marine cages and in coastal areas. The ambition is to deliver a tool to produce online national aquaculture sector overviews using OGC compliant data services also accessible through the Demonstrator #4 integrated Open FAIR Viewer.
  The work plan includes the use of the existing Virtual Lab for the data presentation and editing, including an interactive map-viewer: the Aquaculture Atlas Production System (AAPS). The AAPS already serves to monitor two major aquaculture types: Mariculture in floating cages in the Mediterranean, and coastal pond-based aquaculture tropical areas. For each, a specialized detection algorithm is available, and the output is available in an interactive web-based mapping tool. Both algorithms will be improved.

The work to be undertaken by demonstrators will be conducted in four steps:

- Step 1: Identifying the requested technical requirements.
  The Demonstrators will be analysed in detail with the involved researchers in order to describe their workflows and to use these for designing the Virtual Labs. This is done from the perspective of the Demonstrators.
- Step 2: Implementing the first version of the demonstrators.
  This first version will be based on pre-existing services, but implemented in a standardized common environment and relying, as far as possible, on common components to be set up mainly by WP2 and WP4. While developing, much interaction will take place between the researchers (WP3) and the technical developers (WP2 & WP4) in order to test and fine tune prototypes.
- Step 3: Second version of the demonstrators.
  Focusing on innovative development for the second versions of the Demonstrators, as more Blue-Cloud functionality will become available in WP2 and WP4.
- Step 4: Promoting and demonstrating the achieved demonstrators.
  The final phase leading up to the project completion will be dedicated to promote and to demonstrate the achieved demonstrators at different events which will be organized as part of WP5. Maintenance and support to the users will also be supported.

This document, D3.1 ‘Demonstrator general technical requirements’, is the final deliverable of step one.

3 Technical requirements

3.1 Common requirements between demonstrators

Since the objectives of the demonstrators are derived from their thematic fields, objectives and/or targeted users, the requirements may be very different. For example, some components of the demonstrators are oriented towards research teams that need to explore data using their own scripts and to develop their own algorithms, when some other demonstrators aim to provide finalized indicators to end users. In the first case, scientific teams require flexibility in their developments (e.g., use of notebooks, of API's, etc.) when in the second case, repeatability of results is the objective (e.g. use of well-established processing chains).
However, some requirements are common between several demonstrators

- Most of the demonstrators would like **to access data and** sometimes **products from several different distributed infrastructures** such as CMEMS, several components of EMODnet, SeaDataNet, ERIC –Argo, ELIXIR-ENA, etc.). Studying how the access to and the processing of the datasets provided by those infrastructures might be facilitated and harmonized is a point shared by most of demonstrators. This point includes several topics which may be considered with all project partners (especially infrastructures that provide data) in order to harmonized the approaches and the provided services:
    - Catalogue of data and related services;
    - Access services and API's to these distributed and potentially large datasets
- Since most of the data are **georeferenced data**, all the demonstrators use several features of the **Geographic Information Systems;**
- Almost all demonstrators request **data visualization tools** (e.g. bar chart, pie chart, time series, etc.);
- All the demonstrators, except demonstrator 5, will make use of Artificial Intelligence methods such as **machine learning and/or deep learning.**

Detailed below are these common requirements with some tentative demonstrators' implementation examples.

3.1.1 Data and services catalogue

Accessing to various data sources implies the need of a **catalogue** that provides **discovery metadata and services** and that describes the available datasets and products, their respective access conditions and the associated technical services to access them.
This catalogue should describe **"Data Collections"** (e.g. not each observation, but all observations of the same types that have been harmonized by the data infrastructures) and **"Product Families"** (e.g. a model output but not every temporal occurrences of the output), which means not very fine granules.

Since most of the data are georeferenced data, **ISO 19115 family of standards** seems to be appropriate for implementing this catalogue.

Since some of the **datasets** have **restricted access**, with specific licenses, or controlled access that requires users' identification, access conditions must be detailed in the catalogue.
Descriptions should include **links to available data services in order to allow machine-to-machine data access** (e.g. by using the service description supported by the ISO standards).
This catalogue should be populated by harvesting the existing catalogues already made available by data infrastructures (using CSW – Catalog Service for the Web, or alternatively, OAI-PMH protocols). Demonstrators will make use of GeoNetwork or CKAN for catalogue maintenance.
The descriptions of data sources (discovery metadata) should adopt **common vocabularies. The use of EOV's** appears to be relevant for describing datatypes and adequate for the expecting level of granularity of the catalogue since they are already used by most of the demonstrators.

3.1.2 Data access services and API's

Some data sources are inputs for almost all demonstrators (e.g. CMEMS, SeaDataNet, Argo, EMODnet, etc.) or based on or accessed through different infrastructures (e.g. DIAS WekEO, ELIXIR-ENA. Consequently, these data sources could be considered as reference data to be shared and not duplicated among each demonstrator and users in their virtual labs. In other terms, it is desirable to provide a machine-to-machine Blue-Cloud optimized access framework to these frequently asked data instead of selecting and downloading subsets.

Some demonstrators already have implemented tentative solutions:

- Using data access protocols and API's:
  - Use of common access standards such as OpenDAP, OGC standards (WMS, WFS, WCS).
  - Set up of "data brokers" on top of data sources and repositories. INTAKE (https://www.anaconda.com/intake-taking-the-pain-out-of-data-access/) has been cited as a possible example.
  - Harmonization of data sources by common formatting and possibly using "user oriented" formats (e.g. No-SQL Databases, Data Cubes, etc.);
- Using "**Cloud storage standards**" (such as Virtual File Systems) to access remote files or eventually data objects as if they were located on the user's computer and not remotely accessed (CEPH, iRODS).

Because the proposed solution(s) must be as harmonized as possible, the overall Blue-Cloud architecture should consider this point and propose common guidelines. These guidelines could either be implemented by the demonstrators, by the infrastructures managing these reference datasets and/or by the Blue-Cloud VRE.

As some of these datasets are very large **distributed datasets** - meaning large enough to be very difficult to transfer over networks - several demonstrators run their software and the scripts on IT infrastructures that store the largest datasets, using:

- Synchronize data copies on different IT infrastructures;
- Virtual file systems or cloud storage systems (CEPH, iRODS);

- Thredds servers with OPeNDAP and NCML formats for virtual files;
- OGC-WPS protocol within a GIS context;
- BINDER, https://mybinder.org/, to run a notebook script on a remote computer).

For example, a sustainable and easy access to satellite images is still an issue mainly related to security (e.g. Copernicus satellite images, etc.). Demonstrator 1, 4 and 5 will use satellite images and will need, not only statics images, but also time series to elaborate and identify trends (e.g. reduction of coastline, rising Sea Surface Temperature, etc.).
Finally, some of the datasets have restricted access (e.g. high-resolution satellite imagery) with specific licenses and other datasets are accessible only to registered users (e.g. Copernicus, SeaDataNet). User authentications may vary from one data provider to another (e.g. Protocol CAS, SAML, OAuth2.0, etc.). Consequently, there are issues related to authentication and security (e.g. OCG protocols designed by default to access open data).

These requirements, as collected from the Demonstrators, provide valuable input for the activities in WP2 that has its focus on developing the overarching Blue-Cloud data discovery and access service, giving overview and smart access to the blue data infrastructures that are part of the Blue-Cloud.

3.1.3 Geographic Information systems

Since most of the data are **georeferenced data**, all the demonstrators will use several features of the **Geographic Information Systems**. There are similar needs on the use of GIS protocols for demonstrator 1, 2, 3 and 5. The demonstrator 1, 2, 4 and 5 will make use of GIS visualization for different sources and products. The main functionalities to be used are:

- Data visualization (maps) from different sources (inputs) and visualization of products and indicators (output of demonstrators). In addition, GIS's offer overlaying of different maps to compare data visually, this might require https or CORS Web protocols to be enabled;
- Implementation of interoperability using OGC standards (e.g. access to remote high-resolution satellite images): WMS protocol for images in demonstrator 5, or WCS protocol and standardization of dataflow (WFS being used for vector data);
- Run of processing tools using WPS protocol (demonstrator 4 and 5).

3.1.4 Data Visualization tools

Almost all demonstrators request **data visualization tools** (e.g. bar chart, pie chart, time series, etc.). These tools might be used in different contexts, such as in a script (e.g. using the Pangeo toolbox in Python), to generate a web page (e.g. using R-Shiny) or on a map (species distribution at several locations). Since datatype may differ from one demonstrator to another, several demonstrators suggest offering generic tools, which can accept the entrance of **multi-dimensional data tables.**

3.1.5 Artificial intelligence data processing methods

All the demonstrators will make use of Artificial Intelligence methods such as **machine learning and/or deep learning** for image pattern detection, data analyses and generation of indicators, etc. Providing toolboxes that implement these algorithms, preferably in a notebook context, seems to be very relevant.

That also may involve important storage capacity with **adapted data structures** like **Data Cubes**, even for data samples, specifically for building machine-learning databases. It must also be noticed that, for some demonstrators (e.g. demonstrator #5), the machine-learning phase is not included in the demonstrator itself (computed separately on other e-infrastructures) but imported in the demonstrator as a "knowledge database". Separating this machine-learning phase decreases a lot the need of accessing massive datasets**.**

Using the AI methods requires in addition **fast CPUs / GPUs**.

3.2 High priorities requirements but not common

Some demonstrators require specific tools and contexts:

- Several of them will have a need of interpolation tools, such as **DIVA** (developed by the University of Liège).
- Demonstrator 2 makes use of **Environment Common Workflow Language** and **containerized software components** to run and chain bioinformatics processing.
- Demonstrators 2 will make use of **Ecotaxa** (developed by Sorbonne University) for microscopy image processing of plankton, also demonstrator 1 will check if Ecotaxa is useful for their demonstrator. Improved discovery and access to Ecotaxa is being worked out in cooperation with EurOBIS, which will provide the API('s) towards the Blue-Cloud data discovery and access service.
- Demonstrator 4 will extend the **GRSF semantic data model** to support new content (i.e. Tuna Atlas VRE) and linkage with different data sources (i.e. Food Cloud).
- The **Open Fair Viewer (OFV)** will display GRSF data; already a use case of geoflow+ofv and its inventory spatial datasets can be browsed in fisheries atlas together with other datasets.
- The Tuna Atlas has requirements regarding **Postgres**, **Jupyter**, facilitation of **Shiny apps deployment, RDF & SPARQL endpoint access** to controlled vocabularies, CORS & https protocols.
- Demonstrator 5, on top of 'traditional' satellite images, will need **Very High Resolution (VHR) satellite images**. As they develop their demonstrator, they will need more and more easy and **sustainable access** to time series of VHR satellite image. A sustainable and easy access to satellite images is still an issue (WMS protocols, standardized data flows, etc.) mainly related to security.
- Demonstrator 5 will need **Analysis Data Ready** (ARD) from **Sentinel images** (1 and/or 2) via DIAS platform, provided via WFS.

3.3 Links between demonstrators

Demonstrators 1 and 4 will **need** some marine environmental indicators produced by **demonstrator 3** (common visualization using GIS viewer at a minimum, computing of correlations).
The OpenFAIRViewer further developed in **Demonstrator 4 will be used in Demonstrator 5**, and can be used to display compliant datasets from e.g. Demonstrators 1 and 3, thus providing a holistic view over several Blue-Cloud domains.

3.4 Summary of requirements

3.4.1 Storage and computing capacity estimates / Technical capacity assessments

	Storage	**Computing Capacity**
Demonstrator 1 (D1) specific	Up to 100 GB	• 10-20 fast CPUs • 64GB RAM • 2 NVIDIA GPUs • 16 computing cores
Demonstrator 2 (D2) specific	Up to 10s of PB of data	• 1 TB local scratch disk • 64GB RAM • 32 vCPUs
Demonstrator 3 (D3) specific	• Several 10s of GB of data • Beginning stage: 150GB/product • Later stage: 3.5Tb/product • Maps and trends: 20GB	
Demonstrator 4 (D4) specific	GRSF • Up to 100 GB (for the triplestore with GRSF KB contents) • Up to 100 GB the published GRSF contents (managed by CKAN) Tuna Atlas / SDI • Up to 100 GB / < 1TB	Updating GRSF • 4 vCPUs • 32 GB RAM
Demonstrator 5 (D5) specific	Several 10s of GB of data	NVIDIA GPUs

3.4.2 Operating systems and languages

Most of the algorithms are implemented over UNIX operating systems or rely on services, which are implemented over UNIX operating systems. Several demonstrators make use of “Unix command lines” scripts (e.g. bash).

There is no specific request on the used file systems, more specifically there are no specific requests about object or block storage management. However, the following requests have to be noticed:

- **Data formats**: Extensive use of NetCDF format (CF convention), CSV Ascii format, and more specific file such as georeferenced images (compressed such as JPEG 2000 or not compressed);
- Management of large number of files;
- **Access to distributed data sources**, located on different data repositories such as Copernicus Dias (Wekeo), EUDat (SeaDataNet & some EMODNET thematical lots), etc.

The demonstrators use the following technologies (programming or scripting languages):

- R
- Java
- Python
- Fortran
- Julia
- Perl

These languages are often implemented using notebooks such as Jupyter notebooks, at least in the test phase of the algorithms, and R is used with RStudio IDE. Some software are already provided within containers (Docker). Demonstrators also point out that maintenance of software, associated documents and scripts must be part of the Blue-Cloud Virtual Labs. Tools already in use are mainly Git (GitHub, GitLab). For some demonstrators (e.g. Demonstrator 2 and Demonstrators 3), these software repositories are directly associated with the runtime environment (e.g. Galaxy, Jupyter notebooks) in order to edit, to containerize and to run scripts from one single environment.

3.4.3 Data access and sources (Infrastructures)

	Data Sources	Infrastructures
Common between several demonstrators	• Argo: salinity, oxygen, chlorophyll data • CMEMS: ocean color, altimetry, temperature and salinity field data, ocean and climate variables • EMODnet Chemistry: environmental data • EMODnet Biology: biodiversity data • SeaDataNet: biogeochemistry, physics, biology, environmental data	• Euro-Argo & Argo GDAC • SeaDataNet • WekEO • EcoTaxa
Demonstrator 1 (D1) specific	• EurOBIS: abundance zooplankton data • EcoTaxa: Ecological images of plankton • LifeWatch: BioOracle ecological modelling and observatory sensor data • WORMS • MAREDAT: phytoplankton diversity information • GLODAP V2: nutrient data	• EurOBIS • LifeWatch
Demonstrator 2 (D2) specific	• Tara: Arctic, Oceans (global ocean, plankton), Pacific (pacific ocean, coral reefs) & Mission microplastics • ELIXIR-ENA: genomics data • EcoTaxa: high resolution precision microscopy images • BioImage Archive: high resolution precision microscopy images	• Tara: Tara Arctic, Tara Oceans, Tara Pacific • ELIXIR-ENA • EuroBioImaging
Demonstrator 3 (D3) specific	• ICOS-Marine: inorganic carbon data • CMEMS Med MFC: Temperature and salinity – Mediterranean Sea – daily means 1990-2009 • CMEMS and C3S: ocean and climate variables	• ICOS-Marine

Demonstrator 4 (D4) specific	• FAO: Fisheries statistics (capture, commodities, , etc.) and stocks • uFish2/Infoods/Codex: food composition data • BlueBRIDGE RDB: regional datasets • FIRMS: stocks & fisheries inventories • FishSource • AIS/VMS data and other data related to trajectories (eg FADs, , etc.)	• FoodCloud • FIRMS • INFOODS • FAO Geonetwork • GFW
Demonstrator 5 (D5) specific	• Sentinel 1 (S1) and Sentinel 2 (S2) • VHR images	• FIRMS • WekEO • CLS

3.4.4 Software and services

	Software & services
Common between demonstrators	• Discovery metadata and catalogues • Common Vocabulary server (EOV's, etc.) • API, access protocols or virtual File System for an harmonized access to distributed and potentially large dataset • GIS functionalities (online visualization, services, , etc.) • Data visualization (generic charts, charts on maps, , etc.) • Notebooks (Jupyter, R Studio) for scripting • Data analyses based on IA methods and data structures adapted to these methods (e.g. data cubes) • RShiny • OpenFairViewer
Demonstrator 1 (D1) specific	• Diva interpolation software
Demonstrator 2 (D2) specific	• Ecotaxa software • Environment CWL engine
Demonstrator 3 (D3) specific	• Diva interpolation software • Machine Learning / Deep Learning (TensorFlow, etc.)
Demonstrator 4 (D4) specific	• Species Discovery Service • CSW protocol for data discovery • Geonetwork and Geoserver APIs • SPARQL endpoint for generic public databases (i.e. DBpedia)
Demonstrator 5 (D5) specific	• Machine Learning / Deep Learning (Keras, TensorFlow, , etc.)

4 Conclusions

At this stage of the implementation of the Blue-Cloud, all the requirements recorded in this document should be considered as a “wish list”, developed solely from the point of view of the demonstrators. They have not yet been confronted to the already implemented components provided by the Blue-Cloud partners, especially by WP2 and WP4. Consequently, this document is provided as an input, to be confronted with others, for the elaboration of the overall Blue-Cloud architecture. Consequently, some of the requirements, facing the feasibility principles may be discarded, or alternate solutions proposed.

When the overall architecture of Blue-Cloud will be designed, the implementation of these requirements could then be made alternatively by the technical work-packages of the project, the infrastructures that are partners of the project, or by the demonstrators themselves. Guidelines for this implementation will be provided to demonstrators by the deliverable D3.2 “Guidelines for demonstrators”.

5 Appendix

5.1 Demonstrators – Kick-off synthesis

Blue Cloud – Demonstrators – Kick-off synthesis – V2

Table of content

1. Aim of demonstrator

D1	Plankton EOV demonstrator A machine learning approach using Blue Cloud data and Services to derive zooplankton diversity products and phytoplankton biomass and diversity products from *European seas and the Global Ocean.*
	Zooplankton EOV products This demonstrator will produce new and complete data products on **zooplankton** distributions and abundances, focusing on the most abundant Copepod species (*Calanus h., Calanus f., Acartia spp, Oithona s., Temora l. , Metridia l.*) from the North East Atlantic. The products will be spatio-temporally completed by using environmental parameters (Oxygen, temperature, salinity, chlorophyll, depth, nutrients) in a neural network model. The workflow that uses and combines several large scale marine biological and environmental data from different providers in this neural network model, including near real-time observations will be made available in the cloud.
	Phytoplankton EOV products To generate **global 3D key biogeochemical EOVs, i.e. phytoplankton biomass & diversity, nutrients**, that will contribute to improve our understanding of the marine phytoplankton component and its biogeochemical impact in relation to environmental conditions and to predict its response to climate change
D2	Plankton Genomics & Imaging • Discovering unknown “species” and functions & associated ecosystem functions and services • Predicting the distribution of known & unknown “species” and functions & provide high-resolution EOVs about plankton biodiversity & ecology This demonstrator will bring a thorough understanding of integration of data across microscopy imaging, molecular biology and environmental platforms, including a knowledge of how future data should be collected and structured to allow cross-data type integration while retaining relevance and consistency within each respective data type; The demonstrator will enable scientific exploration of plankton, including correlating plankton concentration and diversity with local environmental variables, deriving known and new indices of ecosystem health, and predicting the distribution of these variables in space and time.
D3	To provide and display information and indicators on the environmental quality of the Mediterranean Sea (later addition of NE Atlantic). To integrate different data sources and exploit scientific based algorithm (included machine learning ones) for the online calculation of environmental indicators. To deliver and document the associated computation routines for re-use as part of the Blue Cloud Virtual Lab.
D4	The Global Record of Stocks and Fisheries (GRSF) Virtual Lab will be expanded to contain the results of approved status assessments of fisheries, including those from the Fisheries Atlas. It will expose relevant information on fish food and nutrition though GRSF, either as static information or dynamically loaded. For monitoring global stocks status (SOFIA, SDG14.4.1, regional monitoring systems) To support traceability in the value chain GRSF aims at supporting the collaborative production and maintenance of a comprehensive and transparent inventory of stocks and fisheries • Integrate records from different data sources • Assign global identifiers to records (both machine and human interpretable) • Expose GRSF records in an homogenous way To strengthen and extend common methodologies and tools in support of Fisheries(#4) and Aquaculture Atlases (#5), by implementing: • International Standards: ISO, OGC, W3C • FAIR Principles (Findable Accessible Interoperable and Reusable)

D5	Expand the remote sensing analytic capabilities of the existing Aquaculture Atlas Production System (AAPS) Virtual Lab in order to provide a robust and replicable environment for monitoring aquaculture in marine cages and in coastal areas.

2. Scientific context / Present situation

D1	Plankton EOV demonstrator • Machine learning-based algorithms are available, have been tested at limited scale as part of EMODnet Biology for ICES-OOPS products, and scientific publications have been produced. • Data sources are existing, but currently not all accessible in an efficient way.
	Zooplankton EOV products
	Phytoplankton EOV products Owing to recent advances in sensor and robotic platform technologies along with an implementation onto global observation programs (e.g. OceanSITES, BGC-Argo), the space-time resolution and coverage of certain EOVs has dramatically increased. Yet, **some EOVs still suffer from critical undersampling** due to technological limitation. **Machine learning methods** represent an **alternative** to derive, from easily remotely (satellite or in situ) measured variables, more complex variables that are not yet amenable to autonomous measurements, i.e. so-called **"virtual variables"**
D2	1. Distinct data archives and infrastructures exist for imaging, molecular and environmental data with variable levels of maturity 2. In the case of Tara Oceans, different data types share common metadata of high quality 3. BOTTLENECK: Lack of platforms to integrate and compute across data types
D3	• Web site with a collection of preliminary set of indicators http://www.marinenvironment.com • Set of pre-calculated datasets and indicators are available on the above mentioned website and at the other partners premises • Algorithm for the online calculation of some MSFD - Marine Quality Indicators were already tested in a cloud environment at limited scale – EU MELODIES Project, https://github.com/ecmelodies/wp06-ges-toolbox • Machine learning-based algorithms are available and have been tested by IFREMER mainly on Argo float data • Data sources are existing, but not accessible in an efficient and fully integrated way and indicators cannot be calculated on line in a dynamic mode • Uncertainty is not associated to each indicator
D4	The Global Record of Stocks and Fisheries (GRSF) has been implemented in the context of H2020 BlueBRIDGE project (GA No. 675680)
D5	

3. Data to be used (Which data ? Format ? Sources ? Existing or yet to be produced ?)

D1	Data sources	• Blue Cloud sources:
		  • Euro-Argo and Argo GDAC: salinity, oxygen, chlorophyll data
		  • SeaDataNet: physics, biogeochemistry, biology data
		  • CMEMS (DIAS-WEkEO): ocean color, altimetry data
		  • EurOBIS: abundance zooplankton data
		  • EuroBioImaging - EcoTaxa: ecological images of plankton
		• Additional sources via existing services:
		  • LifeWatch: BioOracle ecological modelling data
		  • LifeWatch: observatory sensor data
		  • LifeWatch: World Register Marine Species
		• Additional sources via ad-hoc retrieval:
		  • MAREDAT: phytoplankton diversity information
		  • GLODAP V2: nutrient data
		  • Ocean color products

D2	Data Sources	Expeditions of the Tara Ocean Foundation • Tara Arctic • Tara Oceans : global ocean, plankton • Tara Pacific : pacific ocean, coral reefs Mission microplastics : 10 rivers in Europe, Microplastics
		Blue Cloud sources: • ELIXIR-ENA: genomics data • EuroBioImaging – EcoTaxa: high-resolution precision microscopy images • EuroBioImaging BioImage Archive: high-resolution precision microscopy images • SeaDataNet: environmental data • CMEMS (DIAS-WEkEO): environmental data • Euro-Argo and Argo GDAC: environmental data • EMODnet Biology: biodiversity data EMODnet Chemistry: environmental data

D3	Data sources	• CMEMS Med MFC : Temperature and Salinity field, Mediterranean Sea, daily means 1990-2009 • CMEMS and C3S (DIAS-WEkEO): ocean and climate variables • Euro-Argo and Argo GDAC: salinity, oxygen, chlorophyll data • SeaDataNet: physics, biogeochemistry, biology data • EMODnet: physics, biology, chemistry data • ICOS-Marine: inorganic carbon data
D4	Data sources	Fisheries and Resource Monitoring System (FIRMS) RAM Legacy Stock Assessment Database FishSource (program of the Sustainable Fisheries Partnership) Other RDBMS than IRD Sardara / BlueBridge RDB Simple sources (eg CSV, ESRI Shapefiles) OpenDAP / NCML
D5		Fisheries and Resource Monitoring System (FIRMS) RAM Legacy Stock Assessment Database FishSource (program of the Sustainable Fisheries Partnership)

4. Based on which infrastructure ? (ELIXIR-ENA, EuroBioImaging, SeaDataNet, EMODnet, EurOBIS, Euro-Argo, EcoTaxa, DIAS, ICOS-Marine, BlueBRIDGE and/or EUDAT ?)

D1	
D2	ENA (European Nucleotide Archive) Copernicus BioIMAGE EMODnet
D3	• DIAS-WekEO • Euro-Argo and Argo GDAC • SeaDataNet • EMODnet • ICOS-Marine
D4	
D5	

5. Method

D1	Plankton EOV demonstrator
	Zooplankton EOV products The proposed demonstrator will build upon DIVA and (Barth et al., 2014; Troupin et al., 2012) and existing neural network-based methods (Yuret, 2016).Abundance data for selected zooplankton species will be derived from EurOBIS / EMODnet Biology. The neural network will use environmental data from different sources (Bio-ORACLE, Copernicus, SeaDataNet, …). Additionally the position (latitude and longitude) and the year are provided to the neural network.
	Phytoplankton EOV products The proposed demonstrator will build upon existing neural network-based methods (e.g. Sauzède et al. 2016), trained with a range of oceanographic data from multiple streams, made interoperable and integrated: MAREDAT HPLC pigment data, GLODAP v2 nutrient data, BGC-Argo and T/S Argo data (from GDAC), ocean color and altimetry satellite products

D2	Two Notebooks will be constructed that make available and usable the intersect between biomolecular, microscopy image and environmental data types, leveraging workflows around functional genetic inference and image feature classification, made possible through the central infrastructure provided by Blue-Cloud.
	Notebook 1 - Deliverables & perspective
	• List of abundant unknown sequences = markers from marine processes (e.g. carbon export, nitrogen fixation, endemic sequences)
	• A bioinformatic pipeline to link (meta)omics to ecosystems functions and services
	• More inclusive view of marine diversity and keystone « species »
	• Improve models (e.g. communities dynamics, biogeochemical cycles)
	Notebook 2 – Prediction
	• Extract species, morphologies, functions and combination thereof from TaraOceans data
	• Extract fields of environmental data from Copernicus (CMEMS)
	• Correlate them through machine learning models (a.k.a. ecological niche models, habitat models)
	• Project potential distribution of the items identified in Notebook 1
	• Requirements
	  ○ Access to Notebook 1 products
	  ○ Access to CMEMS fields
	  ○ (Limited) computational power
	  ○ R-based (or Python-based) notebook backends
	  ○ (Shiny, R-package, interface for browsing the products)

D3	MARINENVIRONMENT will be a service to provide and display information on the environmental quality of the Mediterranean Sea (in the early stage). The system will be designed on the basis of intermediate users' requirements (e.g environmental protection Agencies) and will be accessible via the website. A set of marine environmental indicators (data products) will be defined and produced. For each indicator an analysis and production workflow will be specified, developed and tested, applying big data analysis and machine learning methods on the multi-source data sets. The user interface will be developed to enable users to perform on line and on the fly operations, such as: selecting portion of dataset for a selected area in horizontal and vertical, to perform statistical analysis (e.g. trends, anomalies) on the selected variable and portion of space and time, to display these indicators by tables, map and graphics visualisations. This task will focus on the scientific and computational challenges of the Demonstrator, while the technological aspects will be covered in other BLUECLOUD WPs.	
D4	Activities	Set-up a fisheries atlas centralized catalogue to harvest from existing fisheries catalogues Foster set-up / exploitation of registries (controlled vocabularies) in support to dataset metadata harmonization and semantic interoperability, including: • ISO 19135-2 / 19139 registries for geographic items (eg INSPIRE) • RDF/SKOS Semantic registries Consolidate R generic workflow and related tools: • Upgrade workflow for Tuna Atlas/Sardara & RDB, in view of their official release • Set-up workflow for other fishery databases, eg RTTP Tuna tagging database, etc. • Set-up workflow for uses cases of raster data cubes in support of fisheries, and underlying tools (eg Thredds, Rasdaman, Geoserver) • Handle simple data formats: CSV, ESRI Shapefiles • Develop R Shiny interface to trigger the workflow in a user-friendly manner Consolidate data visualization & analysis tools to support diversity of data domains, sources and models: • Data viewer - Enhanced geo-visualization, with map export module • Data viewer - Support statistics browsing through tabular & statistical graph methods • Data viewer – Develop citation module • Data analysis tools with R Shiny Document methodologies and tools, with online user manuals and Communication (papers, workshops)
D5		

6. Which are the technical needs to be considered ?

D1	
D2	

D3	• Computing: to be assessed within first three months
	• Storage:
	  ○ input
	  ▪ at the beginning, usage of monthly data: 150GB (1 product) * #products
	  ▪ in a later stage, usage of daily data: 3.5TB (1 product) * #products
	  ○ output:
	  ▪ maps and trends: ~ 20G
	• Web services to access data and results:
	• Platform for computation, access and viewing of analysis performed by users in order to create added value data products

D4	Spatial Data Infrastructure (SDI)
	• made of widely used software components
	  ○ spatial databases (e.g. Postgis or Thredds)
	  ○ geographic data server (e.g. GeoServer)
	  ○ metadata catalogue (e.g. GeoNetwork)
	  ○ processing server (e.g. R / RStudio)
	• deployed in D4Science infrastructure with VREs:
	  ○ FAO Tuna Atlas, Aquaculture Atlas, SDI Lab
	• operated with R Workflow Tools
	  ○ R CRAN & GitHub packages
	• delivering ISO / OGC / INSPIRE compliant services
	  ○ Metadata (ISO 19115/19110/19139, OGC CSW)
	  ○ Data (OGC WMS/WFS/WCS)
	• exposed to data visualization tools
	Data models: in general, to be capable to handle and describe in standard way any type of spatio-temporal arrays (raster and vector data cubes)
	• Vector data models: not only Polygons but also Polylines / Transects / Points (eg fishery surveys, trawls/fishing operations, vessel trajectories, photos)
	• Raster / Imagery data models (EO data, model outputs)
	Operational VREs - Equipped with SDI components
	• FAO Tuna Atlas: https://bluebridge.d4science.org/group/fao_tunaatlas
	• WECAFC-FIRMS Regional Database: https://bluebridge.d4science.org/group/wecafc-firms
	• Aquaculture Atlas: https://bluebridge.d4science.org/group/aquacultureatlasgeneration
	• SDI Lab: https://bluebridge.d4science.org/group/sdi_lab
	Operational R Tools
	• R CRAN Packages for SDI operations:
	  ○ Metadata handing & publication: geometa (ISO/OGC metadata), EML (Ecological metadata), ncdf4 (CF conventions), rsdmx (SDMX), ows4R / geonapi (ISO/OGC Metadata publication)
	  ○ (Meta)Data publication: geosapi (with Geoserver), zen4R (with Zenodo)
	• R Legacy workflows for: Tuna Atlas, WECAFC-FIRMS, Aquaculture Atlas
	• R Generic workflow, enabled by the geoflow R package initiative, under consolidation
	Operational Data Viewer
	• OpenFairViewer initiative
	• Instances: Global Tuna Atlas, FIRMS Tuna Atlas, WECAFC-FIRMS RDB, SDI Lab (sandbox)

D5	Standardize data processes: • To enable CLS satellite data analyses as standard processes (OGC WPS) Apply R generic workflow: • To bind upstream processes with GIS (meta)data publication to OGC services Consolidate data visualization & analysis tools • Data viewer – Features development and use shared with Fisheries Atlas • Enable OpenFairViewer instance for Aquaculture atlas • Data analytics: Geospatial data processes (WPS) involving EO and Environmental monitoring data, and reused to enrich aquaculture atlas

7. What are the development stages ? (Timing)

D1	Year 1 (data collection + algorithm tests) • Collection and integration of global data from multiple sources and testing of the machine learning-based algorithms => generating prototype of global phytoplankton and nutrient EOV products. • Collection and integration of data from multiple sources and testing of the machine learning based algorithms => generating prototype of regional (North Atlantic) zooplankton EOV products. Year 2 (final production) • Phytoplankton and nutrient global EOV products and associated computation routine. • Zooplankton abundance regional (North Atlantic) • EOV products and associated computation routine.
D2	YEAR 1 : Design and test 2 prototype Notebooks (Discovery & Prediction) → • Prototype iterative Notebook design and testing; data and workflow requirements gathering and work with technical partners to provide these; standards and conventions for data integration across the data types • Gather requirements (incl. standards) for data integration and workflows • Work towards their implementation with technical partners YEAR 2 : Deliver 2 science-ready Notebooks → Move to production, science-ready Notebooks; socialisation with the relevant research communities; use by partners and the wider community to demonstrate scientific value; feedback into Notebook design and data sources to iterate further. • Train the relevant research communities (partner SU & wider) • Demonstrate scientific value by users (partners SU & wider) • Feedback into Notebook design YEAR 3 : Outreach and promotion of the 2 Notebooks → outreach and promotion of Notebook functionalities; extrapolation and enumeration of possible future functions enabled by the data integration achieved; community support; capture of experience and white paper highlighting the future of the approach. • Showcase functionalities • Highlight possible future applications (white paper)

	Next activities • Next 6 months: wait for Nb 1 to progress ;-) • Jan 2021: hire post-doc • By June 2021: data aggregated • By Oct 2021: prototype notebook • By Jan 2022: demonstrator ready
D3	Year 1 (data collection + algorithm improvement and tests + dev. of the cloud based online calculation tools) • Collection and integration of data from multiple sources and testing of the machine learning-based algorithms • Development of the cloud based online computational capacities • Development/improvement of the website for calculation and visualization connected to the DIASs => generating prototype of information system. Year 2 (final production) • Website and cloud platform connected to the DIASs allowing the users to online compute and visualize indicators and associated uncertainties integrating multiple different data sources and using machine learning algorithm

D4	Extend Semantic Model / Mappings for including:
	• Status assessments of fisheries → Coming from FAO Tuna Atlas (FAO + IRD)
	• Fish food and nutrition information → Coming from the Food Cloud Project (FAO)
	Re-construction of GRSF KB with the new information. The process includes:
	• Update Semantic Model, Mappings
	• Update services/tools to support new contents (i.e. new harvesters, update publishing clients, etc.)
	• Update GRSF publishing/updating workflow
	  ○ So that already approved and manually merged records are preserved
	  ○ So time-dependent information (i.e. timeseries) are updated for all records (including the approved ones)

D5	Cage detection and support to FAO for workflow implementation
	• Consider added value of using Norwegian offer of free optical data, FAO GEE option
	• Work toward fully automated cage detection using Deep learning technics →Condition for making a workflow to repeat on the same area (e.g. Malta)
	• Reload existing metadata after a new analysis
	Pond / padi identification and support to FAO for workflow implementation
	• Analyse and implement EO requirements from Indonesian users
	• Definition of a 'dynamic' workflow (repeat analysis in same area)
	  ○ 2 possible options: 1) using Copernicus data or 2) VHR optical images (pending above)
	  ○ Towards fully automated classification scheme
	• Enhanced coastal mapping
	  ○ Add vegetation (mangrove) and hydrological features, to be discussed with Indonesian counterparts
	  ○ Consider option to work with Norwegian optical data, but also with other departments (FO) and agencies (WFP)

8. What are the expected impacts ?

D1	Plankton EOV demonstrator Plankton indicators are used within several descriptors of the **MSFD** (D1: Biodiversity; D4: Food webs; D6: Sea Floor Integrity; D5: Eutrophication). Dataproducts showing spatial trends and long term anomalies in abundances of phytoplankton and zooplankton provide exemplary illustration of the dynamics of food availability for commercially exploited fish species. As biological EOVs on phytoplankton and zooplankton are still in conceptual or pilot phase, this demonstrator will contribute directly to their further development and application on large scales. The global description of the abundance and diversity of plankton communities yields an indication of the health of marine ecosystems and their response to anthropogenic stressors
	Zooplankton EOV products At the moment limited operational oceanographic products and services are provided by **EMODnet** biology and ingested by **ICES OOPS**. We anticipate that this workflow could increase the quality and quantity of data products that could be provided to ICES, and actively used in their advisory processes.
	Phytoplankton EOV products We anticipate the virtual EOVs will be ultimately considered as **standard marine products** and distributed as part of **Copernicus** marine services to **end-users from a variety of domains** (fundamental research, biogeochemical/ecological modeling, operational modeling, resource management)
D2	1. Understanding of data integration requirements across microscopy imaging, molecular biology and environmental platforms 2. Best practices and standards for data collection and metadata structure in order to facilitate integration 3. Enabling scientific exploration of plankton data for various applications, including bioprospecting, modelling/monitoring the distribution of novel EOVs in time and space, and ecosystem health assessment This demonstrator will bring a thorough understanding of integration of data across microscopy imaging, molecular biology and environmental platforms, including a knowledge of how future data should be collected and structured to allow cross-data type integration while retaining relevance and consistency within each respective data type; The demonstrator will enable scientific exploration of plankton, including correlating plankton concentration and diversity with local environmental variables, deriving known and new indices of ecosystem health, and predicting the distribution of these variables in space and time.

<table>
<tr><td rowspan="5">D3</td><td colspan="2">MARINENVIRONMENT will provide information concerning environmental indicators, integrating numerical model outputs, in situ data and satellite data shown as maps or time-series.</td></tr>
<tr><td>Output data</td><td>• Monthly Mean Maps
• Monthly Climatology Map
• Seasonal Climatology Timeseries
• Annual Climatology Maps
• Average Annual Mean Timeseries</td></tr>
<tr><td></td><td>Derived Fields, ongoing development plug-in for :
- Density ,
- Kinetic Energy ,
- Upwelling Indicator,
- Mixing Indicator</td></tr>
<tr><td>Expected result</td><td>• Data and products from existing services (CMEMS, C3S, etc) will be integrated into a unique online analysis and distribution service facilitating users to generate added-value data products and to perform extra analytics.
• The service will support stakeholders in the MSFD and Blue Economy: the resulting Marine Environmental Indicators will generate large popularity because the user community constituted by national, European, international stakeholders urgently need a flexible capacity to analyse the quality and characteristics of the marine environment from the Global to the Regional scale.
• The users will be able to perform on line and on the fly operations such as selecting portion of dataset for a selected area in horizontal and vertical.
• The users will be able to perform statistical analysis (e.g. trends, anomalies) on the selected variable and portion of space and time.</td></tr>
<tr><td></td><td>Current data can be downloaded by users in NetCDF or other formats that need an expert knowledge to be used, while in MARINENVIRONMENT the technological advancement will allow to access more user-friendly formatted data (e.g. JSON, HTML).</td></tr>
<tr><td rowspan="2">D4</td><td colspan="2">GRSF
• aims to be the first global record with uniquely identified stocks and fisheries by consolidating content that exists in various sources
• Is considered as the key instrument of global fish stock status monitoring and traceability
Stakeholders involved: Seafood industry, resource managers, companies, marine scientists, analysts, geologists, etc</td></tr>
<tr><td colspan="2">Consolidated data workflow methodologies and open-source tools for data ingestion, harmonization and publication, as technical drivers of Data Management Plans
• Transparency: handling rich metadata
• Reproducibility: using open-source programmatic tools
• Sustainability: enforcing international standards
• Engage the community: online tutorials, workshops with VREs...
FAIR Data for fisheries & aquaculture
• Findable: Schemas, vocabularies, catalogues, search interfaces & standard web protocols
• Accessible: with DOIs, visualization/download interfaces & standard web formats
• Interoperable: with standard web-protocols and services, and semantics (URIs)
• Reusable: with DOIs, dataset/queries URL for visualization / sharing</td></tr>
</table>

D5	

9. Challenges

D1	
D2	
D3	

D4	Extend GRSF while preserving approved & published records, especially
	• Their semantic IDs (e.g. asfis:YFT+fao:27.3)
	  ○ The IDs of the approved published record should be preserved
	• The manually merged records
	  ○ This requires work on the process and the API services ➔ We have to update the workflow and develop the required changes in the APIs
	• The dependency to the publishing services (from CNR) should be considered as well
	  ○ To be discussed with CNR to minimize the changes that have to be performed
	Advance the integration process per se
	• Identify those steps that could be improved, prioritize them and plan what is feasible in the context of this project

D5	Access to Very High Resolution imagery for replicable or repeatable analysis Evaluate suitability and limitations of Sentinel-2 / Norwegian offer for this purpose Advance in the interoperability process

10. Various

D1	Timing – Tasks
	• Zooplankton case:
	  ○ Collect and integrate the necessary ocean data
	  ○ Apply fitted machine learning method, Ground truth model using NRT data from Lifewatch, CPR
	  ○ Develop workflow and application using Blue cloud services
	  ○ Implement workflow in Blue Cloud
	• Phytoplankton case:
	  ○ collect and integrate the necessary global ocean data
	  ○ take in hand and refine existing neural network-based methods to produce the virtual EOVs
	  ○ develop global-scale applications of the methods in order to demonstrate the potential and value of the virtual EOVs in a scientific research perspective
	  ○ collaborate with relevant European infrastructure towards the routine implementation, operationalization and distribution of the produced virtual EOVs.

D2	
D3	

D4	GRSF – Potential users	• Regional Fishery Bodies (RFBs) and their member states • Seafood industry producers • Seafood certifiers processors & retailers National agencies responsible for stocks and fisheries reporting • Researchers and officers working on global analyses on state of fishery resources • NGOs promoting sustainable fisheries • General public
	Expected Contribution from other partners	• Description of the data sources / datasets to be used for the above activities [FAO, IRD] • Specification of policies (e.g. for the generation of values or identifiers) [FAO, IRD]
	Data domains:	• Global & Regional Fisheries statistics: RFMO data collections, FAO / Others • Vessel / Fishing operations (AIS / VMS) • Scientific fisheries monitoring surveys (eg Tagging surveys, biological data) • Ocean & Marine data, Biology & Ecology (eg EMODnet/Seadatanet, OBIS) • Satellite imagery & output models
	Collaborations inside Blue Cloud	WP2 Data discovery and access services • Use discovery and access services to harvest from existing fisheries catalogues • Enrich GRSF and TA with data on local environment • Integrate geoviewers and geo WP4 4.2 Taming services (R-WorkFlows of FAO), 4.3 Analytics (Data Miner), and 4.4 Publishing (Zenodo) • Upgrade workflow for Tuna Atlas/Sardara & RDB, generalize to Fisheries Data • Publishd data, algortithms and reports to Zenodo with DoI for reproducability WP5 Communicate and Uptake of demonstrator and tools: • Propose GRSF and Tuna Atlas cases for hackathon and training workshops • Integrate online manuals with TA and GRSF tools • eCommunication (papers, publications ZENODO datasets and replicale workflows as communication objects) WP6 Governance model • Continue with iMarine community • Enforce iMarine MoU between FAO and CNR-ISTI for exploitation of VREs and data services • FAO-FORTH Collaborative Arrangement for the FIRMS Partnership on GRSF Maintenance and Management • FAO-IRD Collaborative Arrangement for the FIRMS Partnership on TA Maintenance and Management

<table>
<tr><td></td><td>Collaborations outside Blue Cloud</td><td colspan="2">Build on BlueBRIDGE AI and CC Experiments; aim at Fishing Under Climate Change
Work with OGC Consortium on R features
Work with RFB’s on Fisheries Data Harmonization and publish aggregate data
Continue successful GRSF with Ram and SFP, invite new partners for market traceability</td></tr>
<tr><td>D5</td><td colspan="2">Expected Contribution from other partners</td><td>OGC Standard data access (WCS / WFS / WMS) for:
• EO data: satellite images, derivate products (eg CMEMS)
• Environmental monitoring data (eg SeaDaNet/EMODnet chemistry / physical data)</td></tr>
<tr><td></td><td>Collaborations inside Blue Cloud</td><td colspan="2">WP2 Data discovery and access services
• Use discovery and access services to harvest from existing fisheries catalogues
• Enrich GRSF and TA with data on local environment
• Integrate geoviewers and geo
WP4 4.2 Taming services (R-WorkFlows of FAO), 4.3 Analytics (Data Miner), and 4.4 Publishing (Zenodo)
• Upgrade workflow for Tuna Atlas/Sardara & RDB, generalize to Fisheries Data
• Publishd data, algortithms and reports to Zenodo with DoI for reproducability
WP5 Communicate and Uptake of demonstrator and tools:
• Propose GRSF and Tuna Atlas cases for hackathon and training workshops
• Integrate online manuals with TA and GRSF tools
• eCommunication (papers, publications ZENODO datasets and replicale workflows as communication objects)
WP6 Governance model
• Continue with iMarine community
• Enforce iMarine MoU between FAO and CNR-ISTI for exploitation of VREs and data services
• FAO-FORTH Collaborative Arrangement for the FIRMS Partnership on GRSF Maintenance and Management
FAO-IRD Collaborative Arrangement for the FIRMS Partnership on TA Maintenance and Management</td></tr>
<tr><td></td><td>Collaborations outside Blue Cloud</td><td colspan="2">Build on BlueBRIDGE AI and CC Experiments; aim at Fishing Under Climate Change
Work with OGC Consortium on R features
Work with RFB’s on Fisheries Data Harmonization and publish aggregate data
Continue successful GRSF with Ram and SFP, invite new partners for market traceability</td></tr>
</table>

5.2 Demonstrator #1 – “Zoo- and Phytoplankton EOV products”: Technical requirements

Demonstrator #1 – Zoo- and Phytoplankton

–

Technical Requirements

- *Why is Blue Cloud fundamental for your demonstrator?*
  - *Integration of data sources*
  - *Bringing together data and scientists working with the data*
  - *Providing a platform to do our analysis: harmonize the data sources, calculate our products and provide them*
  - *sharing of workflow and final products*

- *Which stakeholders will you involve? Who are your end-users? What is your targeted audience?*

  - *Ecosystem advisory processes: the service will generate operational oceanographic zooplankton products and services being developed as we move towards making ICES ecosystem advice operational — defining and describing all steps of the process that culminate in the final advice. Through this we seek to link data to outputs, meeting the public expectations for a modern science and advisory organization and increasing the accessibility and profile of what we produce. The ICES Ecosystem Overviews are also part of this drive.*

  - *Fundamental research: The proposed EOVs will contribute to the understanding of the environmental conditions and top-down factors that shape the global distribution of phytoplankton community biomass and diversity. The spatio-temporal resolution of the products will also allow developing such understanding at new scales of observation (e.g. regional, seasonal).*

  - *Biogeochemical ecosystem modeling: Phytoplankton diversity is an essential driver of biogeochemical processes and carbon fluxes. The current generation of coupled ocean-biogeochemical models include explicitly several phytoplankton functional types (PFTs) and critically need global 4D data for their initialization and validation. Such models are fundamental tools to predict the ocean response to environmental and climate change.*

  - *Ecological modeling and stakeholder: phytoplankton and zooplankton abundance and community distribution impacts fish stocks, habitats and dynamics. The proposed EOVs thus represent valuable information for the modeling, assessment and management of marine pelagic fisheries.*

  - *·Operational modeling: Global 4D monitoring of the distribution of nutrients along with phytoplankton abundance and diversity is of interest for assimilation in operational models, which provide a description of the current state of the oceans and forecast their evolution.*

- *Previsional workplan of the demonstrators?*
    - *Year 1 (data collection + algorithm tests)*
        - *Collection and integration of global data from multiple sources and testing of the machine learning-based algorithms => generating prototype of global phytoplankton and nutrient EOV products.*
        - *Collection and integration of data from multiple sources and testing of the machine learning based algorithms => generating prototype of regional (North Atlantic) zooplankton EOV products.*
    - *Year 2 (final production)*
        - *Phytoplankton and nutrient global EOV products and associated computation routine.*
        - *Zooplankton abundance regional (North Atlantic) EOV products and associated computation routine.*
- *Which evolutions do you have for your requirements (technical and other) compared to your kick-off presentations?*

  Computing

  Virtual machine with easy access to data sources and 10-20 fast CPUs + ~100 GB of memory for machine learning algorithms (artificial neural networks);

    - *Zooplankton:*
        - *SSH access to a Linux computer with 64 GB and 2 NVIDIA GPUs and 16 computing cores*
        - *The CUDA toolkit (version 10 or later) and CuDNN (version 7 or later) which should be installed by the system administrator.*
        - *Julia and DIVAnd.jl (https://github.com/gher-ulg/DIVAnd.jl) which can be installed by ULiege.*
        -
    - *Phytoplankton:*
        - *For phytoplankton related EOVs : Python for development and implementation of ANN-based algorithms, and R + RStudio for data visualization and further statistical analyses*
        - *For phytoplankton related EOVs + interaction biotic and abiotic variables: R + Rstudio for data gap analyses & time series analyses*
        -

Do we need other data sources than these ones (listed in the demonstrator proposal)?

- *Blue Cloud sources:*
  - *Euro-Argo and Argo GDAC: salinity, oxygen, chlorophyll data*
  - *SeaDataNet: physics, biogeochemistry, biology data*
  - *CMEMS (DIAS-WEkEO): ocean color, altimetry data*
  - *EurOBIS: abundance zooplankton data*
  - *EuroBioImaging - EcoTaxa: ecological images of plankton*
- *Additional sources via existing services:*
  - *LifeWatch: BioOracle ecological modelling data*
  - *LifeWatch: observatory sensor data*
  - *LifeWatch: World Register Marine Species*
- *Additional sources via ad-hoc retrieval:*
  - *MAREDAT: phytoplankton diversity information*
  - *GLODAP V2: nutrient data*
  - *Ocean color products*

- *Do you use pre-existing VRE?*
  - *The University of Liege is currently using clusters from the CECI HPC consortium (http://www.ceci-hpc.be/). However, the current generation of clusters is not optimized for machine-learning workloads.*

- *Which infrastructures do you plan to use?*

With respect to qualified databases, this demonstrator will rely on the following European infrastructures and services:

1. *The BGC-Argo and Argo data available from the Global Data Assembly Service (Coriolis), including data acquired by the ERIC Euro-Argo (temperature/salinity, oxygen, chlorophyll concentration)*

2. *EurOBIS; The Ocean Biogeographic database for plankton abundance data.*

3. *The MAREDAT global database for HPLC-determined phytoplankton pigments (phytoplankton diversity information). We note that the LOV made the largest contribution (22%) to the MAREDAT global database and that this contribution will substantially increase within the timeframe of the project.*

4. *The GOOS-endorsed GLODAP V2 (Global Ocean Data Analysis Project) for nutrient data*

5. *LifeWatch: Bio-ORACLE is a set of GIS rasters providing geophysical, biotic and environmental data for surface and benthic marine realms.*

6. *The SeaDataNet that makes available to the ocean research community in situ data of various types (physics, biogeochemistry, biology) and sources (oceanographic vessels and robotic systems)*

7. *The Copernicus Marine Environment Monitoring Services for satellite data (ocean color and altimetry)*

8. *The Ocean Colour CCI products*

9. *The GlobColour products archive (ocean color)*

10. *EMODnet, The European Marine Observation and Data Network*

11. *LifeWatch: Marine Observatory sensor data through LifeWatch marine services.*

12. *LifeWatch: WoRMS, World Register Marine Species, through LifeWatch marine services*

13. *Ecological images database for real time plankton monitoring (through EuroBioImaging) (access yet to be developed as part of WP2/WP4)*

We anticipate the virtual EOVs proposed in this demonstrator may ultimately be operationalized and distributed as open-access 1) EMODnet Operational services and as 2) Copernicus marine services.

5.3 Demonstrator #2 – “Plankton Genomics”: Technical requirements

Blue Cloud Demonstrator #2
Plankton Genomics & Imaging

Background documents here:
https://drive.google.com/drive/folders/1sBqZ9xMrTB2ntj0dxbMEz71vYTR6MYgm

Aim

The aim is to showcase a deep assessment of plankton distributions, dynamics and fine-grained diversity to molecular resolution. Working across biomolecular, image and environmental data domains, drawing on the outputs of such initiatives as Tara Oceans, it will focus on two areas:

1. ***Species and functions discovery***: discovery of as yet undescribed biodiversity from genetic and morphological signals from the characterisation of their geographical distributions, cooccurrences/exclusions and correlation with environmental contexts.
2. ***Biodiversity and ecology***: exploration of genetic and morphological markers of plankton diversity and abundance, in particular the new ones discovered above, to predict their spatiotemporal distribution and serve as high-resolution EOVs for biological processes.

Current state & challenges

(why we need the Blue Cloud)

1. Distinct data sources and platforms exist for microscopic imaging, molecular biology and environmental domains with typically high levels of maturity
2. Lack of a supporting platform for scientists to integrate and compute across data types for given samples and events presents a very real bottleneck to progress.

Scientific background

Expected impacts

(immediate with our community + for other communities)

This demonstrator will bring a thorough understanding of integration of data across microscopy imaging, molecular biology and environmental platforms, including a knowledge of how future data should be collected and structured to allow cross-data type integration while retaining relevance and consistency within each respective data type.

The demonstrator will enable scientific exploration of plankton, including correlating plankton concentration and diversity with local environmental variables, deriving known and new indices of ecosystem health, and predicting the distribution of these variables in space and time.

Methods

Two Notebooks will be constructed that make available and usable the intersect between biomolecular, microscopy image and environmental data types, leveraging workflows around functional genetic inference and image feature classification, made possible through the central infrastructure provided by Blue-Cloud. Notebook requirements are defined additively as follows:

- Early project
    - facility to declare relevant data sets by identifier/URI in respective data sources
    - facility to load and stage data sets
    - basic metadata reporting (source, size, etc.)
    - facility to download data for local computation
- Mid-project
    - facility to discover data sets by metadata in respective data sources
    - facility to load and stage data sets with full metadata
    - support for containerised workflows to operate on staged data
- End-project
    - facility to apply data metadata search rules to operate data synchronisation for continuous ingress
    - facility to automate scheduled workflow operation upon staged metadata
    - support for computation and results visualisation based on workflows and metadata (e.g. Pandas, ggplot)
    - Unix command-line access and ability to install tools

Data sources / infrastructures

- ELIXIR-ENA: genomics data
- EuroBioImaging – EcoTaxa: high-resolution precision microscopy images
- EuroBioImaging BioImage Archive: high-resolution precision microscopy images
- SeaDataNet: environmental data
- CMEMS (DIAS-WEkEO): environmental data
- Euro-Argo and Argo GDAC: environmental data
- EMODnet Biology: biodiversity data
- EMODnet Chemistry: environmental data

Data infrastructures

Technical requirements - tools/workflows for:

- Distant similarity searches (e.g. sequence similarity networks)
- Co-occurence based on incidence and/or abundance
- Population genomics wihout reference
- Functional community exploration through modelisation
- Proteins structure exploration
- Deep learning techniques to identify phenotypes associated with genes

Notebook general features timeline

Year 1: prototype iterative Notebook design and testing; data and workflow requirements gathering and work with technical partners to provide these; standards and conventions for data integration across the data types.
Year 2: move to production, science-ready Notebooks; socialisation with the relevant research communities; use by partners and the wider community to demonstrate scientific value; feedback into Notebook design and data sources to iterate further.
Year 3: outreach and promotion of Notebook functionalities; extrapolation and enumeration of possible future functions enabled by the data integration achieved; community support; capture of experience and white paper highlighting the future of the approach.

Notebook 1 timeline

Year 1: ability to declare, stage and download relevant data sets and metadata from image and molecular data sources; curation of image data sets.
Year 2: ability to install and operate computational workflows for identification from metagenomics/metabarcoding sequence and image classification; first catalogues of identified taxa available.
Year 3: visualisation and deeper exploration tools to understand the intersect between sequence- and image identified taxa.

Notebook 2 timeline

Year 1: ability to declare, stage and download relevant data sets and metadata.
Year 2-3: workflows covering e.g. co-occurences, population genomics without reference, deep learning to identify gene-phenotype association; tools for modelling of community function and protein structure exploration.

5.4 Demonstrator #3 – “Marine Environment Indicators”: Technical requirements

Work Package 3: Blue Cloud Pilot Demonstrators

Task 3.4 Demonstrator #3 – Marine Environmental Indicators

–

Technical Requirements

Blue-Cloud - Piloting Innovative services for Marine Research & the Blue Economy - has received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue Cloud services, Grant Agreement n. 862409.

Contents

1 Scientific context / Present situation

The EU's ocean agenda is part of the response to the UN SDG 14 dedicated to the health of our oceans, and is also closely related to targets of other key UN SDG goals which concern the sustainability of human activities:

1. SDG 11 dedicated to sustainable cities and communities
2. SDG 12 dedicated to sustainable consumption and production
3. SDG 13 dedicated to climate change

The agenda is therefore part of an international governance initiative for the future of the global marine environment, with the priority of reducing the pressure and strengthening the scientific research on oceans. In this framework, the understanding of oceans has a critical role, and needs a science-based approach with the aim to rise the knowledge, and create the prerequisite for a sustainable blue economy. Following policy actions, including also the MSFD, various initiatives have been taken to develop an international marine data network and the capability to monitor the oceans.

Facts on which this Demonstrator is based are:

- Web site with a collection of preliminary set of marine indicators http://www.marinenvironment.com
- Set of pre-calculated datasets and indicators are available on the above-mentioned website and at the other partners premises
- Algorithm for the online calculation of some MSFD - Marine Quality Indicators were already tested in a cloud environment at limited scale – EU MELODIES Project, https://github.com/ec-melodies/wp06-ges-toolbox
- Machine learning and data mining based algorithms are available and have been tested by IFREMER mainly on Argo float data (https://pyxpcm.readthedocs.io)
- Data sources are existing, but not accessible in an efficient and fully integrated way and indicators cannot be calculated online in a dynamic mode
- Uncertainty is not associated to each indicator

2 Overview Information

2.1 Aims of the demonstrator

- To provide and display information and indicators on the environmental quality of the Mediterranean Sea (later addition of North Atlantic).
- To integrate different data sources and exploit scientific based algorithm (included machine learning ones) for the online calculation of environmental indicators.
- To deliver and document the associated computation routines for re-use as part of the Blue Cloud Virtual Lab.

2.2 Stakeholders

(Which stakeholders will you involve? Who are your end-users? What/Who is your targeted audience?)

In this framework, as described in section “Present Situation”, the four important stakeholders for the demonstrator are:

1. policy makers, involved in the legislation activities and in the launch of initiatives for the economic growth, innovation and scientific development;
2. data source infrastructures, representing the result of past and present initiatives focusing on the development of the marine data network;
3. research community, major actor to rise the knowledge and understanding of the oceans, and for the scientific investigation of new effective methodology to underpin the monitoring activities and to identify climate trends;
4. environmental agencies, directly involved in the monitoring activities;

Among them, the end-users of the online service are the environmental agencies and the research community. It is important to notice that these two actors have very different expectations, since on one hand environmental agencies need an effective service for the monitoring activities, while the research community needs the development of capability and research infrastructures to analyse the growing big amount of environmental data.

2.3 Data to be used

(Which data? Format? Sources? Existing or yet to be produced?)

Data sources:

- CMEMS Med MFC: Temperature and Salinity field, Mediterranean Sea, daily means 1990-2009
- CMEMS and C3S (DIAS-WEkEO): ocean and climate variables
- Euro-Argo and Argo GDAC: temperature, salinity, oxygen, chlorophyll data
- SeaDataNet: physics, biogeochemistry, biology data
- EMODnet: physics, biology, chemistry data
- ICOS-Marine: inorganic carbon data

Update of this paragraph will be based on the content in paragraph: 7 “Which are the technical needs to be considered? / Which evolutions do you have for your requirements compared to your kick-off presentations?”

2.4 Infrastructure to be exploited

(Based on which infrastructure? (ELIXIR-ENA, EuroBioImaging, SeaDataNet, EMODnet, EurOBIS, Euro-Argo, EcoTaxa, DIAS, ICOS-Marine, BlueBRIDGE and/or EUDAT?) / Which infrastructures do you plan to use?)

- DIAS-WekEO
- Euro-Argo and Argo GDAC
- SeaDataNet
- EMODnet
- ICOS-Marine

Update of this paragraph will be based on the content in paragraph: 7 “Which are the technical needs to be considered? / Which evolutions do you have for your requirements compared to your kick-off presentations?”

2.5 Why is Blue Cloud fundamental for your demonstrator?

The Blue-Cloud includes all the relevant aspects to perform the demonstrator. In particular, they are: the discovery service, the access service and the exploitation of clouds based infrastructure.

The challenge of this specific demonstrator is to improve the capability for a user/stakeholder to perform custom computation in a reasonable time and with reasonable effort, and create new value added products from the open data which are already available from services, such as CMEMS.

2.5.1 Do you use pre-existing VRE?

- Not until now: CMCC
- Pangeo VRE at ocean.pangeo.io : IFREMER
- …?

2.6 Challenges

2.6.1 Make it Interoperable

Since ocean data are stored in different formats and using different schemes (space/time structure), it is very complex to transform a multi-source ocean data selection and to make it ready for visualisation and statistical analysis. A challenge is to make this manipulation and transformation fast and flexible.

Ocean data are multi-source, multi-scale, multi-dimensional, multi-variate. It is a key challenge to our community to explore such data and to identify the relevant patterns and correlations that will foster new scientific knowledge of the ocean.

Ad-hoc scientific analysis and methods are developed by few experts. It's another challenge to be able to (i) generalise and/or (ii) make interactive such analysis so that a larger audience can have access to them.

2.6.2 Make it Interactive

In the roadmap of the blue growth strategy, the understanding of our oceans play a very important and central role. Various initiatives are facilitating a science-based approach and the cooperation between business and public authorities. The development of infrastructure started with the focus of creating the data network, and afterwards with the development of cloud based services and research infrastructure. To this end, strengthening the scientific research is one of the pillar of the strategy, with the effort to understand dynamics and trends, and the investigation to probe new patterns. As already highlighted in past report "data should be made available for as wide a range of uses and users as possible. Frequently the unexpected and unforeseen use may be the most valuable", it become a challenge to develop a research infrastructure with the maximum flexibility that the scientific investigation needs, able to perform interactive big data analysis.

2.6.3 Make it Findable

Inside the Marine Knowledge ecosystem are active many institutions in charge of providing environmental data from several observation systems or from numerical models. In addition to that, also other important value added data are present, such as the indicators for the monitoring of the good environmental status, which are derived and extrapolated from the big volume of available observations and simulations data. On one hand the development of effective indicators is a continual research effort, while on the other hand some well-established indicators already exist and support the activities of environmental agencies and decision makers. In this framework, it is important to offer the best available indicators at any time. But this is not an easy task, because the input data are many and continuously under evolution, more data come available, and the quality is improved. Because of this, to facilitate the automated processing and the adoption of semantic technologies, become the enabler to make really possible the access to the correct value added indicators from the decision makers, and finally offer to them the best possible view of our complex environment.

2.7 What are the expected impacts?

Expected Results:

- Data and products from existing services (CMEMS, C3S, etc) will be integrated into a unique online analysis and distribution service facilitating users to generate added-value data products and to perform extra analytics.
- The service will support stakeholders in the MSFD and Blue Economy: the resulting Marine Environmental Indicators will generate large popularity because the user community constituted by national, European, international stakeholders urgently need a flexible capacity to analyse the quality and characteristics of the marine environment from the Global to the Regional scale.
- The users will be able to perform on line and on the fly operations such as selecting portion of dataset for a selected area in horizontal and vertical.
- The users will be able to perform simple (e.g. mean, variance, histograms, trends, anomalies) and advanced (e.g. regression, clustering) statistical analysis and visualization on the selected variable and portion of space and time.
- The users will be able to access an online research environment (e.g. Jupyter lab) with full access to the selected variables and portion of space and time to be freely analysed by users. Data selection and notebooks will persist from one session to another to let users fully adopt the VRE as a working environment.
- The selected variables and portion of space and time will be saved by the VRE and made available to the users as a "cloud" resource, so that it could be accessed from any other VRE.

Current data can be downloaded by users in NetCDF or other formats that need an expert knowledge to be used, while in MARINENVIRONMENT the technological advancement will allow to access more user-friendly formatted data.

3 Method and Planning Information

The demonstrator #3 will develop an online service with associated cloud-based analytical computing framework and dedicated web interface to provide and display environmental indicators and information on the environmental quality of the ocean. The online service will exploit already available data sources, integrating numerical model outputs, in situ data and satellite.

3.1 Method

MARINENVIRONMENT will be an online service to provide and display information on the environmental quality of the Mediterranean Sea (in the early stage).

The service will be designed on the basis of intermediate users' requirements (e.g environmental protection Agencies, researchers,…) and will be accessible via the website.

A new set of marine environmental indicators (data products) will be defined and produced within Blue Cloud project. For each indicator an analysis and production workflow will be specified, developed and tested, applying big data analysis and machine learning methods on the multi-source data sets.

The online service user interface will be developed to enable users to perform on line and on the fly operations, such as

- selecting portion of dataset for a selected time and area in horizontal and vertical,
- to perform statistical analysis (e.g. trends, anomalies, clustering) on the selected variable and portion of space and time,
- to display these indicators by tables, map and graphics visualisations.

WP3/Demonstrator#3 task will focus on the scientific and computational challenges of the Demonstrator, and also on the user interface development, while other the technological aspects will be covered in other BLUECLOUD WPs.

Overview of the output indicators :

Output Indicator	Type of output				
	Map	**Timeseries**	**Profile**	**Hovmöller [TBC] ?**	**Compass rose**
Temperature	X	X	X	X	
Salinity	X	X	X	X	
Currents	X				X
Mixing indicator	X				
Upwelling index	X				
Density	X	X			
Kinetic energy	X	X			
Sst trend	X	X			
Sst anomaly	X	X			
Swh anomaly	X	X			
TRIX	X				
Wave period anomaly	X				
Sea-level trend	X	X			
Chl-a trends	X	X			
Chl-a	X	X	X	X	
Plancton biomass and diversity	X				
Inorganic carbon indicators	X				
Heat content trend	X	X			
Wave	X				X
Wind	X	X			X
Ocean patterns (based on ML performed online), for temperature and salinity. By Ifremer: GM,KB	X		X		
Ocean regimes (based on ML performed online) for mixed layer and surface properties. By Ifremer: GM,KB	X	X			

In the table, ″Map″ refers to:

- Monthly, Seasonal, Yearly Mean and anomaly Maps
- Monthly, Seasonal, Annual Climatology Maps

and ″Timeseries″ refers to:

- Average Monthly/Seasonal/Annual Mean Timeseries
- Average Monthly/Seasonal/Annual Climatology Timeseries

3.2 Workplan

3.2.1 General Approach

The general plan has been defined during the proposal, and consists in a first year in which the focus is the (re-)implementation of a consolidated set of indicators in the new Blue Could infrastructure. Afterward, the focus of the second year will be to exploit the working environment that was set up during the first year, and proceed further with the development of the online service and output indicators.

Year 1 M4-M14 (data collection + algorithm improvement and tests + dev. of the cloud based online calculation tools)

- Collection and integration of data from multiple sources and testing of the machine learning-based algorithms
- Development of the cloud based workflow/s for the initial set of marine indicators
- Initial development/improvement of the online service user interface (website), allowing the users to compute and visualize environmental indicators connected to the DIASs

Year 2 M15-M27 (final production)

- Online service and cloud based computational capabilities connected to the DIASs, allowing the users to compute and visualize indicators and associated uncertainties integrating multiple different data sources and exploiting scientific based algorithms

3.2.2 Detailed Workplan

The following table shows the expected development status of a selected set of indicators at the end of Year 1 development stage. The first implementation will take into consideration the initial reference Marine Area (it depends on the specific indicators, details in table below *"Reference Marine Area in first year activities"*):

Output Indicator	Online service Type of output		Prototype based on Python notebook
	Map	Timeseries	
Temperature	X	X	
Salinity	X	X	
Currents	X		
density	X	X	
kinetic energy	X	X	
Ocean patterns			X
Ocean regimes			X

In Year 2 development stage, the above indicators will be extended to all the EU Marine Area, and, as input, when available, will be possible to adopt also daily mean fields from model and observational datasets. Furthermore, in the second year will be also implemented the complete list of environmental indicator in table "Output Indicators". The indicators that were implemented as python notebook in first year, will be integrated into the online service user interface.

3.3 Technical needs

Needs related to the online service:

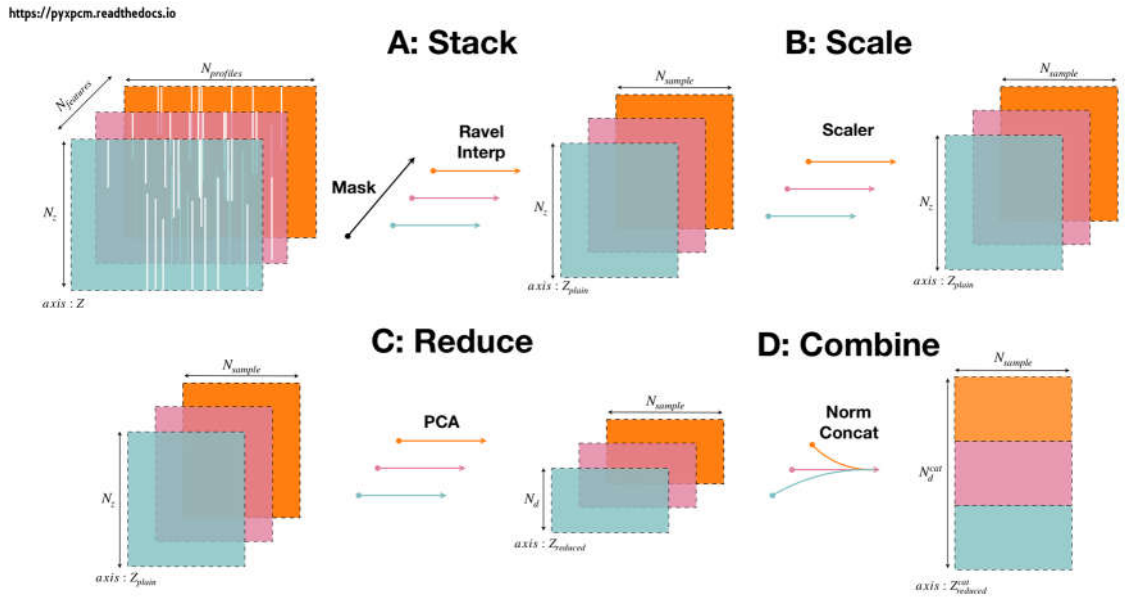
- The possibility to use, extend or reuse already defined and create new workflow
- Given a workflow, the possibility to create an instance by providing the input data sources and processing parameters
    - As input data source, the possibility to adopt the suggested input data sources, or different data sources with consistent semantic and compatible data format
- Given the instance of a workflow, the possibility to submit its execution
- Given the instance of a workflow, the possibility to establish a triggering mechanism, such that in case of changes or update of the input data sources, the automated processing of the workflow will start and provides the updated version of the output indicators
- The possibility to perform interactive big data analysis

Needs related to the workflow of indicators: Temperature, Salinity, Currents, Density, Kinetic Energy

- Discovery of input data (WP2)
  - in the 1st year implementation, the input data source is the product MEDSEA_REANALYSIS_PHYS_006_004 in the CMEMS catalogue; this input data source provides several input fields; the temporal extension is from 1987 to 2018; the available temporal resolution are: "daily mean" and "monthly mean"; in the 1st year will be used only fields with temporal resolution "monthly mean"; the dimension of one time record of one field is about 50MB
  - in the 2nd year there will be multiple input products, with different horizontal resolution, and covering different marine area: for example, a user can be interested in a specific area with the higher available resolution of input data
- **Data Access and Taming (WP2,WP4)** - to prepare the submission of a workflow, it is expected :
  - the possibility to select and access a sub-region (lon range, lat range, depth range) from the complete domain; in the worst case, the computation will be done on the complete input domain (i.e. each time record will cover the complete domain and have a dimension of about 50MB)
  - the possibility to specify and access a specific temporal range (depends on the type of output); in the worst case, the computation will cover the complete temporal extension of the input dataset (i.e. about 30 years, one time record for each month – about 360 time records)
- **Data Processing and Analytics (WP4)** - the workflow needs to receive one (i.e. case of output indicators: *temperature, salinity*) or two (i.e. case of output indicators: *currents, density, kinetic energy*) input fields for the selected sub-region and temporal range;
  - the first task of the workflow can be very demanding in terms of amount of input data to access, and consists in the access and processing of the input monthly mean fields; the processing related to each calendar month is independent and it is expected a high level of parallelism, in order to reduce the elapsed execution time of the workflow
  - the following task/s of the workflow are less demanding in terms of amount of input data
  - in the 2nd year, the temporal resolution of the input field will be "daily mean"
  - EXAMPLE : if we consider the output indicator *density*, its workflow needs to receive as input the monthly mean fields of temperature and salinity; if we select the complete domain, and the complete temporal extension of the input data source, the available months are about 30*12=360; it means the first stage of the workflow, consists in 360 independent task, each expecting to receive 2 input fields, each fields having a dimension of 50MB; the total amount of input is 50MB * 2 * 360 = 36GB; the computation is expected to be done less than 1 hour; in the 2nd year it is expected to start the processing from daily mean fields instead of monthly mean fields, and also the possibility to reduce the processing time.

Needs related to the workflow of indicators: *Ocean patterns*, *Ocean regimes*

- Data Access and Taming (WP2, WP4)
  - All data should be accessible through a webAPI as a collection of profiles localised in space and time (2D collection of samples)
  - All data should be accessible through a webAPI as a collection of time series localised in space
  - Several 10s of Gb (millions to billions of samples, eg: Argo is 2 millions of profiles and about 50Gb, any CMEMS high-resolution re-analysis or forecast could be much larger)
- Data Processing and Analytics (WP4):
  - Multi-dimensional dataset must be ravelled along a sampling dimension and then processed following the pipeline described in Figure below.
  - Once data are pre-processed, patterns and regimes can be determined through unsupervised classification algorithm (GMM, Kmeans, etc …)
  - More statistics must be computed on samples from classification results (maps, profiles, etc …)



- Publishing (WP4)
  - Data and selection of data produced online should be made persistent (at least to the user who generated then) and available to other VRE.
  - Data and selection of data produced online should be made accessible with a Pangeo approach (Jupyter notebooks, xarray datasets, zarr format cloud storage, intake catalogue)
  - Classification models should be made available as an online resource in order to be used with other datasets than the training one.

Reference Marine Area in first year activities:

Output Indicator	**Reference Marine Area in 1st year**
Temperature	Med
Salinity	Med
Currents	Med
density	Med
kinetic energy	Med
Ocean patterns	Glo/North
Ocean regimes	Glo/North

Detailed list of data to be used in first year:

Data Source	We need the Data Source for the following Output Indicator:
CMEMS catalogue / MEDSEA_REANALYSIS_PHYS_006_004	All indicators having **Med** as reference Marine Area in 1st year
CMEMS catalogue / INSITU_GLO_NRT_OBSERVATIONS_013_030 (Euro-Argo data)	Ocean patterns
CMEMS catalogue / GLOBAL-ANALYSIS-FORECAST-PHY-001-024	Ocean patterns and Ocean regimes
CMEMS catalogue / GLOBAL_REANALYSIS_PHY_001_030	Ocean patterns and Ocean regimes
CMEMS catalogue / GLOBAL_REANALYSIS_PHY_001_025	Ocean patterns and Ocean regimes
CMEMS catalogue / GLOBAL_REANALYSIS_BIO_001_029	Ocean patterns and Ocean regimes
CMEMS catalogue / SST_GLO_SST_L4_REP_OBSERVATIONS_010_024	Ocean patterns and Ocean regimes
CMEMS catalogue / SST_GLO_SST_L4_NRT_OBSERVATIONS_010_014	Ocean patterns and Ocean regimes
CMEMS catalogue / OCEANCOLOUR_GLO_CHL_L4_NRT_OBSERVATIONS_009_033	Ocean regimes

The following tables shows the kind of input that each output indicator expects as input. This information can help in the early stage to have a complete preview of the dependencies and needs concerning the input data to be used.

Table of input data source by indicator:

Output Indicator	Input Data Source		
	Model Data		Observation
	3D	2D	
Temperature	X		
Salinity	X		
Currents	X		
Mixing indicator		X	
Upwelling index		X	
Density	X		
Kinetic energy	X		
Sst trend		X	satellite
Sst anomaly		X	
Swh anomaly		X	
TRIX		X	
Wave period anomaly		X	
Sea-level trend		X	satellite
Chl-a trends	X		
Chl-a	X		
Plancton biomass and diversity			
Inorganic carbon indicators			
Heat content trend		X	
Wave		X	
Wind		X	
Ocean patterns	X		insitu
Ocean regimes	X		satellite

5.5 Demonstrator #4 – “Fish, a matter of scale”: Technical requirements

Blue Cloud D4 Requirements - Fisheries Atlas

=========First draft==========

Authors: FAO,

Last Update: Nov 5, 2019

Table of Contents

Requirements Analysis

Purpose

According to the Description of Work, the current Tuna Atlas will be extended to become a Fisheries Atlas (FiAtlas) Virtual Lab. It will be **expanded** to contain the results of **approved status assessments of fisheries**, including those from the Fisheries Atlas. It will **expose** relevant information on **fish food** and **nutrition** though FiAtlas, either as static information or dynamically loaded. To support traceability in the value chain.

FiAtlas is developed to serve the needs of

- FAO and IRD Staff working on fisheries data management, including statisticians, data managers, scientists and stakeholders.
- FAO Projects

Audience

The audience for the RAD includes the Fisheries Atlas team, future users, WP2 and WP4 Leadership, and the system designers (i.e., the developer(s) who participate in the system design). The first part of the document, including use cases and nonfunctional requirements, is written during requirements elicitation. The formalization of the specification in terms of object models is written during analysis.

1. Introduction

Scope, and references to the development context (e.g., reference to the problem statement written by the client, references to existing systems, feasibility studies).

The Global Tuna Atlas has been implemented in the context of H2020 BlueBRIDGE project (GA No. 675680) with the aim to mutualize FAO and IRD efforts to collate and harmonize tuna fisheries datasets at global level.

In Blue Cloud, the demonstrator is described as:

Task 3.5 Demonstrator #4 – Fish, a matter of scale (FAO) [M4-M27] Lead: FAO; Particip.: FORTH, IRD This demonstrator will result in improved data management and analytic capabilities for fisheries. It will be based on the solid and proven technologies behind the FAO Tuna Atlas Virtual Lab and the GRSF Virtual Lab. It will aim at expanding the Virtual Lab for the FAO Tuna Atlas (tuna and billfish catch data) into the Fisheries Atlas: a global vertically integrated toolset that will manage public fisheries statistical data from ingestion, through harmonization, to publication. The Atlas should have features for data analysis using state-of-the-art assessment models. The Global Record of Stocks and Fisheries (GRSF) Virtual Lab will be expanded to contain the results of approved status assessments of fisheries, including those from the Fisheries Atlas. Finally it will provide an option for synergy with the thematic Food Cloud initiative by development of exposure of relevant information on fish food and nutrition through the GRSF, either as static information, or dynamically loaded.

Here, we aim at making the system suitable to load, standardize, publish, access, query, vizualize, and analyze time series and spatial information related to fisheries, support the collaborative production and maintenance of a comprehensive and transparent database on fisheries statistics, and support global and regional standards and

- Load and standardize fisheries records from different data sources
- Standardize time series using reference data from global (FAO) or local (RDF / IRD) reference data
- Assign global identifiers to datasets (DoIs) when they are made sharable
- Expose datasets in a homogenous way, including through CSW for the geospatial representation in a catalogue, and through the GRSF

Operational VREs

- Global Tuna Atlas (with legacy VRE name “FAO Tuna Atlas” from i-Marine project)
- WECAFC-FIRMS
- Fisheries Atlas - If in the future there is a need to include non-tuna and tuna-like species or if the VRE has to be replicated to serve teh needs of a confined area, similar to teh WECAF-FIRMS needs
- SDI Lab - as a proxy for a development environment.

Partners involved in BlueCloud Fisheries Atlas:

· FAO, IRD

1.1 Purpose of the system

Fisheries Atlas will enable the online management of fisheries time series including as geographic explicit data; It will support the data “from catch to Consumption” Food at the source, at the sales point, consumed regularly, and consumed in emergency situations. Different from the current situation, it will be an Open system, implementing internationally-recognized standards (W3C, ISO, OGC) and FAIR principles, and with full control over data by the contributors. Data contributors are now the organizations managing the VRE content (FAO and IRD), yet this list can grow in the future.

1.2 Scope of the system

Fisheries Atlas will be a reporting system where authorized users of 3-4 organizations can submit fisheries data, and work with registered users in a VRE on the standardization of catch data to regional or global standards.

1.3 Objectives and success criteria of the project

The first phase (M4) will be successful if the current VRE is accessible by the task participants, and if the relevant Deliverables are ready.

D3.1 “Demonstrator general technical requirements” [IFREMER; M4, PU, R] The report defines the technical requirements that have to be considered and implemented by WP2 and WP4.
D3.2 “Demonstrator Implementation guidelines” [IFREMER; M4, PU, R] The report summarised the technical guidelines to be endorsed by the demonstrators developers.

Specific activities until M4 therefore include:

1. Develop QA specs on data, e.g. similar to QUACK climateservice
2. Describe the international metadata and data exchange standards across the fisheries data-chain
3. Describe publishing requirements including use of identifiers such as DoIs
4. Describe exposing data through catalogues
    a. Geospatial catalogues and CSW
    b. Targeted catalogues such as CKAN (for GRSF. See GRSF section)
5. Access data through several interfaces, including the OpenFairViewer and R shiny applications
6. Connect the data services to analytical services including the DataMiner, but also RShiny applications

Specific objectives for after M4 will be added later.

2.Current situation

The second section, Current situation, describes the current state of affairs. If the new system will replace an existing system, this section describes the functionality and the problems of the current system. Otherwise, this section describes how the tasks supported by the new system are accomplished now.

- Tuna Atlas VRE
- WECAFC VRE
- R-Tools for geospatial (meta)data management
- SDI and OpenFairViewer
- Data analytics with RShiny
- Data analytics with DataMiner

3.Proposed system

The third section documents the requirements and the analysis and introduces a model of the new system.

3.1 Overview

The overview presents a functional overview of the system.

The overall system is a stack of software components containerized as Virtual Research Environment, including:

- Spatial Data infrastructure components:
    - Spatial databases/schema(s) (depending on use cases)
    - OGC CSW metadata catalogue(s) (Geonetwork to start, other if needed)
    - Geographic data server(s) (GeoServer to start, other if needed eg Thredds)
- Processing components
    - RStudio server for both i) geospatial (meta)data management to feed the SDI and ii) data mining & analytics
    - DataMiner instance to support data processings deployment as web-service
- User Applications servers
    - An HTTP(S) server with OpenFairViewer front-end (client) application deployed
    - An R Shiny server with the capacity to deploy R Shiny applications

3.2 Functional requirements

Functional requirements describes the high-level functionality of the system.

3.2.1 Related to OpenFairViewer or related Javascript clients

Two crucial prerequisites are needed to operate atlases in D4Science and be able to find and access data (the 2 first FAIR principles) from existing remote (meta)data protocols:

- support of HTTPS (some providers do) for their web (meta)data protocols (eg Geonetwork, Thredds, Geoserver, Mapserver, QGIS server, etc.). The D4science e-infrastructure is running under secure (HTTPS) protocol. Mixed content is blocked by default in all web browsers.
- support of CORS (most of data providers do not implement CORS, preventing third party applications to query data (through AJAX requests) and even WMS images for some of them)

The main functional requirements of OpenFairViewer are articulated around and extend the FAIR principles based on geospatial standards (ISO and OGC):

- Capacity to **find** datasets by browsing OGC CSW catalogue, fetch ISO 19115/19139 metadata documents and exploit core business metadata (eg spatial coverage, temporal coverage, etc.)
- Capacity to **access** OGC data protocols through the ISO 19115/19139 distribution information / online resources

- Capacity to **query** datasets, through the OGC data protocols, according to different query strategies such as filters, dimensions or view parameters; and the availability (or not) of structural metadata (ISO 19110).
- Capacity to **visualize** datasets/queries on maps and DOWNLOAD data/subsets
- Capacity to **export** maps for reporting purpose
- Capacity to **share** datasets/queries based on uniquely identified datasets and URL sharing/resolving mechanisms
- Capacity to **cite** datasets/queries for referencing purpose.

3.3 Nonfunctional requirements

Nonfunctional requirements describes user-level requirements that are not directly related to functionality. This includes Feasibility, Usability, Reliability, Performance, Supportability (FURPS), and implementation, interface, operational, packaging, and legal requirements.

3.3.4 Performance

Performance can be assessed on:

- Capacity to publish quickly dataset metadata, structural metadata and associated data p
- Capacity to search/browse quickly large metadata catalogue (enabling pagination)
- Capacity to access/query/visualize quickly data

3.4 System models

System models describes the scenarios, use cases, object model, and dynamic models for the system. This section contains the complete functional specification, including mock-ups illustrating the user interface of the system and navigational paths representing the sequence of screens. The subsections Object model and Dynamic model are written during the Analysis activity.

4. Work plan

Inbound services (from WP2 to Fisheries Atlas)

- Identify statistical datasets and services related to the objectives of Fisheries Atlas
    - From demonstrator partners
        - FAO:
            - Global database on fishery statistics (capture, commodities, etc.)
            - Draft Regional datasets produced through RDB
            - FIRMS stocks & fisheries inventories placemarks (from FIRMS stocks & fisheries map viewer)
        - IRD:
            - To ask Julien if we could reasonably enable some views on top of RTTP database (if allowed)
            - Other Tuna fisheries datasets?
        - Ifremer: ?
    - In EMODnet scope: ?
    - In Blue CLoud scope:
        - Global stocks & fisheries placemarks (from GRSF)
    - Outside / new :
        - AIS /VMS maps?
- What step(s) of the Fisheries Atlas workflow process to exploit WP2 services
    - Harvesting of OGC CSW endpoints of interest
    - Browsing of ISO 19139 compliant metadata sheets (through CSW)
    - Accessing standard data protocols (especially OGC WxS) through ISO 19139 metadata sheets
- Identify geospatial datasets and services
    - OGC (meta)data service protocols (in particular CSW, WMS, WFS, WCS)
- What steps (if any) are needed to access metadata and data through WP2 services
    - Support of HTTPS secure protocol
    - CORS is enabled on (meta)data provider's servers

Outbound services (From Fisheries Atlas to WP2 - FAIRy Tales)

- Feedback on (meta)data interoperability issues
- Best practices/Advice for metadata fixing / fine-tuning to enable or foster further (meta)data interoperability

Processing and transformation services (WP4)

- VRE operation including provision/deployment of software components stack (SDI components, processing servers, end-user application servers)

Extend Model / Fisheries to include:

- Status assessments of fisheries; from FAO Tuna Atlas (FAO + IRD)
- Fish food and nutrition information; from the Food Cloud Project (FAO)

General Plan

Duration: Jan-2020 (M4) – Dec 2021 (M27): 24 months

Partners:

FAO (lead): 22 PMs (*for WP3*) | FORTH: 12.5 PMs (GRSF) | IRD: 10 PMs (TA)

Deliverables:

D3.3 – Blue Cloud Demonstrators Users Handbook V1 – [IFREMER] [M14]

D3.3 – Blue Cloud Demonstrators Users Handbook V2 – [IFREMER] [M27]

Milestones:

MS16 – Blue Cloud Demonstrator 4 release #1 – [FAO] [M14]

MS20 – Blue Cloud Demonstrator 4 release #2 – [FAO] [M27]

Detailed Plan for Fisheries Atlas

Period	Description	Participants	Status
M1-M4	Collect requirements (for WP2) and sketch a scenario showing how BC will bring benefits to This requires Consultation from the target	FAO, rest participants	
M5-M14	Objective: Release #1		
M17-M27	Objective: Release #2		

5. Glossary

A glossary of important terms, to ensure consistency in the specification and to ensure that we use the client's terms. A precursor to the Data Dictionary

D4Science An EU INfrastructure for data management and analysis
EU European Union
FAO Food and Agriculture Organisation of the United Nations
GEOMETA
IRD
SDI

5.6 Demonstrator #5 – “Aquaculture Monitor”: Technical requirements

Blue Cloud D5 Requirements - Aquaculture Atlas

=========First draft==========

Authors: FAO, CLS

Last Update: Nov 21, 2019

Table of Contents

Requirements Analysis

Purpose

Task 3.6 Demonstrator #5 – Aquaculture Monitor (FAO) [M4-M27] Lead: FAO; Participants: CLS

According to the Description of Work, this demonstrator will expand the remote sensing analytic capabilities of the existing Aquaculture Atlas Production System (AAPS) Virtual Lab in order to provide a robust and replicable environment for monitoring aquaculture in marine cages and in coastal areas. The ambition is to deliver a tool to produce national aquaculture sector overviews whereby a country can make use of OGC compliant data services to monitor its aquaculture sector, not in an isolated way, but built on interoperable services where teams can compute and publish reproducible experiments.

The **Blue Cloud is fundamental** for this demonstrator as it provide the continuity of the BlueBRIGE VRE; it provides the cloud infrastructure to host, share with a larger and growing “blue growth community” and communicate on potential and utility of added value products based on EO data already accessible through a VRE OpenMapViewer UI.

In the specific BlueCloud context, further study how the other contributing infrastructures (CMEMS, WEkEO, Seadatanet,…) can be used to broaden the demonstrator scope and maximize utility/ impact for targeted stakeholders/end users.

The key **stakeholders** will be European aquaculture and maritime spatial planning actors; maritime, coastal and fisheries authorities of ASEAN (in particular KKP in Indonesia) and in developing maritime countries. Staff of these organization will be data managers and **end-users**. The **targeted audience** are data managers in spatial advisory units for maritime spatial monitoring and planning, fisheries and aquaculture policy makers, and monitoring and control agencies.

The **Provisional workplan** of the demonstrators was presented and is included in the Kick Off presentations as baseline. The target is to effectively start Feb 1st as planned with an updated baseline based on

1. additional inputs (FAO and end user/ stakeholders high level requirements) and
2. integration or potential evolutions suggest by CLS @ TCom.

The **evolution** foreseen of above requirements (technical and other) compared to your kick-off presentations, include potential new inputs (from FAO/end user stakeholders) and an on-going CLS analysis by an already appointed project team based on the Kick Off slides presented by other partners. This collaborative effort (CLS/FAO and WP2 and 4) includes (but will evolve)

3. Improved UI functioning and interactivity
4. Improved management of data exchange between external systems and the VRE
5. Improved data management in the Spatial Data Infrastructure in D4Science (list some ...)

The Kick Off feedback showed a strong interest from EU and technical project lead to maximize use of the Blue Cloud data sources; suggestions here include here

6. Use WEkEO for the access to Sentinel 2 (mainly) / Sentinel 1 (complementary) imagery instead of ESA Scientific Hub (BlueBridge solution). This will depend of WEkEO deployment schedule. Could be tentatively planned in Year 2 for example.
7. Focus on non European use case (Indonesia) if no sustainable access to VHR is realistic in the mid-term (project timeframe)
8. Use CMEMS (in particular Sentinel 3) and Sentinel 1 and 2 data in non European countries for other coastal environment studies of interest to end users and FAO: In the VRE Map Viewer and as synchronized (over space and time) statistics on sea level rise impact, vulnerability studies, impact of changing oceanographic parameters (SST, salinity, currents, …)
9. Use CMEMS data to assess impact of pollutions close to fish farms in Med Sea
10. Explore use of SeaDataNet or biogeochemical data

The **existing VRE** for the Aquaculture Atlas Production System, a D4 Science VRE, will be further elaborated to serve the needs of this demonstrator.

This implies continued use of the iMarine.D4Science infrastructures and the VRE. In addtion, WEkEO is scheduled for exploitation, and the CLS internal processing facilities (at least to support the transition towards optimal use of the BlueCloud infrastructure, data and resources)

Audience

The audience for the RAD includes the Aquaculture Atlas team, future users, WP2 and WP4 Leadership, and the system designers (i.e., the developer(s) who participate in the system design). The first part of the document, including use cases and nonfunctional requirements, is written during requirements elicitation. The formalization of the specification in terms of object models is written during analysis.

1. Introduction

Scope, and references to the development context (e.g., reference to the problem statement written by the client, references to existing systems, feasibility studies).

The Aquaculture Atlas has been implemented in the context of H2020 BlueBRIDGE project (GA No. 675680). Here, we aim at making the system suitable to load, harmonize, publish, vizualize, and analyze Aquaculture time series and spatial information, support the collaborative production and maintenance of a comprehensive and transparent database on Aquaculture statistics, and support global and regional standards and

- Load and harmonize geospatial products from different data sources
- Harmonize using reference data from global (FAO) or local (CLS) reference data
- Assign global identifiers to datasets (DoIs?) when they are made sharable
- Expose records in a homogenous way, including through the GRSF

Operational VREs

- Aquaculture Atlas

Partners involved:

- FAO, OpenFairViewer, Geoserver, Geonetwork
- CLS; Image analysis,

1.1 Purpose of the system

AquacultureAtlas will enable the online management of Aquaculture Time series including as geographic explicit data; It will support the data “from catch to Consumption” Food at the source, at the sales point, consumed regularly, and consumed in emergency situations. Different from the current situation, it will be an Open system, implementing FAIR principles, and with full control over data by the contributors.

1.2 Scope of the system

AquacultureAtlas will be a reporting system where authorized users can submit Aquaculturedata, and work with other registered users on the harmonization of catch data to regional or global standards.

1.3 Objectives and success criteria of the project

The first phase (M4) will be successful if a dummy entry system with sufficient references to

Specific objectives:

2. Current situation

The current state of affairs.

- Aquaculture Atlas - Cage detection
- Aquaculture Atlas - Coastal Aquaculture
- SDI - D4Science Side
- CLS Infrastructure

3. Proposed system

The requirements and analysis, and introduction of new system.

3.1 Overview

The overview presents a functional overview of the system.

3.2 Functional requirements

Functional requirements describes the high-level functionality of the system.

3.2.1 Related to OpenFairViewer

D4science: 2 crucial requirements are needed to deliver atlases in D4Science:

- support of HTTPS (some providers do)
- support of CORS (most of data providers, eg ifremer, emodnet do not implement CORS, they even prevent third party application to query their WMS images)

EMODnet:

3.2.2 Related to Cage detection

3.2.3 Related to Coastal pond detection

3.2.4 Related to SDI

3. Work plan

Inbound services (from WP2 to AquacultureAtlas)

- Identify statistical datasets and services related to the objectives of AA
    - In EMODnet scope
    - In Blue CLoud scope
    - Outside / new
- What step(s) of the AA workflow process to exploit WP2 services
    - Develop
- Identify geospatial datasets and services
    - OGC services
- What steps (if any) are needed to access data through WP2 services
    - Https
    - CORS

Outbound services (From AquacultureAtlas to WP2 - FAIRy Tales)

Processing and transformation services (WP4)

-

.

General Plan

Duration: Jan-2020 (M4) – Dec 2021 (M27): 24 months

Partners:

FAO (lead): 22 PMs (*for WP3*) | CLS: 12.5 PMs

Deliverables:

D3.3 – Blue Cloud Demonstrators Users Handbook V1 – [IFREMER] [M14]

D3.3 – Blue Cloud Demonstrators Users Handbook V2 – [IFREMER] [M27]

Milestones:

MS16 – Blue Cloud Demonstrator 4 release #1 – [FAO] [M14]

MS20 – Blue Cloud Demonstrator 4 release #2 – [FAO] [M27]

Detailed Plan for AquacultureAtlas

Period	Description	Participants	Status
M1-M4	Collect requirements (for WP2) and sketch a scenario showing how BC will bring benefits to This requires Consultation from the target	FAO, rest participants	
M5-M14	Objective: Release #1		
M17-M27	Objective: Release #2		

Sujet : Draft structure for Demonstrator#5 - Aquaculture Atlas
De : "Ellenbroek, Anton (FIAS)" <Anton.Ellenbroek@fao.org>
Date : 21/11/2019 14:23
Pour : Ifremer - BLUE CLOUD <Bluecloud.irsi@ifremer.fr>, NYS <Cecile.Nys@ifremer.fr>, Lebras Jean-yves <jlebras@cls.fr>, Longepe Nicolas <nlongepe@groupcls.com>
Copie à : Gilbert MAUDIRE <Gilbert.Maudire@ifremer.fr>, "Gentile, Aureliano (FIAS)" <Aureliano.Gentile@fao.org>, "Blondel, Emmanuel (FIAS)" <Emmanuel.Blondel@fao.org>

From: Ellenbroek, Anton (FIAS)
Sent: Wednesday, November 20, 2019 6:10 PM
To: Ifremer - BLUE CLOUD <Bluecloud.irsi@ifremer.fr>; NYS <Cecile.Nys@ifremer.fr>
Cc: Gilbert MAUDIRE <Gilbert.Maudire@ifremer.fr>; Gentile, Aureliano (FIAS) <Aureliano.Gentile@fao.org>; julien.barde@ird.fr; Blondel, Emmanuel (FIAS) <Emmanuel.Blondel@fao.org>
Subject: RE: Draft structure for Demonstrator#4 - Fisheries Atlas and GRSF

Hi Cécile,

Please find attached the description of demonstrator #4; the part of the fisheries atlas.

I included your questions with links to where the first answers can be found.

Tomorrow I will send the second key element of this demonstrator; the GRSF description, and try to have content in the Demonstrator #5 Aquaculture.

Anton Ellenbroek.

Requirements Analysis

Purpose

Task 3.6 Demonstrator #5 – Aquaculture Monitor (FAO) [M4-M27] Lead: FAO; Participants: CLS

According to the Description of Work, this demonstrator will expand the remote sensing analytic capabilities of the existing Aquaculture Atlas Production System (AAPS) Virtual Lab in order to provide a robust and replicable environment for monitoring aquaculture in marine cages and in coastal areas. The ambition is to deliver a tool to produce national aquaculture sector overviews whereby a country can make use of OGC compliant data services to monitor its aquaculture sector, not in an isolated way, but built on interoperable services where teams can compute and publish reproducible experiments.

The **Blue Cloud is fundamental** for this demonstrator as it provide the continuity of the BlueBRIGE VRE; it provides the cloud infrastructure to host, share with a larger and growing “blue growth community” and communicate on potential and utility of added value products based on EO data already accessible through a VRE OpenMapViewer UI.

In the specific BlueCloud context, further study how the other contributing infrastructures (CMEMS, WEkEO, Seadatanet,…) can be used to broaden the demonstrator scope and maximize utility/ impact for targeted stakeholders/end users.

The key **stakeholders** will be European aquaculture and maritime spatial planning actors; maritime, coastal and fisheries authorities of ASEAN (in particular KKP in Indonesia) and in developing maritime countries. Staff of these organization will be data managers and **end-users**. The **targeted audience** are data managers in spatial advisory units for maritime spatial monitoring and planning, fisheries and aquaculture policy makers, and monitoring and control agencies.

The **Provisional workplan** of the demonstrators was presented and is included in the Kick Off presentations as baseline. The target is to effectively start Feb 1st as planned with an updated baseline based on

1. additional inputs (FAO and end user/ stakeholders high level requirements) and
2. integration or potential evolutions suggest by CLS @ TCom.

The **evolution** foreseen of above requirements (technical and other) compared to your kick-off presentations, include potential new inputs (from FAO/end user stakeholders) and an on-going CLS analysis by an already appointed project team based on the Kick Off slides presented by other partners. This collaborative effort (CLS/FAO and WP2 and 4) includes (but will evolve)

3. Improved UI functioning and interactivity
4. Improved management of data exchange between external systems and the VRE
5. Improved data management in the Spatial Data Infrastructure in D4Science (Version, time-aware maps, metadata driven classifiers, computed spatial statistics such as area covered by item / by admin area, change in area from previous map)
6. Improved download and sharing features of Analytics (e.g. a pdf with a regional Aquaculture Monitoring Analysis)

The Kick Off feedback showed a strong interest from EU and technical project lead to maximize use of the Blue Cloud data sources; suggestions here include here

7. Use WEkEO for the access to Sentinel 2 (mainly) / Sentinel 1 (complementary) imagery instead of ESA Scientific Hub (BlueBridge solution). This will depend of WEkEO deployment schedule. Could be tentatively planned in Year 2 for example.
8. Focus on non European use case (Indonesia) if no sustainable access to VHR is realistic in the mid-term (project timeframe)
9. Use CMEMS (in particular Sentinel 3) and Sentinel 1 and 2 data in non European countries for other coastal environment studies of interest to end users and FAO: In the VRE Map Viewer and as synchronized (over space and time) statistics on sea level rise impact, vulnerability studies, impact of changing oceanographic parameters (SST, salinity, currents, …)
10. Use CMEMS data to assess impact of pollutions close to fish farms in Med Sea
11. Explore use of SeaDataNet or biogeochemical data

The **existing VRE** for the Aquaculture Atlas Production System, a D4 Science VRE, will be further elaborated to serve the needs of this demonstrator.

This implies continued use of the iMarine.D4Science infrastructures and the VRE. In addtion, WEkEO is scheduled for exploitation, and the CLS internal processing facilities (at least to support the transition towards optimal use of the BlueCloud infrastructure, data and resources)