Blue-Cloud

Piloting innovative services for Marine Research & the Blue Economy

D1.2 Data Management Plan

Work Package	WP1, Project Management & Coordination
Lead Partner	MARIS
Lead Author (Org)	MARIS
Contributing Author(s)	Dick M.A. Schaap (MARIS), Pasquale Pagano (CNR)
Reviewers	Christopher Ariyo (CSC), Merret Buurman (DKRZ)
Due Date	31-03-2020, M6
Submission Date	26-05-2020
Version	1.0

Dissemination Level

X	PU: Public
	PP: Restricted to other programme 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	20.04.2020	MARIS	First version
0.2	28.04.2020	CNR	Revision and contribution
0.3	06-05-2020	CSC and DKRZ	Internal Review
0.4	16-05-2020	CNR	Revision
0.5	18-05-2020	MARIS	Finalisation
1.0	26-05-2020	Trust-IT	Acceptance and EU submission

Contents

Executive summary 4
1 Introduction 5
2 Data summary 6
2.1 Purpose of the data gathering and generation 6
2.2 Relation to the objectives of the project 7
2.3 The types and formats of data gathered and generated 7
2.4 Re-use of existing data 11
2.5 Origin of the data 11
2.6 Expected size of the data (if known) 12
2.7 Outline of the data utility and its uses 12
3 FAIR data 14
3.1 FAIR principles 14
3.2 Making data findable, including provisions for metadata 14
3.3 Making data openly accessible 16
3.4 Making data interoperable 17
3.5 Increase data re-use (through clarifying licenses) 20
4 Allocation of resources 22
5 Data Security 23
6. Ethical aspects 25

Executive summary

The Blue-Cloud project aims to pilot a cyber platform bringing together and providing access to: 1) multidisciplinary data from observations and models, 2) analytical tools, & 3) computing facilities essential to support research to better understand and manage the many aspects of ocean sustainability.

Data management will be an integrated activity throughout the project lifetime and should be deployed and supported by the Blue-Cloud services as planned.

An initial Blue-Cloud Data Management Plan (DMP) has been compiled to serve as a guidance giving directions. This DMP follows the Horizon 2020 guidelines for Data Management Plans[1] as published by the European Commission. It is organised around solutions and approaches aiming at making data findable, accessible, interoperable and re-usable (FAIR). For that purpose, an overview is given of data (to be) managed by the Blue-Cloud project and solutions, both 'existing/already planned' and 'possibly to be developed' are given for applying FAIR principles. In addition, the DMP considers other aspects of data management, such as resources, data security and ethical aspects.

This initial DMP will be updated over the course of the project whenever significant changes arise and/or more insights become available as part of the ongoing developments.

[1] Open Access and Data Management: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

1 Introduction

The Blue-Cloud project is part of 'The Future of Seas and Oceans Flagship Initiative' and is undertaken by a consortium of 20 organisations. It aims:

- To build and demonstrate a Pilot Blue Cloud by combining distributed marine data resources, computing platforms, and analytical services;
- To develop services for supporting research to better understand & manage the many aspects of ocean sustainability;
- To develop and validate a number of demonstrators of relevance for marine societal challenges;
- To formulate a roadmap for expansion and sustainability of the Blue Cloud infrastructure and services.

The project will federate leading European marine data management infrastructures (SeaDataNet, EurOBIS, Euro-Argo, Argo GDAC, EMODnet, ELIXIR-ENA, EuroBioImaging, CMEMS, C3S, and ICOS-Marine), and horizontal e-infrastructures (EUDAT, DIAS, D4Science) to capitalise on what exists already and to develop and deploy the "Blue Cloud" framework. The federation will be realized at the levels of data resources, computing resources and analytical service resources.

A Blue Cloud data discovery and access service will be developed to facilitate sharing with users of multidisciplinary datasets. A Blue Cloud Virtual Research Environment (VRE) will be established to facilitate that computing and analytical services can be shared and combined for specific applications. This innovation potential will be explored and unlocked by developing five dedicated Demonstrators as Virtual Labs together with excellent marine researchers.

Data management will be an integrated activity throughout the project lifetime and should be deployed and supported by the Blue-Cloud services as planned. To serve as a guideline when developing these services, an initial Blue-Cloud Data Management Plan (DMP) is drafted in this Deliverable D1.2. This initial DMP will follow the Horizon 2020 guidelines for Data Management Plans as published by the European Commission. It will give an overview of data (to be) managed by the Blue-Cloud project and identify solutions, both 'existing/already planned' and 'possibly to be developed' for meeting the FAIR principles, for as far as feasible and required within the scope of the Blue-Cloud project.

This initial DMP will be updated over the course of the project whenever significant changes arise and/or more insights become available as part of the ongoing developments.

2 Data summary

2.1 Purpose of the data gathering and generation

The Blue-Cloud will develop a technical framework and will evaluate and showcase the potential of this framework by five dedicated scientific Virtual Lab Demonstrators which are relevant for marine environment and blue economy. The technical framework of the pilot Blue-Cloud will feature:

1) **the Blue Cloud data discovery and access service** component to serve federated discovery and access to blue data infrastructures
2) **the Blue Cloud Virtual Research Environment (VRE)** component to provide a Blue Cloud VRE as a federation of computing platforms and analytical services. This VRE will provide the platform for developing and running the Virtual Labs for the scientific Demonstrators.

The technical framework will support the main flow of marine data in the VRE from input to output. Input will be provided by retrieving data sets (including metadata) from the blue data infrastructures that will be connected to the Blue Cloud discovery and access service. Additional input will be provided by users ingesting data sets from other external sources, including their own data sources. The data input can be in-situ data, earth observation data, and model outputs.

Selected input will be used in each of the Virtual Labs as to be developed for the five scientific demonstrators for further analysis, combining, and processing input using analytical services. These analytical services can be various algorithms and services. The results of the analytical processes in the form of data products will be considered as output, that might be published and made available for viewing and downloading.

The Blue-Cloud project itself does not gather observation data about the oceans, e.g. by measurements or sensors. As a federation of existing marine data infrastructures aiming to enable data analyses by its end users, it brings together data from various existing sources. The main goal of Blue-Cloud being to provide users the possibility to process data, the focus is on providing input data, and handling the resulting derived data, which may then be made public. Thus these groups of data can be identified:

(A) Input data from existing known marine infrastructures, i.e. the project partners. This can be divided into:
    (1) Input data that is being kept at the originating marine data infrastructures and is transferred to the Blue-Cloud servers only on demand, when a user requests the data to perform processing on it. The originating infrastructures remain responsible for this data and their DMP apply.
    (2) Input data that was copied from the originating infrastructures and stored (and possibly modified) by the Blue-Cloud project, to match the demonstrators.

(B) Other input data, uploaded by the users.

(C) Resulting datasets/derived datasets, created by users as a result of their usage of Blue-Cloud analytical services.

In the following DMP, these different groups of data will be addressed. The groups (A2), (B), (C) reside on and are managed by the Blue-Cloud infrastructure. They make up the "VRE data pool". For the (A1) data, the data management itself (e.g. preservation) is handled by the originating sources, the blue infrastructures. Their FAIRness, and the FAIRness of the federated access in Blue-Cloud, is described in section 3.

2.2 Relation to the objectives of the project

One major aim of the Blue-Cloud project is to evaluate and demonstrate the potential added-value for science of combining multi-disciplinary data sets. Another major aim is to promote web-based science, promoted as a potential benefit of EOSC. Both aims will be supported by the Blue-Cloud technical framework and should become manifest by developing and showcasing the scientific Virtual Labs, facilitating researchers to combine multi-disciplinary data input from multiple sources and to run efficiently processing pipelines on the cloud, generating useful and interesting data products for several user communities.

2.3 The types and formats of data gathered and generated

Data input for the Virtual Labs will be largely gathered and retrieved from the blue data infrastructures that are included in the Blue-Cloud project. Additional input might be added by researchers ingesting data sets from other external sources. Thereafter, researchers may generate data of various kinds using the Blue-Cloud VRE services.

Data retrieved from existing blue infrastructures:

The following blue data infrastructures will be pillars under the initial **Blue Cloud data discovery and access service**:

- SeaDataNet (physics, bathymetry, geology, chemistry, biology, geophysics);
- EMODnet Bathymetry (bathymetry);
- EMODnet Chemistry (chemistry);
- EurOBIS – EMODnet Biology (marine biodiversity);
- Euro-Argo and Argo GDAC (ocean physics and marine biogeochemistry);
- ELIXIR-ENA (biogenomics);
- EuroBioImaging (microscopy);
- EcoTaxa (bio images);
- WekEO (CMEMS ocean analysis and forecasting and C3S climate analysis and forecasting);
- ICOS-Marine (carbon).

These infrastructures are developed and operated by research, governmental, and industry organizations from European states, and they have established links to data originators and their data collections, facilitating to oversee and engage in the process from collection to validation to storage and distribution. Several are also increasingly involved in generating data products and models, which are run by the infrastructure teams or made available as services for external users from research, government and industry. These blue data infrastructures are also mostly complementary to each other, dealing with other data originators and/or different stages in the processing chains from data acquisition to data products to knowledge.

The infrastructures represent a wide range of data types and formats which are summarised in the following concise table:

Blue data infrastructure	**Types of Data**	**Formats**
SeaDataNet	physical, geological, chemical, biological, bathymetric, and geophysical observation data sets and climatology products	Common Data Index (CDI) metadata format; SeaDataNet ODV4 ASCII, NetCDF (CF), and MedAtlas data formats; Sextant Products Catalogue metadata format; ODV4 ASCII data collection and NetCDF (CF) data product formats
EMODnet Bathymetry	Bathymetric observation data sets and generated DTM data products	Common Data Index (CDI) metadata format; SeaDataNet ODV4 ASCII, and NetCDF (CF) data formats; XYZ, ESRI ASCII, EMODnet EMO, and GeoTIFF formats for DTM products
EMODnet Chemistry	Chemical observation data sets and generated harmonised chemical data collections	Common Data Index (CDI) metadata format; SeaDataNet ODV4 ASCII, and NetCDF (CF) data formats; Marine Litter dedicated ASCII data formats; ODV4 ASCII data collection format
EurOBIS – EMODnet Biology	biogeographic data with focus on taxonomy and distribution records in space and time	IMIS metadata format; Darwin Core data model
Euro-Argo – Argo GDAC	salinity/temperature and biogeochemical observation profiles	NetCDF (CF) data format incl metadata
ELIXIR-ENA	nucleotide sequencing information: covering raw sequencing data, sequence assembly information, functional annotation and a host of further data types	TSV and JSON compatible metadata formats; EMBL Flatfile data format, FASTA data format for sequences; XML data Format

EcoTaxa	images of plankton	TSV and ODV compatible metadata formats; image formats. EcoTaxa (meta)data will become available through EurOBIS
WekEO	Sentinel satellite data sets: S1, S2, S3 Marine, S3 Land, S5P; main Copernicus Service Data products from Copernicus Marine Service (CMEMS), Copernicus Atmospheric Service (CAMS), Copernicus Climate Service (C3S), and Copernicus Land Service (CLMS)	Details not yet available. Planned during 2020.
ICOS-Marine	long-term oceanic atmospheric observations for carbon	ICOS metadata format; ODV and NetCDFdata formats
EuroBioImaging	biological images for life-science researchers	TSV, JSON, and XML metadata formats; MRC, MRCS, TIFF, DM4, IMAGIC, SPIDER, MRC FEI and RAW FEI image formats

Table 2.1: Blue data infrastructures with data types and formats (based upon Blue-Cloud D2.1)

Selected input will be used in each of the Virtual Labs as to be developed for the five scientific demonstrators. The following table gives an overview of the demonstrators and a preliminary indication which types of data they will use.

VRE Demonstrator	**Data input (expected)**	**Blue data source (expected)**
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

Zoo- and Phytoplankton EOV products	• 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
Plankton Genomics	• 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
Marine Environmental Indicators	• 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
Fish, a matter of scales	• 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
Aquaculture Monitor	• Sentinel 1 (S1) and Sentinel 2 (S2) • VHR images	• FIRMS • WekEO • CLS

Table 2.2: Blue-Cloud demonstrators and their expected data input requirements (based upon Blue-Cloud D3.1)

Data inputs from other external sources, retrieved by users:

The users of the Blue-Cloud services will be able to upload their own input datasets to complement existing datasets as retrieved from the Blue-Cloud data management infrastructures. As this is completely under the responsibility of the users, no information about the types can be given.

However, to be useful to the users, the own input datasets will have to adopt formats that the Blue-Cloud services are able to handle, thus likely to be of one of the above described formats.

Data generated by users during and after the Blue-Cloud project:

The researchers using and running the five Virtual Labs at the Blue-Cloud VRE will generate data products that they will want to store for further use and possibly make available for discovery and access by viewing and downloading. The details of these data products itself are not known yet and will become clearer as part of WP3 activities for the development and operation of the scientific demonstrators.

Summary

The Blue-Cloud VRE will be largely based upon the existing D4Science e-infrastructure, which successfully provided the analytical and computing cloud platform for the BlueBridge project in which multiple Virtual Labs were deployed. The existing D4Science e-infrastructure will be adopted and adapted, and new services will be added, where needed and as part of WP4. The earlier BlueBridge experience will be very useful for these evaluations which will be done in dialogue with WP3 and WP2. As a result of the BlueBridge activities, already a number of facilities for data management are available in D4Science which will be described further in this DMP, where relevant.

2.4 Re-use of existing data

As indicated in paragraph 2.3, the Blue-Cloud will make use on a major scale of marine data sets which are to be retrieved from blue data infrastructures, complemented with marine data sets from other sources. The data input can be in-situ data, earth observation data, and model outputs. This way, on a large-scale use and re-use will be made of existing data: use as input for the analyses in the Virtual Labs and possible re-use for validating and calibrating the Virtual Labs.

The data is subject to the data stewardship, the data policies and DMPs of the originating data repositories, the blue data infrastructures. These are mostly in favour of open use and re-use of data sets for various applications, while the licenses are used to specify disclaimers and to encourage users to acknowledge used data sources, when publishing their results. Some of the data policies also require registration of users, which is implemented OR realized by logon procedures, not to hamper access but to facilitate tracking and tracing of use and users in respect of the GDPR regulation. A small percentage of marine data sets has access restrictions, such as for very costly data sets such as bathymetry and seismic, and sensitive data sets, such as for marine pollutants in selected countries. In those cases, a form of negotiation is required between user, indicating and motivating its requests, and the data originator.

2.5 Origin of the data

Most data input will comprise marine data to be retrieved from the blue data infrastructures that are represented in the Blue-Cloud project, complemented with marine data sets from other sources (see paragraph 2.3 and 2.4). These blue data infrastructures are leading infrastructures in the European marine data management landscape and they manage and maintain repositories of data and data products originating from a large number (thousands) of research institutes, national governmental agencies and departments, European and international associations, and industry parties. Their coverage of data and networks of data originators are steadily expanding.

2.6 Expected size of the data (if known)

The data offerings from the blue data infrastructures represent large data volumes, which are steadily increasing. However, details about their sizes are not easily to be assessed due to the large numbers and heterogeneity of the data collections. Data set sizes can range from a few Kilobytes (CTD profiles) to many Gigabytes (e.g. satellite images, model outcomes). In a later stage of the project, in particular as part of WP2 activities for developing the Blue-Cloud data discovery and access service, more insights will be gained. This will also apply to the assessment of the sizes of the generated data products, which will become more known during the WP3 and WP4 activities for deploying the Virtual Labs.

2.7 Outline of the data utility and its uses

Federated discovery and access to the blue data infrastructures will be arranged by the **Blue Cloud data discovery and access service** which is under development in WP2 and is planned for release end M17 (early March 2021). This service will interact with the Blue-Cloud VRE and will also serve as a standalone service for users interested in discovering and retrieving data sets from the connected blue data repositories. The overall concept is that the Blue-Cloud data service utility will regularly harvest metadata on data collections from each blue data infrastructure in order to build and maintain a common Blue-Cloud metadata catalogue by means of a metadata brokerage mechanism. The resulting metadata catalogue will be made available by means of web services for machine-to-machine interactions, as well as by GUI for human users. The catalogue will feature a common metadata model with a limited number of metadata tags, including URLs for retrieving additional metadata from the blue infrastructures, and an URL to the Blue-Cloud data brokerage mechanism, to be developed for handling requests by users for accessing the associated data collections as managed at the blue infrastructures. The data brokerage mechanism will interact with API's (already existing or to be developed) at each of the blue data infrastructures. In practice, the data brokerage might have to deal with direct download links from fully open data repositories next to dealing with local shopping mechanisms, requiring authentication and authorization.

This way, the Blue Cloud data discovery and access service will provide a common interface, both by web services as by GUI, for discovery and retrieval of data collections from the federated blue data infrastructures. Users can download and store the retrieved data collections on their own machines or at the data pool environment that is implemented through the Workspace service provided as part of the Blue-Cloud VRE system.

It should be noted that the delivery of data collections will be in the existing formats and using vocabularies as offered by each of the blue data infrastructures. In a later stage, a further harmonisation at data level will be considered, where possible.

For analytical processes and for generating data products, activities are ongoing in WP3 and WP4 for analysing, specifying and developing the five Demonstrator Virtual Labs. These Virtual Labs will be deployed on the Blue-Cloud Virtual Research Environment (VRE). The VRE will be largely based upon the existing D4Science infrastructure, which was adopted and used earlier successfully for supporting the BlueBridge project with many Virtual Labs. The existing D4Science infrastructure will be adopted and adapted for the Blue-Cloud, and new services will be added, where needed. The D4Science infrastructure already has a utility for building and maintaining provenance information about the processes as applied for generating specific data products in the Virtual Labs. These includes storing information on data input, use and settings of algorithms, and data output into standard PROV records (the PROV standard[2] defines a data model, serializations, and definitions to support the interchange of provenance information on the Web). These provenance records support to document data products and provide very useful information about the data products for any potential use and re-use performed both by end-users and machine algorithms. They are also very useful for researchers that want to analyse and re-run the analytical processes, possibly changing input or settings for comparisons. Moreover, the D4Science infrastructure also has a catalogue service for all data sets and data products as managed for the Virtual Labs. This concerns two catalogues:

- a **CSW-compliant catalogue,** based on GeoNetwork technology, enabling users to browse and search for geospatial items by relying on the accompanying metadata;
- an **SDMX-compliant catalogue,** based on Fusion Registry technology – for searching statistical data by relying on their structural metadata;

which are joint in one **overall catalogue,** based on CKAN technology, which enables users to perform faceted search on the entire set of resources managed by the Virtual Labs. In the Blue-Cloud data input from the blue data infrastructures will be arranged by bridging between the planned Blue-Cloud data discovery and access service and the planned Blue-Cloud VRE data pool, which also can be filled by users with data sets from other sources. The existing D4Science VRE data catalogues might be adopted for giving VRE users overview and access to all data sets in the VRE input data pool as well as to all data products generated by them in the Virtual Labs, while an additional catalogue (virtual subsetting) configuration might be set-up as a public catalogue to give metadata and access to selected VRE data products. This set-up will be further analysed and specified as part of the WP2 activities for the overall Blue-Cloud architecture.

Users can be external persons from science, government, and industry, who want to discover and access data sets from the blue data infrastructures and/or data products as resulting from the Virtual Labs. They might want to use these data and data products for a wide range of applications. Internal users are researchers that will have an account on the VRE for specific Virtual Labs and that will undertake activities for deploying, fine tuning, and running the Virtual Labs for different use cases.

[2] PROV standard: https://www.w3.org/TR/prov-overview/

3 FAIR data

3.1 FAIR principles

The FAIR concept relates to "Data and services that should be Findable, Accessible, Interoperable, and Re-usable, both for machines and for people." The emphasis is on machine FAIRness. In the following paragraphs, the current status for the Blue-Cloud will be described, and reference will be made to activities which are already planned as part of the Blue-Cloud Work Plan and to activities which might be added, for improving FAIRness.

3.2 Making data findable, including provisions for metadata

Data are **Findable** when they are described by sufficiently rich metadata and registered or indexed in a searchable resource that is known and accessible to potential users. Additionally, a unique and persistent identifier should be assigned such that the data can be unequivocally referenced and cited in research communications. The identifier enables persistent linkages to be established between the data, metadata and other related materials in order to assist data discovery and reuse. Related materials may include the code or models necessary to use the data, research literature that provides further insights into the creation and interpretation of the data and other related information. [[3]]

As described in Chapter 2, the data input for the Blue-Cloud VRE and its Virtual Labs will be derived mostly from the blue data infrastructures which are represented in the Blue-Cloud project. They are leading in the European marine data management landscape, managing and giving discovery and access to data and data products originating from a large number (thousands) of research institutes, national governmental agencies and departments, European and international associations, and industry parties. For that purpose, each has developed and is operating their dedicated discovery and access services, applying community standards and principles. Moreover, several of the blue data infrastructures are engaged in activities for analysing and improving their FAIRness, for instance as part of the ENVRI-FAIR project.

However, in practice different metadata formats and different vocabularies are being applied, which is largely due to existing community differences and history and background of the different initiatives. This situation is gradually changing, under influence of increased cooperation and the push for FAIRness, but also in the future differences will remain and have to be overcome by interoperability, because 'one size fits all' is not feasible and adopting common standards can be quite disruptive and costly for the existing infrastructures, their networks and best practices. Also, there is no room in the Blue-Cloud project itself for working on improving the FAIRness of the constituent blue cloud infrastructures. Those improvements have to come from other projects, while the Blue-Cloud has to maintain its relations and interactions with the blue data infrastructures and their possible innovations in time.

[3] TURNING FAIR INTO REALITY, Final Report and Action Plan from the European Commission Expert Group on FAIR Data, European Union 2018, doi: 10.2777/1524

As outlined in paragraph 2.7, the Blue Cloud data discovery and access service will be developed in WP2 to give federated discovery and access to the blue data infrastructures. This will be done by building and maintaining a common Blue-Cloud metadata catalogue, which will feature a common metadata model with a limited number of metadata tags, to be completed by mappings with each of the metadata services of the blue data infrastructures. The Blue-Cloud common data model will focus on describing data collections by means of fields for What, When, Where, Who, and additional fields for querying more detailed metadata at each source and for facilitating access by means of downloading of the data collections from the source infrastructures. The resulting Blue-Cloud metadata catalogue will be established adopting the GEODAB brokerage service principles. The Blue-Cloud discovery service will then be made available by means of web services for machine-to-machine interactions (CSW; OAI-PMH; and SPARQL RDF endpoint), as well as by GUI for human users.

Identified additional actions for FAIRness:

- The common Blue-Cloud metadata catalogue will be populated and maintained by regularly harvesting and applying dedicated mapping profiles. Initially, these will be filled with the information as provided from the sources, applying international standards for When and Where; in a later stage, once more experience has been built up, semantic mappings might be applied against the well-developed and maintained SeaDataNet vocabularies for What and Who.

- Increasingly, blue data infrastructures are adopting DOI identifiers with landing pages for providing metadata and possible download options of their data resources. Wherever available, these DOIs will be used in the Blue-Cloud common catalogue for providing users additional metadata from the sources.

As also indicated in paragraph 2.7, the Blue-Cloud VRE will be largely based upon adopting and adapting the existing D4Science e-infrastructure for the purpose of the Blue-Cloud functionalities as originally planned and as to be identified when further analysing and developing the Blue-Cloud Demonstrator Virtual Labs. As part of the Blue-Cloud, a bridge will be developed between the Blue Cloud data discovery and access service and the Blue-Cloud VRE, facilitating to store retrieved data collections in a VRE data pool. Users should also be able to import data sets from other sources into this VRE data pool. The VRE data pool is realised through a family of technologies leveraging cloud-based technologies to deliver high scalability, security, reliability, and availability. To any object stored in it, a unique and persistent identifier is generated and a unique and persistent web locator (PURL) is generated. Versioning is automatically enabled to guarantee persistence also when new versions are incrementally generated and stored.

Identified additional actions for FAIRness:

- On top of the VRE data pool, a data catalogue will be configured as part of WP4 activities to give Virtual Lab users access to (parts) of the data pool for using the selected data in their analytical processes. It is considered to do this by adapting the existing D4Science catalogue service (see paragraph 2.7) which enabled users to perform faceted search on the entire set of resources managed by the Virtual Labs. For the Blue-Cloud exploitation, an instance of this catalogue will be created for each Virtual Lab and it will cover the contents of the VRE data pool, which is filled with data sets from the Blue-Cloud data discovery and access service and from other sources. It will also cover data products generated in the Virtual Lab and algorithms and analytical methods used to analyse those data products. For each of the managed products in the Virtual Lab catalogue, a persistent and unique URL will be generated and provenance information will be linked.

- An additional VRE catalogue will provide an integrated view on the content provided in each Virtual Lab. It will have an internal interface, giving registered VRE users discovery and access to data and generated data products on the VRE platform, differentiated depending on their accounts, and an external interface, giving public users discovery and access to published data products as generated with the Virtual Labs. This potential set-up will be further analysed and specified as part of the WP2 activities for the overall Blue-Cloud architecture, while its possible deployment will be done by WP4 in cooperation with WP2. In this process, also attention will be given to adoption of DOIs for published data products, adoption of use of SeaDataNet vocabularies, and versioning.

- The D4Science infrastructure already has a utility for building and maintaining provenance information for data products as generated in the Virtual Labs. This should be combined with the VRE catalogue as planned above.

3.3 Making data openly accessible

Accessible data objects can be obtained by humans and machines upon appropriate authorisation and through a well-defined and universally implementable protocol. Anyone should be able to access at least the metadata. It is important to emphasise that Accessible in FAIR does not mean Open without constraint. Accessibility means that the human or machine is provided - through metadata - with the precise conditions by which the data are accessible and that the mechanisms and technical protocols for data access are implemented such that the data and/or metadata can be accessed and used at scale, by machines, across the web.[[4]]

As outlined in paragraph 2.7 and 3.2 above, the **Blue Cloud data discovery and access service** will be developed in WP2 to give federated discovery and access to the blue data infrastructures. The data access part will be handled by a Blue-Cloud data brokerage mechanism, to be developed for handling requests by users for access to the associated data collections as managed at the blue infrastructures. For that purpose, D2.1 gives an overview of current data delivery mechanisms at each of the blue data infrastructures, which are currently further analysed. In practice, some blue data infrastructures feature fully open data download links, while some other feature a shopping mechanism, with user login, and possibly making a clear distinction between unrestricted and restricted data. For restricted data, requests require decisions by data originators. While unrestricted data sets are released for

[4] TURNING FAIR INTO REALITY, Final Report and Action Plan from the European Commission Expert Group on FAIR Data, European Union 2018, doi: 10.2777/1524

downloading immediately, only requiring login for track and tracing purposes, and consenting to the license of the data.

The planned approach is to adopt or formulate API's together with the technical representatives of the blue data infrastructures. These should be configured by each to be fit for interacting with the Blue-Cloud data brokerage service. The API's should deal with the particulars of the local set-ups and also should arrange that data requests can be handled and responded at the agreed collection levels.

The integrated Blue-Cloud metadata - data brokerage service will allow users (humans and machines) to discover and browse metadata without any restrictions, while requesting access will require for both user types a login by which the Blue-Cloud service can keep track of transactions in a ledger and can interact with login systems that some of the blue data infrastructures are operating. The ledger will allow the repositories to oversee transactions as made for them, while users can oversee their transactions, perform downloading, and follow requests that are delayed because of access restrictions.

The delivered data will follow the formats as provided by the blue data infrastructures and as indicated in paragraph 2.3. These formats range from general international standards to specific community standards. In the latter case, these will be fit for specific software as recommended by the blue data infrastructures, while international standards can be handled by a wider range of software tools and services. Information about the formats and appropriate software and services can be found at the portals of the blue data infrastructures.

As indicated in paragraph 3.2 it is considered to adapt the current VRE catalogue system at the D4Science e-infrastructure to give Virtual Lab users discovery and access to (parts) of the VRE data pool and stored data products as generated with the Virtual Labs. A virtual subsetting might be applied on this catalogue for offering external users, public discovery and access to selected data products as generated with the Virtual Labs.

Identified additional actions for FAIRness:

- Consider to outfit the public VRE data products catalogue with a logon for downloading so that information can be gathered on the users and their interests. This list of users can then also be used for follow-up activities towards these users. Moreover, the download procedure should include users agreeing to a license for the data products, including a disclaimer and an encouragement for acknowledging the products by their DOIs in publications.

3.4 Making data interoperable

Interoperable data and metadata are described by community and/or domain standards for technical interoperability and vocabularies for semantic interoperability, and they include qualified references to other data or metadata. It is this that allows the data to be 'machine-actionable'. Interoperability is an essential feature in the value and usability of data. Legal interoperability of data has to be considered as well. In FAIR, legal interoperability falls under the principle that data should be 'Reusable'. [[5]]

The delivered data will follow the formats as provided by the blue data infrastructures and as indicated in paragraph 2.3. These formats range from general international standards to specific community standards. In the latter case, these will be fit for specific software as recommended by the blue data infrastructures, while international standards can be handled by a wider range of software tools and services. The blue data infrastructures also make use of specific vocabularies for marking up metadata and data. Information about the formats, vocabularies and appropriate software and services can be found at the portals of the blue data infrastructures.

As part of WP4 activities are planned to improve the interoperability between these data sets originating from the different blue data infrastructures, where possible. WP4 activities will be undertaken aiming at designing and developing cloud services enacting aggregation and harmonizing of the possible heterogeneous datasets as discovered and delivered to the VRE data pool. The goal of the so-called Taming service is to make the ingested and available data sets better suitable for specific data analytics tasks and related algorithms (blue services) the user is willing to perform. The harmonization will include making a priority list of possible formats and types of data as provided by blue data infrastructures, and compiling and deploying a set of existing conversion software packages in cooperation with Blue Cloud partners. In addition, possible developments might be undertaken for integration and upgrading of existing software for new formats. The taming service activity will be built upon existing tools and developments for such software in SeaDataNet, EurOBIS, EBI, NOAA, and others.

Use of common vocabularies in all metadata and data formats is an important prerequisite towards consistency and interoperability. Common vocabularies consist of lists of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic and wider community. Using standardised sets of terms solves the problem of ambiguities associated with data markup and also enables records to be interpreted by computers. This opens up data sets to a whole world of possibilities for computer aided manipulation, distribution and long-term reuse.

Therefore, as part of the Blue-Cloud WP4 Taming service activities, it will be explored if common vocabularies can be adopted and deployed in the Blue-Cloud catalogues which should include setting up semantic mappings between sources and targets. A very good candidate for targets is the set of common vocabularies which have been initiated and are maintained by SeaDataNet for the marine and oceanographic data community.

SeaDataNet (https://www.seadatanet.org), one of the Blue-Cloud blue data infrastructures, is a major pan-European infrastructure for managing, indexing and providing access to marine data sets and data products, acquired by European organisations from research cruises and other observational activities in European coastal marine waters, regional seas and the global ocean. Founding partners are National

[5] TURNING FAIR INTO REALITY, Final Report and Action Plan from the European Commission Expert Group on FAIR Data, European Union 2018, doi: 10.2777/1524

Oceanographic Data Centres (NODCs), major marine research institutes, UNESCO-IOC, ICES, and EC-JRC. The SeaDataNet network was initiated in the 1990s and its network of data centres and infrastructure with standards, tools, and services has expanded, inter alia with support of many EU projects. SeaDataNet is also a major infrastructure in the European Marine Observation and Data network (EMODnet) and has an MoU and close cooperation with Copernicus Marine Environmental Monitoring Service (CMEMS). SeaDataNet develops, governs, and promotes common standards, vocabularies, software tools, and services for marine data management, which are freely available from its portal and widely adopted and used.

The SeaDataNet common vocabularies contains multiple lists of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic and wider community. It follows the W3C SKOS specification for encoding data dictionaries and taxonomies served. The SeaDataNet vocabulary services are technically managed and hosted by the British Oceanographic Data Centre (BODC) at the NERC Vocabulary Server (NVS2.0). Content governance of the vocabularies is very important and is done by a combined SeaDataNet and MarineXML Vocabulary Content Governance Group (SeaVoX), moderated by BODC, and including many European and international experts. Documentation about the SeaDataNet controlled vocabularies can be found at:
https://www.seadatanet.org/Standards/Common-Vocabularies

Moreover, the SeaDataNet network of data centres maintains and publishes a series of European directory services which are also widely used for marking up metadata and data:

- European Directory of Marine Organisations (EDMO)
- European Directory of Marine Environmental Data Sets (EDMED)
- European Directory of Marine Research Projects (EDMERP)
- European Directory of Ocean-Observing Systems (EDIOS)
- Cruise Summary Reports (CSR)

Documentation about the SeaDataNet Directories can be found at:
https://www.seadatanet.org/Metadata

Both the SeaDataNet controlled vocabularies and SeaDataNet Directories are made available as web services for machines and by means of client interfaces for end-users. The client interfaces provide end-users options for searching, browsing and CSV-format export of selected entries. The machine interfaces are provided via a SOAP API and as SPARQL endpoints.

Identified additional actions for FAIRness:

- The adoption of SeaDataNet vocabularies will be considered where mappings are possible and not too extensive in order to contribute to harmonisation of metadata and data, improving the interoperability. This will apply to the planned VRE catalogues both for input data and data products resulting from Virtual Labs.

3.5 Increase data re-use (through clarifying licenses)

For data to be **Reusable**, the FAIR principles reassert the need for rich metadata and documentation that meet relevant community standards and provide information about provenance, reporting how data was created and information about consecutive data reduction or transformation processes to make data more usable, understandable or 'science-ready'. The ability of humans and machines to assess and select data on the basis of criteria relating to provenance information is essential to data reuse, especially at scale. Reusability also requires that the data be released with a 'clear and accessible data usage license': in other words, the conditions under which the data can be used should be transparent to both humans and machines. [6]

The data input for the Blue-Cloud VRE and its Virtual Labs will be derived mostly from the blue data infrastructures which are represented in the Blue-Cloud project. Each has developed and is operating their dedicated discovery and access services, applying community standards and principles. Moreover, several of the blue data infrastructures are engaged in activities for analysing and improving their FAIRness. In a number of cases, this might include activities for enriching metadata by adding more details about observations and quality control methods applied, which might be done by linking to other repositories such as the SeaDataNet directories (see paragraph 2.3), Ocean Best Practices, published literature repositories, and others.

Anyway, the Blue-Cloud data discovery and access service will use the metadata from the blue data infrastructures as it is offered and there is no budget nor activity in the Blue-Cloud project itself for improving their richness.

This is different for the Blue-Cloud VRE catalogue aiming at publishing data products as generated by the researchers with the Virtual Labs. Their metadata will be enriched with the provenance documentation, this way providing users detailed information about the way that the data products were generated, which will be very useful for wider use. And the provenance information will give scientific users detailed information how the data products might be reproduced with the Virtual Labs which is again very useful as it allows VRE users to repeat the same analyses, but for instance changing input and/or settings.

The Blue-Cloud data discovery and access service will function with a shopping basket mechanism, whereby each user has to register once. All users can freely query and browse the data catalogue; however, submitting requests for data access via the shopping basket requires that users are logged on. Data requests can concern unrestricted and/or restricted data sets. Requests for unrestricted data sets will be processed immediately after submission and requested data sets, once retrieved from the blue data infrastructures, will be made ready for download from a Blue-Cloud data cloud. While requests for restricted data sets (where applicable for the blue data infrastructures) are forwarded to the managers of connected blue data infrastructure for their consideration, most of the cases deliberating with data originators. The processing of all data requests will be controlled

[6] TURNING FAIR INTO REALITY, Final Report and Action Plan from the European Commission Expert Group on FAIR Data, European Union 2018, doi: 10.2777/1524

by a Shopping Ledger component which will be integrated in the catalogue interfaces. The Shopping Ledger administers and processes all transactions, communicating with the connected blue data infrastructures through the data brokerage service. Users will receive confirmation e-mails of their data set requests and subsequent processing, and can also check progress and undertake data downloading from the Shopping Ledger which is part of their personal dashboard. On their turn, managers from blue data infrastructures will be able to follow via the Shopping Ledger all transactions for their data sets online and can also handle requests for restricted data, which require their mediation.

Each of the blue data infrastructures might have their own data policy and associated licenses for use. Those licenses will be mentioned in the metadata and will be propagated to users, so that they can look up the prevailing licenses.

Identified additional actions for FAIRness:

- Considering that in most cases there is fully open access offered by the blue data infrastructures, it will be explored if in those cases a common Blue-Cloud license might be applied, based upon CC BY principles. However, use of other licenses by data originators has to be respected.

- For the Blue-Cloud VRE catalogue for published data products as generated by the Virtual Labs, a comparable Shopping Basket mechanism with logon, shopping ledger and CC BY license might be considered. This will be further discussed.

4 Allocation of resources

There are a number of categories of costs to be considered for making Blue-Cloud data and services FAIR. These include:

- Data input costs: the input data will be retrieved from the connected blue data infrastructures. The costs for getting connected, which can include content and service adaptations for metadata and data retrieval, will be carried by the Blue-Cloud product budget. The cost for operation of the blue data infrastructures will fall outside the Blue-Cloud and is to be carried by the blue data infrastructures themselves. The costs of the operation of the Blue-Cloud data discovery and access service, both service and staff resources, will be part of the exploitation of the Blue-Cloud infrastructure.
- Data curation costs and costs of data products generation: the input data as retrieved from the blue data infrastructures and other sources might require additional processing and conversions to make the data sets fully fit for purpose of the analytical services in the Virtual Labs. Part of these efforts and costs might relate to staff costs of the managers and operators of the Blue-Cloud infrastructure and will be included in the exploitation of the Blue-Cloud infrastructure. While, another part will be carried out by researchers that are performing analyses using the Virtual Labs. The activities of the Blue-Cloud scientific teams will be covered by the Blue-Cloud project budget. External users, if allowed within the project stage, will have to bear these costs themselves. The same applies to the costs of performing analyses with the Virtual Labs to generate data products. The activities of the Blue-Cloud scientific teams within the project duration will be covered by the Blue-Cloud project budget. External users, if allowed within the project stage, will have to bear these costs themselves.
- Operation costs of Blue-Cloud infrastructure: these include the costs for operating the different service components of the Blue-Cloud infrastructure such as for storage, catalogue services, analytical processes, and publishing. These are part of the exploitation of the Blue-Cloud infrastructure and will be covered by the Blue-Cloud project budget during the project lifetime. After the end of the project and for the duration of two years, these same costs will be covered by the D4Science infrastructure operated by CNR.

Assessing the costs at this stage of the project is not possible, as the overall Blue-Cloud architecture, the dedicated services, and the organisation of the operations with specific roles for selected Blue-Cloud partners are not known yet at the required detail.

The same applies to giving information about future exploitation of the Blue-Cloud infrastructure beyond the Blue-Cloud project. That will be included in the activities in WP6 for preparing an exploitation and business plan as well as a roadmap to 2030, exploring opportunities both for wider application and expansion of the initial Blue-Cloud infrastructure and for funding.

In the project phase, the responsibilities for data management will be carried jointly by the partners that are involved in WP2, WP3, and WP4.

5 Data Security

The Blue-Cloud technical framework will feature:

1) the Blue Cloud data discovery and access service component to serve federated discovery and access to blue data infrastructures
2) the Blue Cloud Virtual Research Environment (VRE**)** component to provide a Blue Cloud VRE as a federation of computing platforms and analytical services. This VRE will provide the platform for developing and running the Virtual Labs for the scientific Demonstrators.

It will be arranged that data will be safely handled and stored in both components. This will be done by relying on repositories operated by MARIS and EUDAT for the Blue-Cloud discovery and access service, whereby MARIS will run the metadata catalogue and related services, while retrieved datasets will be temporarily stored at EUDAT for downloading by users or transfer and ingestion into the VRE data pool. That data pool will be safely handled in the Blue-Cloud VRE platform, by relying on repositories operated by CNR in the underlying D4Science infrastructure. This way the safety of the data and the accompanying metadata will be secured. Standard practices are in place for this including transparent replication of content across several machines and systematic backup of the content.

In particular, the following strategies for data security are planned for the Blue-Cloud components:

- Use of high availability storage systems which keep both the data and their enacting system
- Replicated on-site and off-site, enabling continuous access to systems and data, even after a disaster;
- Use of Hybrid Cloud solutions that replicate both on-site and off-site the main storage systems. This solution provides the ability to instantly fail-over to local on-site hardware all of the storage systems, but in the event of a physical disaster, storage systems can be brought up in additional data centers;
- Backups made at regular intervals of all storage systems (the maximum interval for two consecutive backups is one day);
- Replication of service to an off-site location, which overcomes the need to restore the service (only the data need to be restored or synchronized).

In addition to preparing for the need to recover systems, also precautionary measures can be implemented with the objective of preventing a disaster in the first place. These include:

- local mirrors of systems and/or data and use of disk protection technology such as RAID;
- surge protectors — to minimize the effect of power surges on delicate electronic equipment;
- use of an uninterruptible power supply (UPS) and backup generator to keep systems going in the event of a power failure;
- fire prevention/mitigation systems equipped with alarms and fire extinguishers;
- firewall and network frameworks to avoid intrusion and attacks.

- Address data recovery as well as secure storage and transfer of sensitive data

Finally, the connection between the Blue-Cloud components is ***secured with Transport Level Security*** (TLS) that provides communications security over the computer network. It ensures privacy and data integrity between two communicating computer applications. In particular, any connections between a client (e.g., a web browser) and a Blue-Cloud server have the following properties:

- The connection is *private* (or *secure*) thanks to the adoption of the symmetric cryptography to encrypt the data transmitted. The keys for this symmetric encryption are generated uniquely for each connection and are based on a shared secret negotiated at the start of the session. The server and client negotiate the details of which encryption algorithm and cryptographic keys to use before the first byte of data is transmitted. The negotiation of a shared secret is both secure (the negotiated secret is unavailable to eavesdroppers and cannot be obtained, even by an attacker who places themselves in the middle of the connection) and reliable (no attacker can modify the communications during the negotiation without being detected);
- The identity of the communicating parties can be *authenticated* using public-key cryptography. This authentication can be made optional at client side, but is ensured at the server side;
- The connection ensures *integrity* because each message transmitted includes a message integrity check using a message authentication code to prevent undetected loss or alteration of the data during transmission;
- The connection ensures forward secrecy, ensuring that any future disclosure of encryption keys cannot be used to decrypt any TLS communications recorded in the past.

These measures will be detailed further during the project when the development of the infrastructure components has progressed.

6. Ethical aspects

In this stage of the project it is not possible to identify legal or ethical issues concerning the data that will be handled as input for the Blue-Cloud and the data products that will be generated using the Blue-Cloud Virtual Labs.

The Blue-Cloud project is not explicitly dealing with any activity related with personal data collection. Only as part of the registration of users for the operation of the Blue-Cloud catalogues and for creating accounts at the Blue-Cloud VRE information will be gathered of users and their activities. These uses and measures for securing these personal data is clearly explained to users that register for these services (see https://www.iubenda.com/privacy-policy/441050) in accordance with the GDPR. The Privacy Policy at the Blue-Cloud portal includes a Data Protection Notice and a reference to the Cookies Policies (https://blue-cloud.d4science.org/cookie-policy) is also noticed at the homepage. Finally, the Terms of Use (https://blue-cloud.d4science.org/terms-of-use) clarify all the policies regulating the usage and exploitation of the Blue-Cloud services. It will be annotated and extended along the time with the specification of the policies associated with the new Blue-Cloud services.

Moreover, specific surveys might be undertaken by the Blue-Cloud project, which will collect personal data. These will be treated again in the most appropriate and secure way and users will be well informed about this approach.