Project Acronym	***BlueBRIDGE***
Project Title	***Building Research environments for fostering Innovation, Decision making, Governance and Education to support Blue growth***
Project Number	***675680***
Deliverable Title	***BlueBRIDGE Data Management Plan: Final Version***
Deliverable No.	***D2.3***
Delivery Date	***November 2017***
Authors	***Leonardo Candela, Donatella Castelli, Pasquale Pagano***

BlueBRIDGE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 675680

DOCUMENT INFORMATION

PROJECT	
Project Acronym	BlueBRIDGE
Project Title	Building Research environments for fostering Innovation, Decision making, Governance and Education to support Blue growth
Project Start	1st September 2015
Project Duration	30 months
Funding	H2020-EINFRA-2014-2015/H2020-EINFRA-2015-1
Grant Agreement No.	675680
DOCUMENT	
Deliverable No.	D2.3
Deliverable Title	BlueBRIDGE Data Management Plan: Final Version
Contractual Delivery Date	November 2017
Actual Delivery Date	April 2018
Author(s)	Leonardo Candela (CNR), Donatella Castelli (CNR), Pasquale Pagano (CNR)
Editor(s)	Leonardo Candela (CNR)
Reviewer(s)	Sara Garavelli (TRUST-IT)
Contributor(s)	J. Barde (IRD), A. Ellenbroek (FAO), D. Katris (CITE), Y. Marketakis (FORTH)
Work Package No.	WP2
Work Package Title	Project Governance, Exploitation and Sustainability
Work Package Leader	CNR
Work Package Participants	ERCIM, ENG, UOA, FAO, ICES, IRD, FORTH, TRUST-IT, I2S, CITE, CLS, GRID Arendal, PMBret
Estimated Person Months	3
Distribution	Public
Nature	Report
Version / Revision	1.0
Draft / Final	Final
Total No. Pages (including cover)	35
Keywords	Datasets; Management; Data Storage; Data Repository; Data dissemination; Data preservation;

DISCLAIMER

BlueBRIDGE (675680) is a Research and Innovation Action (RIA) co-funded by the European Commission under the Horizon 2020 research and innovation programme

The goal of BlueBRIDGE*, Building Research environments for fostering Innovation, Decision making, Governance and Education to support Blue growth*, is to support capacity building in interdisciplinary research communities actively involved in increasing the scientific knowledge of the marine environment, its living resources, and its economy with the aim of providing a better ground for informed advice to competent authorities and to enlarge the spectrum of growth opportunities as addressed by the Blue Growth societal challenge.

This document contains information on BlueBRIDGE core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as BlueBRIDGE Board members. 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 BlueBRIDGE Consortium and its experts, 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 the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
CC	Creative Commons
CC-BY	Creative Commons Attribution Licence
CC-BY-NC	Creative Commons Attribution-Non Commercial Licence
Compound information object	Is an information object formed by either a dataset or of a set of related compound objects semantically forming a single entity
CSV	Comma Separated Values
CSW	OGC Catalogue Service
DMP	Data Management Plan
FCR	Feed Conversion Ratio
ISO 19115	Geographic information -- Metadata
ISO 19139	Geographic information -- Metadata -- XML schema implementation
NetCDF	Network Common Data Form
QAO	Quality Assurance Office
SFR	Specific Feeding Ratio
SGR	Specific Growth Ratio
VRE	Virtual Research Environment
WCS	OGC Web Coverage Service
WFS	OGC Web Feature Service
WMS	OGC Web Map Service
XDR	External Data Representation

TABLE OF CONTENT

DOCUMENT INFORMATION **2**

DISCLAIMER **3**

GLOSSARY **4**

TABLE OF CONTENT **5**

DELIVERABLE SUMMARY **7**

EXECUTIVE SUMMARY **8**

1 Introduction **9**

2 Data summary **10**

2.1 Purpose of the data collection/generation **10**

2.2 Relation to the objectives of the project **10**

2.3 Types and formats of data generated/collected **12**

2.4 Existing data re-use **14**

2.5 Origin of the data **16**

2.6 The expected size of the data **17**

2.7 Data utility: to whom will it be useful **18**

3 FAIR data **20**

3.1 Making data findable **20**

3.1.1 Discoverability of data (metadata provision) 20

3.1.2 Identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? 21

3.1.3 Naming conventions used 22

3.1.4 Approach towards search keyword 22

3.1.5 Approach for clear versioning 22

3.1.6 Standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how 23

3.2 Making data openly accessible **23**

3.2.1 Which data will be made openly available? If some data is kept closed provide rationale for doing so 24

3.2.2 How the data will be made available 25

3.2.3 What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? 25

3.2.4 Where the data and associated metadata, documentation and code are deposited 25

3.2.5 How access will be provided in case there are any restrictions 26

3.3 Making data interoperable **26**

3.3.1 Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability 26

3.3.2 Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies? 27
3.4 Increase data re-use 27
3.4.1 How the data will be licenced to permit the widest reuse possible 27
3.4.2 When the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed 27
3.4.3 Whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why 28
3.4.4 Data quality assurance processes 28
3.4.5 The length of time for which the data will remain re-usable 28
4 Allocation of resources 29
4.1 Costs for making your data FAIR. Describe how you intend to cover these costs 29
4.2 Responsibilities for data management in your project 29
4.3 Costs and potential value of long term preservation 30
5 Data security 32
6 Ethical aspects 33
7 Conclusion 34
REFERENCES 35

DELIVERABLE SUMMARY

This deliverable documents the BlueBRIDGE data management strategy. The strategy was developed by producing three successive versions of this deliverable: a preliminary version at M6, an intermediate version at M18, a final version, this one, at M27. No major modifications have been done to this version since plans established at M18 have mostly been confirmed. The deliverable extends the content of D2.2 primarily by adding details on data management responsibilities and data security. The deliverable is organised around solutions and approaches aiming at making BlueBRIDGE data / datasets findable, accessible, interoperable and re-usable (FAIR). In particular, the deliverable provides an overview of the data managed by the project by describing the purpose of the data collection / generation and its relation to the objectives of the project, the types and formats, any re-use of existing data as well as giving hints on "data utility", i.e. scenarios and stakeholders that might benefit from BlueBRIDGE data. Then the deliverable describes the specific solutions and approaches for making BlueBRIDGE data FAIR, namely (i) the provisioning of an array of catalogues including an "overall" one making it possible to associate suitable metadata to the datasets, (ii) the provisioning of several repositories for data thus to deal with the heterogeneity characterising the VREs as well as preserving the peculiarities of the diverse typologies, (iii) the support for standard and controlled vocabularies, and (iv) the promotion of practices favouring the re-use, e.g. licences making the data as open as possible. Finally, the deliverable concludes by discussing costs and resources underlying the data management and the data security, legal and ethical issues.

EXECUTIVE SUMMARY

Data is a key asset for the economy and our society similar to the classic categories of human and financial resources. In particular, making data publicly available will contribute to offer a series of opportunities, e.g., creating jobs, spurring growth, boosting research productivity and creativity, helping people, engaging citizens [10].

The European Commission is promoting data availability through a series of actions including a specific policy (Art. 29.3 "Open access to research data") requesting funded projects to "*(a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate – free of charge for any user – the following: (i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible; (ii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan'; (b) provide information – via the repository – about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and – where possible – provide the tools and instruments themselves).*". Such a policy goes in tandem with a new element in Horizon 2020 which is requesting projects to develop and use Data Management Plans.

This Data Management Plan outlines how the data "produced" (either generated or collected) during the BlueBRIDGE project are planned to be managed (during the project and after the project completion). This final version of the deliverable focuses on the solutions and approaches put in place to make BlueBRIDGE data findable, accessible, interoperable and re-usable (FAIR). The deliverable is organised according to the template prepared by the European Commission for "FAIR Data Management in Horizon 2020" and responds to the set of questions aiming at capturing a summary of the data collected / generated by BlueBRIDGE, the approaches for making data FAIR, the resources needed to implement what is planned and the data security and ethical aspects addressed.

In the context of the BlueBRIDGE project a rich array of datasets is expected to be managed including geospatial data, species data, tabular data and software. However, the primary goal of BlueBRIDGE is to support the creation and development of Virtual Research Environments rather that the systematic production of data of any sort. Because of this heterogeneity BlueBRIDGE is requested to deal with many standards and formats. Methodologies for data management are diverse yet built by relying on a shared data infrastructure (D4Science.org). This deliverable describes how VRE members are allowed to produce data and make them compliant with the FAIR principles by relying on a set of catalogues (including an "overall" one), repositories and other supporting facilities.

1 INTRODUCTION

A Data Management Plan (DMP) is a document outlining how the data "produced" (generated / collected) during a research project are planned to be handled (during the project and after the project completion).

This deliverable is the final version of the BlueBRIDGE Data Management Plan. Its format is compliant with the expectations emerging from the willingness to make the datasets findable, accessible, interoperable, and re-usable (FAIR) [10].

For the sake of this deliverable the following definition of dataset is used:

A dataset is any set of data (no matter how many files it materialises) that is worth to be considered as a unit for data management activities [1][15]

Examples of possible datasets are the following:

- The set of files (and references to files) stored in a VRE workspace. These may include the "experiments" executed by VRE members, the (reference to) data analysed and the results obtained;
- The set of posts and comments produced by the VRE members;
- Any dataset produced by aggregating data from data providers for the sake of building an integrated dataset out of the aggregated data (e.g. Knowledge Bases);
- The material of a training course;
- A dataset documenting and providing evidence for either a report or a publication produced in the context of project (and VREs supported) activities.

The rest of the deliverable is organised as follows. Section 2 provides the reader with a summary of the data falling under the BlueBRIDGE Data Management Plan, including information on data types and formats as well as on the purpose leading to them. Section 3 describes the practices and solutions BlueBRIDGE put in place in order to make the data FAIR. Section 4, 5 and 6 discuss challenging issues like costs and resources underlying the data management as well as data security, legal and ethical issues. Finally, Section 7 concludes the report.

2 DATA SUMMARY

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of a Data Management Plan is expected to give answers to the following questions:

- What is the purpose of the data collection/generation and its relation to the objectives of the project?
- What types and formats of data will the project generate/collect?
- Will you re-use any existing data and how?
- What is the origin of the data?
- What is the expected size of the data?
- To whom might it be useful ('data utility')?

2.1 PURPOSE OF THE DATA COLLECTION/GENERATION

The BlueBRIDGE project overall objective is to support the development of a series of *Virtual Research Environments (VREs)* facilitating communities of scientists, innovators from SMEs and educators operating in different domains (e.g. fisheries, biology, economics, statistics, environment, mathematics, social sciences, natural sciences, computer science) in their knowledge production chain, from the initial phases, data collection and aggregation, to the production of indicators for competent authorities and investors.

From the end-user perspective, Virtual Research Environments are web-based working environment providing them with a set of facilities for seamlessly accessing and processing the data of interest and producing new ones. BlueBRIDGE, by relying on the D4Science infrastructure, enacts the development and operation of the VREs by relying on common / shared services. This has a number of implications from the data management perspective:

- Data / datasets are expected to be “integrated” one time only within the infrastructure and be re-usable, according to their re-use policy, in as many VREs as possible / worth to have;
- Whenever a data / dataset is produced in a VRE, it is automatically “integrated” into the underlying infrastructure thus becoming a possible constituent to re-use in the context of another VRE (in accordance to the specific re-use policy).

Thus the purpose of BlueBRIDGE data collection and generation activities is twofold:

- To enact the creation and development of the envisaged VREs. This is mainly related with mechanisms aiming at making existing data / datasets available for VRE users;
- To guarantee that new data / datasets resulting from VRE exploitation activity are managed thus to maximize their FAIRness and openness yet making findability, accessibility, interoperability and re-usability compliant with any per data / dataset policy and license accompanying them.

2.2 RELATION TO THE OBJECTIVES OF THE PROJECT

The BlueBRIDGE overall objective of creating and operating VREs discussed in 2.1 is further organised in a set of detailed objectives:

- ***Blue Assessment*** [3]: *Developing and deploying VREs for supporting the collaborative production of scientific knowledge required for* ***assessing the status of fish stocks and producing a global record of stocks and fisheries.*** In such a context there are two typologies of target VREs:

- *Stock Assessment VREs*: providing on-line collaborative environments for Stock Assessment for Blue Growth practitioners with the long-term strategy to produce evidence-based understanding of the status of marine fisheries. In these VREs there is an important re-use of existing data / datasets (cf. Sec. 2.4, e.g. environmental parameters, occurrence points) and the production of specific methods for stock assessment and datasets resulting from the exploitation of such methods (cf. Sec. 2.3).

- *Global Record of Stocks and Fisheries VREs*: providing for the development and consumption of an on-line knowledge base on the Global Record of Stocks and Fisheries for a Blue Growth audience of ecologists, resource managers, market parties, and the general public with the long-term objective to provide evidence-based information on the status of marine stocks and fisheries and promote responsible consumption. In these VREs there is an important re-use of existing data / datasets (cf. Sec. 2.4) and the production of a completely new dataset (actually a knowledge base of stocks and fishery records), i.e. the ***Global Record of Stocks and Fisheries*** (GRSF).

- ***Blue Economy*** [8]: *Developing and deploying VREs for supporting the production of scientific knowledge for* ***analysing socio- economic performance in aquaculture.*** In such a context there are two typologies of target VREs:

  - *Performance Evaluation in Aquaculture VREs*: providing a service that focuses on increasing aquaculture productivity, while minimizing impacts on the environment by providing capacities for aqua-farming companies for performance estimation, benchmarking, decision making and strategic investment analysis. In these VREs the main datasets collected are ***aquafarms production statistics*** (cf. Sec. 2.4) that are going to be exploited to produce ***aquaculture production KPIs*** (cf. Sec. 2.3).

  - *Strategic Investment Analysis and Scientific Planning and Alerting VREs*: Providing an on-line environment, for the probing of investment cases in aquaculture, and to scientists and policy makers for the detection of locations for scientific, environmental or socioeconomic attention. In these VREs there is an important re-use of existing data (cf. Sec. 2.4, e.g. socio-economic indicators, aquaculture farm production statistics) to produce ***investment analysis results*** (cf. Sec. 2.3).

- ***Blue Environment*** [11]: *Developing and deploying VREs for supporting the production of scientific knowledge for* ***fisheries & habitat degradation monitoring***. In such a context there are two typologies of target VREs:

  - *Aquaculture Atlas Generation VREs*: providing scientists with an innovative environment supporting the effective and efficient production of aquaculture products (maps of human activity and natural zones) contributing to an aquaculture atlas compliant with NASO standards. In these VREs there is a re-use of existing datasets (cf. Sec. 2.4, e.g. satellite images) to produce ***aquafarms and cages locations*** (cf. Sec. 2.3).

  - *Protected Area Impact Maps VREs*: providing scientists with an integrated environment supporting the efficient and effective production of maps of vegetation types and human impacts on them and enabling ecosystem degradation analysis. In these VREs there is a re-use of existing datasets (cf. Sec. 2.4, e.g. satellite images) to produce ***marine protected areas maps*** (cf. Sec. 2.3).

- ***Blue Skills***: *Developing and deploying VREs for boosting* ***education and knowledge bridging between research and innovation in the area of protection and management of marine resources, giving them a new volume and thematic and geographical reach.*** In such a context the VREs are very diverse each other depending on the peculiarities of the specific course they are conceived for.

Very often the VREs are conceived to exploit datasets and facilities that are already made available by the underlying infrastructure because they have been selected, integrated or developed for other purposes (namely, Blue Assessment / Economy / Environment VREs). Other datasets and supporting material are provided by course instructors by using the shared workspace. No systematic generation of data / datasets is envisaged for the VREs in the Blue Skills domain.

- ***Blue Commons***: *Developing and deploying a* ***service and resource commons*** *across VREs contributing to the implementation of the* ***"Infrastructure Commons"*** *vision and facilitating* ***the exploitation of existing infrastructure resources*** *and* ***re-use of scientific outcomes.*** This objective is mainly related with the development and operation of the technology needed for the VREs discussed above to work. The major outcome related with it is the comprehensive set of software artefacts contributing to implement the gCube technology [2][7].
- ***Blue Uptake***: *Ensuring uptake of the BlueBRIDGE tools and services* ***within and beyond the scientific and academic communities*** *addressed by the planned VREs, with a particular focus on* ***industry including SMEs,*** *and on* ***other scientific domains & policy making contexts.*** There is no systematic production of datasets stemming from this activity apart from the production of dissemination and outreach material.

2.3 TYPES AND FORMATS OF DATA GENERATED/COLLECTED

BlueBRIDGE VREs deals with a rich array of data typologies and formats (cf. Table 1). The complete list of data / datasets managed by BlueBRIDGE is lively available through the BlueBRIDGE Catalogue[1].

Table 1. BlueBRIDGE data / datasets main types and formats per Context

Data / Dataset Type	Format	Nature	Context
Research object – any collection of files equipped with rich and detailed metadata.	Any collection of files and URLs with associated metadata.	Generated	Any
Catch statistics – any dataset reporting on catch statistics with various coverages (spatial extent, temporal extent).	Various formats (CSV, MS Excel, NetCDF, SDMX)	Collected	Blue Assessment
Codes and controlled vocabularies – Including Fishing gears classification, marine species codes, countries codes, water areas code and geographic locations.	Various formats	Collected	Blue Assessment
Species occurrence points and taxonomies – any dataset reporting on species occurrence.	DarwinCore, EML, JSON, CSV	Collected / Generated	Blue Assessment
Stock assessment methods – a rich array of scripts and methods integrated in DataMiner (e.g. ICCAT VPA for eastern bluefin tuna, SS3 for WECAFC and IOTC).	Multiple formats (R scripts, Fortran, …)	Collected / Generated	Blue Assessment
Stock assessment outputs – time series	More than one (NetCDF,	Collected /	Blue Assessment

[1] https://bluebridge.d4science.org/catalogue

on multiple parameters (e.g. Total Allowable Catch, abundance by age and year, number of individuals caught by age and year, Fishing mortality, Spawning Stock Biomass)	SDMX, CSV)	Generated	
Stocks and Fishery Records – information on fish stocks and fishery including time-independent fields (e.g. identifiers, descriptions) and time-dependent indicators (e.g. state and trend for stocks).	Records. The collected are in various formats (e.g. XML, JSON, XLS, AccDB). The generated ones are RDF and Catalogue records.	Collected / Generated	Blue Assessment
Tuna Atlas – gridded fisheries datasets (e.g. (nominal) catches datasets, fishing effort datasets, size frequencies datasets)	Many formats (NetCDF, CSV, and other formats managed by PostGIS and GeoServer (GML, shapefiles, GeoJSON))	Collected / Generated	Blue Assessment
Environmental Parameters – time series including Sea Surface Temperature, Currents Speed, Dissolved Oxygen Level, etc.	NetCDF, NcML, Shapefiles (WMS, WFS), GeoJSON	Collected	Blue Assessment, Blue Economy, Blue Environment
Aquaculture farms production statistics – Datasets with sample data collected by aquafarms to record the number and the average weight of fishes in cages.	Tabular data (CSV)	Collected	Blue Economy
Aquaculture production KPIs – FCR (Biological / Economical Feed Conversion Rate), GPD (Growth Rate per Day), SGR (Specific Growth Rate), SFR (Suggested Feeding Rate) MR (Mortality Rate).	Tabular data (RDBMs tables)	Generated	Blue Economy
Investment Analysis Results – IRR (Internal Rate of Return), NPV (Net Present Value), EBIDTA (Earnings Before Interest, Taxes, Depreciation and Amortization), EBIAT (Earnings Before Interest After Taxes), Cumulative Profit / Loss, Yearly net profit margin	Tabular data (for single site analysis), geospatial data (WMS, WFS) (for investment analysis in a region).	Generated	Blue Economy
Socio-economic Indicators – time series on several aspects including Population, Labour costs, etc.	Tabular data	Collected	Blue Economy
Aquafarms and cages locations – geospatial data / maps reporting on assets locations.	Shapefiles, WMS, WFS, Database content (SpatialLite)	Generated	Blue Environment
Marine protected areas maps – maps of vegetation types and human impacts on them enabling ecosystem degradation analysis.	Shapefiles, WMS, WFS	Generated	Blue Environment

Satellite images, RS and geospatial and model outputs (e.g. OSCAR, SST)	Copernicus SAR and optical, BING and GE WMS, Shapefiles and raster data, NetCDF	Collected	Blue Environment
Course supporting material – any material supporting a BlueBRIDGE course / training event.	Many formats	Collected / Generated	Blue Skills
Software	Many formats (e.g. Java, R, Rshiny, Docker Images, OpenCPU, Rnotebooks, ShareLatex)	Generated	Blue Commons, Blue Assessment, Blue Economy, Blue Environment

2.4 EXISTING DATA RE-USE

The data / datasets collected or generated in the context of BlueBRIDGE VREs are often made available to external users and/or they are exploited in the development and operation of others VREs. A number of re-use scenarios are reported below (cf. Table 2). For each class of datasets that can be reused the table concisely presents the context and the activities they are reused for. The table provide necessarily partial information. Given the dynamicity of the VREs (they can be created and be dismissed), the re-use of many data products generated in the context of VREs is not fixed and cannot be anticipated. The following are examples of re-use relationships active at the time of the writing of this deliverable: (i) the climate change related data produced in the BiodiversityLab VRE are re-used in Protected Areas Impact Maps; (ii) the maps produced in the AquaMaps VRE that are re-used in VREs like BiodiversityLab and ScalableDataMining; (iii) datasets produced in the ScalableDataMining VRE that are reused in Tuna Atlas and Blue Fin tuna Assessment. These are only examples and the situation is continuously evolving as new VREs are created and new activities are carried out within existing ones.

Table 2. BlueBRIDGE existing data re-use

Data / Dataset Type	Re-use cases
Catch statistics	These data are mainly collected from several sources for the needs of stock assessment tasks.
Codes and controlled vocabularies	Fishing gears classification codes are used like standard abbreviation and the international (with respect to ISSCFG) code of a fishing gear. Marine species codes are used like international codes (with respect to ASFIS, APHIA) of marine species. Countries code are used like international codes (with respect to ISO2, ISO3, UN) of countries. Water areas codes and geographic locations are used like international codes (with respect to FAO coding system) of water areas.
Species occurrence points and taxonomies	These data are mainly collected from several sources for the needs of stock assessment tasks.
Stock assessment methods	These data are either collected or generated by BlueBRIDGE and represent a valuable asset to be used in any stock assessment task.
Stock assessment outputs	Stock data collected in Sardara and FAO projects are used in stock assessment models in other VREs. Codelists and other artefacts extracted from these datasets are used to harmonize other datasets; this leads to more uniform stock assessment models as they use the

	same datatypes at input level. The harmonized time series, where possible for ownership and copyright reasons, are to be published in a single SDMX registry, facilitating discovery and re-use. The outputs of the stock assessment models are to be few into the scope of the GRSF for validation of GRSF content.
Stocks and Fishery Records	Existing Stock and Fishery records provided by FIRMS, FishSource, and RAM (cf. Sec. 2.5) are exploited to build and develop the GRSF records.
Tuna Atlas	These datasets are expected to be used in research and fishery management cases. Examples of (re-)use include assessment of the size, spatial extent, temporal evolution, and characteristics of tuna fisheries across oceans; comparisons of nominal CPUE time series between fleets; discovery of patterns in tuna ecology, e.g. associative behaviour with floating objects and animals, behaviour in the vicinity of seamounts.
Environmental Parameters	These data are mainly collected from several sources for the needs of various VREs.
Aquaculture farms production statistics	There are several potential re-use of such data yet farms are reluctant in sharing them. These data are mainly collected for enabling the analysis in the specific VRE developed for the needs of the aquaculture farm.
Aquaculture production KPIs	These data are of primary use for the aquaculture farm related assessment applications.
Investment Analysis Results	These data are of primary use for the aquaculture farm assessment applications.
Socio-economic Indicators	These data are mainly collected from several sources for the needs of various VREs (mainly Blue Economy)
Aquafarms and cages locations	The overviews produced by the CLS based remote sensing tools are merged with the FAO inventories, and the resulting product matches field observations (farm names, types, activity) with the layers. This provides a comprehensive inventory for FAO to publish as NASO maps The same map will be overlaid with the Marine Protected Areas Maps (from PAIM VRE) to visualize the spatial relations between aquaculture and MPA's. These datasets might be used by the investment opportunity algorithms, during the identification of optimal sites for aquafarm construction, in order to potentially exclude some areas from the final result.
Marine protected areas maps	PAIM Maps are expected to be used in other projects to illustrate the impact of human and environmental processes on the MPAs. A first example is the BIOPAMA project[2].
Satellite images, RS and geospatial and model outputs (e.g. OSCAR,	These data are mainly collected from several sources for the needs of various VREs.

[2] http://www.biopama.org/

SST)	
Course supporting material	These data are expected to be (re-)used in cases including courses and training events other that the originator ones.

2.5 ORIGIN OF THE DATA

The data / datasets collected by BlueBRIDGE originate from a series of organisations and data providers (cf. Table 3).

Table 3. BlueBRIDGE Collected Data Origins

Data / Dataset Type	Provider(s)
Catch statistics	Stock assessment data is sources from various fisheries management bodies, and stored in databases for FAO and IRD Tuna Atlases; Timeseries for stock assessment data are contributed through Ram, ICES Datras, and by regional projects stock assessment teams.
Codes and controlled vocabularies	Originated from various sources/organizations including Copernicus, EmodNet, ICES, FAO, NASA, OBIS, GBIF, WoRMS and EuroStat.
Species occurrence points and taxonomies	Through the BiodiversityLab, GBIF, OBIS, WoRMS and other data providers can be accessed to obtain species data. In addition, the D4S infrastructure stores tens of thousands of species distribution maps and hosts several databases (OBIS, FISHBASE, SARDARA) that are copies or extended versions of the source databases.
Stock assessment methods	Almost open ended, meaning that stock assessment methods can be provided by institutions or single researchers in an almost continuous way by the dedicated facilities enabling to easily integrate and share a new method.
Stock assessment outputs	These are expected to be among the results produced by BlueBRIDGE Stock Assessment Virtual Research Environments.
Stocks and Fishery Records	The primary providers for such typology of data are: FIRMS[3], FishSource[4], and RAM[5]. FIRMS collects data from 14 intergovernmental organizations and contains information for more than 600 stocks and 300 fisheries. FishSource contains information for more than 2,000 fishery profiles. RAM offers assessments records assembled from 21 national and international management agencies for a total of 331 stocks.
Tuna Atlas	Tuna Regional Fisheries Management Organizations (RFMOs)
Environmental Parameters	Copernicus, GRID-Arendal, NOAA, and WOD are only some of the providers of geospatial data that can be used; in theory any dataset expose through a web service is a valid origin.
Aquaculture farms production statistics	Aquafarms
Aquaculture production KPIs	These are expected to be among the results produced by

[3] http://firms.fao.org/firms/en
[4] http://www.fishsource.com
[5] http://ramlegacy.org

	BlueBRIDGE Blue Economy Virtual Research Environments.
Investment Analysis Results	These are expected to be among the results produced by BlueBRIDGE Blue Economy Virtual Research Environments.
Socio-economic Indicators	Several providers including Eurostat and regional bodies.
Aquafarms and cages locations	These data expected to be among the results produced by BlueBRIDGE Blue Environment Virtual Research Environments. Some
Marine protected areas maps	These data expected to be among the results produced by BlueBRIDGE Blue Environment Virtual Research Environments.
Satellite images, RS and geospatial and model outputs (e.g. OSCAR, SST)	Several providers including World Ocean Atlas, EMODnet, Copernicus Marine Environmental Monitoring System, Planet OS, GEBCO, NASA (Ocean Surface Current Analyses Real-time - OSCAR), NOAA (Ocean Currents Data)
Course supporting material	Almost open ended, meaning that there is no systematic provision of these data other that the specific course.

2.6 THE EXPECTED SIZE OF THE DATA

The data collected / generated by BlueBRIDGE is very heterogeneous in size ranging from few megabytes per item (e.g. aquafarm production statistics) to many gigabytes (e.g. satellite images). Estimating the number of items to be managed is very challenging since it depends from the number of Virtual Research Environments supported.

Table 4. BlueBRIDGE data / datasets size

Data / Dataset Type	Expected size (if known)
Catch statistics	N/A
Codes and controlled vocabularies	N/A
Species occurrence points and taxonomies	N/A
Stock assessment methods	Difficult to estimate yet the code realising a stock assessment method is not a huge artefact to manage.
Stock assessment outputs	Difficult to estimate.
Stocks and Fishery Records	The size of the raw data is 52 MBs for datasets derived from FIRMS (in XML format), 19 MBs for datasets derived from FishSource (in JSON format) and 260 MBs for datasets derived from RAM (in AccDB and XLS formats). The transformed data has been ingested into an RDF triplestore.
Tuna Atlas	The Tuna Atlas database (for Sardara) now measures several GBs, and equivalent size is required for the Geoserver storing Tuna Atlas information.
Environmental Parameters	The main data are managed as temporary files in interoperable information systems from GRID-Arendal and CLS. The storage on the infrastructure requires several GB's From few GBs (e.g. 30GBs for all OSCAR images) to few TBs (other satellite products or model outputs.

Aquaculture farms production statistics	These tabular datasets are collected per aquafarm. FCR (Biological / Economical Feed Conversion Rate) datasets, GPD (Growth Rate per Day) datasets, SGR (Specific Growth Rate) datasets, SFR (Suggested Feeding Rate) datasets, MR (Mortality Rate) datasets are each 30 columns by 20-50 rows up to 1500 rows.
Aquaculture production KPIs	Small object mainly consisting of few numbers and minimal metadata.
Investment Analysis Results	The size of the generated datasets depends on the size of the region in which an analysis is performed as well as the detail level of the analysis. In common use cases the result of an analysis can be a few MBs.
Socio-economic Indicators	Estimated up to 100 GB
Aquafarms and cages locations	A typical layer of a country set of features measures 5 MB for farms and cages, and much bigger files for environmental features. Several hundred layers are expected to be stored.
Marine protected areas maps	Estimated up to 1 TB
Satellite images, RS and geospatial and model outputs (e.g. OSCAR, SST)	Estimated up to 300 TB
Course supporting material	Difficult to estimate, it mainly depends on the specific course.

2.7 DATA UTILITY: TO WHOM WILL IT BE USEFUL

Given the great variety of the datasets collected / generated by BlueBRIDGE reporting on their reuse potential is quite challenging. Moreover, the fact that completely new and unconstrained /unknown datasets can be produced in the context of VREs (e.g. Research Objects including datasets, methods and any other information worth for representing an entire research activity) makes the exploitation scenarios almost open ended. The table below reports some of the known potential exploitations. More on this is in the exploitation and sustainability plan [9].

Table 5. BlueBRIDGE Datasets re-use

Data / Dataset Type	Potential use/re-use
Catch statistics	Stock assessment experts and fisheries scientists who rely on these data to make analyses and suggest management actions.
Codes and controlled vocabularies	Reporting institutes or RFMO that want to transmit their data to a central organization like FAO.
Species occurrence points and taxonomies	Biologists, scientists building ecological models and ecological niche models. Studies on invasive species, biodiversity, and habitat.
Stock assessment methods	Fisheries scientists and fisheries management organizations that want to assess the health status of a stock or verify the results produced by other scientists.
Stock assessment outputs	Fisheries Management Organizations that have to decide strategic plans for fish repopulation and exploitation.
Stocks and Fishery Records	Fisheries resource managers, stock assessment scientists that can find reports in standardized fashion. Industry that can use these data as a skeleton for understanding fisheries provenance, a key component of traceability.

	Governments that find standardized information on the management of stocks and fisheries, essential for international collaboration.
Tuna Atlas	Scientists and fisheries management organizations that want to perform fisheries management and marine ecosystems studies at global scale, in particular to regulate the fishing pressure on tuna stocks.
Environmental Parameters	The potential use of these data is wide. For example, by scientists and organizations that monitor climate change and produce ecological models.
Aquaculture FARMs production statistics	Aquafarmers or companies that want to understand the potential gain of an aquafarm in a certain area, based on previous experience.
Aquaculture production KPIs	These datasets (FCR: Biological / Economical Feed Conversion Rate, GPD: Growth Rate per Day, SGR: Specific Growth Rate, SFR: Suggested Feeding Rate, MR: Mortality Rate) are primary oriented to Aquafarms owners exploiting the BlueBRIDGE VREs to produce what-if scenarios about the expected growth. These are sensitive datasets to be kept private / not disclosed unless the specific aquafarm that own them agrees on their sharing.
Investment Analysis Results	The financial indicators generated by a techno-economic/socio-economic analysis are primarily valuable to existing aquafarm owners as well as to potential investors trying to estimate the profitability of a new investment.
Socio-economic Indicators	Organizations and scientists who want to estimate the socio-economic impact of an aquafarm in a certain area.
Aquafarms and cages locations	Organizations that want to regulate or monitor the ecological impact of aquafarming in a certain area.
Marine protected areas	Scientists or organizations that want to evaluate possible impact of human activity on marine protected areas.
Satellite images, RS and geospatial and model outputs (e.g. OSCAR, SST)	Scientists who want to build environmental-based analyses, e.g. ecological models, disaster prevention, ecological niche models etc.
Course supporting material	Students, scientists, organizations who want to reuse best practices and theoretical explanations from the BlueBRIDGE expertises.

3 FAIR DATA

This section describes the approaches BlueBRIDGE put in place to make datasets findable (cf. Sec. 3.1), accessible (cf. Sec. 3.2), interoperable (cf. Sec. 3.3), and re-usable (cf. Sec. 3.4).

3.1 MAKING DATA FINDABLE

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of the Data Management Plan is expected to give answers to the following questions:

- Is the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
- What naming conventions are followed?
- Will search keywords be provided to optimize possibilities for re-use?
- Do you provide clear version numbers?
- What metadata will be created?
- In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

3.1.1 DISCOVERABILITY OF DATA (METADATA PROVISION)

All the data / datasets described in the data summary part of this deliverable are going to be discoverable and accessible by several catalogues including:

- 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 / datasets by relying on their structural metadata;
- An ***overall catalogue*** – based on CKAN technology – enabling users to perform faceted search on the entire set of resources managed by the BlueBRIDGE VREs.

These three catalogues are not disjoint, in the reality the overall catalogue is configured to contain the items of both the CSW catalogue and the SDMX catalogue. It is equipped with dedicated harvesting mechanisms that automatically collect the items (actually their metadata) from the other two catalogues. This solution thus requires a single publication of the item even if it its metadata are accessible through diverse catalogues.

The development and enhancement of these catalogues is part of the Blue Commons activities carried out in WP9. There is a strong dependency between the data management plan and the technology put in place to support it: any data management plan imposes requirements on the management technology and, at the same time, it is strongly influenced by the available technology. This consideration is particularly relevant for the discoverability of data.

Another factor that strongly characterises and influences discoverability is the metadata associated with items in each of the catalogues envisaged above:

- For CSW Catalogue, every item published in it is endowed with metadata compliant with the ISO19115. The provision of such metadata is done at item publication time, i.e. when the item is added to the specific catalogue.

- The SDMX Catalogue disseminates a specific type of metadata called 'Structural Metadata'. The SDMX standard defines 19 different structure types, that are made available through message formats based on two basic expressions, SDMX-ML (using XML syntax) and SDMX-EDI (using EDIFACT syntax and based on the GESMES/TS statistical message).
- For the overall catalogue, every item published in it is endowed with specific metadata. The basic technology (i.e. CKAN) provides for one typology of item only (i.e. Dataset) and metadata consisting of basic fields (e.g. title, description, tag, licence, author, maintainer) and an open-ended list of <key, value> pairs for adding any additional information. We extended this basic offering thus to explicitly enact the creation of "item typologies" with an associated metadata format carefully defining the metadata fields, the allowed values (including controlled vocabularies) and some directive to automatically transform the values of certain fields in cataloguing actions. The provision of metadata is done: (a) at publishing time for the items natively published in overall catalogue, (b) at harvesting time for the items collected by the other two typologies of catalogues by relying on their metadata. Finally, we could mention that we have set-up and configured the required plugins that allow exporting the contents of the catalogue in a semantically rich format. More specifically, resources and their metadata can be exported in RDF format; these data descriptions can be exploited for discovering resources based on complex query answering (i.e. by submitting SPARQL queries that include multiple metadata fields).

3.1.2 IDENTIFIABILITY OF DATA AND REFER TO STANDARD IDENTIFICATION MECHANISM. DO YOU MAKE USE OF PERSISTENT AND UNIQUE IDENTIFIERS SUCH AS DIGITAL OBJECT IDENTIFIERS?

All the items published by the catalogues have a unique identifier:

- For the CSW catalogue, every item is automatically provided with a UUID enacting a web-based identification mechanism (the URL is composed by the catalogue base URL plus the UUID). By using such an identifier, it is possible to access metadata records about the dataset (including a web page for human users, and protocols and metadata in various formats for machines). There is no overall agreement or constraint on UUID generation, in different contexts diverse strategies are used for UUID production ranging from human-friendly ones to randomly generated ones;
- For the SDMX Catalogue, a Unique Resource Name is generated for each structural metadata. Thus, a unique Identifier is associated not only for the Agencies, consumers, providers, code lists, and data flows but also for the scheme used to model those resources.
- For the overall catalogue, every item is automatically provided with a web-based unique identifier. By using such an identifier, it is possible to access metadata records about the dataset (including a web page for human users, and protocols and metadata in various formats for machines). There is no overall agreement or constraint on identifiers generation. Exploited data sources and content providers use different strategies for identifier production ranging from human-friendly ones to randomly generated ones.

In the context of GRSF [3], identification of records is of paramount importance. As a consequence, a specific approach for their production has been developed and implemented. In particular, two typologies of identifiers are envisaged[6]: *record identifiers* and *semantic identifiers*. Record identifiers are automatically generated and associated with every GRSF record by relying on UUIDs (they are just an alphanumeric value

[6] https://support.d4science.org/projects/stocksandfisherieskb/wiki/GRSF_database_overview#GRSF-record-identifiers

having no informative power). The semantic identifiers are built by relying on record attributes (e.g. species name, various area(s)) to enable "semantic" driven identification of the records. The main purpose of semantic identifiers is to enable the fast and reliable identification of the significant properties of a record (stock or fishery) and is intended to be comprehensive by humans as well.

Overall, there is no requirement for systematic exploitation of DOIs in BlueBRIDGE but for the software case. In the case of software, BlueBRIDGE is systematically exploiting Zenodo to publish every version of the artefacts contributing to a gCube release. A specific community has been created[7] and all the software artefacts produced are automatically equipped with a DOI.

3.1.3 NAMING CONVENTIONS USED

The great variety of datasets typologies and originating contexts makes quite challenging to define project-wise naming conventions, e.g. in the context of a VRE every single user is almost free when defining object names. However, this is not considered to be a limitation nor it implies that in certain contexts naming conventions can be agreed and implemented.

3.1.4 APPROACH TOWARDS SEARCH KEYWORD

All the catalogues offer the keyword search, e.g. this is among the main facilities supported by such a typology of service. Figure 1 depicts the home page of the BlueBRIDGE Overall Catalogue where the keyword search is a prominent part of the GUI.

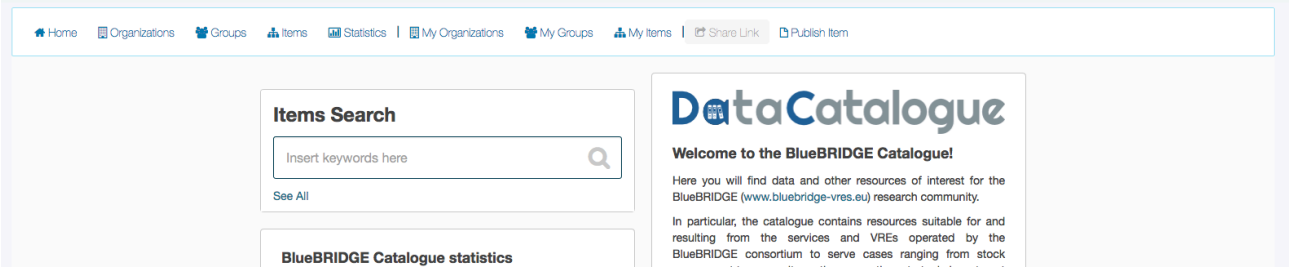

Figure 1. BlueBRIDGE Overall Catalogue main page

Keywords specified by the users are matched against the entire metadata record associated with a catalogue item.

In addition to that, catalogues support item annotation with free keywords (the overall catalogue provides for tags) for enhancing classification and discovery purposes.

There is no agreed "standard" for keywords selection and use yet there are common patters and strategies used in the various contexts including:

- add the name of the species the item is about both in the title and as a keyword;
- add the name of the geographic area representing the coverage the item.

3.1.5 APPROACH FOR CLEAR VERSIONING

A version number is associated with data to represent its various manifestations. However, at the moment none of the VRE is actually conceived to support a systematic production of data in multiple versions.

[7] https://zenodo.org/communities/gcube-system

In some cases, versioning is automatically managed by the service responsible for the storage of a given item, e.g. this is the case of workspace items having the same name. Whenever a user uploads a workspace item in a given folder and the item has the same name (and other properties) of an existing one a new version is created.

In the majority of cases "versioning" is managed by the producer of the data that really knows to what extent a certain item is semantically a new version of an existing one or simply a new data on its own. For example, only the producer of a research object can take decisions on whether the item resulting from its activity is worth being a new version of an existing one or a new artefact.

Versioning is complemented by provenance / lineage, i.e. whenever possible the system automatically produces a provenance record and associates it with the data to capture the process leading to the data.

3.1.6 STANDARDS FOR METADATA CREATION (IF ANY). IF THERE ARE NO STANDARDS IN YOUR DISCIPLINE DESCRIBE WHAT METADATA WILL BE CREATED AND HOW

Given the variety of the BlueBRIDGE data, the project is facing with a hybrid scenario where standards may or may not exist. The approach has been very pragmatic, i.e. to rely on standard whenever they exist.

All the data managed by the system as a whole have minimal metadata, semi-automatically generated, and represented as Dublin Core record. This format has been selected because of its simplicity and genericity.

All the data having a relevant geospatial characterisation are equipped with geospatial metadata in standard formats, namely ISO 19115.

To support the construction of the Global Record of Stocks and Fisheries, the harvested data and their corresponding metadata are being stored in semantically-rich warehouse, the GRSF Knowledge Base. There data are described with respect to the CIDOC CRM ontology which is an ISO standard (ISO 21127:2006) capable of describing the implicit and explicit concepts and relationships used in cultural heritage documentation, and its domain-specific extension MarineTLO [16]. In addition, the RDA Working Group, "Fisheries Data Interoperability Working Group" has been set up to devise a global data exchange and integration framework to support scientific advice on stock status and exploitation that build on fisheries data[8].

3.2 MAKING DATA OPENLY ACCESSIBLE

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of the Data Management Plan is expected to give answers to the following questions:

- Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
- Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
- How will the data be made accessible (e.g. by deposition in a repository)?
- What methods or software tools are needed to access the data?

[8] https://www.rd-alliance.org/group/fisheries-data-interoperability-wg/case-statement/fisheries-data-interoperability-working

- Is documentation about the software needed to access the data included?
- Is it possible to include the relevant software (e.g. in open source code)?
- Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
- Have you explored appropriate arrangements with the identified repository?
- If there are restrictions on use, how will access be provided?
- Is there a need for a data access committee?
- Are there well described conditions for access (i.e. a machine readable license)?
- How will the identity of the person accessing the data be ascertained?

3.2.1 WHICH DATA WILL BE MADE OPENLY AVAILABLE? IF SOME DATA IS KEPT CLOSED PROVIDE RATIONALE FOR DOING SO

BlueBRIDGE is not mainly conceived to produce data in a systematic way, rather its main goal is to enact the development of a series of virtual research environments. As a consequence, the policy governing the decision to make openly available the data produced in a given VRE is up to its administrators / managers. Among the possible alternatives s/he can also decide that it is author-specific, i.e. the decision is up to the specific author / data producer.

The overall approach is to promote the publication and the release of all the data that is generated, yet the action is in the hand of their producers since not the entire pool of data is deemed worth publishing and maintaining (e.g. test data, data under revision / approval) and there may be also sensitive data.

For the data collected, the approach is to be compliant with their licence. Whenever the data are originally openly available they will continue to be openly available by the BlueBRIDGE services. Whenever the data are characterised by access restrictions, the availability of the data will be announced by the BlueBRIDGE managing catalogue with the associated restrictions. Particularly relevant is the case of the aquaculture farms involved in the Blue Economy VREs. The managed data are collected for the sole purpose to generate the KPIs of the specific aquafarm, it is not expected to be shared and made publicly available for obvious reasons.

For the data generated by BlueBRIDGE:

- ***Research objects***: the decision to make the produced objects public and openly available is up to the producer / author;
- ***Species occurrence points and taxonomies***: the decision to make the produced objects public and openly available is up to the producer / author;
- ***Stock assessment methods***: the default option is to make all the methods publicly available. In some cases, availability is for VRE members only;
- ***Stock assessment outputs:*** the decision to make the produced objects public and openly available is up to the producer / author;
- ***Tuna atlas***: this typology data will be openly available;
- ***Aquaculture production KPIs***: this typology of data is of primary interest for the associated aquafarm, they remain private to the aquafarm;

- ***Investment analysis results***: this typology of data is of primary interest for the associated aquafarm, they remain private to the aquafarm;
- ***Aquafarms and cages locations:*** this typology of data results from an analytics process that is still under assessment. Moreover, it is performed on selected regions. The policy governing the availability of such data is still under discussion and it is likely that differs from region to region;
- ***Marine protected areas:*** this typology of data results from an analytics process that is still under assessment. Data are produced on-demand (per EEZ or Ecoregion) yet it is possible to automatically produce reports in PDF with links to the accompanying data;
- ***Course supporting material***: the default option is to make data produced for the courses publicly available. However, there might be some material that is restricted / made available for course participants only;
- ***Software***: the software is mainly part of the gCube platform and it is publicly available. For scripts and methods provided by various users and integrated into the data analytics platform, the decision to make them publicly available or not is up to the owners.

3.2.2 HOW THE DATA WILL BE MADE AVAILABLE

BlueBRIDGE is committed to maximize the exploitation and use of the data that are collected and / or generated in the supported Virtual Research Environments. Given the diversity of the data, a rich array of approaches and standards is needed and the specific data availability is announced by the associated catalogue entries. Whenever possible standards are exploited and web-based programmatic access is supported. For instance, all the geospatial data are made available by OGC standards like WMS, WFS, and WCS.

Every research object residing in the Workspace is accessible via a URL. Two diverse URIs can be associated with any of these objects: (a) a restricted one allowing only authorized users to actually access linked the object and (b) a public one enabling any users to access the linked object. This enables a URI-based dissemination mechanism that can be used in several contexts, including social networking ones where the norm is to disseminate content by posting a URL.

3.2.3 WHAT METHODS OR SOFTWARE TOOLS ARE NEEDED TO ACCESS THE DATA? IS DOCUMENTATION ABOUT THE SOFTWARE NEEDED TO ACCESS THE DATA INCLUDED? IS IT POSSIBLE TO INCLUDE THE RELEVANT SOFTWARE (E.G. IN OPEN SOURCE CODE)?

No specific software tool is actually needed to access BlueBRIDGE data other than one suitable for the specific data format.

3.2.4 WHERE THE DATA AND ASSOCIATED METADATA, DOCUMENTATION AND CODE ARE DEPOSITED

BlueBRIDGE data are in several repositories depending on their typologies, e.g. research objects are stored on the workspace, geospatial objects are stored on geospatial repositories (THREDDS DS or GeoServer), software is stored on gCube SVN[9] and Zenodo[10].

[9] http://svn.research-infrastructures.eu/public/d4science/gcube/
[10] https://zenodo.org/communities/gcube-system

The primary entry point for each data is expected to be the associated catalogue entry that contains all the links deemed necessary to actually make the data accessible, i.e. catalogue entries contain links to the data, links to accompanying documentation and services related with the data plus metadata as rich and detailed as possible.

3.2.5 HOW ACCESS WILL BE PROVIDED IN CASE THERE ARE ANY RESTRICTIONS

Every data has an author and a maintainer (at least). Authors and maintainers are either individuals or groups that can be contacted (their name is actionable) in order to start a dialogue for accessing the data in case they are restricted.

In order to make this process as simple and organised as possible, it is envisaged to enrich catalogue entries with facilities for requesting access to identified datasets. Requests are expected to be recorded and managed by the ticketing system.

3.3 MAKING DATA INTEROPERABLE

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of a Data Management Plan is expected to give answers to the following questions:

- Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
- What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
- Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
- In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The issues related with legal interoperability are discussed in the project exploitation and sustainability plan [9].

3.3.1 ASSESS THE INTEROPERABILITY OF YOUR DATA. SPECIFY WHAT DATA AND METADATA VOCABULARIES, STANDARDS OR METHODOLOGIES YOU WILL FOLLOW TO FACILITATE INTEROPERABILITY

Because of the almost open-ended set of data that potentially results from the BlueBRIDGE VREs it is nearly impossible to produce an exhaustive list of vocabularies and standards exploited. However, the project is promoting the exploitation of state-of-the-art solutions and approaches whenever available:

- All the datasets having a geospatial extent are made available by OGC W*S protocols with metadata in ISO 19115;
- All the datasets referring to species report species names and codes clearly indicating the classification system exploited, e.g. ASFIS;
- All the catalogue items are made available by RDF DCAT format and through the OAI-PMH protocol;
- Provenance information for artefacts resulting from the data analytics platform is in PROV-O records;

- All the contents of the Global Record of Stocks and Fisheries Knowledge Base are modelled as instances of the ISO 21127:2006 standard and its extensions.

3.3.2 SPECIFY WHETHER YOU WILL BE USING STANDARD VOCABULARY FOR ALL DATA TYPES PRESENT IN YOUR DATA SET, TO ALLOW INTER-DISCIPLINARY INTEROPERABILITY? IF NOT, WILL YOU PROVIDE MAPPING TO MORE COMMONLY USED ONTOLOGIES?

Whenever known standard vocabularies are expected to be exploited when compiling datasets and their accompanying metadata, these vocabularies will be properly exploited and referred.

ISO 19115 topic categories are exploited to categorize geospatial datasets.

No mapping is yet planned to be provided in order to transform data from a proprietary format to commonly used ontologies.

3.4 INCREASE DATA RE-USE

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of a Data Management Plan is expected to give answers to the following questions:

- How will the data be licensed to permit the widest re-use possible?
- When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
- Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
- How long is it intended that the data remains re-usable?
- Are data quality assurance processes described?

3.4.1 HOW THE DATA WILL BE LICENCED TO PERMIT THE WIDEST REUSE POSSIBLE

The default licence promoted within the project is the CC BY-SA, however every data producer is free to use the licence he/she considers the most suitable one.

Every catalogue item clearly indicates the licence associated with the catalogue entry. The project Quality Assurance Office constantly monitors the distribution and use of licences and whenever notice restrictive licences starts a dialogue with the data producer(s).

3.4.2 WHEN THE DATA WILL BE MADE AVAILABLE FOR RE-USE. IF APPLICABLE, SPECIFY WHY AND FOR WHAT PERIOD A DATA EMBARGO IS NEEDED

In BlueBRIDGE VREs, there is neither a project-wise nor a data-wise embargo period envisaged. The default project option is to make the data available “as soon as possible”.

There are cases where data availability depends from an assessment / approval phase, e.g. this is the case of Stocks and Fisheries Records that deserve an explicit approval from designated experts before being disseminated.

3.4.3 WHETHER THE DATA PRODUCED AND/OR USED IN THE PROJECT IS USEABLE BY THIRD PARTIES, IN PARTICULAR AFTER THE END OF THE PROJECT? IF THE RE-USE OF SOME DATA IS RESTRICTED, EXPLAIN WHY

The exploitation of the data depends on the licence accompanying them. No major restrictions are envisaged / currently known for data that are “public”.

Some data are restricted. For instance, the aquaculture farm production statistics represent sensitive data the farms are not willing to share.

3.4.4 DATA QUALITY ASSURANCE PROCESSES

BlueBRIDGE is not planned to produce data in a systematic way, rather data are expected to result from the VRE it is supporting and developing.

VREs represent the ideal environment for putting in place collaborative approaches for data quality assessment. Whenever a data is published into the overall catalogue a post announcing its availability is automatically produced in the VRE. VRE members can then comment on the post and use their comments to report on any issue with the data.

The process leading to the data is, whenever possible, (a) captured by a PROV-O record, and (b) made repeatable. For instance, this is the case of all the data produced by the data analytics platform.

3.4.5 THE LENGTH OF TIME FOR WHICH THE DATA WILL REMAIN RE-USABLE

Data are expected to remain available in the format they have been published for 4 years after the end of the project. Format migration actions are planned to occur in accordance with the evolution of the standards and technologies adopted by the project. Such evolution can be estimated in 5 years from the latest version adopted by the project.

4 ALLOCATION OF RESOURCES

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of a Data Management Plan is expected to give answers to the following questions:

- What are the costs for making data FAIR in your project?
- How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
- Who will be responsible for data management in your project?
- Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

4.1 COSTS FOR MAKING YOUR DATA FAIR. DESCRIBE HOW YOU INTEND TO COVER THESE COSTS

There are various typologies of costs to be considered when discussing on making the BlueBRIDGE data FAIR including:

- Data production costs: these mainly include the costs for collecting and collating the data. In the case of BlueBRIDGE data these costs are primarily related with human activity (with the support of specific services) and they mainly coincide with the exploitation of the Virtual Research Environments.
- Data curation and publishing costs: these include the costs for selection, organisation and presentation of the data. In the case of BlueBRIDGE data these costs are related with human activity and they mainly coincide with VREs exploitation.
- Supporting services operation costs: these include the costs for operating the services supporting the storage, dissemination and curation of the data, e.g. the costs for operating the catalogues and the costs for operating the repositories.

All in all, the overhead resulting from the decision to make a given data FAIR is very limited with respect to the standard cost of activities data providers are called to sustain when interfacing with and exploiting the VREs for their tasks. In fact, such actors are usually only requested to compile some basic metadata (e.g. choose the licence) when they decide to make the data they produced known to either the rest of VRE members or to the general public. The rest of the process (e.g. deposition of data in repositories thought for long term availability, the population of catalogues) is outsourced to specific services and processes.

The strategies for covering these costs are actually intimately related with the strategies and approaches the project will put in place for the exploitation and sustainability of the entire project results [9].

4.2 RESPONSIBILITIES FOR DATA MANAGEMENT IN YOUR PROJECT

The responsibilities for data management in BlueBRIDGE are shared among the data producer(s), the data maintainer(s), and the service manager(s).

Data producer(s) and maintainer(s) are called to respond to any issue possibly affecting the data ranging from potential flaws in the process leading to them up to technical issues in accessing and consuming the data. They are the front-end data consumers will interface with. In particular, the maintainer(s) are responsible for taking care that data, once published, continue to be “healthy”. For the ***Stocks and Fishery Records of the Global Record of Stocks and Fisheries*** the following is planned [9]. The Global Record of Stocks and Fisheries enables the sharing and harmonization of information on global fisheries. The Public -

not-for-profit Private partnership business model was selected as a FIRMS expanded partnership, as FIRMS and GRSF share one goal, to gather and disseminate data on stocks and fisheries, and use the same data sources primarily provided by countries. It was preferred over a Public only partnership because the involvement of organizations already connected to the seafood industry will provide opportunities to develop commercial data products and services to ensure the sustainability of the GRSF VRE. The choice of the governance model was guided by FIRMS, as a partnership controlled by FAO rules. FIRMS will be the owner of the GRSF VRE and its datasets. The data standardization and content activities will be completely integrated to the FIRMS organization, SFP and UWA becoming FIRMS Members (with no voting rights), while the technological activities will operated under a Service Level Agreement by CNR and FORTH that will participate to technical committees as experts: this organization allows to produce revenues from the development of data products and services that are to ensure the GRSF sustainability without infringing FIRMS and FAO principles. For ***Aquafarms and cages locations*** and ***Marine protected areas maps*** (cf. Sec. 2.3), the following is planned [9]. The business model is characterised by the following steps: (i) To establish a Joint Venture (JV) similar team, with shared commitment to promote exploitation of services developed under BlueBRIDGE, where the FAO-CNR MoU can provide a starting point; (ii) Establish a governance mechanism for the JV to establish the legal framework, the sharing of information, the data policy and where MoU's and SLA's are established. These include: an Engineering team, outside FAO; an Operations team, outside FAO premises; a Product team, under responsibility of FAO; a Secretariat in FAO-FIAS for the delivery and management of geospatial services. (iii) to establish a business process within FAO where units, in collaboration with CIO, may flexibly engage on-demand with a consultancy service (e.g. in the form of PSA or SME) that has the responsibility to mediate between prospect end users of these geospatial services and the Engineering and Operations team. The same approach can be replicated by other interested entities. Specific actions that can already start center around the two VREs: PAIM and AAPS.

Service manager(s) are called to guarantee that the service they are operating works according to the service level agreement established with service consumer(s). Data producer(s) and maintainer(s) relies on the facilities offered by the infrastructure and its services, they took into account the SLA when decided to exploit a given facility for their data. In order to support the sustainability of the BlueBRIDGE services the partners have discussed the maintenance and support of the infrastructure [9]. Below are the main plans:

- CNR commits to maintain operational the infrastructure until at least 2023;
- Establishment of an iMarine Advisory Board under a FAO Partnership agreement; this Advisory Board will promote a public data catalogue; practices facilitating harmonization and sharing; sharing and reuse of community software and expertise; promotion of the BlueBRIDGE services through outreach and broker role;
- Setting of a series of bilateral MoUs between core partners including:
    - Public-private MoU for the maintenance of the underlying management software framework, i.e. gCube, and the build and integration system between CNR and ENG;
    - MoUs between CNR and one or two SMEs for responses to limited development needs;
    - Strategic MoU between FAO (representing the Community of Practice) and CNR (assuring the maintenance of the D4Science data infrastructure), including the set-up of a small Steering Board involving the few strategic partners involved in the above MoUs.

4.3 COSTS AND POTENTIAL VALUE OF LONG TERM PRESERVATION

BlueBRIDGE is not primarily called to produce datasets on its own, datasets are expected to result from the activities performed in the context of the various VREs by their communities. The datasets produced in a

VRE can be reused by other VREs yet it is almost impossible to figure out to what extent this is going to happen given the variety and dynamicity of the datasets falling under the BlueBRIDGE VREs. Overall, the project consortium is committing to maintain the datasets in the format they have been published for 4 years after the end of the project. Format migration actions are planned to occur in accordance with the evolution of the standards and technologies adopted by the project. Such evolution can be estimated in 5 years from the latest version adopted by the project. Specific plans exceeding these setting can be established for datasets whose exploitation plan is known / envisaged [9].

In the case of the Global Record of Stocks and Fisheries, it should be taken into account that it does not generate new data, rather it collates information coming from the underlying database sources. When these sources generate new data they can be ingested in GRSF by relying on the resources that already exist (i.e. mappings, software, processes, etc.). However, if the underlying sources change the representation of their data (i.e. their structure or format) or they enrich them with more information that was not previously existed (i.e. addition of fishing gears), then it is required to update certain parts of the entire process for adding those in GRSF, like for example the mappings, the merging and dissection rules, etc. In addition, it is required to dedicate some effort for alleviating with the new requirements that emerge. The aforementioned activities are crucial for preserving the high quality of the record.

5 DATA SECURITY

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of a Data Management Plan is expected to give answers to the following questions:

- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
- Is the data safely stored in certified repositories for long term preservation and curation?

No certified repository is expected to be used for satisfying the needs emerging in BlueBRIDGE Virtual Research Environments. However, data are safely stored by relying on repositories operated by the underlying D4Science infrastructure thus to guarantee the safety of the data and the accompanying metadata. 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 have been planned and implemented:

- 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 D4Science storage systems. This solution provides the ability to instantly fail-over to local on-site hardware all of the D4Science storage systems, but in the event of a physical disaster at the main D4Science data center, storage systems can be brought up in two additional data centers that are federated to D4Science;
- 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, D4Science also implements precautionary measures 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.

6 ETHICAL ASPECTS

According to the Guidelines on FAIR Data Management in Horizon 2020 [10] this section of a Data Management Plan is expected to give answers to the following questions:

- Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).
- Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Re ethical and legal issues, some data produced by BlueBRIDGE Virtual Research Environments might be affected by legal issues (e.g. depending on the typology of data and because of regional policies or regulations) while no ethical issues are known. The issues related with legal issues have been analysed and discussed in the exploitation and sustainability plan [9]. In particular, the discussion focused on the challenges arising when combining multiple datasets from multiple sources to build a new dataset (e.g. attribution vs citation). Per dataset discussion are planned to take place to carefully identify the licence accompanying every dataset produced by BlueBRIDGE.

Re personal data collection, the project is not explicitly dealing with any activity related with personal data collection. However, it is possible that in the context of specific VREs such an activity took place, e.g. VRE managers / users can set up specific surveys and questionnaire possibly collecting personal data. It is a responsibility of the data collector(s) to put in place appropriate strategies for the management of such sensitive data.

7 CONCLUSION

This deliverable is the final version of the BlueBRIDGE Data Management Plan. This deliverable is following the guidelines issued by the European Commission and has been organised around the solutions and approaches aiming at making BlueBRIDGE data / datasets findable, accessible, interoperable and re-usable (FAIR). It consolidates and extends the intermediate version (D2.2). With respect to the intermediate version, details on data management responsibilities and data security strategies have been added. Overall, this deliverable presented a comprehensive summary of the data collected or generated by BlueBRIDGE highlighting the great variety of typologies, formats and contexts characterising the project and its Virtual Research Environments. Then, it has described the solutions and approaches the project has set up to promote BlueBRIDGE data FAIRness, namely (a) a systematic exploitation of catalogues (both generic and specific) to expose available data, (b) the concurrent use of several repositories each suitable for certain data, (c) the promotion and use of standards and controlled vocabularies whenever they exist, and (d) the connection between the data and the environments supporting their development (the VREs) and use.

REFERENCES

[1] Assante, M., Candela, L., Castelli, D. and Tani, A. (2016) Are Scientific Data Repositories Coping with Research Data Publishing? Data Science Journal, 15:6, pp. 1–24, DOI: 10.5334/dsj-2016-006

[2] Assante, M., Candela, L., Cirillo, R., Coro, G., Koltsida, P., Marioli, V., Perciante, C., Sinibaldi, F., and Pagano, P. (2016) BlueBRIDGE VRE Commons Facilities. BlueDRIDGE Project Deliverable. D9.1

[3] Barde, J., Ellenbroek, A., Formisano, C., Large, S., and Marketakis, Y. (2017) Blue Assessment VRE Specification: Revised Version. BlueBRIDGE Project Deliverable. D5.3

[4] Candela, L., Castelli, D., Michel Assoumou, J., Pagano, P. and Zoppi, F. (2015) Quality Plan. BlueDRIDGE Project Deliverable. D1.1

[5] Candela, L., Castelli, D., Pagano, P. and Zoppi, F. (2017) BlueBRIDGE Data Management Plan: Intermediate version. BlueDRIDGE Project Deliverable. D2.2

[6] Candela, L., Castelli, D., Pagano, P. (2016) BlueBRIDGE Data Management Plan: Preliminary version. BlueDRIDGE Project Deliverable. D2.1

[7] Candela, L., Coro, G., Frosini, L., Galante, N.A., Giammatteo, G., Kakaletris, G., Lelii, L., Marioli, V., and Sinibaldi, F. (2016) BlueBRIDGE Resources Federation Facilities. BlueDRIDGE Project Deliverable. D10.1

[8] Dimitrakopoulos, C., Antzoulatos, G., and Kakaletris, G. (2016) Blue Economy VRE Specification: Revised Version. BlueBRIDGE Project Deliverable. D6.3

[9] Ellenbroek, A.; Fabriani, P.; Matranga, I.; Nardi, N.; Pagano, P. (2018) BlueBRIDGE Exploitation and Sustainability Plan. BlueBRIDGE D2.5

[10] European Commission (2016) Guidelines on FAIR Data Management in Horizon 2020. Version 3.0 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[11] Formisano, C., Ellenbroek, A., Longépé, N., Macmillian-Lawler, M., Westerweld, L., and Lebras, J.-Y. (2017) Blue Environment VRE Specification: Revised version. BlueBRIDGE Project Deliverable. D7.3

[12] Genova, F., Hanahoe, H., Laaksonen, L., Morais-Pires, C., Wittenburg, P. and Wood, J. (2014) The Data Harvest Report: How sharing research data can yield knowledge, jobs and growth. RDA Europe Report

[13] Piwowar, H.A., Vision, T.J. and Whitlock, M.C. (2011) Data archiving is a good investment. Nature, Vol. 473, 285/285; 10.1038/473285a

[14] Piwowar, H.A., Day, R.S. and Fridsma, D.B. (2007) Sharing detailed research data is associated with increased citation rate. PLoS ONE 2(3): e308; 10.1371/journal.pone.0000308

[15] Renear, A.H., Sacchi, S. and Wickett, K.M. (2010) Definitions of dataset in the scientific and technical literature. Proceedings of the American Society for Information Science and Technology, 47(1): 1–4. DOI: 10.1002/meet.14504701240

[16] Tzitzikas, Y., Allocca, C., Bekiari, C., Marketakis, Y., Fafalios, P., Doerr, M., Minadakis, N., Patkos, T. and Candela, L. (2016). Unifying Heterogeneous and Distributed Information about Marine Species through the Top Level Ontology MarineTLO. Emerald Group Publishing Limited, Vol. 50 Issue: 1, pp.16 - 40, DOI: http://dx.doi.org/10.1108/PROG-10-2014-0072