Project Acronym	***iMarine***
Project Title	***Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources***
Project Number	***283644***
Deliverable Title	***Virtual Research Environments Activity Report***
Deliverable No.	***D6.5***
Delivery Date	***October 2014***
Author	***Leonardo Candela, Anton Ellenbroek, Pasquale Pagano***

Abstract*: This report documents the activity needed to provide the iMarine Ecosystem Approach Community of Practice with a set of Virtual Research Environments aiming at serving the scenarios and requirements discussed by such community. In particular, the report describes (a) the set of Virtual Research Environments that have been deployed and (b) the development of specific applications and tools that are needed to realize the expected Virtual Research Environments in tandem with the rest of gCube technology.*

iMarine (RI – 283644) is a Research Infrastructures Combination of Collaborative Project and Coordination and Support Action (CP-CSA) co-funded by the European Commission under the Capacities Programme, Framework Programme Seven (FP7).

DOCUMENT INFORMATION

PROJECT	
Project Acronym	iMarine
Project Title	Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources
Project Start	1st November 2011
Project Duration	30 months
Funding	FP7-INFRASTRUCTURES-2011-2
Grant Agreement No.	283644
DOCUMENT	
Deliverable No.	D6.5
Deliverable Title	Virtual Research Environments Activity Report
Contractual Delivery Date	August 2014
Actual Delivery Date	October 2014
Author(s)	Leonardo Candela (CNR), Anton Ellenbroek (FAO), Pasquale Pagano (CNR)
Editor(s)	Leonardo Candela
Reviewer(s)	Leonardo Candela
Contributor(s)	Panagiota Koltsida (NKUA), Yannis Marketakis (FORTH)
Work Package No.	WP6
Work Package Title	Virtual Research Environments Deployment and Operation
Work Package Leader	CNR
Work Package Participants	ERCIM, NKUA, FAO, FIN, UNESCO, CRIA, IRD
Estimated Person Months	1.5
Distribution	Public
Nature	Report
Version / Revision	1.0
Draft / Final	Final
Total No. Pages (including cover)	43
Keywords	Virtual Research Environment, Community Tools, Applications

DISCLAIMER

iMarine (RI – 283644) is a Research Infrastructures Combination of Collaborative Project and Coordination and Support Action (CP-CSA) co-funded by the European Commission under the Capacities Programme, Framework Programme Seven (FP7).

The goal of iMarine*, Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources*, is to establish and operate a data infrastructure supporting the principles of the Ecosystem Approach to Fisheries Management and Conservation of Marine Living Resources and to facilitate the emergence of a unified Ecosystem Approach Community of Practice (EA-CoP).

This document contains information on iMarine core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as iMarine 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 iMarine 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/).

DELIVERABLE SUMMARY

Virtual Research Environments are among the key products that the iMarine project is requested to deliver to serve the needs of the Ecosystem Approach Community of Practice. The deployment and operation of a Virtual Research Environment is a task that involves the exploitation of technologies that have been developed as well as the development of new technologies aiming at offering new facilities. This report describes the activities performed to provide the Ecosystem Approach Community of Practice with the set of Virtual Research Environments hosted by the iMarine portal up to September 2014.

EXECUTIVE SUMMARY

Virtual Research Environments (VRE's) are "systems" that provide their users with a web-based set of facilities (including services, data and computational facilities) to accomplish a set of tasks by dynamically relying on the underlying infrastructure. VRE's are among the products to be delivered by the iMarine project to support the business cases of the iMarine Ecosystem Approach Community of Practice (EA-CoP).

The development of VRE's is based on three main activities: *(i)* the development of software artefacts that realise a set of functions, *(ii)* the deployment of these artefacts in an operational infrastructure, and *(iii)* the deployment and operation of these Virtual Research Environments that exploit the facilities offered by the deployed infrastructure to achieve objectives specified in the Business Cases of the Community of Practice.

This report documents the first and the last of the above three activities – i.e. dedicated software development and VREs deployment – as implemented in the context of the iMarine project as per September 2014. The second aspect – i.e. infrastructure deployment – is captured by another document, deliverable D5.4 "iMarine Data e-Infrastructure Operation Report". Thus, this report describes the set of software artefacts that, in addition to the core technology, has been developed to serve the business cases identified by the Ecosystem Approach Community of Practice (EA-CoP).

As of September 2014, ***20 VREs*** were active, with several new VREs added to support specific collaborative activities since the previous report. In particular, ***BiodiversityLab*** was deployed to support the needs of scholars willing to perform experiments (e.g. species distribution maps production, species data inspection) on single individuals or groups of marine species[1], ***BiOnym*** was deployed to provide scientists dealing with species names with a service (ByOnym) to compare a set of scientific names against taxonomic reference lists including recognised ones, ***iSearch*** was deployed to provide its users with information retrieval and semantic web facilities for seamlessly discovering information objects from heterogeneous data sources, ***MarineSearch*** was deployed to provide its users with information retrieval and semantic web facilities for seamlessly discovering information objects from heterogeneous yet marine-oriented data sources, ***TBTI*** was deployed to provide the members of the Too Big to Ignore initiative with an environment for experiencing with the gCube services, ***TabularDataLab*** was deployed to provide its users with a working environment supporting the management of tabular data, i.e. to import, curate, analyse, visualize and publish tabular data resources in a collaborative way. All VREs have been substantially enriched with new generic services, and the number of users continues to grow. In average, it has been observed an increase

[1] BiodiversityLab was exploited to support two courses overall serving 66 students. The first course takes place on 5-6 December 2013 at the UPMC - Sorbonne Université in Paris. The second course takes place the following week, 12-13 December 2013 at the Observatoire Océanologique de Villefranche sur mer, Villefranche sur mer. Both courses are part of the **Master in Océanographie Environnements Marins (OEM)**.

of ***+41%*** (197 at Jun 2013 vs 278 at Sep 2014) in the number of ***users served by FARM VREs*** and of ***+105%*** (240 at Jun 2013 vs 492 at Sep 2014) in the number of ***users served by gCubeApps VREs***.

During the reporting period, examples of software artefacts that were newly released, or substantially altered following recommendations resulting from EA-CoP validation include ***BiOnym***, i.e. a taxonomic naming matching application, ***COTRIX***, i.e. a collaborative code lists management system, ***Tabular Data Manager***, i.e. a comprehensive environment for tabular data management, and ***TrendyLyzer***, i.e. an environment for discovering and visualising trends in species databases.

The deliverable captures the activities between July 2013 and September 2014. In order to do that, the deliverable builds upon D6.4 and updates it thus to describe the state of the art at September 2014. The activities are dynamic and reach across the boundaries of WP6. In fact, the development and operation of Virtual Research Environments is performed in the context of WP6 in close cooperation with: *(a)* WP3 for what concerns the interaction with the Ecosystem Approach Community of Practice, *(b)* WP5 for what concerns the deployment and operation of the underlying infrastructure, and *(c)* WP8, WP9, WP10 and WP11 for what concerns the development of the core technology needed to enable the deployment of the infrastructure and the Virtual Research Environments. In addition to the WP6 activities, the deliverable documents such a network of cooperation and provides the reader with a list of related documents, Wiki pages and TRAC tickets allowing to build a comprehensive understanding of the overall process and related activities.

TABLE OF CONTENTS

1 INTRODUCTION 10

2 VIRTUAL RESEARCH ENVIRONMENTS RESOURCES AND TOOLS 12

2.1 DATA RESOURCES 12

2.1.1 Species Data 12

2.1.2 GeoSpatial Data 13

2.1.3 Statistical Data 14

2.1.4 Other Data 14

2.2 TOOLS 15

2.2.1 BIOnym Portlets 15

2.2.2 COTRIX Portlet 17

2.2.3 Tabular Data Manager portlets 17

2.2.4 TrendyLyzer Portlet 19

2.2.5 Mobile Applications 19

3 VIRTUAL RESEARCH ENVIRONMENTS DEPLOYMENT AND OPERATION 23

3.1 DESIGN, DEPLOYMENT AND OPERATION 23

3.1.1 The AquaMaps VRE 24

3.1.2 The BiodiversityLab VRE 25

3.1.3 The BiodiversityResearchEnvironment VRE 26

3.1.4 The BiOnym VRE 27

3.1.5 The DocumentWorkflow VRE 28

3.1.6 The EcologicalModelling VRE 28

3.1.7 The FCPPS – Fisheries Country Profiles Production System VRE 29

3.1.8 The FishFinderVRE 30

3.1.9 The ICIS – Integrated Capture Information System VRE 31

3.1.10 The iMarine Board VRE 32

3.1.11 The iSearch VRE 34

3.1.12 The MarineSearch VRE 34

3.1.13 The Scalable Data Mining VRE 35

3.1.14 The TabularDataLab VRE 36

3.1.15 The TBTI VRE 37

3.1.16 The TimeSeries VRE 39

3.1.17 The VesselActivitiesAnalyser VRE 40

3.1.18 The VME-DB VRE 41

3.1.19 The VTI – Vessel Transmitted Information VRE 41

4 CONCLUDING REMARKS 43

TABLE OF FIGURES

Figure 1. BiOnym Portlet Screenshot 16
Figure 2. COTRIX Portlet Screenshot 17
Figure 3. Tabular Data Manager Portlet Screenshot 18
Figure 4. TrendyLyzer Portlet Screenshot 19
Figure 5. iMarine Search Mobile App Screenshots 21
Figure 6. The AquaMaps VRE Homepage 24
Figure 7. The BiodiversityLab VRE Homepage 25
Figure 8. The BiodiversityResearchEnvironment VRE Homepage 26
Figure 9. The BiOnym VRE Homepage 27
Figure 10. The DocumentsWorkflow VRE Homepage 28
Figure 11. The EcologicalModelling VRE Homepage 29
Figure 12. The FCPPS VRE Homepage 30
Figure 13. The FishFinderVRE Homepage 31
Figure 14. The ICIS VRE Homepage 31
Figure 15. The iMarine Board VRE Homepage 33
Figure 16. The iSearch VRE Homepage 34
Figure 17. The MarineSearch VRE Homepage 35
Figure 18. The Scalable Data Mining VRE Homepage 36
Figure 19. The TabularDataLab VRE Homepage 37
Figure 20. The TBTI VRE Homepage 38
Figure 21. The TimeSeries VRE Homepage 39
Figure 22. The VesselActivitiesAnalyser VRE Homepage 40
Figure 23. The VME-DB VRE Homepage 41
Figure 24. The VTI VRE Homepage 41

GLOSSARY

Code list: A predefined list from which some (statistical) coded concepts take their values.

EA-CoP: Ecosystem Approach Community of Practice.

Community of Practice: A term coined to capture an "activity system" that includes individuals who are united in action and in the meaning that "action" has for them and for the larger collective. The communities of practice are "virtual", *i.e.*, they are not formal structures, such as departments or project teams. Instead, these communities exist in the minds of their members, are glued together by the connections they have with each other, as well as by their specific shared problems or areas of interest. The generation of knowledge in communities of practice occurs when people participate in problem solving and share the knowledge necessary to solve the problems.

OAI-PMH: An HTTP based protocol consisting of a set of six verbs or services that make it possible for (i) *Data Providers* to expose structured metadata on their resources and (ii) *Service Providers* to harvest that metadata and offer enhanced services on that.

Occurrence data: A set of observations of the presence (sometimes also absence) of a species or other taxonomic entity, usually in a specified location. These data are often contained in a repository that can be queried using de-facto standards for species observational data.

SDMX: An international Standard for the exchange of statistical (mainly aggregated) datasets.

Virtual Research Environment: A "*system"* with the following distinguishing features: *(i)* it is a Web-based working environment; *(ii)* it is tailored to serve the needs of a Community of Practice; *(iii)* it is expected to provide a community of practice with the whole array of commodities needed to accomplish the community's goal(s); *(iv)* it is open and flexible with respect to the overall service offering and lifetime; and *(v)* it promotes fine-grained controlled sharing of both intermediate and final research results by guaranteeing ownership, provenance, and attribution.

VRE: see Virtual Research Environment.

1 INTRODUCTION

Virtual Research Environments [5] are "systems" aiming at providing their users with web-based working environments offering the entire spectrum of facilities (including services, data and computational facilities) needed to accomplish a given task by dynamically relying on the underlying infrastructure. VRE's are the key products to be delivered by the iMarine project to meet the needs of the iMarine Ecosystem Approach Community of Practice.

This deliverable – D6.5 *'Virtual Research Environments Activity Report*' – details the deployed Virtual Research Environments in terms of community tools integrated, resources involved, and user exploitation. It describes the set of software artefacts that, in addition to the core technology, has been developed to serve the specific needs identified by Ecosystem Approach Community of Practice (EA-CoP). It build upon D6.4 [10] released in the previous period and updates it to describe the state of art at September 2014.

All tools related to EA-CoP activities have been discussed by the iMarine Board. The WP3 Wiki page summarizes the expected VREs in terms of functional requirements and on the EA-CoP exploitation planning, requirements and desiderata:

http://wiki.i-marine.eu/index.php/Ecosystem_Approach_Community_of_Practice:_VRE_planning

The described VREs were expected to evolve (and have) for the entire duration of the project. New VREs, community tools and applications and data sources were described and discussed, and existing community tools were subject to change and review. At a later stage, delivered community tools were also be validated and commented.

Once the EA-CoP desiderata have reached a stable state, they are thoroughly assessed from the technology perspective. The results are captured in wiki-pages representing the Virtual Research Environment development plan [3] and the Applications and Tools development plan [9]. Here, planning and implementation goes through 3 well identified steps; Analysis, Development, and Deployment. The requirements can be implemented as any of the following functional components: VREs, (Community) Tools, or integration of Community Data and Data Tools.

The e-infrastructure is equipped with tools that facilitate Community tools integration. Here the opportunity is to operate in the e-infrastructure existing frameworks, applications, or components that leverage services or support use-cases that the community wants to port to the e-infrastructure. These expand on already integrated e-infrastructure resources such as PostgreSQL/PostGIS and Geoserver. Examples of integrated tools are: OpenSDMX for statistical data access and transport, THREDDS for geospatial data access, R for tabular data analysis, AquaMaps for projection of species distribution algorithms, GeoNetwork for geospatial data discovery and access, FiMES schema support for publication of species fact-sheets species information, Hadoop and WPS for data processing, and X-Search for semantic search.

Finally, the planning has to cope with the integration or interoperation of Community data and data tools that will be critical for the exploitation by the EA-CoP of the e-infrastructure. Most of this integration is performed dynamically, with data only loaded on demand, and respecting the data providers' access and sharing policies. These cover several domains, including:

- Species Data: OBIS: the Ocean Bio-geographic Information System, GBIF: the Global Biodiversity Information Facility[2], CoL: Catalogue of Life[3], WoRMS: World Register of Marine Species[4], IRMNG, WoRDSS;
- Statistical: FAO Code lists exposed through the FAO SDMX registry[5], code lists exposed through any SDMX registry like IRD and Eurostat, and FAO capture data sets;
- Geospatial data exposed via OGC standards like Web Map Service for map images, and WFS for features for any compliant source the EA-CoP wishes to access.

The remainder of this report is organised as follows. Section 2 describes the technologies and resources that have been developed and integrated to support the deployment of Virtual Research Environments. Section 3 describes the Virtual Research Environments that have been deployed and operated during the reporting period. For each Virtual Research Environment the deliverable describes the goal, the data available through such an environment and the set of facilities supported.

[2] http://www.gbif.org/
[3] http://www.catalogueoflife.org/
[4] http://www.marinespecies.org/
[5] http://www.fao.org/figis/sdmx/

2 VIRTUAL RESEARCH ENVIRONMENTS RESOURCES AND TOOLS

In order to realize Virtual Research Environments it is fundamental to equip these working environments with an effective mix of facilities and data. This section reports how *(a)* the data sources that have been integrated in the infrastructure can be used in a Virtual Research Environment (cf. Section 2.1), and *(b)* the set of tools that, in addition to the gCube technology [1][8][12][14], has been developed to provide Virtual Research Environment users with facilities supporting their tasks (cf. Section 2.2).

2.1 DATA RESOURCES

In many applications users have to be provided with data. The effectiveness of a Virtual Research Environment might be affected if the user must acquire the data with no support from the Virtual Research Environment.

In the majority of cases, the data are of diverse types and are scattered among a number of heterogeneous data sources. Moreover, data evolve over time and the user should be provided with an up to date version or dynamically accessed data. Because of this, a number of facilities have been developed (namely in the context of the Data Management area [8]) with the aim to act as mediators between the data sources and services/clients aiming at consuming and making available such data in a seamless way, e.g. Species Products Discovery [4]. Very often – actually whenever possible – such mediators are built by relying on standards and protocols for data access and discovery including OAI-PMH[6], TAPIR[7], SDMX[8], and DarwinCore[9].

By relying on the gCube technologies, the most important data sources needed to support the applications discussed in the context of the EA-CoP have been linked to the iMarine infrastructure and made available in various Virtual Research Environments (see Section 3 for a list of data made available in the context of every VRE).

2.1.1 SPECIES DATA

The following data bases and information systems have been integrated to give access to species data, both taxonomic data and occurrence data:

- Catalogue of Life: The data source offers an integrated checklist and a taxonomic hierarchy of more than 1.3 million species of animals, plants, fungi and micro-organisms.
- FAO ASFIS: The List of Species for Fishery Statistics Purpose includes 12,000+ species of interest or relations to fisheries and aquaculture; www.fao.org/fishery/collection/asfis/en

[6] www.openarchives.org/pmh/
[7] wiki.tdwg.org/TAPIR/
[8] sdmx.org
[9] rs.tdwg.org/dwc/

- GBIF: The data source offers more than 430 million of records on species and more than 14,000 datasets aggregated from 580+ publishers; www.gbif.org
- Fishbase: The data source offers access to 32,700 Species, 302,900 Common names, 53,600 Pictures, 49,700 References aggregated thanks to the effort of thousand collaborators.
- IRMNG: The Interim Register of Marine and Nonmarine Genera data source offers access to over 465,000 genus names and 1.6 million species names; www.obis.org.au/irmng
- ITIS: The Integrated Taxonomic Information System data source offers authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world; http://www.itis.gov
- NCBI Taxonomy: The National Center of Biotechnology Information data source offers a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet; www.ncbi.nlm.nih.gov/taxonomy
- OBIS: The Ocean Biogeographic Information System data source offers more that 40 million records on species and 1,600+ datasets; www.iobis.org
- SeaLifeBase: The data source offers access to 126,000 Species, 27,300 Common names, 11,900 Pictures, 18,200 References aggregated thanks to the effort of hundred collaborators.
- WoRMS: The World Register of Marine Species data source offers species “names” for more than 200,000 species including 300,000+ species names and synonyms and 400,000+ taxa; http://www.marinespecies.org
- WoRDSS & The World Register of Deep-Sea Species data source offers species “names” for deep-sea species based on WoRMS. http://www.marinespecies.org/deepsea

2.1.2 GEOSPATIAL DATA

The following data bases and information systems have been integrated to give access to geospatial data:

- FAO GeoNetwork: The data source exposes spatial data maintained by FAO and its partners; http://www.fao.org/geonetwork
- World Ocean Atlas: The data source gives access to a number of environmental variables. In particular, iMarine focuses on some indicators including Apparent Oxygen Utilisation, Dissolved Oxygen, Nitrate, Oxygen Saturation, Phosphate, Sea Water Salinity, Sea Water Temperature, and Silicate; www.nodc.noaa.gov/OC5/WOA09/pr_woa09.html
- Marine Regions: The data source gives access to a standard list of marine georeferenced place names and areas including EEZ; www.marineregions.org
- myOceans: The data source gives access to a number of environmental variables. In particular, D4Science focuses on some indicators including ice concentration, ice thickness, ice velocity, mass concentration of chlorophyll in sea water, meridional velocity, mole concentration of dissolved oxygen in sea water, mole concentration of nitrate in sea water, mole concentration of phosphate in sea water, mole concentration of phytoplankton expressed as carbon in sea water, net primary production of carbon, salinity, sea surface height, temperature, zonal velocity, wind speed, and wind stress. http://www.myocean.eu

2.1.3 STATISTICAL DATA

The following data bases and information systems have been integrated to give access to statistical datasets:

- IRD Datasets: The UMR EME/Observatoire Thonier SDMX Registry and Repository exposes the Sardara database that contains tuna captures data from several countries, aggregated according to CWP statistical squares (1'x1' or 5'x5') and the ObServe database that contains tuna and bycatches captures observed by scientific observers on-board of French industrial purse seiners.
- Codelists: A set of SDMX Codelists either directly accessed from the FAO Registry, or manually uploaded through the facility developed in the context of ICIS.

2.1.4 OTHER DATA

The following data bases and information systems have been integrated to give access to other resources including papers:

- Aquatic Commons: The data source offers access to thematic material covering natural marine, estuarine/brackish and fresh water environments; aquaticcommons.org
- BHL: The Biodiversity Heritage Library data source offers access to legacy literature of biodiversity held by a consortium of natural history and botanical libraries; www.biodiversitylibrary.org
- Bioline: The Bioline International data source offers access to open access quality research journals published in developing countries; www.bioline.org.br
- CEEMar: The Central and Eastern European Marine Repository data source offers material covering marine, brackish and fresh water environments; www.ceemar.org/dspace
- DataCite: The data source offers access to the same service whose mission is to give access to research data; www.datacite.org
- DBPedia: The knowledge base results from Wikipedia. It contains over 4 millions things including persons, places, creative works, organisations, species and diseases; dbpedia.org/About
- DRS: The data source at National Institute of Oceanography offers institutional publications including journal articles and technical reports; drs.nio.org/drs
- Dryad: The data source offers access to the same service whose mission is to give access to research data underlying research publications; datadryad.org
- FactForge: The knowledge base results from the integration of a number of datasets including DBPedia, WordNet, Geonames, and Freebase; factforge.net
- FAO Factsheets: The data source gives access to the Aquatic Species Fact Sheets developed by the same FAO programme; www.fao.org/fishery/fishfinder
- FAO FLOD: A semantic knowledge base hosted by FAO containing a dense network of relationships among the major entities of the fishery domain, including marine species, water areas, land areas, and exclusive economic zones; www.fao.org/figis/flod
- iMarine TLO Warehous: The warehouse integrates information from FishBase, WoRMS, ECOSCOPE, FLOD and DBPedia by using the same top-level ontology developed for the marine domain. It currently contains approximately 3 millions of triples about more than 40,000 entities including marine species, ecosystems, water areas, and vessels; http://www.ics.forth.gr/isl/MarineTLO

- Nature: The data source offers access to the articles published by nature.com.
- OceanDocs: The data source offers research and publication materials in Marine Science by aggregating content form 256 repositories; www.oceandocs.net
- OpenAIRE: The data source give access to the publications aggregated by the same European funded project; www.openaire.eu
- PANGAEA: The data source offers georeferenced data from earth system research via OAI-PMH. The system guarantees long-term availability of its content through a commitment of the operating institutions. The aggregated repositories are 475; www.pangaea.de
- PenSoft Journals: The data source gives access to a number of open-access journals. In particular, iMarine focuses on BioRisk, Comparative Cytogenetics, International Journal of Myriapodology, Journal of Hymenoptera Research, MycoKeys, Nature Conservation, NeoBiota, PhytoKeys, Subterranean Biology, and ZooKeys.
- SmartFish: The SmartFish Chimaera knowledge base offers a unified and integrated view on three marine fisheries information sources, i.e. FIRMS - an international knowledge base including fisheries and resource from West Indian Ocean; StatBase - a statistical database containing statistics provided by West Indian Ocean countries; and WIOFish - a regional knowledge base on West Indian Ocean Fisheries.
- WHOAS: The data source offers the production of Woods Hole scientific community including articles and data sets; www.mblwhoilibrary.org/services/whoas-repository-services
- YAGO2: The knowledge base extends the YAGO knowledge base by anchoring entities, facts and events in time and space. The knowledge base is built from Wikipedia, GeoNames and WordNet and contains more than 440 million facts about 9.8 million entities.

2.2 TOOLS

In addition to the rest of gCube technology [1][8][12][14] that is mainly conceived to offer core facilities, a number of common tools and user interfaces have been developed by relying on such core facilities to provide Virtual Research Environment users with instruments supporting their tasks. These facilities are mainly oriented to realise specific environments supporting well defined tasks (e.g. BiOnym, Sec. 2.2.1). In the remainder of this section, the major tools that have either be developed or reinforced between July 2013 and September 2014 are reported.

2.2.1 BIONYM PORTLETS

The *BiOnym* portlets are a set of user interfaces conceived to enable its users to use the homonymous facility [16]. The BiOnym facility is a taxonomic naming matching system that is flexible and customisable with respect to the species names to use as references and the lexical matching criteria.

The user is offered a choice of several taxonomic reference lists, including the option to upload his/her own list to the iMarine infrastructure. Where possible, internationally recognized references are dynamically linked to the iMarine infrastructure; this avoids issues with intellectual property rights, and eliminates the inconvenience of keeping the reference lists up to date. The following lists are available in the infrastructure: the Catalogue of Life, World Register of Marine Species, Interim Register of Marine and Non-marine Genera, National Center for Biotechnology Information, and the Integrated Taxonomic Information System.

The matching process follows a workflow approach, starting with a pre-processing step, followed by series of operators to do the actual matching, concluding with a post-processing step. The pre-processing includes a parser, to split a taxonomic name in its atomized components (e.g. splitting the string in the name proper and the authority field), and a resolver to settle common spelling variations (e.g. replacing all occurrences of 'var.' to 'v.'). In the post-processing step, the modalities governing how the results of the matching process are presented to the user is defined. The matching is performed through a series of operators acting as switches. Each switch decides, on the basis of customizable criteria, whether a pair of names should be considered as 'matches', and splits the input list in 'matched' and 'non-matching' names. The matches go, with the criteria that were used to establish the match, to post-processing; the non-matching names are sent to the next switch. Two broad categories of switches are considered. A first type uses some kind of distance, such as the Levenshtein or soundex distance. Another type of switch applies a transformation to both test- and reference names (e.g. strip off gender-specific suffix of specific epitheton), and then look for matches. The switches are configurable and it is possible to upload customized character/string substitutions to configure the pre-processing step and transformations used by switches.

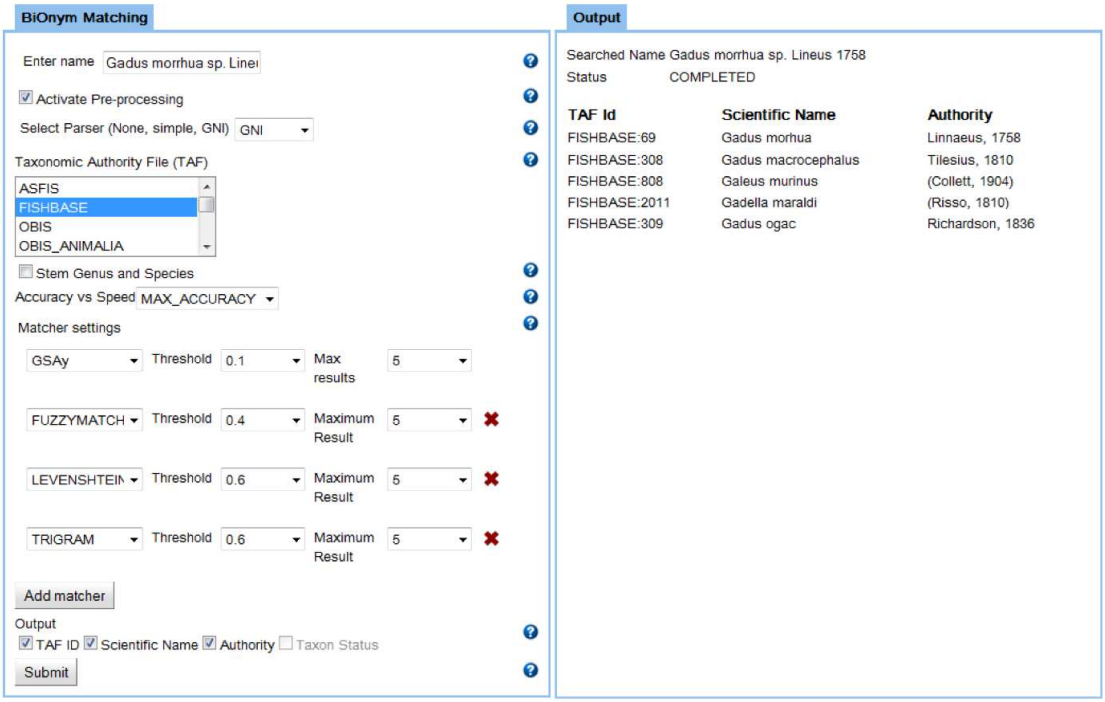

Figure 1. BiOnym Portlet Screenshot

Figure 1 depicts one of the web user interfaces that have been implemented to allow users configure the complete chain. The user can choose the sequence of matchers, the maximum length of the list produced by each matcher and the recognition threshold for each matcher score, under which the transcription will be not reported. At the end of the process, the system returns a table with the list of possible transcriptions for the input string. This list can contain a maximum number of entries corresponding to the sum of the list lengths allowed for the matchers. This is the theoretical case of a sequence of matchers that were absolutely complementary.

Two additional interfaces have been developed: (a) a simple version where the user is allowed to specify only the species name to compare (the rest of parameters are pre configured), and (b) a Statistical Manager

version, i.e. a portlet enabling to execute BiOnym via the Statistical Manager environment [4] on pair with other algorithms.

2.2.2 COTRIX PORTLET

The COTRIX portlet is specifically conceived to act as the front end for the COTRIX facility, a collaborative oriented tool for Code Lists management specifically devised to deal with multiple roles, code lists versioning and evolution. The portlet is developed to work both as a standalone application as well as in a gCube based portal. When running in the gCube based portal, the portlet is nicely integrated with many gCube facilities including the Workspace and Social Networking.

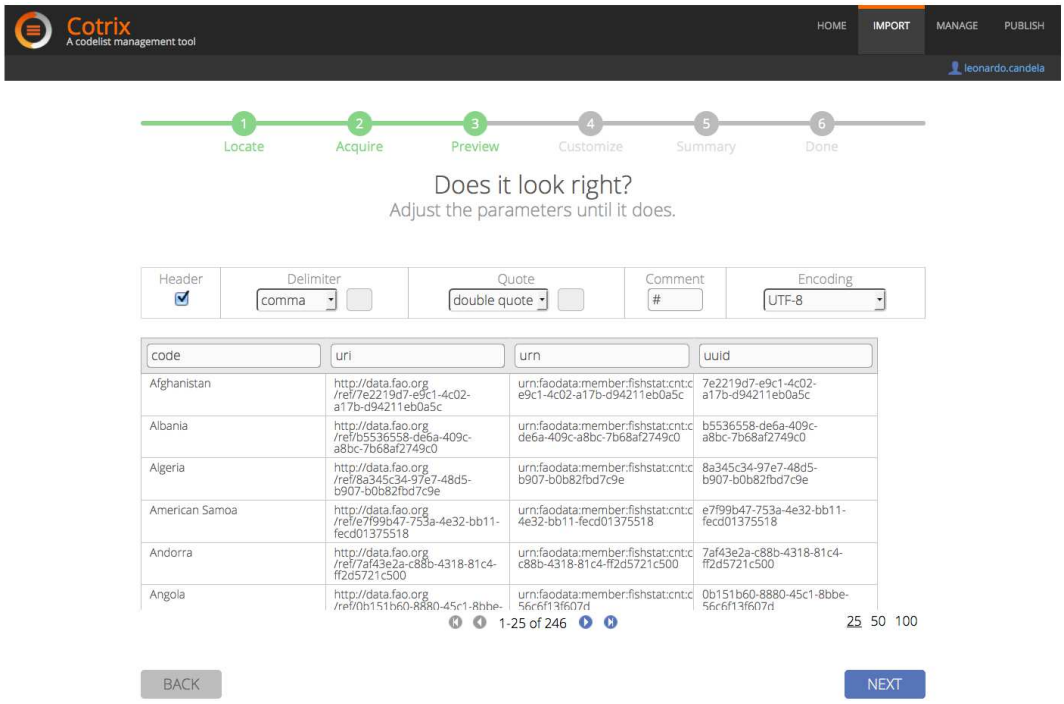

Figure 2. COTRIX Portlet Screenshot

Figure 2 depicts a COTRIX portlet screenshot. This portlet enables authorised users (i) to import a new Code List, either from the user machine or from a series of recognised repositories including SDMX compliant ones, (ii) to collaboratively manipulate Code Lists, e.g. add new entries, merge two entries, inspect code lists values, and (iii) to publish Code Lists either by producing a file (CSV, SDMX, COMET) to be stored on the user machine or by publishing it in a recognised repository including an SDMX compliant one.

2.2.3 TABULAR DATA MANAGER PORTLETS

The Tabular Data Manager environment is a new set of portlets and widgets offering a feature rich set of facilities for the management of tabular resources, i.e. datasets that can be represented in a table format. This set of portlets is expected to replace the environment previously developed for the management of Time Services since it is more flexible and richer than the previous.

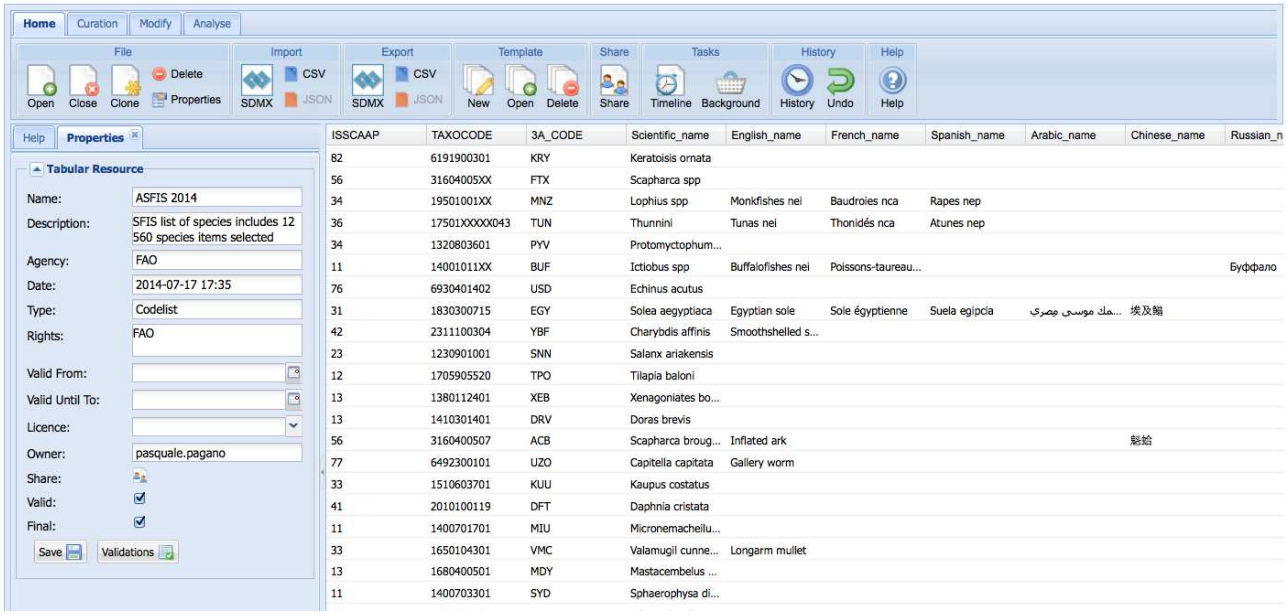

Figure 3. Tabular Data Manager Portlet Screenshot

Figure 3 depicts a screenshot of the entire environment. From this picture it emerges that the environment is oriented to offer facilities organised according to four major areas: (i) *Home*, collecting all the facilities for browsing the available tabular resources and their detailed metadata, importing new ones and exporting existing ones by relying on CSV, SDMX and JSON formats; (ii) *Curation*, collecting all the facilities for manipulating the tabular resource format and content, e.g. associating code lists to columns, merging columns; (iii) *Modify,* collecting all the facilities for modifying the content of a tabular resource, e.g. adding or removing entire rows, replacing values; and, (iv) *Analyse*, collecting all the facilities for inspecting the content of a tabular resource and extract information from it, e.g. produce graphs and maps, analyse the data via Statistical Manager algorithms [7].

2.2.4 TRENDYLYZER PORTLET

The TrendyLyzer portlet is specifically dedicated to realise a user friendly front end for the homonymous application [2], i.e. an environment to enable users to discover and visualise trends in species databases like the identification of most common species across time.

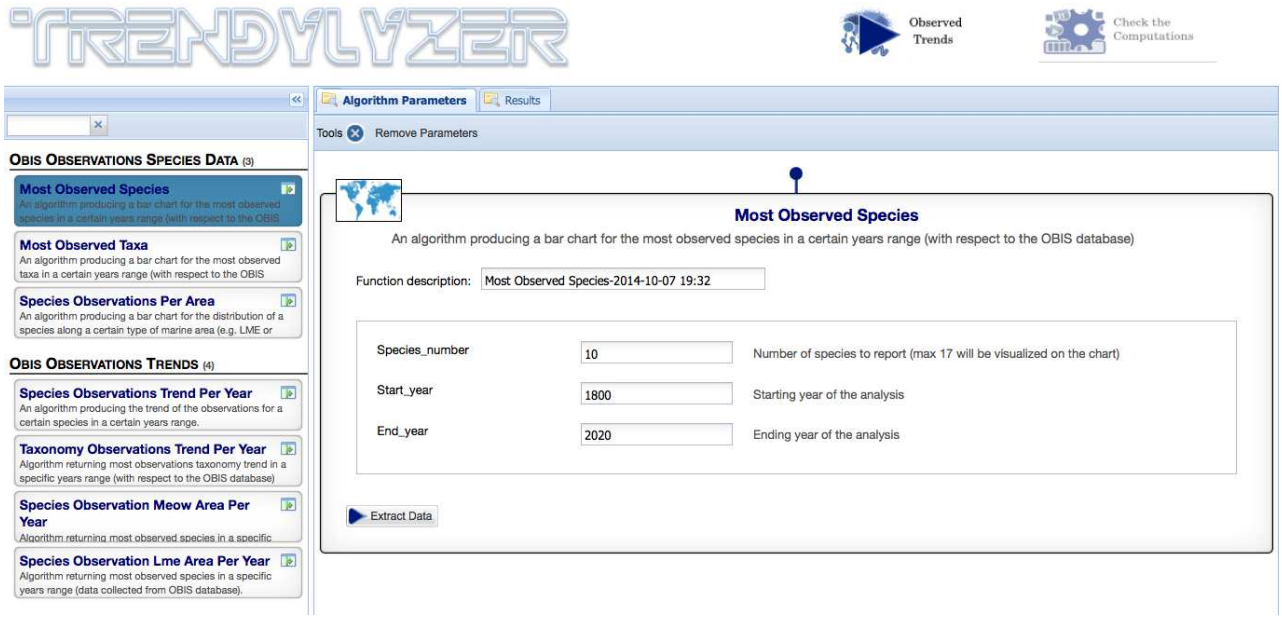

Figure 4. TrendyLyzer Portlet Screenshot

Figure 4 depicts a screenshot of the TrendyLyzer portlet. The portlet is inspired by the Statistical Manager one [4], actually builds upon it. On the left side there is the list of supported algorithms analysing data to extract the trends, while on the right side there is a form to specify the parameters characterising the algorithm behaviour and execute the computation. Once a computation is complete, the portlet shows the results (graphs and other indicators) resulting from it and make it possible to save these resources into the workspace.

2.2.5 MOBILE APPLICATIONS

Three mobile applications have been developed to make is possible for users to exploit the iMarine services via their mobile phones.

The **AppliFish** mobile app for Android and iOS combines species fact sheets with information from other resources. Over 8,000 downloads are reported since its release. The more than 550 fact-sheets inform on species and their distribution, common names, maps and images. They can be used by consumers to better understand the status of a species and make a conscious choice when choosing fish.

The specific facilities this App offers are:

- Species Fact Sheets: AppliFish is built from five major data components: FAO Species Fact-sheets, OBIS and WoRMS common names, AquaMaps, FishBase / SeaLifeBase, and FAO Capture Statistics. This data is

combined in an input process for acquiring data, and brought together in a database which is then incorporated in the mobile application framework;

- Map Download: To reduce the size of the App, maps are dynamically downloaded from the e-Infrastructure.

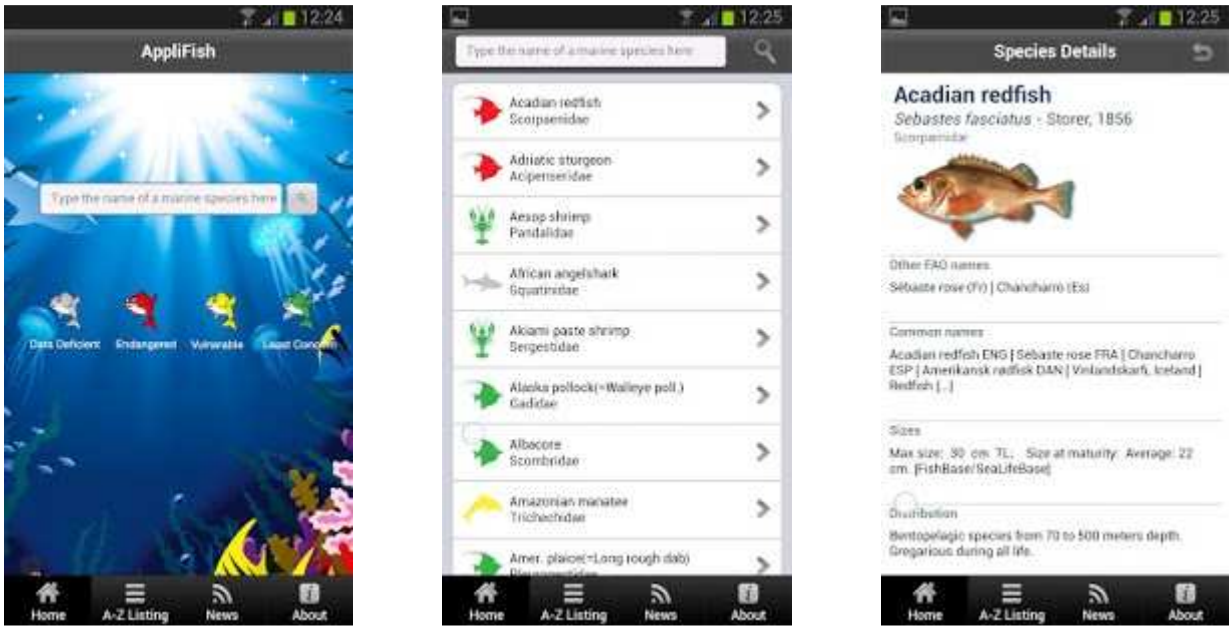

Figure 5. AppliFish Mobile App Screenshots

The ***iMarine Mobile Search App*** is an application that brings the federated search facilities of gCube System, that empowers iMarine Project, into the reach of the mobile user. Through this application, scientists can use their familiar mobile device to easily locate data coming from sources registered in D4Science infrastructure. For the anonymous user, the mobile application offers services over the publicly available subset of the data sources registered in the infrastructure. For the anonymous user, the application offers services over the publicly available subset of the data sources registered in the infrastructure: free text search, results formulation and content retrieval. In addition iMarine registered users can access more features including the selection of the VRE to search though, an advanced search interface for searching, including specific collections selection and/or fields selection.

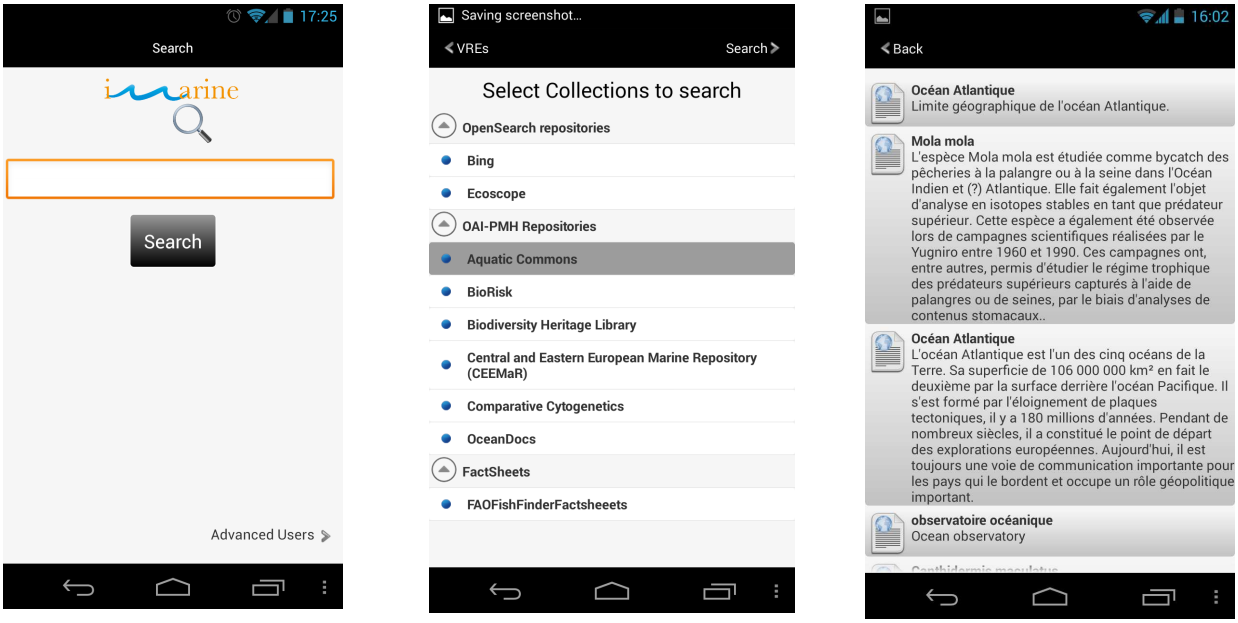

Figure 6. iMarine Search Mobile App Screenshots

The ***Ichtys mobile app*** is an application aiming at providing information about marine species by querying the MarineTLO-based warehouse. In brief, the user enters some basic information about what he is looking for (e.g. the common name of a species in Greek, or the scientific name of a species, etc.) and the application is responsible for constructing the appropriate SPARQL queries, submit them to the warehouse, retrieve the results, and integrate them and present them to the user.

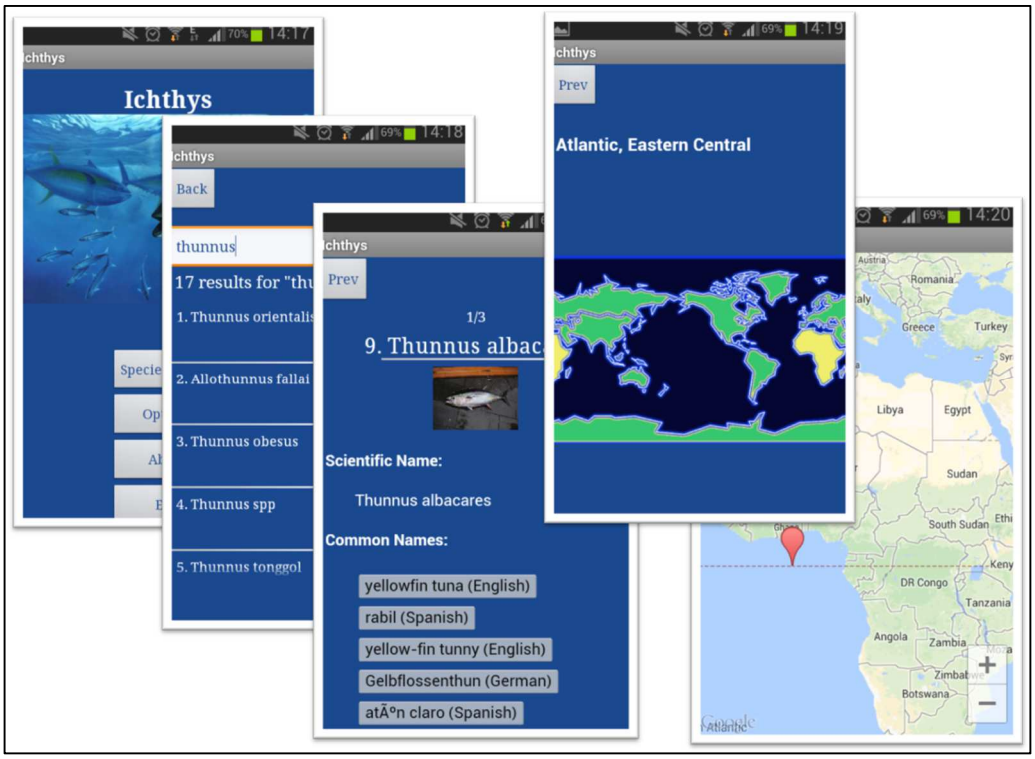

Figure 7. Icthys Mobile App Screenshots

The Icthys mobile app has not been released yet. Currently the application is under testing. From the analysis so far, it became clear that such applications should have an advanced mechanism for caching in order to avoid sending SPARQL queries for every information card. It is worth to highlight that the development of this application was not directly supported by iMarine and will be continue after the project ends.

3 VIRTUAL RESEARCH ENVIRONMENTS DEPLOYMENT AND OPERATION

The iMarine Board Work Plan contains requests for facilities in support of managing EA-CoP data workflows spanning the statistical, geospatial, and biodiversity domains. Some of these facilities can only be defined after others have been released, as not only the technology requirements will evolve, but also the exploitation scenario may change, e.g. when the Board identifies new exploitation opportunities in the Business Cases. This implies that the description will only be finalized after a facility is released.

In order to provide the EA-CoP members with concrete realisation of the expected facilities, the iMarine project deployed and operated a number of Virtual Research Environments as described in the remainder of this section.

3.1 DESIGN, DEPLOYMENT AND OPERATION

The activity leading to the deployment and management of a Virtual Research Environment is driven by a dedicated development plan [3]. According to such a plan which is dynamic and evolving, the WP6 team is requested to analyse the requirements posed by the EA-CoP on various data management workflows and to put in place three types of activities: *(i)* the development of a new or enhanced technology (Services, software libraries, portlets) needed to support the specific need; *(ii)* the modification of an existing Virtual Research Environment to make available the new facility and/or the data that are needed in a given facility; and *(iii)* design and deploy a new Virtual Research Environment. Table 1 reports the list of the Virtual Research Environments operational in September 2014. Some of these Virtual Research Environments have been inherited by previous projects (namely D4Science-II) and because of this are operational since the beginning of the project.

Table 1. Virtual Research Environments

VRE Name	VO	Start Date	# Users
AquaMaps	FARM	Nov 2011	57
BiodiversityLab	gCubeApps	Dec 2013	118
BiodiversityResearchEnvironment	gCubeApps	Jul 2012	41
BiOnym	gCubeApps	Apr 2014	16
DocumentWorkflow	gCubeApps	Nov 2011	44
EcologicalModelling	gCubeApps	Nov 2011	73
FCPPS – Fisheries Country Profiles Production System	FARM	Nov 2011	35
FishFinderVRE	FARM	Mar 2013	15
ICIS – Integrated Capture Information System	FARM	Nov 2011	44
iSearch	gCubeApps	Aug 2013	35
iMarineBoardVRE	FARM	Feb 2013	28
MarineSearch	FARM	Mar 2014	16
ScalableDataMining	FARM	Oct 2012	27
TabularDataLab	gCubeApps	Jun 2014	17
TBTI	FARM	Jul 2013	11
TCom	gCubeApps	Apr 2013	41
TimeSeries	gCubeApps	Nov 2011	60
VesselActivitiesAnalyser	gCubeApps	Nov 2011	47

VME-DB	FARM	Jul 2012	17
VTI – Vessel Transmitted Information	FARM	Nov 2011	28

A brief description of each available VRE is reported in the following sections.

All the VREs are provided with:

- A *shared workspace* to enable every user to store and organise the information objects he/she is interested to work with. In addition to that, the user is allowed to collaborate with other users by sharing objects and messages;
- A *VRE Management facility* to enable authorized users (i.e. VRE Managers) to manage other users using or willing to access the VRE. VRE Managers can (i) authorize users in accessing the VRE, (ii) assign or withdraw roles to users, (iii) remove users, and (iv) send a communication to the current users;
- A *social networking facility* to enable users to use the common facilities promoted by social networks – e.g., posting news, commenting on posted news – yet adapted to settings of the iMarine environment. News can be posted by users as well as applications;
- A *notification facility* to alert users on an as-it-happens basis. These notifications offer a sense of anticipation and create a productivity boost. Users receive an alert (through a priori selected channels, e.g., email, web portal, twitter) notifying them when something of interest has happened in their VRE(s);
- A *messaging facility* to provide users with a common email environment as-a-Service. The distinguishing feature is represented by its integration with the rest, e.g., it is possible to send as attachment any information object residing in the workspace (although "big" and "complex") without consuming bandwidth.

3.1.1 THE AQUAMAPS VRE

The AquaMaps Virtual Research Environment is for providing fisheries and aquaculture scientists with facilities for producing and accessing species predictive distribution maps showing the likelihood that a certain species or a combination of species will live in specific regions or areas.

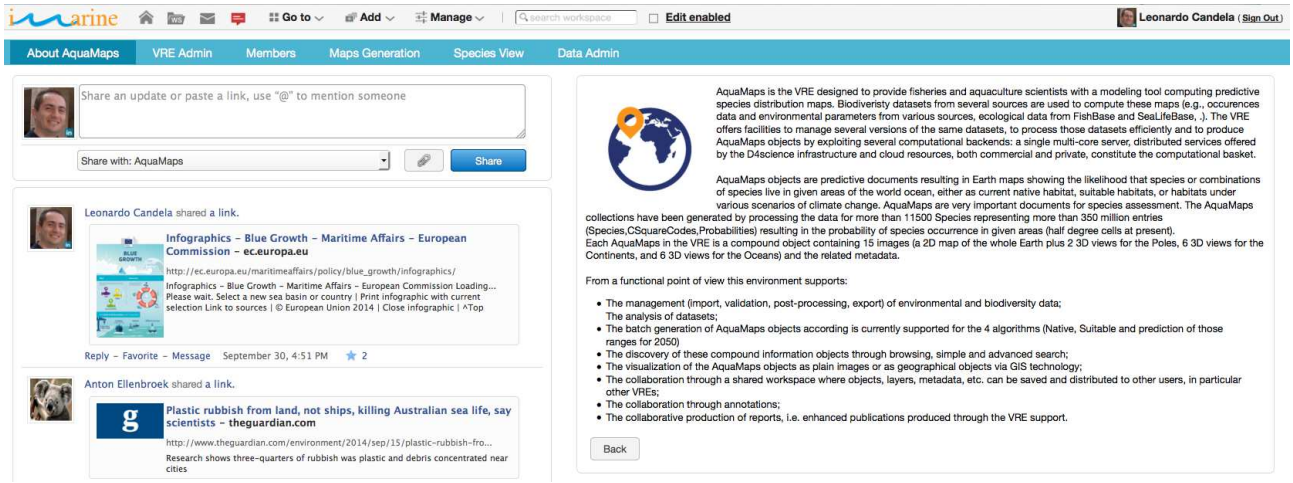

Figure 8. The AquaMaps VRE Homepage

The main facilities this VRE offers are:

- AquaMaps Data Administration: to enable VRE Data Managers to produce new versions of the AquaMaps datasets (the datasets exploited by the AquaMaps service to produce species distribution maps). Such datasets include HSPEC (an estimation of species occurrence by species and cell), HSPEN (an envelope representing the preference of species for environmental ranges), and HCAF (environmental parameters by cell) while the algorithms that can be used to produce new versions of them include Linear or Parabolic Interpolations, Native and Suitable Range, Native 2050 and Suitable 2050 Range;
- AquaMaps Maps Generation: to enable users to produce species and biodiversity predictive distribution maps. The portlet make it possible to select the set of species to analysed, to define the data to use and to submit massive generation tasks leading to the production of AquaMaps objects representing the maps eventually including their GIS version;
- Species View: to enable users to discover and browse species products (namely distribution maps) produced via the AquaMaps Maps Generation facility in an innovative way. This facility supports discovery mechanisms ranging from simple search based on species names to very specific search criterion and it offers a comprehensive set of products visualisation approaches;
- Search: to enable users to discover information objects over a number of collections via a keyword-based Google-like search, an advanced query consisting of diverse criteria per field, or via browsing. Collections are either materialised set of information objects or virtual collections resulting from the interaction with existing data sources (e.g. via harvesting or via query);

The main datasets that are available via the services hosted by this VRE include datasets needed to the AquaMaps algorithm (i.e. HSPEC, an estimation of species occurrence by species and cell; HSPEN, an envelop representing the tolerance of species wrt environmental parameters; and HCAF, environmental parameters by cell), time series graphs produced by FAO (aquaculture, capture, production, and trade), fact sheets produced by FAO on introduced species and cultured aquatic species, maps produced by FAO on country, Current National Legislation Overview (NALO) and National Aquaculture Sector Overview (NASO).

3.1.2 THE BIODIVERSITYLAB VRE

The BiodiversityLab Virtual Research Environment is for scholars willing to perform complete experiments about single individuals or groups of marine species.

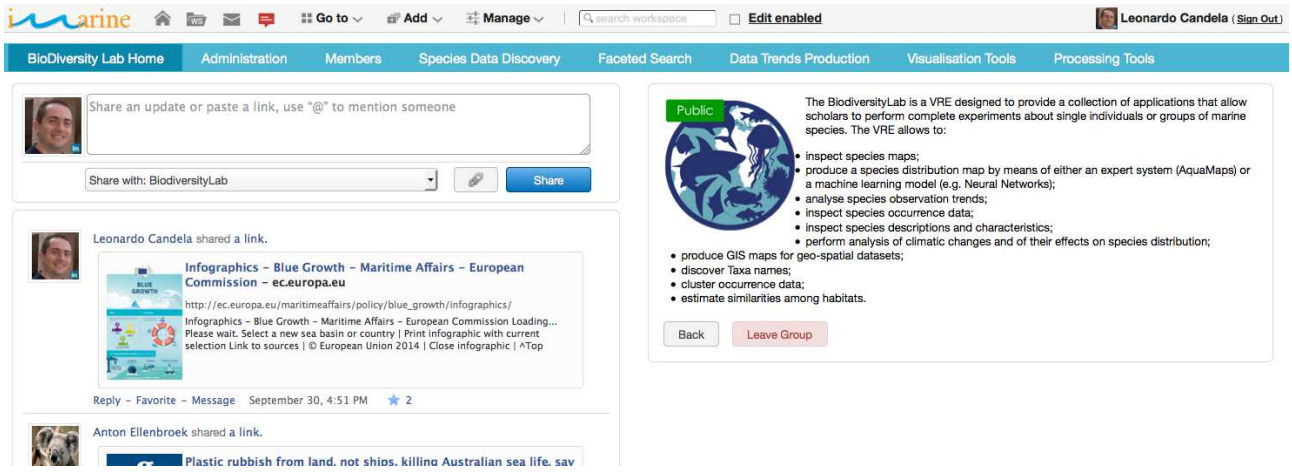

Figure 9. The BiodiversityLab VRE Homepage

The main facilities this VRE offers are:

- Species Data Discovery: to enable users to discover and manage species products (occurrence data and taxonomic data) from a number of heterogeneous providers in a seamless way. Once discovered, objects can be stored in the workspace for future uses;
- Data Trends Identification: to enable users to discover and visualise trends in species databases, e.g. identification of most common species across time (cf. Sec. 2.2.4) [1];
- Species View: to enable users to discover and browse species products (namely distribution maps) produced via the AquaMaps Maps Generation facility in an innovative way. This facility supports discovery mechanisms ranging from simple search based on species names to very specific search criterion and it offers a comprehensive set of products visualisation approaches;
- Geospatial Data View: to enable users to discover and visualize GIS layers, e.g. species distribution maps, that have been generated and published in the iMarine infrastructure. This facilities relies on the GeoExplorer portlet and make it possible to effectively exploit the generated maps and perform comparisons and analysis of the diverse distributions by enabling maps overlay, transects production and values inspection;
- Data Analytics at scale: to enable users to benefit from the offering of the Statistical Manager [4][7] service and interactively execute a large array of data analytics algorithms on datasets. These algorithms range from approaches to produce a species distribution map by means of either an expert system (AquaMaps) or a machine learning model (e.g. Neural Networks) to approaches for analysing climatic changes and their effects on species distribution, approaches for estimating similarities among habitats.

3.1.3 THE BIODIVERSITYRESEARCHENVIRONMENT VRE

The BiodiversityResearchEnvironment Virtual Research Environment is conceived to provide biodiversity scientist with facilities for seamless access to a rich array of biodiversity data including occurrence points and taxa records from established providers including GBIF, Catalogue of Life, and OBIS. This VRE was dismissed during September 2014 since its facilities largely overlap with the BiodiversityLab VRE. Its users have been automatically subscribed to the BiodiversityLab VRE.

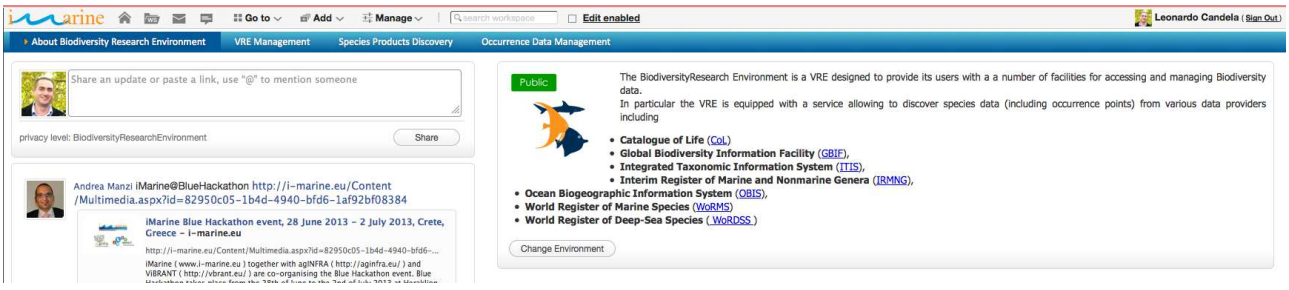

Figure 10. The BiodiversityResearchEnvironment VRE Homepage

The main facilities this VRE offered were:

- Species Products Discovery: to enable users to discover and manage species products (occurrence data and taxa names) from a number of heterogeneous providers in a seamless way [4]. Once discovered, objects can be stored in the workspace for future uses;

- Occurrence Data Management: to enable users to perform data processing tasks on Occurrence Data [4]. In particular, tasks like duplicates detection and removal, datasets union, intersection, and subtraction are supported;

The main datasets that are available via the services hosted by this VRE include:

- Catalogue of Life – offering information on known species of animals, plants, fungi and micro-organisms, namely taxonomic information (cf. Sec. 2.1);
- GBIF – offering information on species, namely occurrence data (cf. Sec. 2.1);
- OBIS – offering information on marine species from all of the world's oceans, both occurrence and taxonomic information (cf. Sec. 2.1);
- SpeciesLink – offering information on species, namely occurrence data;
- WoRMS – offering information on marine species, namely taxonomic information (cf. Sec. 2.1);

3.1.4 THE BIONYM VRE

The BiOnym Virtual Research Environment is for taxonomists and biodiversity scientists dealing with species taxonomic information. The main facility is BiOnym (cf. Sec. 2.2.1), an application specifically conceived to compare species scientific names against reference lists.

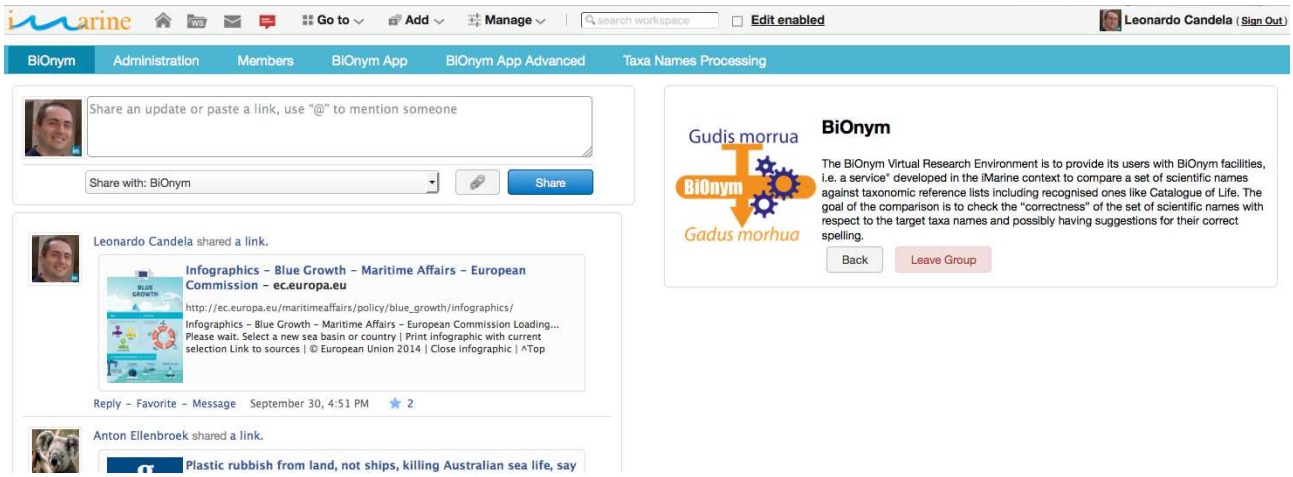

Figure 11. The BiOnym VRE Homepage

The main facilities this VRE offers are:

- BiOnym: to enable users to use a taxonomic naming matching system that is flexible and customisable with respect to the species names to use as references and the lexical matching criteria (cf. Sec. 2.2.1) [16];
- Data Analytics at scale: to enable users to benefit from the offering of the Statistical Manager [4][7] service and interactively execute a large array of data analytics algorithms on taxonomic data. The supported algorithms enables to acquire taxonomic data from well known data providers to customised versions on BiOnym workflows, e.g. BiOnym Biodiv that applies in sequence the following Matchers: GSay (thr:0.6, maxRes:10), FuzzyMatcher (thr:0.6, maxRes:10), Levenshtein (thr:0.4, maxRes:10), Trigram (thr:0.4, maxRes:10).

3.1.5 THE DOCUMENTWORKFLOW VRE

The DocumentsWorkflow Virtual Research Environment is conceived to provide its users with a working environment focused on the gCube facilities for managing Document life-cycles. It exploit the facilities offered by the gCube Business Documents Workflow Management Suite enabling the production of reports that require a collaborative activity of several actors.

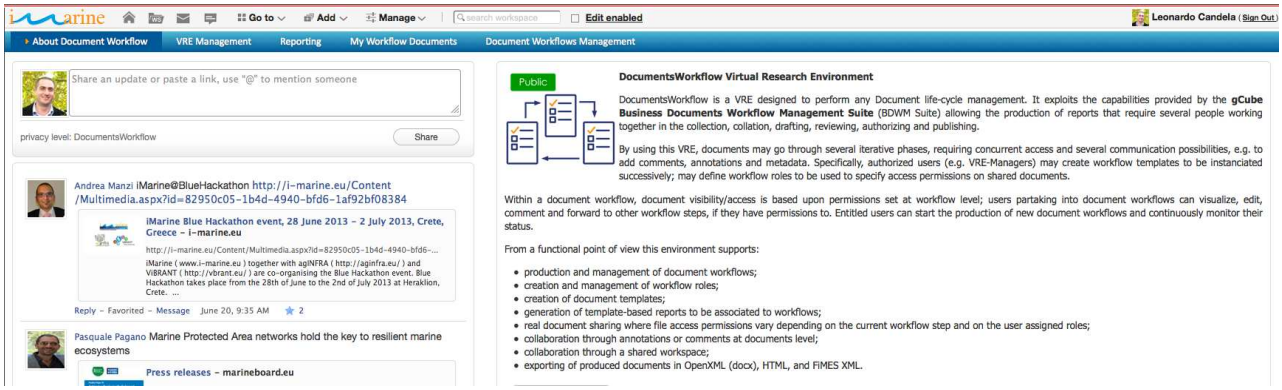

Figure 12. The DocumentsWorkflow VRE Homepage

The main facilities this VRE offers are:

- Reporting facilities: to enable users to collaboratively produce reports consisting in complex documents characterised by well defined structures (templates). Via this facility, users can define new templates as well as collaboratively create new reports compliant with defined templates. Reports might contain diverse elements ranging from texts to images and tables, and such constituents can result from objects stored in the user workspace. Reports can be materialised in multiple formats including PDF, HTML and OpenXML;
- Documents Workflow facilities: to enable users (i) to define complex workflows (including steps and roles users should have to perform certain steps) governing the production of gCube documents, (ii) to instantiate such workflows to actual documents to be collaboratively created, and (iii) to monitor workflow execution;

3.1.6 THE ECOLOGICALMODELLING VRE

The EcologicalModelling Virtual Research Environment is conceived to provide its users with a working environment focused on the gCube facilities for producing species distribution maps resulting from the processing of data on species characteristics and environmental observations. The resulting maps are actually rich information objects containing PNG images, GIS layers as well as metadata.

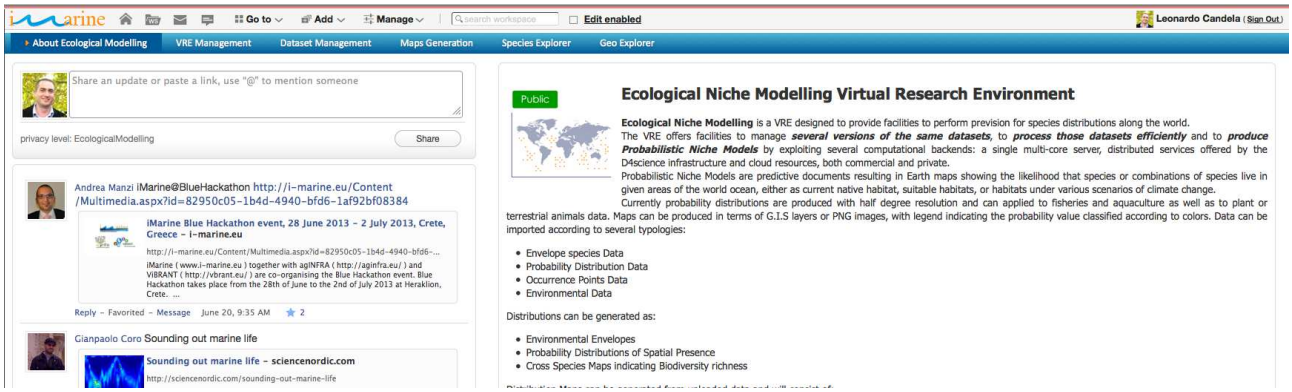

Figure 13. The EcologicalModelling VRE Homepage

The main facilities this VRE offers are:

- AquaMaps Data Administration: to enable VRE Data Managers to produce new versions of the AquaMaps datasets (the datasets exploited by the AquaMaps service to produce species distribution maps). Such datasets include HSPEC (an estimation of species occurrence by species and cell), HSPEN (an envelop representing the tolerance of species with respect to environmental parameters), and HCAF (environmental parameters by cell) while the algorithms that can be used to produce new versions of them include Linear or Parabolic Interpolations, Native and Suitable Range, Native 2050 and Suitable 2050 Range;
- AquaMaps Maps Generation: to enable users to produce species and biodiversity predictive distribution maps. The portlet make it possible to select the set of species to analysed, to define the data to use and to submit massive generation tasks leading to the production of AquaMaps objects representing the maps eventually including their GIS version;
- Species View: to enable users to discover and browse species products (namely distribution maps) produced via the AquaMaps Maps Generation facility in an innovative way. This facility supports discovery mechanisms ranging from simple search based on species names to very specific search criterion and it offers a comprehensive set of products visualisation approaches;
- Species Maps Discovery: to enable users to discover and visualize GIS layers corresponding to species distribution maps that have been generated and published in the iMarine infrastructure. This facilities relies on the GeoExplorer portlet and make it possible to effectively exploit the generated maps and perform comparisons and analysis of the diverse distributions by enabling maps overlay, transects production and values inspection;

The main datasets that are available via the services hosted by this VRE include datasets needed to the AquaMaps algorithm (i.e. HSPEC, an estimation of species occurrence by species and cell; HSPEN, an envelop representing the tolerance of species with respect to environmental parameters; and HCAF, environmental parameters by cell).

3.1.7 THE FCPPS – FISHERIES COUNTRY PROFILES PRODUCTION SYSTEM VRE

The Fisheries Country Profiles Production System (FCPPS) Virtual Research Environment is for fisheries and aquaculture authors, managers and researchers who produce reports containing country-level data. It

provides seamless access to multiple data sources, including their annotation and versioning and permits production of structured text, tables, charts and graphs from these sources to be easily inserted into custom reporting templates that can support multiple output formats.

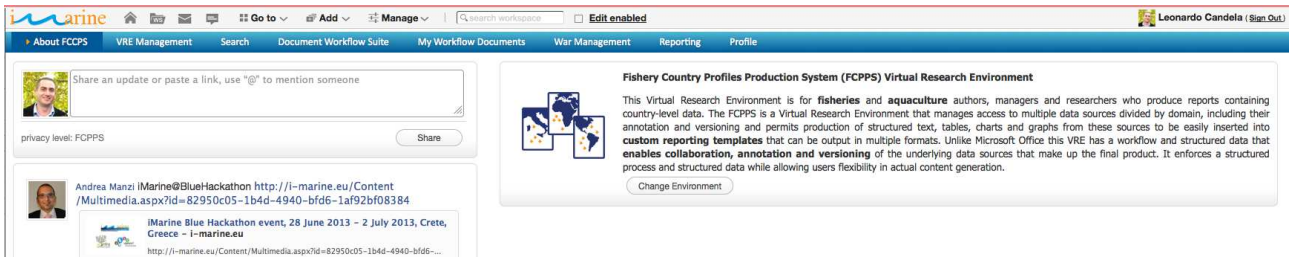

Figure 14. The FCPPS VRE Homepage

The main facilities this VRE offers are:

- Reporting facilities: to enable users to collaboratively produce reports consisting in complex documents characterised by well defined structures (templates). Via this facility, users can define new templates as well as collaboratively create new reports compliant with defined templates. Reports might contain diverse elements ranging from texts to images and tables, and such constituents can result from objects stored in the user workspace. Reports can be materialised in multiple formats including PDF, HTML and OpenXML;
- Documents Workflow facilities: to enable users (i) to define complex workflows (including steps and roles users should have to perform certain steps) governing the production of gCube documents, (ii) to instantiate such workflows to actual documents to be collaboratively created, and (iii) to monitor workflow execution;

The main datasets that are available via the services hosted by this VRE include material acquired via external repositories like AquaticCommons and WHOAS (cf. Sec. 2.1), time series graphs produced by FAO (aquaculture, capture, production, and trade), fact sheets produced by FAO on introduced species and cultured aquatic species, maps produced by FAO on country, Current National Legislation Overview (NALO) and National Aquaculture Sector Overview (NASO).

3.1.8 THE FISHFINDERVRE

The FishFinderVRE is established to elaborate Species Fact Sheets, fill / view their metadata, and select data for download and-or display in the stand-alone version of this VRE. The VRE extends the previously released FCPPS VRE.

The explicit purpose of this VRE is to enable some 50 authors to prepare hundreds of species fact sheets. The tools will be operated by one VRE manager in FAO, with globally distributed authors. The FishFinderVRE allows defining, pre-filling, and elaborating a fact sheet using a template. Fact-sheet compilation will be marshaled by a work flow, with one VRE manager communicating with dozens of authors globally distributed that can access their work-flow to find the tasks assigned to them.

A fact-sheet will be given a scientific species name from a controlled list. The fact-sheet structure will rely on a template structure that is identical for all individual fact-sheet / reports. The user fills this template,

for taxonomic names, biological characteristics, etc. Data can be either visible data or invisible metadata. The users will be assigned their tasks on particular fact-sheets through the work-flow features of the reporting environment.

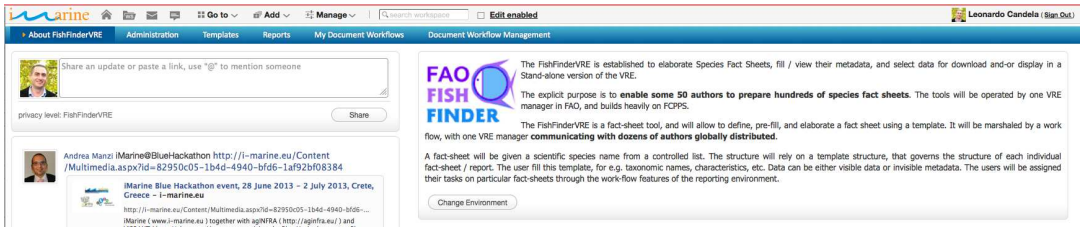

Figure 15. The FishFinderVRE Homepage

The specific facilities this VRE offers are:

- Reporting facilities: to enable users to produce reports as complex documents characterised by well defined structures (templates). Via this facility, users can define new templates as well as collaboratively create new reports compliant with defined templates. Reports might contain diverse elements ranging from texts to images and tables, and such constituents can result from objects stored in the user workspace. Reports can be materialised in multiple formats including PDF, HTML and OpenXML;
- Documents Workflow facilities: to enable users (i) to define complex workflows (including steps and roles users should have to perform certain steps) governing the production of gCube documents, (ii) to instantiate such workflows to actual documents to be collaboratively created, and (iii) to monitor workflow execution;

3.1.9 THE ICIS – INTEGRATED CAPTURE INFORMATION SYSTEM VRE

The Integrated Capture Information System (ICIS) Virtual Research Environment offers fisheries statisticians a set of tools to manage their data. Statisticians produce statistics from often very different data sources, and need a controlled process for the ingestion, validation, transformation, comparison and exploitation of statistical data for the fisheries captures domain.

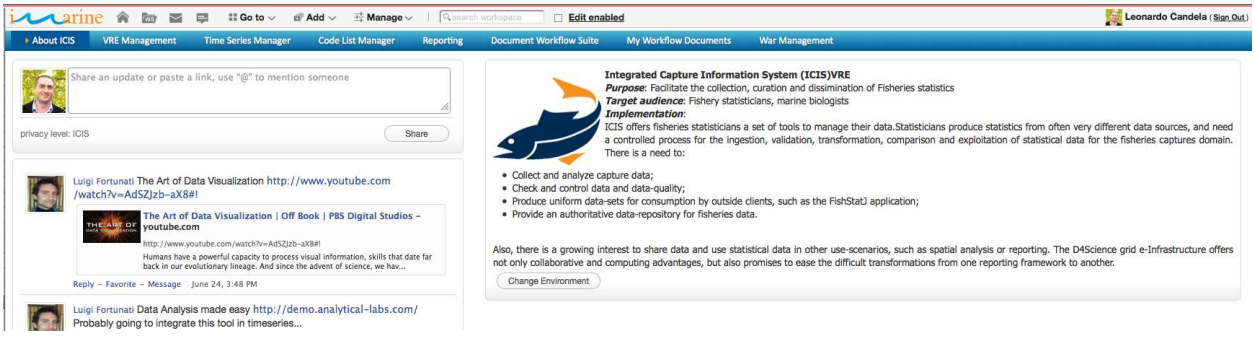

Figure 16. The ICIS VRE Homepage

The main facilities this VRE offers are:

- Time Series Management: to enable users to import, curate and manage time series. This is a comprehensive and feature-rich environment that supports data managers during the whole life cycle of

data management from capture to publishing and visualisation. In enable data managers to import and transform CSV files in time series, i.e. tabular data having proper types associated with columns eventually referring to code lists – reference datasets representing recognized value instances of the elements the dataset is about, e.g., species, zones, countries. The environment guarantees that the time series are compliant with the defined types and code lists. Besides the curation, the environment supports the analysis of the data by enabling a user to (i) perform operations like grouping and filtering, (ii) producing charts and GIS maps (if the data have geographic features) and (iii) analysing the data via an R[10] environment. Finally, the environment supports the publishing of time series in the infrastructure by equipping them with rich metadata so that such resources can be used in other application contexts;

- Code Lists Management: to enable the users to import and manage code lists, i.e. reference datasets representing recognised value instances of the elements the dataset is about. Such environment enable users to import CSV files or existing code lists from SDMX[11] repositories, curate them when needed, inspect the current values, and produce and publish new versions that can be used during the curation phase of a time series. Code lists are annotated with rich metadata capturing attribution and lineage;
- Reporting facilities: to enable users to collaboratively produce reports consisting in complex documents characterised by well defined structures (templates). Via this facility, users can define new templates as well as collaboratively create new reports compliant with defined templates. Reports might contain diverse elements ranging from texts to images and tables, and such constituents can result from objects stored in the user workspace. Reports can be materialised in multiple formats including PDF, HTML and OpenXML;

The main datasets that are available via the services hosted by this VRE include a series of code lists including FAO and IRD SDMX repositories (cf. Sec. 2.1).

3.1.10 THE IMARINE BOARD VRE

The iMarineBoardVRE is designed to provide the members of the iMarine Board with collaboration tools, project information and publications, and as a demonstrator of infrastructure facilities.

The iMarine Board Members are involved in the decision making process for the implementation of a robust digital framework to support the Ecosystem Approach to Fisheries and Natural resources Management. The iMarine e-infrastructure offer a several high-capacity facilities for accessing, sharing, managing and analyzing data, deploying entire applications and services, and this VRE is intended to assist with real-life facilities this decision making process.

It is a responsibility of the Ecosystem Approach Communities of Practice (EA-CoP) to help develop a sustainable governance model, and the guidelines and policies regulating the management and operation of the e-infrastructure. This VRE therefore includes those services that put into effect iMarine governance models and policies, such as (i) the collaboration suite including a shared workspace and messaging system,

[10] http://www.r-project.org/
[11] sdmx.org

(ii) services for accessing biodiversity data from several major databases, and (iii) services for managing tabular data (e.g. catch statistics) and code lists.

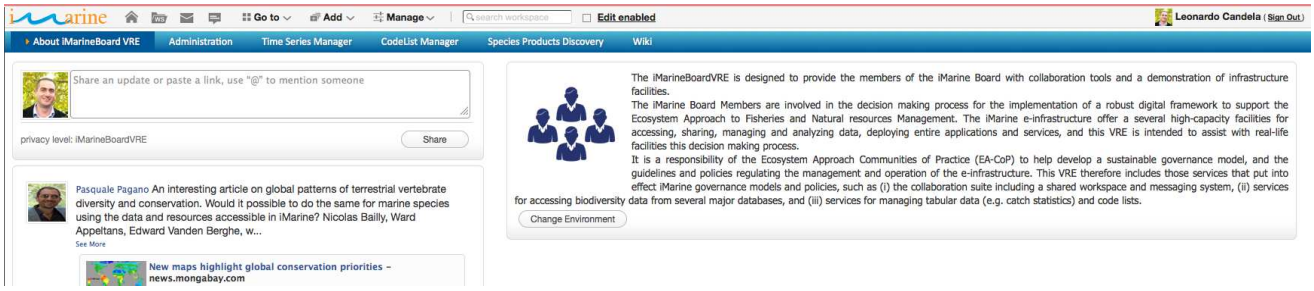

Figure 17. The iMarine Board VRE Homepage

The main facilities this VRE offers are:

- Access through the workspace of all iMarine Board documentation. Previously, these were fragmented in several locations. Through the VRE, communication with and within the Board is facilities by mail, messaging, sharing, and learning facilities.
- Time Series Management: to enable users to import, curate and manage time series. This is a comprehensive and feature-rich environment that support data managers during the whole life cycle of data management from capture to publishing and visualisation. In enable data managers to import and transform CSV files in time series, i.e. tabular data having proper types associated with columns eventually referring to code lists – reference datasets representing recognized value instances of the elements the dataset is about, e.g., species, zones, countries. The environment guarantees that the time series are compliant with the defined types and code lists. Besides the curation, the environment supports the analysis of the data by enabling a user to (i) perform operations like grouping and filtering, (ii) producing charts and GIS maps (if the data have geographic features) and (iii) analysing the data via an R[12] environment. Finally, the environment supports the publishing of time series in the infrastructure by equipping them with rich metadata so that such resources can be used in other application contexts;
- Code Lists Management: to enable the users to import and manage code lists, i.e. reference datasets representing recognised value instances of the elements the dataset is about. Such environment enable users to import CSV files or existing code lists from SDMX[13] repositories, curate them when needed, inspect the current values, and produce and publish new versions that can be used during the curation phase of a time series. Code lists are annotated with rich metadata capturing attribution and lineage;
- A wiki for elaborating Board specific documentation. This wiki allows the board to develop a shared position transparent to all members, while ensuring confidentiality and security. For instance, reviews of iMarine policy documents that are initially for circulation in the Board, or proposals for collaboration at board level are proposed as wiki pages.

[12] http://www.r-project.org/
[13] sdmx.org

3.1.11 THE ISEARCH VRE

The iSearch Virtual Research Environment is for providing its users with an environment dedicated to showcase the data discovery facilities, in particular the IR and semantic based ones.

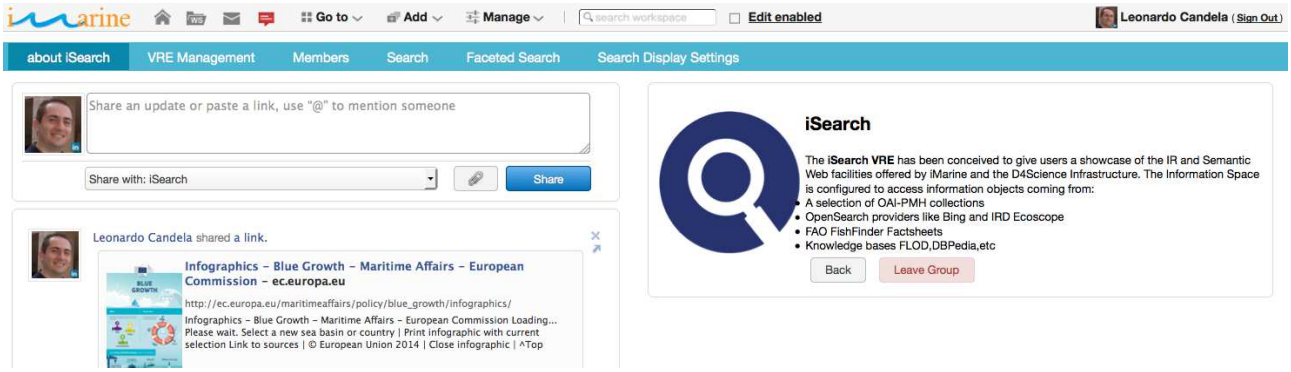

Figure 18. The iSearch VRE Homepage

The main facilities this VRE offers are:

- A keywords-based search facility: to enable users to specify their Google-like queries, search for objects in the VRE information space, visualise the resulting objects, and eventually store them in the workspace;
- An advanced search: to enable to specify complex queries (e.g. per field criteria, set of collection to search in), search for objects in the VRE information space, visualise the resulting objects, and eventually store them in the workspace;
- A faceted search: to enable users to search and browse the MarineTLO-based Data Warehouse [15].

The datasets contributing to form the VRE Information Space include:

- Information objects including research papers and datasets collected by OAI-PMH data providers, e.g. Pensoft Journals like Biodiversity Data Journal, repositories like arXiv, Aquatic Commons, Biodiversity Heritage Library, DataCite and Dryad;
- Information objects resulting from services like Bing and Ecoscope;
- Information objects resulting from Knowledge bases including FLOD, DBPedia, Marine-based TLO;

3.1.12 THE MARINESEARCH VRE

The MarineSearch Virtual Research Environment is for providing the members of the EA-CoP with an environment dedicated to showcase the data discovery facilities, in particular the IR and semantic based ones. The information space, is tailored to marine related data sources.

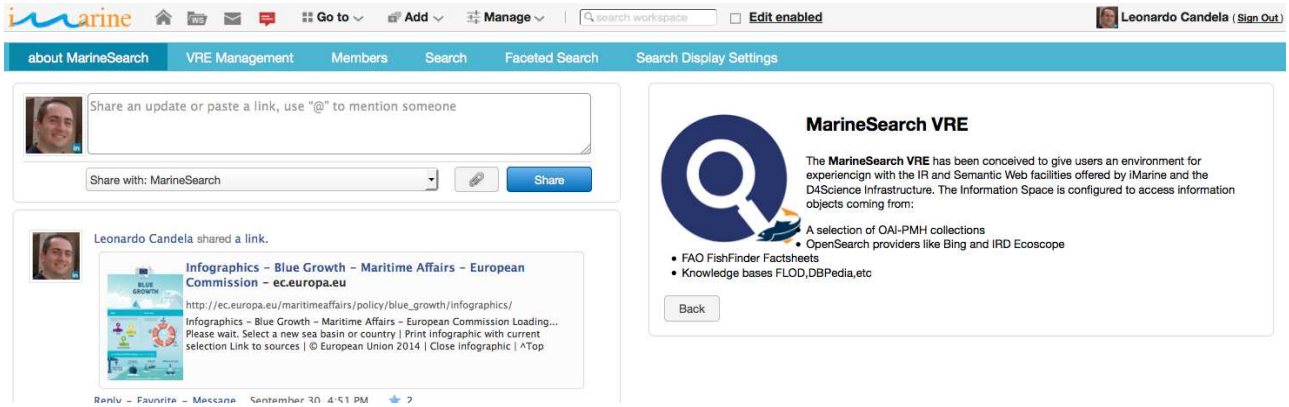

Figure 19. The MarineSearch VRE Homepage

The main facilities this VRE offers are:

- A keywords-based search facility: to enable users to specify their Google-like queries, search for objects in the VRE information space, visualise the resulting objects, and eventually store them in the workspace;
- An advanced search: to enable to specify complex queries (e.g. per field criteria, set of collection to search in), search for objects in the VRE information space, visualise the resulting objects, and eventually store them in the workspace;
- A faceted search: to enable users to search and browse the MarineTLO-based Data Warehouse [15].

The datasets contributing to form the VRE Information Space include:

- Information objects including research papers and datasets collected by OAI-PMH data providers, e.g. Pensoft Journals like Biodiversity Data Journal, repositories like arXiv, OceanDocs, Aquatic Commons Biodiversity Heritage Library, DataCite and Dryad;
- Information objects resulting from services like Bing and Ecoscope;
- Information objects resulting from Knowledge bases including FLOD, DBPedia, Marine-based TLO;

3.1.13 THE SCALABLE DATA MINING VRE

Scalable Data Mining is a Virtual Research Environment designed to apply Data Mining techniques to biological data. The algorithms are executed in a distributed fashion on the e-Infrastructure nodes or on local multi-core machines. Scalability is thus meant as distributed data processing but even as services dynamically provided to the users. The system is scalable in the number of users and in the size of the data to process. Statistical data processing can be applied to perform Niche Modelling or Ecological Modelling experiments. Other applications can use general purpose techniques like Bayesian models. Time series of observations can be managed as well, in order to classify trends, catch anomaly patterns and perform simulations. The idea under the distributed computation for data mining techniques is to overcome common limitations that can happen when using statistical algorithms.

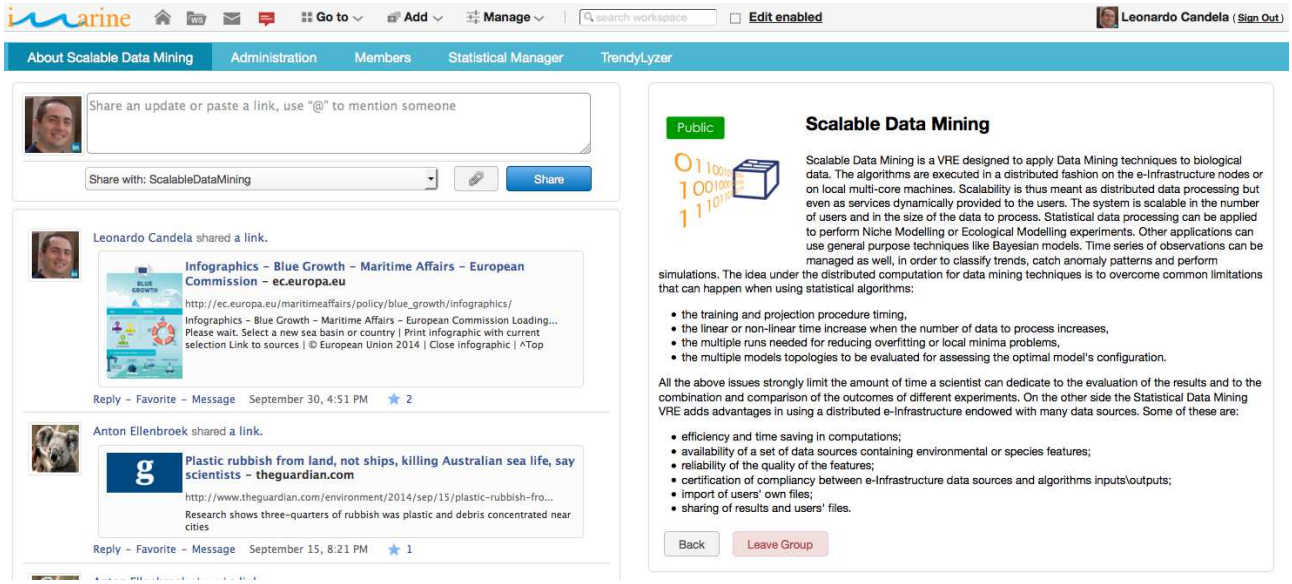

Figure 20. The Scalable Data Mining VRE Homepage

The main facilities this VRE offers are:

- Data Analytics as scale: to enable users to benefit from the offering of the Statistical Manager [4][7] service and interactively execute a large array of data analytics algorithms on datasets. These algorithms range from approaches to produce a species distribution map by means of either an expert system (AquaMaps) or a machine learning model (e.g. Neural Networks) to approaches for analysing climatic changes and their effects on species distribution, approaches for estimating similarities among habitats. Moreover, the environment is configured to enable the execution of algorithms made available via the WPS standard.
- TrendyLyzer: to enable users to benefit from the TrendyLyzer application [2], an application specifically conceived to to discover and visualise trends in species databases, e.g. identification of most common species across time.

3.1.14 THE TABULARDATALAB VRE

The TabularDataLab Virtual Research Environment is for providing data managers with a feature rich environment for supporting the management of tabular data, i.e. any dataset that can be represented in a table format, and code lists. In practice, this environment is aiming at replacing and reinforcing the facilities offered by the TimeSeries VRE (cf. Sec. 3.1.16) with new services.

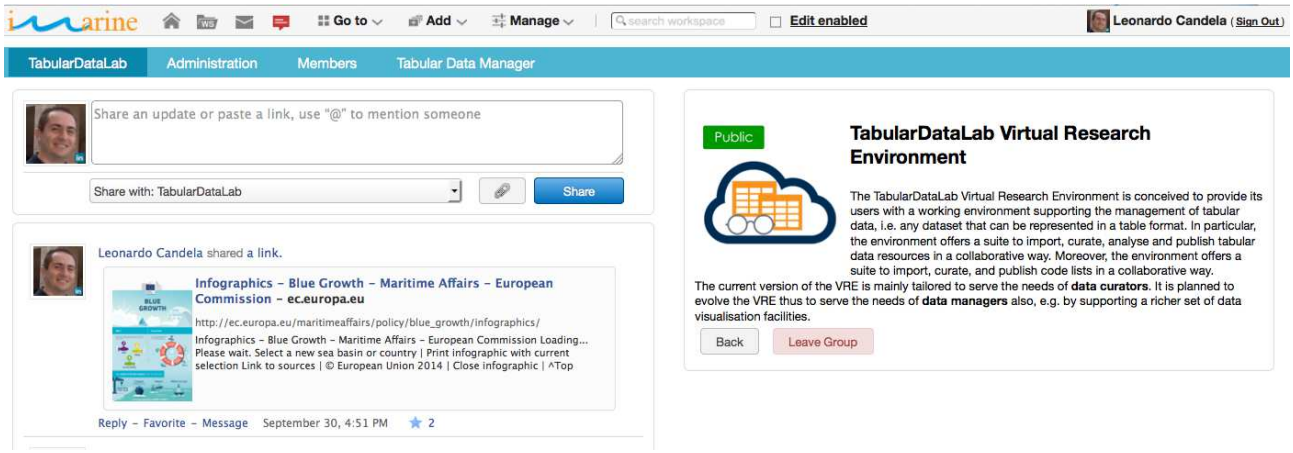

Figure 21. The TabularDataLab VRE Homepage

The main facilities this VRE offers are:

- Tabular Data Management: to enable users to import, curate and manage tabular data (cf. Sec. 2.2.3). This is a comprehensive and feature-rich environment that support data managers during the whole life cycle of data management from capture to publishing and visualisation. In enable data managers to import and transform datasets (CSV, SDMX, JSON) into *tabular resources*, i.e. tabular data having proper types associated with columns eventually referring to code lists – reference datasets representing recognized value instances of the elements the dataset is about, e.g., species, zones, countries. The environment guarantees that the tabular resource are compliant with the defined types and code lists. Besides the curation, the environment supports the analysis of the data by enabling a user to (i) perform operations like grouping and filtering, (ii) producing charts and GIS maps (if the data have geographic features), (iii) analysing the data via an R[14] environment as well as via the data analytics facilities of the Statistical Manager [4]. Finally, the environment supports the publishing of tabular resources in the infrastructure by equipping them with rich metadata so that such resources can be used in other application contexts.
- COTRIX Code Lists Management: to enable the users to experience with the COTRIX environment for Code List Management (cf. Sec. 2.2.2), i.e. a collaborative oriented tool for Code Lists specifically devised to deal with multiple roles, code lists versioning and evolution.

3.1.15 THE TBTI VRE

The TBTI Virtual Research Environment is for the members of the Too Big to Ignore[15] initiative. In particular, the VRE is oriented to provide the members of this community with an environment demonstrating how the iMarine set of facilities can be used to serve the needs of this community that is established to rectify

[14] http://www.r-project.org/
[15] http://toobigtoignore.net Too Big To Ignore Initiative website

the marginalization of small-scale fisheries in national and international policies, and to develop research and governance capacity to address global fisheries challenges.

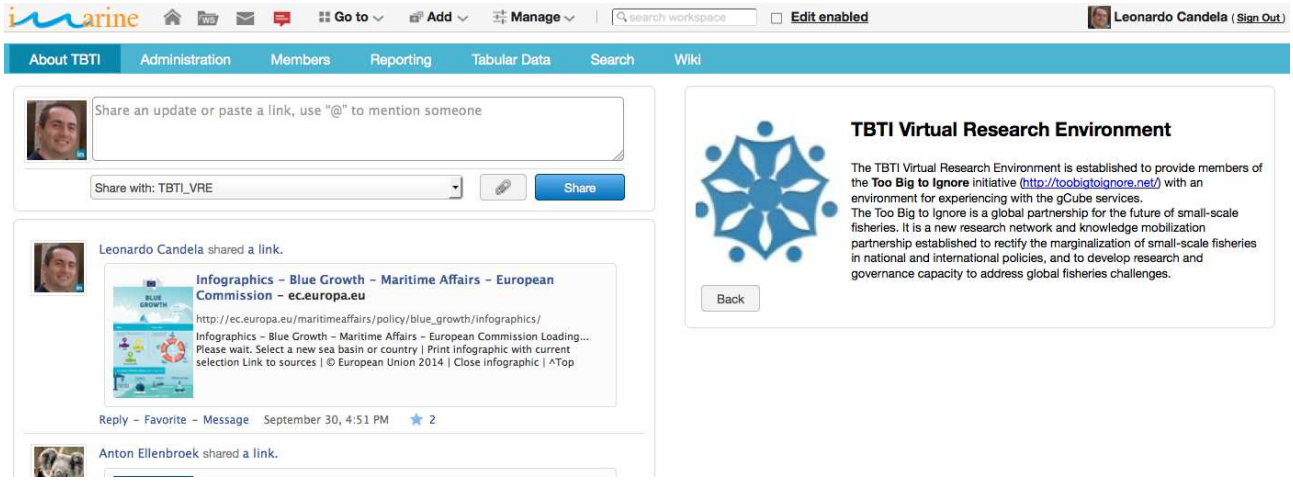

Figure 22. The TBTI VRE Homepage

The main facilities this VRE offers are:

- Reporting facilities: to enable users to collaboratively produce reports consisting in complex documents characterised by well defined structures (templates). Via this facility, users can define new templates as well as collaboratively create new reports compliant with defined templates. Reports might contain diverse elements ranging from texts to images and tables, and such constituents can result from objects stored in the user workspace. Reports can be materialised in multiple formats including PDF, HTML and OpenXML;
- Documents Workflow facilities: to enable users (i) to define complex workflows (including steps and roles users should have to perform certain steps) governing the production of gCube documents, (ii) to instantiate such workflows to actual documents to be collaboratively created, and (iii) to monitor workflow execution;
- Species Products Discovery: to enable users to discover and manage species products (occurrence data and taxa names) from a number of heterogeneous providers in a seamless way [4]. Once discovered, objects can be stored in the workspace for future uses;
- A keywords-based search facility: to enable users to specify their Google-like queries, search for objects in the VRE information space, visualise the resulting objects, and eventually store them in the workspace;
- An advanced search: to enable to specify complex queries (e.g. per field criteria, set of collection to search in), search for objects in the VRE information space, visualise the resulting objects, and eventually store them in the workspace;
- Geospatial Data View: to enable users to discover and visualize GIS layers, e.g. species distribution maps, that have been generated and published in the iMarine infrastructure. This facilities relies on the GeoExplorer portlet and make it possible to effectively exploit the generated maps and perform comparisons and analysis of the diverse distributions by enabling maps overlay, transects production and values inspection.

3.1.16 THE TIMESERIES VRE

The TimeSeries Virtual Research Environment is conceived to provide its users with a working environment focused on gCube facilities for managing time series. This environment supports the load of time series objects, the curation and validation by relying on authoritative code lists, the sharing of such objects with co-workers, the production of graphs, the visualization through a GIS service;

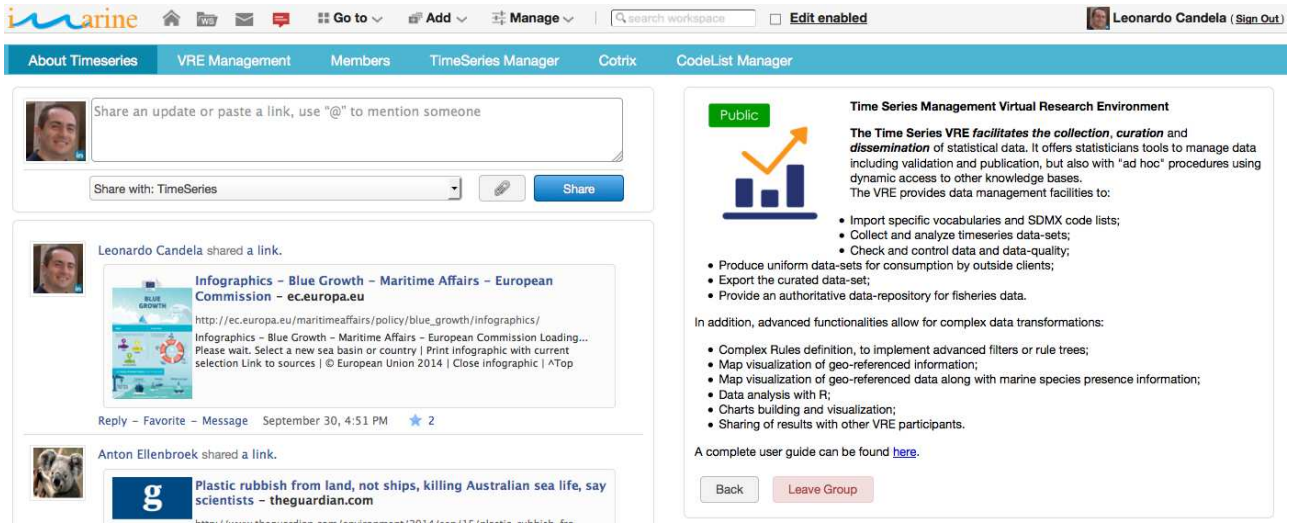

Figure 23. The TimeSeries VRE Homepage

The main facilities this VRE offers are:

- Time Series Management: to enable users to import, curate and manage time series. This is a comprehensive and feature-rich environment that support data managers during the whole life cycle of data management from capture to publishing and visualisation. In enable data managers to import and transform CSV files in time series, i.e. tabular data having proper types associated with columns eventually referring to code lists – reference datasets representing recognized value instances of the elements the dataset is about, e.g., species, zones, countries. The environment guarantees that the time series are compliant with the defined types and code lists. Besides the curation, the environment supports the analysis of the data by enabling a user to (i) perform operations like grouping and filtering, (ii) producing charts and GIS maps (if the data have geographic features) and (iii) analysing the data via an R[16] environment. Finally, the environment supports the publishing of time series in the infrastructure by equipping them with rich metadata so that such resources can be used in other application contexts;
- Code Lists Management: to enable the users to import and manage code lists, i.e. reference datasets representing recognised value instances of the elements the dataset is about. Such environment enable users to import CSV files or existing code lists from SDMX[17] repositories, curate them when needed,

[16] http://www.r-project.org/
[17] sdmx.org

inspect the current values, and produce and publish new versions that can be used during the curation phase of a time series. Code lists are annotated with rich metadata capturing attribution and lineage;

The main datasets that are available via the services hosted by this VRE include a series of code lists including FAO and IRD SDMX repositories (cf. Sec. 2.1).

3.1.17 THE VESSELACTIVITIESANALYSER VRE

The VesselActivitiesAnalyzer Virtual Research Environment is conceived to provide its users with a working environment focused on gCube facilities for managing vessel trajectories. This environment support users in loading and curating vessel trajectories, enriching such data with bathymetry and FAO Area, sharing with co-workers, analysing such objects by producing maps on vessel activities and fishing monthly effort.

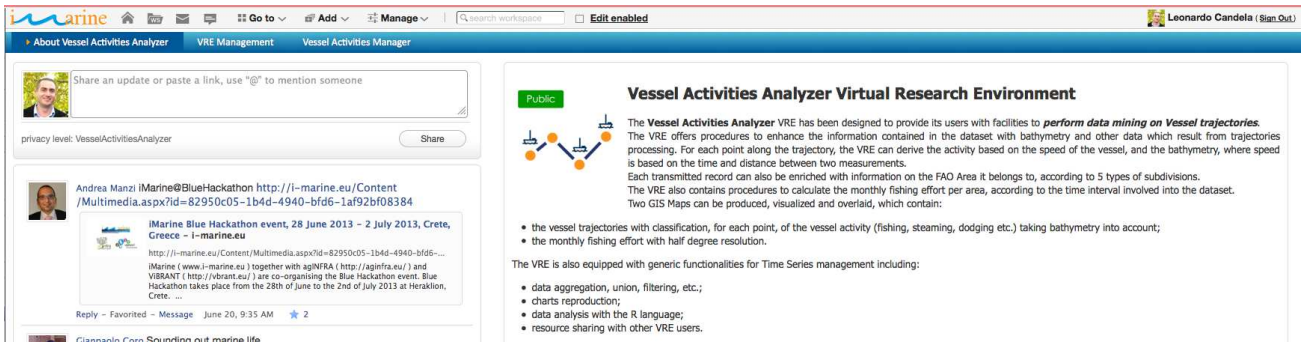

Figure 24. The VesselActivitiesAnalyser VRE Homepage

The specific facilities this VRE offers are:

- Vessel Activities Management: to enable users to perform data mining tasks on Vessel trajectories. It enables data managers to import and transform CSV files representing trajectories into well-defined tabular data, to enrich such tabular data with information on FAO areas and bathymetry, to perform mining tasks aiming at deriving the vessel activity by relying on vessel speed and bathymetry. Besides these specific operations, the environment supports the analysis of the data by enabling a user to (i) perform operations like grouping and filtering, (ii) producing charts and GIS maps (if the data have geographic features) and (iii) analysing the data via an R[18] environment.

The VRE offers a base for extension for EA-CoP members in need of a geospatial explicit analysis of their often confidential data. At the time of writing this document, several initiatives were aware of the potential, but with limiting data access and sharing policies could not yet contribute data and algorithms for further development, implementation and use of the facility.

[18] http://www.r-project.org/

3.1.18 THE VME-DB VRE

The **Vulnerable Marine Ecosystem Database (VME-DB)** Virtual Research Environment is for fisheries and aquaculture authors willing to collaboratively produce Fact Sheets on Vulnerable Marine Ecosystems (VME).

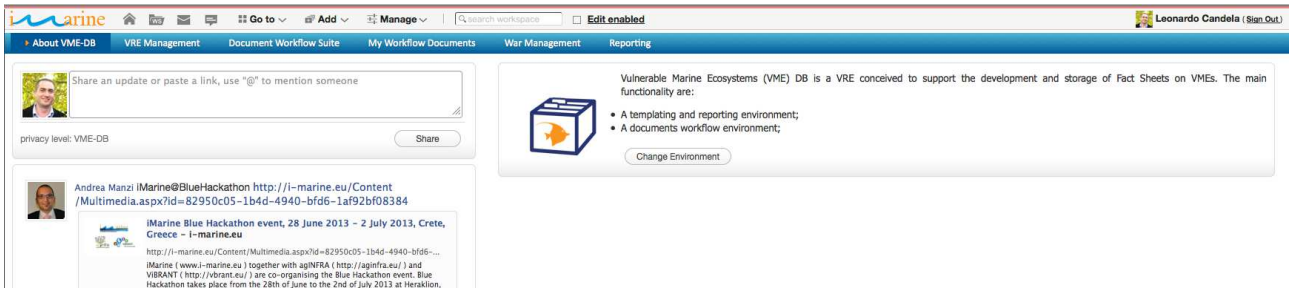

Figure 25. The VME-DB VRE Homepage

The specific facilities this VRE offers are:

- Reporting facilities: to enable users to collaboratively produce reports consisting in complex documents characterised by well defined structures (templates). Via this facility, users can define new templates as well as collaboratively create new reports compliant with defined templates. Reports might contain diverse elements ranging from texts to images and tables, and such constituents can result from objects stored in the user workspace. Reports can be materialised in multiple formats including PDF, HTML and OpenXML;
- Documents Workflow facilities: to enable users (i) to define complex workflows (including steps and roles users should have to perform certain steps) governing the production of gCube documents, (ii) to instantiate such workflows to actual documents to be collaboratively created, and (iii) to monitor workflow execution.

3.1.19 THE VTI – VESSEL TRANSMITTED INFORMATION VRE

The Vessel Transmitted Information (VTI) Virtual Research Environment enables the analysis of vessel activities over space and time by taking into account environmental data.

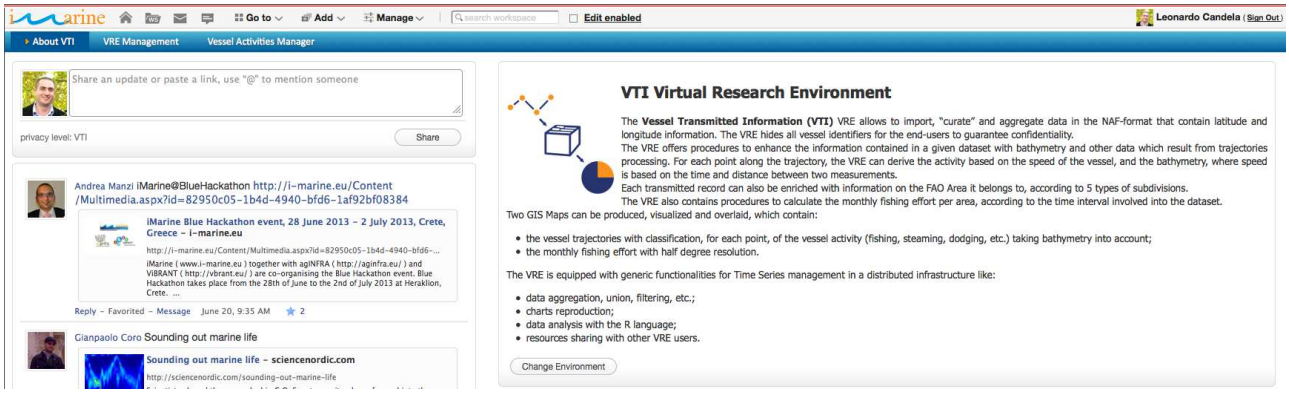

Figure 26. The VTI VRE Homepage

The specific facilities this VRE offers are:

- Vessel Activities Management: to enable users to perform data mining tasks on Vessel trajectories. In enable data managers to import and transform CSV files representing trajectories into well-defined tabular data, to enrich such tabular data with information on FAO areas and bathymetry, to perform mining tasks aiming at deriving the vessel activity by relying on vessel speed and bathymetry. Besides these specific operations, the environment supports the analysis of the data by enabling a user to (i) perform operations like grouping and filtering, (ii) producing charts and GIS maps (if the data have geographic features) and (iii) analysing the data via an R[19] environment.

[19] http://www.r-project.org/

4 CONCLUDING REMARKS

Virtual Research Environments are among the key products to be delivered by the iMarine project to meet the needs of the iMarine Ecosystem Approach Community of Practice. They are “systems” aiming at providing their users with web-based working environments offering the entire spectrum of facilities (including services, data and computational facilities) needed to accomplish a given task by dynamically relying on the underlying infrastructure.

This deliverable detailed the deployed Virtual Research Environments in terms of community tools integrated, resources involved, and user exploitation. It described the set of software artefacts that, in addition to the core technology, has been developed to serve the specific needs identified by Ecosystem Approach Community of Practice. Moreover, it listed the entire offering of iMarine in terms of Virtual Research Environments as of September 2014. Overall ***20 VREs*** have been deployed and operated to serve a total of ***770 users***. The infrastructure is also serving the mobile apps community, e.g. as of today there are more than 8,000 instances of AppliFish.

REFERENCES

[1] A. Antoniadis, R. Cirillo, C. Formisano, L. Fortunati, J. Gerbesiotis, L. Lelii, P. Liakos, A. Savini, F. Simeoni, E. Travaglino. iMarine Data Infrastructure Enabling Software. iMarine D8.3 Project Deliverable. October 2013

[2] W. Appeltans, P. Pissierssens, G. Coro, A. Italiano, P. Pagano, A. Ellenbroek, T. Webb. Trendylyzer: a Long-Term Trend Analysis on Biogeographic Data. *In Proceedings of the International Conference on Marine Data and Information Systems (IMDIS)*. Lucca, Italy, 2013

[3] L. Candela. *Virtual Research Environments Plan.* iMarine D6.1 Project Deliverable. January 2012

[4] L. Candela, D. Castelli, G. Coro, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano. An Infrastructure-oriented Approach for supporting Biodiversity Research, *Ecological Informatics*, DOI: 10.1016/j.ecoinf.2014.07.006, 2014

[5] L. Candela, D. Castelli, P. Pagano. Virtual Research Environments: An Overview and a Research Agenda. *Data Science Journal 12: GRDI75-GRDI81*, 2013

[6] L. Candela, R. Cirillo, A. Manzi. iMarine Data e-Infrastructure Operation Report. iMarine D5.4 Project Deliverable. September 2014.

[7] G. Coro, L. Candela. gCube Statistical Manager: the Algorithms. Technical Report 2014-TR-27. Istituto di Scienza e Tecnologie dell'Informazione, CNR. 2014

[8] G. Coro, L. Lelii, A. Manzi, N. Drakopoulos, V. Marioli, R. Cirillo, F. Simeoni, A. Antoniadis, F. Brito, H. Caumont. iMarine Data Management Software. iMarine D9.3 Project Deliverable. January 2014

[9] A. Ellenbroek, L. Candela. *Applications and Tools Development Plan*. iMarine D6.2 Project Deliverable. May 2012

[10] A. Ellenbroek, L. Candela, P. Pagano. Virtual Research Environments Activity Report. D6.4 iMarine Project Deliverable. July 2013

[11] A. Gentile. FCPPS Assessment for VME-DB Use-Case. iMarine internal document. August 2012

[12] J. Gerbesiotis. iMarine Data Consumption Software. iMarine D10.5 Project Deliverable. January 2014

[13] Y. Laurent. ICIS Assessment for Tuna Atlas Use-Case. iMarine internal document. July 2012

[14] M. Simi, R. Tsantouli. Application Programming Interface Software. iMarine D11.3 Project Deliverable. October 2012

[15] Y. Tzitzikas, C. Allocca, C. Bekiari, Y. Marketakis, P. Fafalios, M. Doerr, N. Minadakis, T. Patkos and L. Candela. Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology, *7th Metadata and Semantics Research Conference, MTSR 2013*, Thessaloniki, Greece. DOI 10.1007/978-3-319-03437-9_29

[16] Vanden Berghe E., Bailly N., Coro G., Fiorellato F., Aldemita C., Ellenbroek A., Pagano P. *BiOnym: a flexible workflow approach to taxon name matching.* Technical Report 2014-TR-22. Istituto di Scienza e Tecnologie dell'Informazione, CNR. 2014