Project Acronym	***iMarine***
Project Title	***Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources***
Project Number	***283644***
Deliverable Title	***iMarine Data Consumption Software***
Deliverable No.	**D10.5**
Delivery Date	***January 2014***
Author	***John Gerbesiotis – NKUA***

Abstract: *This document describes the novelties within the iMarine Data Consumption Software which were achieved from the 13th to the 27th month of the project and provide pointers to the documentation and artifacts of the related components.*

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.	D10.5
Deliverable Title	iMarine Data Consumption Software
Contractual Delivery Date	January 31st 2014
Actual Delivery Date	January 29th 2014
Author(s)	Alex Antoniadis – NKUA, Fabrice Brito – Terradue, Gianpaolo Coro – CNR, John Gerbesiotis – NKUA, Nikolas Laskaris – NKUA, Yannis Marketakis – FORTH
Editor(s)	John Gerbesiotis – NKUA
Reviewer(s)	Andrea Manzi – CERN
Contributor(s)	
Work Package No.	WP 10
Work Package Title	Data Consumption Facilities Development
Work Package Leader	John Gerbesiotis – NKUA
Work Package Participants	NKUA, CNR, FORTH, Terradue, FAO
Estimated Person Months	26.00
Distribution	Public
Nature	Other
Version / Revision	1.0
Draft / Final	Final
Total No. Pages (including cover)	15
Keywords	Data Manipulation, Information Retrieval, gCube

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/).

The information contained in this document represents the views of the iMarine Consortium as of the date they are published. The iMarine Consortium does not guarantee that any information contained herein is error-free, or up to date. THE IMARINE CONSORTIUM MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, BY PUBLISHING THIS DOCUMENT.

GLOSSARY

ABBREVIATION	DEFINITION
iMarine	Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources
gDTS	gCube Data Transformation Service
GWT	Google Web Toolkit
OGC	Open Geospatial Consortium
PE2ng	gCube Process Excecution Engine
SM	Statistical Manager
WCS	Web Coverage Service
WPS	Web Processing Service

DELIVERABLE SUMMARY

1.1 INTRODUCTION

iMarine Data Consumption software is the outcome of development activities took place in the context of iMarine Work Package 10 – Data Consumption Facilities Development. The main objective of this work package is to develop a set of facilities for supporting the data processing tasks the EA-CoP faces with.

These facilities include services for:

- data discovery and retrieval;
- generation and manipulation of data;
- mining and extraction of knowledge from raw data;
- generation of provenance information and the link of this information to the data.

This deliverable describes the novelties within the iMarine Data Consumption Software from M13 (Nov. '12) up to M27 (Jan. ’14).

1.2 TARGET RELEASE(S)

The gCube releases that contain software produced by WP10 activities are:

- gCube 2.11.1
- gCube 2.12.0
- gCube 2.13.0
- gCube 2.14.0
- gCube 2.15.0
- gCube 2.16.0
- gCube 2.16.1
- gCube 2.17.0
- gCube 2.17.1

Announcements containing information about the features, enhancements and fixes contained in the above releases can be found at [1].

1.3 OBJECTIVES

The new versions of components belonging to Data Consumption Software cover the following objectives:

- Search System and FullTextIndex optimizations:

  Caching techniques and code optimizations have been used in order to optimize the overall performance of the Search System and FullTextIndex. Furthermore, the data transferred from the index to the search system have been reduced to further improve the performance of the search process.

- Replacement of underlying Index mechanisms:

  Both FullTextIndex and ForwardIndex have been integrated with new backend components which provide important built-in capabilities, such as replication and load-balancing. This integration improved the overall performance and eliminated the need of low level index management. FullTextIndex has been integrated with ElasticSearch[20], which is a distributed document-oriented database on top of lucene, while ForwardIndex has been integrated with Couchbase[21], also a document-oriented database offering key-value store capabilities.

- Statistical Manager enhancements:

  The Statistical Manager platform has been enhanced, in the aspects of the gCube Service, the web interface and the algorithms running on the system.

  - Statistical Manager Interoperability improvement:

    In particular, the Statistical Manager has been enhanced to handle heterogeneous types of outputs, including tables, images and files at the same time. Experimental evaluation of distributed modelling has been made towards the usage of PE2ng and the Executor components of the gCube platform, for niche modelling experiments. Their respective performances were evaluated and compared.

  - Statistical Manager distributed execution:

    Enhancements were made regarding the multi-tenancy operation of the system and the robustness of the interface. As a result, multiple experiments' "executions" can be managed at the same time on the same SM machines or on the distributed computing facilities (currently based on the Executor). The interface allows to upload tables, files and is able to properly manage Darwin Core Archives.

  - General purpose modeling:

    Well known algorithms were introduced for general purpose modelling. An example is the Feed Forward Neural Networks, employed to perform simulation of real valued functions. Furthermore, algebraic operations for species occurrence datasets have been added and managed in a distributed fashion.

  - Time Series analysis:

    Short-Time Fourier transform has been applied to environmental trends and to catch statistics in order to detect periodic phenomena. Investigation on new methods for taxa matching, based on Latent Semantic Analysis has been performed too. Taxa names matching has been addressed by integrating several techniques, coming from the FIN and FAO experience, which rely on several databases and biodiversity data sources. The resulting system is a flexible workflow engine (BiOnym) which is able to combine a sequence of taxa matching algorithms. Furthermore, in collaboration with Dr. Rainer Froese of the GEOMAR institute resulted in running R algorithms in parallel fashion. Finally, the algorithm for estimation of length-weight relationships for marine species was run in parallel, significantly reducing the execution time.

- Data processing of OGC maps:

  A maps comparison functionality performs discrepancy analysis among species distributions at different time instants and spatial resolutions, by means of Kappa statistics, has been implemented. Work was also placed on the Trendylyzer system, which is able to extract trends for species observations from the OBIS database and to extract unbiased assumptions about species distributions from their observation trends. To such aim, a clustering-based approach was adopted to the definition of common species from such trends. Finally, the problem of "Stock Biomass Spawning vs Recuitments Production" was approached, by means of a Bayesian model to estimate the function that regulated the relationship between an estimated amount of spawned recruits and a certain fish biomass.

- Enriching gCube search results with semantic information:

  In order to enrich the results, a generic meta-search service has been developed, named **xsearch-service** [13]. xsearch-service provides advanced services for satisfying recall-oriented information needs and for semantically enriching the results. These services are:

  (a) textual clustering, which is performed over the textual snippets or the contents of the search results
  (b) textual entity mining, that can be performed over textual snippets or the contents of the results
  (c) provision of gradual faceted search, which allows the user to quickly explore the results space by exploiting the identified entities that have been mined and the clustering results
  (d) connection of the information derived from semantic knowledge bases (for the case of iMarine the MarineTLO warehouse [14] is being exploited) with the extracted entities, which allows the user to retrieve more information about specific entities and start exploring them
  (e) exploitation of the above functionalities in any web page through bookmarklets. It is possible to apply these services over the entire answer returned by the underlying system or only over the top-K hits returned

  In addition, the **xsearch-portlet** has been developed. It is responsible for presenting the semantically enriched results and interacting with the user. Apart from enriching the web search results [16], the above functionalities are also applicable in other domains as well, e.g. in patent search [17].

- Enhancing data by linking them with semantic knowledge bases:

  X-Link it is a fully configurable named entity extraction tool which can analyze the contents of a document, identify entities of interest, map them with URIs derived from a Knowledge Base (which is accessible through a SPARQL endpoint) and enrich them with semantic information (e.g. properties and related entities). It supports a plethora of documents; HTML pages, Microsoft Word and Powerpoint files (.doc, .docx, .ppt and .pptx), PDF files, and XML-based files (e.g. XML and RDF files). X-Link is fully configurable in terms of the supported categories of entities, the underlying Knowledge Bases and the way the system queries the knowledge bases.

- Linking distributed semantic knowledge bases:

  As there are many marine-related datasets available in various formats distributed in different locations/systems, there is the need for fetching and assembling pieces of information coming from these sources, for exploiting them and creating domain-specific applications. Then one can use them for constructing a domain-specific warehouse, offering thereby more complete browsing or query services. For this reason a process for constructing a marine-based warehouse has been defined [18] based on the requirements of the iMarine communities. An ontology called MarineTLO was defined and is used as the conceptual backbone [14]. The warehouse is created automatically by fetching information coming from different sources (the entire process is described also in [19]) while preserving their provenance (more than one graph spaces are used for this purpose). Furthermore several metrics have been defined for quantifying the connectivity of the resulted warehouse. This allows quantifying the value of the warehouse and eases the process of monitoring the warehouse after each reconstruction. The current MarineTLO-based warehouse contains information from several sources (namely FLOD, Ecoscope, FishBase, Worms and DBpedia)[1] and is exploited from several applications including xsearch-service, x-link and FactSheetGenerator[2].

- Extending the gCube Data Transformation Service:

  Exploiting the extensible design of gCube Data Transformation Service (gDTS) two new Data Sources have been created, adding value to system's interoperability capabilities. The one Data Source is a Tree Manager Data Source that is able to fetch objects (content and metadata) from Tree Manager Library and then transform their xml representation. The other is an HTTP Data Source able to ingest data coming in XML format through HTTP protocol. In addition, a prefetching mechanism has been added for those Data Sources in order to allow smooth operation under common over-the-internet use cases, avoiding producer's timeouts. Data can be efficiently consumed as they become available. A file-backed memory queue has been implemented for this reason.

  Tree Manager integration with gCube Data Transformation had as a result to support OAI, SPD and FIGIS Tree Manager plugins. All those plugins that publish information through Tree Manager now be indexed for Search exploitation.

  Interoperability was extended as a new client library for gDTS was developed. The new client library exposes a simplified, easy to use API, while hiding system specific complexity from the client side.

  Finally, the gDTS has been also enhanced with useful additions in the Data Transformation Programs library. Gnuplot and Graphviz visualisation libraries were integrated and wrapped as transformation programs. The former is mostly used to generate two and three dimensional plots of functions and data, while the latter is used for graph visualization.

- Environmental reconciliation of occurrences

  A web service for environmental reconciliation of occurrences has been developed. The occurrences are retrieved from iMarine occurrences services through a WPS service. The first

[1] http://www.fao.org/figis/flod/, http://www.ecoscopebc.ird.fr/EcoscopeKB/ShowWelcomePage.action, http://www.fishbase.org/, http://www.marinespecies.org/, http://dbpedia.org/

[2] http://www.ecoscopebc.ird.fr/

environmental enrichment use the environmental variables from the BioOracle dataset: e.g. salinity, sea surface temperature and chlorophyll. This dataset has been made available on the Hadoop resources feeding the WPS-Hadoop iMarine service.

On the other hand, Environmental reconciliation is also possible including time variable. An analysis of datasets that have the time dimension and used the Sea Surface Temperature Data from data.nodc.noaa.gov (NAVO-L4HR1m-GLOB-K10_SST), daily SST (level 4) served as OGC WMS, WCS and Thredds took place.

The occurrences can often be split into clusters. Data access from data.nodc.noaa.gov has been optimized by performing a DBSCAN clustering algorithm on the occurrences to identity the geographical areas of the clusters. Those areas are then being used to build the queries on the OGC WCS SST and occurrences are enriched with the SST time-series for the period 1990 to 2005.

- Provide Interactive Data Visualisation:

  The current infrastructure is able to manipulate huge amounts of geospatial information through its current GeoExplorer setup. A plugin has been developed to provide advanced interactive visualisations of the underlying geospatial timeseries data, utilizing geoexplorer's Web Map and Feature Services (WMS and WFS). User selections of arbitrary map regions initiate interactive visualisations of the region's underlying data in various forms, such as bubble plots and cumulative plots for quantitative properties (i.e species population) or timelines for time series data. Interactive time-based filtering of the visualised timeseries data is also supported for all of the visualisation types. The aforementioned functionality is offered by the "InteractiveLayers" component as a plugin for GCube's GeoExplorer and GisViewer components. An additional standalone component, under the name "QuatitativePropertiesVisualisation", was also developed to produce bubble chart visualisations of given data within various GWT widgets (windows, popups, panels, etc).

- Data fusion operator implementation:

  In order to provide global ranking in results that are retrieved from multiple datasources, the data fusion operator has been implemented and integrated into the search system. The data fusion operator merges and reranks the search results originating from different datasources and sorts them by their updated score, providing a better quality of results.

- Added distinct and orderby support to ForwardIndex:

  ForwardIndex has been extended to support “distinct” and “orderby” operators in order to remove the duplicates and sort the results by various aspects, respectively.

- gCQL parser extensions:

  gCQL Parser has been extended to support unicode characters as well as add the support for the data fusion reserved keyword in search queries.

- Keep software up-to-date:

Maintaining backwards compatibility is the target not only of Data Consumption Software, but gCube Software in general. However in the process of platform evolution, new components become available, providing advanced functionalities, while other components become obsolete. It is then that might be necessary for code to be aligned with the latest version of the new components. As for example, after Content Manager Library deprecation, many components were affected. All those dependant components were refactored in order to use another storage system, namely Storage Manager. Components that were released for this reason reside in Data Transformation, Indexing and Personalization subsystems.

1.4 COMPONENTS

In the target releases, the following components have been updated or newly introduced:

- Search System and FulltextIndex optimizations:
  - org.gcube.search.searchsystemservice.2-1-0
  - org.gcube.search.searchsystemservice-stubs.2-1-0
  - org.gcube.search.searchsystem.3-3-0
  - org.gcube.index-management.common-library.3-4-0
  - org.gcube.index-management.elasticsearch-gcube.1-2-0
  - org.gcube.index-management.fulltextindexnode.1-2-0
  - org.gcube.index-management.fulltextindexnode-stubs.1-2-0
  - org.gcube.index-management.fulltextindexnode-client-library.1-2-0

- Replacement of underlying Index mechanisms:
  - org.gcube.index-management.elasticsearch-gcube.1-0-0
  - org.gcube.index-management.fulltextindexnode.1-0-0
  - org.gcube.index-management.fulltextindexnode-stubs.1-0-0
  - org.gcube.index-management.couchbase-gcube.1-0-0
  - org.gcube.index-management.forwardindexnode.1-0-0
  - org.gcube.index-management.forwardindexnode-stubs.1-0-0

- Statistical Manager enhancements:
  - org.gcube.data-analysis.statistical-manager-gcubews.1.1.4
  - org.gcube.data-analysis.statistical-manager-cl.1.1.3
  - org.gcube.data-analysis.statistical-manager-stubs.1.1.4
  - org.gcube.data-analysis.ruserservice.1.1.0
  - org.gcube.data-analysis.ecological-engine-executor.1.4.0
  - org.gcube.data-analysis.ecological-engine.1.7.1
  - org.gcube.data-analysis.ecological-engine-geospatial-extensions.1.1.0

- Enriching gCube search system results with semantic information:
  - xsearch-service.1-1-1
  - org.gcube.portlets-user.xsearch-portlet.1-1-1

- Enhancing data by linking them with semantic knowledge bases:
    - gr.forth.ics.isl.x-link-1.0

- Extending the gCube Data Transformation Service:
    - org.gcube.data-transformation.data-transformation-handlers.2-5-1
    - org.gcube.data-transformation.data-transformation-programs.1-5-2
    - org.gcube.data-transformation.data-transformation-client-library.3-0-0

- Provide Interactive Data Visualisation:
    - org.gcube.application.interactivelayers.2-1-0
    - org.gcube.application.quantitativepropertiesvisualisation.1.1.0

- Data fusion operator implementation:
    - org.gcube.search.data-fusion.1-0-0
    - org.gcube.search.G_CQLParser.1-2-0

- Added distinct and orderby support to ForwardIndex:
    - org.gcube.index-management.couchbase-gcube.1-2-0
    - org.gcube.index-management.forwardindexnode.1-1-1
    - org.gcube.index-management.forwardindexnode-stubs.1-1-1
    - org.gcube.index-management.forwardindexnode-client-library.1-2-0

- gCQL parser extensions:
    - org.gcube.search.G_CQLParser.1-2-1

- Keep software up-to-date:
    - org.gcube.data-transformation.data-transformation-handlers.2-5-0
    - org.gcube.data-transformation.data-transformation-programs.1-6-0
    - org.gcube.data-transformation.data-transformation.2-2-4
    - org.gcube.personalisation.profileadministration.2-0-0
    - org.gcube.personalisation.userprofileaccess.3-0-0

1.5 DOCUMENTATION

A comprehensive overview of the subsystem(s) the described components belong to, is available at

- Milestone 41: Data Retrieval Facilities [3]
- Milestone 42: Data Manipulation Facilities [4]
- Milestone 42: Data Mining Facilities [5]
- Milestone 42: Data Visualisation Facilities [6]
- Milestone 42: Semantic Data Analysis [7]

Technical documentation covering all the aspects of the described software is available at:

- Admin’s Guide [8]

- Developer’s Guide [9]
- User’s Guide [10]

For development purposes, javadoc documentation for each component, along with a direct link to the associated section in Developer’s Guide, is available at [12].

Finally, the related publications in conferences [16][17][18][19] as well as [14] also provide useful documentation

1.6 DOWLOAD

The components described in this deliverable are available for download at [11].

Direct links to each component are available at [12].

REFERENCES

[1] gCube News
http://www.gcube-system.org/index.php?option=com_content&view=category&id=2&Itemid=6

[2] gCube Query Language
http://bscw.research-infrastructures.eu/bscw/bscw.cgi/240983

[3] Milestone 41: Data Retrieval Facilities:
https://gcube.wiki.gcube-system.org/gcube/index.php/Data_Retrieval_Facilities

[4] Milestone 42: Data Manipulation Facilities:
https://gcube.wiki.gcube-system.org/gcube/index.php/Data_Manipulation_Facilities

[5] Milestone 43: Data Mining Facilities:
https://gcube.wiki.gcube-system.org/gcube/index.php/Data_Mining_Facilities

[6] Milestone 44: Data Visualisation Facilities:
https://gcube.wiki.gcube-system.org/gcube/index.php/Data_Visualisation_Facilities

[7] Milestone 45: Semantic Data Analysis:
https://gcube.wiki.gcube-system.org/gcube/index.php/Semantic_Data_Analysis

[8] Administrator’s Guide:
https://gcube.wiki.gcube-system.org/gcube/index.php/Administrator%27s_Guide

[9] Developer’s Guide:
https://gcube.wiki.gcube-system.org/gcube/index.php/Developer%27s_Guide

[10] User’s Guide:
https://gcube.wiki.gcube-system.org/gcube/index.php/User%27s_Guide

[11] gCube Maven Repository RELEASES:
http://maven.research-infrastructures.eu/nexus/index.html#view-repositories;gcube-releases~browsestorage

[12] gCube Distribution Site:
http://www.gcube-system.org/index.php?option=com_distribution&view=distribution&itemid=23

[13] XSearch Component:
https://gcube.wiki.gcube-system.org/gcube/index.php/X-Search

[14] Marine Top Level Ontology:
http://wiki.i-marine.eu/index.php/Top_Level_Ontology

[15] XLink Library:
http://wiki.i-marine.eu/index.php/XSearchLink

[16] P. Fafalios and Y. Tzitzikas. X-ENS: Semantic Enrichment of Web Search Results at Real-Time. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (demo paper), SIGIR 2013, Dublin, Ireland.

[17] P. Fafalios, M. Salampasis and Y. Tzitzikas, Exploratory Patent Search with Faceted Search and Configurable Entity Mining. In Proceedings of the 1st International Workshop on Integrating IR technologies for Professional Search in conjunction with the 35th European Conference on Information Retrieval (ECIR’13), Moscow, Russia.

[18] Y. Tzitzikas, C. Allocca, C. Bekiari, Y. Marketakis, P. Fafalios and N. Minadakis, Ontology-based Integration of Heterogeneous and Distributed Information of the Marine Domain, ERCIM News vol. 96, Special Theme on Linked Open Data, January 2014.

[19] 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, November 2013.

[20] http://www.elasticsearch.org/

[21] http://www.couchbase.com/