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.	**D10.4**
Delivery Date	***10 2012***
Author	***Gerasimos Farantatos – NKUA***

***Abstract**: This document describes the novelties within the iMarine Data Consumption Software which were achieved from the 7th to the 12th 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.	D10.4
Deliverable Title	iMarine Data Consumption Software
Contractual Delivery Date	October 30th 2012
Actual Delivery Date	October 29th 2012
Author(s)	Alex Antoniadis – NKUA, Fabrice Brito – Terradue, Gianpaolo Coro – CNR, John Gerbesiotis – NKUA, Yannis Marketakis – FORTH
Editor(s)	Gerasimos Farantatos – NKUA
Reviewer(s)	Fabrice Brito – Terradue
Contributor(s)	
Work Package No.	WP 10
Work Package Title	Data Consumption Facilities Development
Work Package Leader	Gerasimos Farantatos – 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)	12
Keywords	Data Manipulation, Data Mining, Data Visualization, Ecological Niche Modeling, gCube, Information Retrieval, Semantic Data Analysis

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/).

GLOSSARY

ABBREVIATION	DEFINITION
iMarine	Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources
DTS	Data Transformation Service
EM	Ecological Modeling
ENM	Ecological Niche Modeling
GDAL	Geospatial Data Abstraction Library
NetCDF-CF	Network Common Data Form – Climate and Forecast metadata
PE2ng	gCube Process Execution Engine

DELIVERABLE SUMMARY

1.1 INTRODUCTION

iMarine Data Consumption software is the outcome of development activities taking 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;
- semantic data analysis

This deliverable describes the novelties within the iMarine Data Consumption Software from M7 (May '12) up to M12 (Oct. ’12).

1.2 TARGET RELEASE(S)

The releases connected to WP10 are:

- gCube 2.10.0 – September 2012
- gCube 2.11.0 – October 2012

Articles 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:

- Improve the flexibility of Resource Registry

  In order to make Resource Registry more flexible and reduce development costs, all pre and post-processing tasks within the logic of repository providers which should be independent to the latter, have moved to separate plug in modules. Plug in logic is executed in designated points within bridging cycles or triggered in a periodic fashion and is capable of being enabled or disabled easily via configuration. Various administrative or value adding features such as automatic field creation or data source management are also encapsulated in plugins.
  Plugins are packaged in two new software components, one containing environment specific plugins for the gCube environment and the other containing model independent plug-ins used in the gCube environment.
  Direct support for remote data stores has been added; remote data stores for read and update operations are supported in addition to the existing local ones. This enables the Resource Registry

to function in a variety of modes and environments with minimal work in terms of development. Also, new read and write policies are supported since gCube release 2.10.0. As far as the underlying repository provider provides such support, the Resource Registry provides the option to contact the remote data store directly for store and retrieve operations, instead of relying to the periodic bridging iterations. Repository providers are free to provide both or just one of the two modes of operation.
Furthermore, the data source model has been revised for greater flexibility. The model no longer makes any assumption regarding the binding of services to specific endpoints. This revision has given the ability to the model to represent all kinds of data sources: Purely stateless, replicated cluster-bound stateful and purely stateful. Old DataSources have been separated into the DataSource and DataSourceService notions. Each DataSource optionally points to a set of bound service endpoints

- Enhance Resource Registry to satisfy PE2ng requirements

  A number of enhancements have been made in the Resource Registry set of components. The Resource Registry elements for Execution Servers and Execution Services and hosting node adapter for node selection library have been added.

- Support XSearch features and improve the configuration capabilities of Resource Registry

  In order to support entity mining over content, the Resource Registry provides support for the field annotations with specific keywords by providing the corresponding queries and propagates annotations for presentable fields corresponding to data sources newly introduced into the Information Retrieval process. This functionality is useful for software such as XSearch (also released in the context of WP10 and described below in the present document), as it allows the latter to mark search fields as capable or needed to drive their features. Field annotations are included in a more general static configuration construct, which can be extended as needed in the future. The task of propagating annotations to new data sources is handled by a special plug-in developed for this reason.

- Support Dynamic Resource Registry configuration

  In gCube infrastructure, many components, such as Search, Index, DTS etc. depend on and use Resource Registry. Since every component may need a different configuration for Resource Registry as well as deployment on different environments, dynamic configuration for Resource Registry has been introduced, so that every component using it can have its own configuration component depending on its needs and loaded at runtime.

- Enhance Search System and adopt the guidelines of the project

  In updating and enhancing the gCube search service in order to be generalized and follow the project's guidelines the Java Logger has been replaced with slf4j in SearchSystem, OpenSearch Library and OpenSearch Data Source. This will facilitate the configuration of logging at deployment time and help overcome significant configuration obstacles presented by the standard Java logging environment. Also the following projects have been moved to Maven: Registry, SearchSystem, SearchSystemService, OpenSearch Data Source, Open Search Library and Index Common Library.

  SearchSystem and SearchSystemService have been updated in order to be aligned to new environmental hint based interface for resource registry and workflow adaptors.

- To integrate gCube Data Transformation Service with latest environment changes

  The gCube Data Transformation Service is part of Data Consumption Software. In order to increase its scalability and to reduce the response times experienced by the end-users, effort has been placed on the integration of the new node selection library which is part a library developed in the context of WP8 into Data Transformation Service Workflow Adaptor. In such a way, during the construction of a data transformation execution plan, the node selection library is responsible for finding the appropriate execution nodes that will be used. During the exploration of the available nodes, consideration has been taken into account for jobs instructing high resource needs, preserving locality when possible while distributing evenly to all available nodes. Moreover, Resource Registry is being used to further reduce the response times of plan creation. Information about all Execution Nodes that support data transformation are periodically harvested by Resource Registry and replicated to a local database, resulting in instant access to the information describing available nodes.

- To support the indexing of species data collections

  Species Products Discovery, a Tree Manager plug-in, has been integrated with Data Transformation to export out of the new source the desired rowsets that can be used for indexing procedure. For that reason, a new DataSource has been created, general enough for every Tree Manager plug-in, that will model the specific source along with the corresponding XSLT that will be used during transformation process.

- Provide Data Mining Facilities

  The activity on Data Mining and Manipulation Facilities started from an investigation on Ecological Niche Modeling (ENM) techniques. The term ENM refers to a set of methods that use algorithms for predicting the distribution of biological species in a geographical area. An overview of common techniques for biological data processing and maps production was made to understand the state-of-the-art of already implemented systems for ENM and Data Mining. Investigation was made both for interfaces to ENM systems and for libraries providing algorithms. Eventually a set of desktop tools and web interfaces (like Open Modeler and Yabi) were identified which allow users to perform experiments in several biological fields. On the other side a set of state-of-the-art libraries were integrated, including Rapid-Miner and Weka. These contain mathematical algorithms commonly used in computational biology. Other data mining techniques came from CNR previous experiences. The result of such analysis was the release of the gCube Ecological Engine library containing a common layer of mathematical techniques. On top of these techniques a set of methods was developed in the form of plug-ins performing ENM. The procedures were compared to literature well known algorithms in order to assess their effectiveness.
  The Ecological Engine library has then been extended in order to address Ecological Modeling (EM) problems. EM is a set of procedures for managing complex phenomena like to predict the impact of climate changes on biodiversity, prevent the spread of invasive species, identify geographical and ecological aspects of disease transmission, help in conservation planning, guide field surveys, among many other uses. A set of techniques have been developed in this direction by integrating algorithms and suggestions coming from the i-Marine Community of Practice.
  A first version of the Ecological Engine library has been released in gCube 2.10.0 containing two algorithms for Niche Modeling and 20 algorithms for biological data analysis and transformation. From M4 to M11 the library was endowed with capabilities for executing algorithms on the D4Science infrastructure nodes, by using a plug-in of the Executor gCube component. This functionality is included in gCube release 2.11.0.
  On top of the Ecological Engine Library, a gCube Service called Statistical Manager was built. This

gCube Service is responsible for (i) forwarding requests to the library, (ii) managing multi-users requests, (iii) monitoring the occupancy of the resources by the computations. The Statistical Manager currently adopts a distributed architecture for balancing the load due to local computations. It uses an Active Message Queue instance for dispatching messages. The Statistical Manager has been designed and implemented and is included in the gCube 2.11.0 release.

- Provide Data Visualization facilities for geospatial data

  TiffUploader algorithm is based on GDAL native library, a translator library for raster geospatial data formats that is released under an X/MIT style Open Source license by the Open Source Geospatial Foundation. As a library, it presents a single abstract data model to the calling application for all supported formats. It also comes with a variety of useful command line utilities for data translation and processing. The org.gcube.data-analysis.tiff-uploader.1-0-0 component uses the GDAL Java bindings. The library offers the access to a myriad of file formats supported by GDAL and thus extends the support beyond netCDF-CF. The component includes an example that takes a netCDF-CF (e.g. from MyOCEAN) that, after the download of the netCDF file on the local filesystem, the component splits it in as many GeoTIFF files as the number of data layers and publishes them in a new GeoServer Workspace just created from the netCDF file itself. Since org.gcube.data-analysis.tiff-uploader.1-0-0 is based on GDAL Java bind, it needs to be delivered with the right dependencies, according with the architecture that will run the application. For this purpose, multiple library artifacts where created, each for the most popular architectures: linux-i386, linux-x86_64 and mac-x86_64.

- Enhance the results derived from the gCube Search System.

  The gCube Search System is an integral part of Data Consumption Software. In order to enrich the results, a generic meta-search service has been developed, named xsearch-service [2]. xsearch – service provides advanced services for satisfying recall-oriented information needs and for semantically enriching the results. These services include: results clustering, named entity mining, semantic enrichment (by exploiting appropriate Knowledge Bases). It will be possible to apply these services over the entire answer returned by the underlying system or only over the top-K hits returned (in addition ability to analyze only the textual snippets or the full contents).

- Fix observed issues

  iMarine Data Consumption Software is actively maintained and tested in order to discover and fix bugs discovered during development and incidents occurring in the production environment. In particular, the following issues have been successfully resolved:
  - The initialization of the PE2ng environment at the level of the Search System Service.
  - A fix for *"any"* (simple search) queries at the level of Lucene, so that the *allIndexes* field which corresponds to this functionality is not erroneously included in the constructed Lucene query.
  - Fixes of issues in the entity persistence layer and in-memory cache of Resource Registry.
  - Fixes in XSearch portlet UI and its interface with the corresponding service.

1.4 COMPONENTS

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

- To improve the flexibility of Resource Registry
    - ResourceRegistry 1.3.1
    - RRModel 1.3.0
    - RRGCubeBridge 1.3.1
    - RRPlugins 1.0.0
    - RRGCubePlugins 1.0.0

- To enhance Resource Registry to satisfy PE2ng requirements
    - RRModel 1.3.0

- To support XSearch features and improve the configuration capabilities of Resource Registry
    - RRModel 1.3.0
    - RRGCubeBridge 1.3.1

- To support Dynamic Resource Registry configuration
    - ResourceRegistry-configuration-default 1.0.0
    - ResourceRegistry-configuration-dts 1.0.0
    - ResourceRegistry-configuration-execution 1.0.0
    - ResourceRegistry-configuration-index 1.0.0
    - ResourceRegistry-configuration-search 1.0.0
    - ResourceRegistry-configuration-workflow 1.0.0
    - ResourceRegistry-configuration-portal 1.0.0

- To enhance Search System and adopt the guidelines of the project
    - ResourceRegistry 1.3.1
    - RRGCubeBridge 1.3.1
    - RRModel 1.3.0
    - opensearch-library 1.6.0
    - opensearchdatasource 1.6.0
    - opensearchdatasource-stubs 1.6.0
    - searchsystem 3.1.0
    - searchsystemservice 2.0.2
    - searchsystemservice-stubs 2.0.2
    - index-management.common-library 3.3.3

- To integrate gCube Data Transformation Service with latest environment changes and to support indexing of species data collections
    - data-transformation-handlers 2.4.0
    - data-transformation-library 2.0.2
    - data-transformation-programs 1.4.0
    - data-transformation 2.2.2
    - WorkflowDTSAdaptor 1.0.2

- To provide Data Mining Facilities
    - ecological-engine 1.4.0
    - ecological-engine-executor 1.0.0
    - statistical-manager-cl 1.0.0
    - statistical-manager-cl 1.0.0
    - statistical-manager-stubs 1.0.0

- To provide Data Visualization facilities for geospatial data

  - tiff-uploader 1.0.0

- To enhance the results derived from the gCube Search System.
  - xsearch-service 1.0.1

- To fix observed issues
  - searchsystemservice 2.0.2
  - index-management.common-library 3.3.3
  - RRModel 1.3.0
  - RRGCubeBridge 1.3.1
  - xsearch-service 1.0.1

1.5 DOCUMENTATION

A comprehensive overview of the subsystem(s) the described components belong to is available at [3], [4], [5], [6] and [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]

Finally, 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].

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

[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