
| | |
|---|---|
| *Project Acronym* | ***iMarine*** |
| *Project Title* | ***Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources*** |
| *Project Number* | ***283644*** |
| *Deliverable Title* | ***Data Management Software*** |
| *Deliverable No.* | ***D9.2*** |
| *Delivery Date* | ***10 2012*** |
| *Author* | ***Gianpaolo Coro (CNR), Federico De Faveri (CNR), Lucio Lelii (CNR), Andrea Manzi (CERN), Nikolaos Drakopoulos (CERN), Valentina Marioli (CNR), Fabio Simeoni (FAO), Alexandros Antoniadis (NKUA)*** |

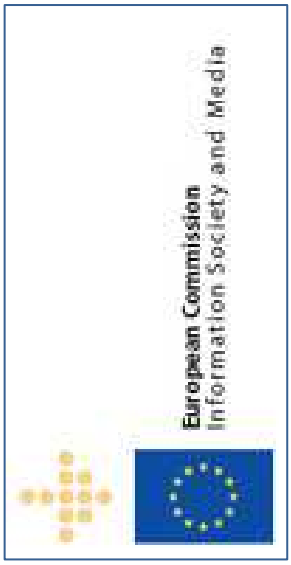
***Abstract***: *The iMarine Data Management Software comprises a number of components and subsystems offering facilities for accessing, transferring and harmonising a rich array of data typologies. This document briefly describes the novelties included in the iMarine Data Management Software from M7 (Apr. '12) to M11 (Sept. '12) and offers a series of links to the accompanying documentation.*

**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.** | D9.2 |
| **Deliverable Title** | iMarine Data Management Software |
| **Contractual Delivery Date** | 09 2012 |
| **Actual Delivery Date** | 10 2012 |
| **Author(s)** | Gianpaolo Coro (CNR), Federico De Faveri (CNR), Lucio Lelii (CNR), Andrea Manzi (CERN), Nikolaos Drakopoulos (CERN), Valentina Marioli (CNR), Fabio Simeoni (FAO), Alexandros Antoniadis (NKUA) |
| **Editor(s)** | Gianpaolo Coro (CNR) |
| **Reviewer(s)** | Julien Barde (IRD) |
| **Contributor(s)** | N/A |
| **Work Package No.** | WP 9 |
| **Work Package Title** | iMarine Data Management Facilities Development |
| **Work Package Leader** | Gianpaolo Coro (CNR) |
| **Work Package Participants** | CNR, CERN, NKUA, FAO, Terradue |
| **Estimated Person Months** | <15 |
| **Distribution** | Public |
| **Nature** | Other |
| **Version / Revision** | 1.0 |
| **Draft / Final** | Final |
| **Total No. Pages (including cover)** | 10 |
| **Keywords** | Data Transfer, Data Storage, Data Harmonization, Data Access |

# 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 |
| WCS | Web Coverage Service |
| WMS | Web Map Service |
| WFS | Web Feature Service |
| SOAP | Simple Object Access Protocol |
| HTTP | HyperText Transfer Protocol |
| FTP | File Transfer Protocol |
| GHN | gCube Hosting Node |
| SDMX | Statistical Data and Metadata Exchange |
| DARWIN CORE | A standard designed to facilitate the exchange of information about the geographic occurrence of species and the existence of specimens in collections |
| ETL | Extract Transform and Load |

# DELIVERABLE SUMMARY

## 1. INTRODUCTION

The iMarine Data Management Software comprises a number of components and subsystems offering facilities for accessing, transferring and harmonising a rich array of data typologies. This document describes the novelties within the iMarine Data Management Software from M7 (May.'12) to M11 (Sept.'12). It complements D9.1 [10] which describes iMarine Data Management Software up to M6 (Apr.'12). This deliverable is intended for documentation purposes only. The actual deliverable is represented by the software artifacts and their accompanying documentation.

## 2. TARGET RELEASE(S)

The iMarine Data Management Software release described by this document and thus realizing D9.2 contribute to the following release:

- gCube 2.9.1
- gCube 2.10.0

## 3. OBJECTIVES

The new version of components belonging the Data Management Software released as part of gCube 2.9.1 and 2.10.0 covers the following objectives:

- ***Data Access Facilities***:
  In this task activities mainly focused on data sets indexing and maintenance by means of automatic procedures, as explained in the previous deliverable [10]. In particular effort was spent on the Tree Manager Service, a gCube service holding smart representations of sets of numeric or textual data. The activity regarded only the implementation of such component, because the design phase had been managed in the previous months. A rich documentation was produced and several enhancements were added to the previous gCube Content Manager in order to realize this new component. Connections to a Semantic knowledge base (FLOD) were added for storing the results coming from SPARQL queries. The result of a SPARQL query is transformed into an internal representation of the Tree Manager and indexed. The aim is to publish common queries onto the infrastructure. The activity then moved towards the creation of an interface to the Tree Manager for exchanging contents with other services by means of the gRS2 gCube library. Finally, the Tree Manager was made compliant to the WP11 requirements. A Client Library interface was developed for such scope.
  On the other side Data Access was intended as the creation of a gateway towards datasets containing information about marine species. A gCube Service was created, containing connectors for the major providers of marine species information like GBIF, OBIS, Catalogue of Life, WORMS and ITIS. Such data sources are remotely hosted and maintained by their respective providers. A caching mechanism is used for storing only the results of a user's query. Results are saved locally to a GHN and refreshed once a week. The e-infrastructure doesn't have dumps or mirrors of remote

data sources. Data referring to a single query are downloaded on demand and stored for a week. The plug-in based architecture is expandable and can be applied to other kinds of species. A generic Darwin Core exporter was created for storing and exporting the results of a query. By means of this, a list of species occurrences is exportable in standard Darwin Core format, furthermore synonyms and common names are managed by cross-referencing the data sources. Finally, a query language interface was added in order to let external systems retrieve information as if it lied on a single database.

- ***Data Transfer Facilities***

  The Data Transfer Agent service and library defined in the previous deliverable [10] was released with enhancements and bug fixes. The Agent Service was firstly extended to support asynchronous transfer handling by means of a local database. This allowed transfer submission and asynchronous monitoring and retrieving of transfer's outcomes. Moreover the service implemented a Virtual File System over the GHN file system in order to better handle transfers having the local GHN as target. Performances and configurability were also enhanced. A new thread manager was implemented in order to perform parallel transfers. Thread pools and transfer timeouts are now configurable. The Client Library (CL) interface for the Service was enhanced as well, with asynchronous operations handling.

  Regarding the transfer scheduling activity a new service was implemented, named Data Transfer Scheduler. This Service manages three types of scheduling: direct, manual and periodic. The transfer logic is delegated to the Data Transfer Agents. Two libraries were included in the service, that is DB and Information System interfaces. The DB interface defines the database schema for the scheduler and abstracts over possible databases implementations, while the IS interface is responsible for contacting the gCube IS to retrieve information about (i) the Agent deployed on the infrastructure and (ii) the available Data Source and Storages.

  The first version of the service integrates the File Transfer logic while future versions will also implement Tree Based Transfers.

  Activity on the gRS2 library was also performed. The library was enhanced with fault transportation capabilities. The aim was to provide a way to developers for raising and catching exceptions at the transportation level and the APIs were then extended to support fault transportation capabilities. Exceptions are now thrown at the consumer's side as Unchecked Exceptions so it is consumer's responsibility to handle them. Furthermore, performances optimizations were done in order to improve the overall performance of the TCP transfer method via sophisticated serialization techniques and compression.

- ***Data Assessment, Harmonization and Certification Facilities***

  This task activity focused on datasets harmonization and assessment, on the basis of the work made during previous projects about fishery catch statistics. The gCube Time Series Service and Interface were studied and extended in order to become more generic. The service is now able to perform aggregations and database processing on tabular data representing time series of fishing catch statistics. The performed enhancements involved a preliminary investigation of data warehouse techniques for generic data import, harmonization and certification. The ETL framework was firstly studied and at the same time the Time Series environment was extended and fixed. Finally, a pure "gCube" approach was chosen as it was more flexible for being used by the many web interfaces of gCube. The underlying data model is based on a pure relational database schema.

Inputs can come in the format of SDMX or CSV and are then transformed into tables on an internal database. Export can be in the format of CSV files on the gCube workspace.
The activity moved towards the transformation of the Time Series environment into a more generic framework. This was called Tabular Data Widget and the base idea was to distribute it as a set of functionalities to perform aggregations or database processing on Tabular Data. A Widget solution was chosen for being rapidly integrated into other gCube user interfaces. The activity on this gCube component is still ongoing but the design phase was finished and the implementation was started. During M10, the activity focused on the assessment of tabular data representing occurrence points of species. A library was developed for "reconciliating" occurrences sets coming from different data sources. This activity involved the development of a probabilistic algebra, where two occurrences were declared to be similar on the basis of a complex distance calculation acting on spatial coordinates and metadata. The resulting service and interface will be officially released in M12.

## 4. COMPONENTS

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

- to support the ***Data Access Facilities***:
  - streams-2.0.0
  - trees-1.1.0
  - tree-manager-2.0.0
  - tree-manager-stubs-2.0.0
  - tree-manager-library-2.0.0
  - tree-manager-framework-2.0.0
  - tree-repository-2.0.0
  - data-access.obis-spd-plugin.1-1-0
  - brazilianflora-spd-plugin.1-1-0
  - catalogueoflife-spd-plugin.1-1-0
  - gbif-spd-plugin.1-1-0
  - itis-spd-plugin.1-1-0
  - irmng-spd-plugin.1-0-0
  - ncbi-spd-plugin.1-0-0
  - specieslink-spd-plugin.1-1-0
  - worms-spd-plugin.1-1-0
  - spd-client-library.1-0-0
  - spd-plugin-framework.1-0-0
  - species-products-discovery.1-1-0
  - species-products-discovery-stubs.1-1-0
  - spql-parser.1-0-0
  - org.gcube.content-management.storagelayer.1-3-2
  - org.gcube.content-management.storagelayer-servicearchive.1-3-2
  - org.gcube.content-management.content-management-library.1-7-0
  - org.gcube.content-management.content-management-library-servicearchive.1-7-0

- to support the ***Data Transfer Facilities***:

  - org.gcube.data-transfer.agent-library.1-1-0
  - org.gcube.data-transfer.agent-service.1-1-0
  - org.gcube.data-transfer.agent-stubs.1-1-0
  - org.gcube.data-transfer.common.1-0-1
  - org.gcube.data-transfer.scheduler-library.1-0-0
  - org.gcube.data-transfer.scheduler-service.1-0-0
  - org.gcube.data-transfer.scheduler-stubs.1-0-0
  - org.gcube.data-transfer.scheduler-is-interface.1-0-0
  - org.gcube.data-transfer.scheduler-db.1-0-0
  - org.gcube.execution.gRS2.2-0-0
  - org.gcube.execution.gRSBridge.1-2-1
  - org.gcube.execution.gRS2Broker.1-0-1

- to support the ***Data Assessment, Harmonization and Certification Facilities***:
  - Components will be officially released in production environment in the gCube 2.11 release

## 5. DOCUMENTATION

A detailed specification of the services documented by this report is in a number of dedicated Wiki pages [6][7][8]. Moreover, technical documentation covering all the aspects of the software is available at:

- Admin's Guide [3]
- Developer's Guide [4]
- User's Guide [5]

Finally, for development purpose, each component is provided with a Javadoc documentation along with a direct link to the associated section in Developer's Guide. This artifact is available at [9].

This documentation is an integral part of the actual deliverable.

## 6. DOWNLOAD

This document describes a deliverable of type "other". The actual deliverable consists of the software artifacts briefly discussed. Such artifacts are available for download via a Maven repository [1] or via the ETICS repository [2].

# REFERENCES

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

[2] gCube Etics Repository RELEASES:
https://grids16.eng.it/BuildReport/

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

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

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

[6] Milestone 37: Data Access and Storage Facilities:
https://gcube.wiki.gcube-system.org/gcube/index.php/Data_Access_and_Storage_Facilities

[7] Milestone 38: Data Transfer Facilities Specification:
https://gcube.wiki.gcube-system.org/gcube/index.php/Data_Transfer_Facilities

[8] Milestone 39: Data Assessment, Harmonisation, and Certification Facilities:
https://gcube.wiki.gcube-system.org/gcube/index.php/Data_Assessment,_Harmonisation,_and_Certification_Facilities

[9] gCube Distribution Site:
https://www.gcubesystem.org/index.php?option=com_distribution&view=distribution&Itemid=23

[10] G. Coro, F. De Faveri, L. Lelii. *iMarine Data Management Software*. iMarine D9.1 Project Deliverable. June 2012