

[www.ecomanage.info](http://www.ecomanage.info)

e-mail: [coordination@ecomanage.info](mailto:coordination@ecomanage.info)

## Deliverable 4.7

Final report on: Data exchange criteria; spatial harmonization of modelling work; data formats

ecomanage  
INTEGRATED ECOLOGICAL COASTAL  
ZONE MANAGEMENT SYSTEM



Ramiro Neves, PhD.  
MARETEC — Marine & Environment Technology Center  
Instituto Superior Técnico  
Secção de Ambiente e Energia — Dep. De Mecânica

Avenida Rovisco Pais  
1049 - 001 Lisboa PORTUGAL  
Tel: (+351) 218 417 397  
Fax: (+351) 218 419 423



| DOCUMENTATION FORM  |                                 |   |
|---|---------------------------------|---|
| <b>DISSEMINATION LEVEL</b><br>PU  | <b>DISTRIBUTION</b><br>Partners | <b>OBSERVATIONS</b>   |
| <b>TITLE</b><br>Deliverable 4.7: Final report on: Data exchange criteria; spatial harmonization of modelling work; data formats   |                                 |   |
| <b>KEYWORDS</b><br>Data organization, metadata base, formats  |                                 |   |
| <b>ABSTRACT</b><br>This report starts the discussion on the definitions for data organization, exchange and retrieval   |                                 |   |
| <b>TASK LEADER</b><br>IST - Instituto Superior Técnico<br>Secção de Ambiente e Energia - Dep. de Mecânica<br>Av. Manuel da Maia n° 36 3ºEsq., 1000-201 Lisboa<br>PORTUGAL |                                 |   |
| <b>Funding</b><br>This project received research funding from European Commission's Six Framework Programme – Contract n° INCO-CT-2004-003715 (Dec2004-Nov2007)           |                                 |  |
| <b>AUTHOR(S)</b><br>Pedro Pina  |                                 |   |
| <b>VERIFICAÇÃO</b><br>Coordination  |                                 |   |
| <b>DATE</b><br>9/06/2005  | <b>NUMBER OF PAGES</b><br>18    | <b>REFERENCE NUMBER</b>   |



## Table of Contents

|   |           |
|---|-----------|
| <b>TABLE OF CONTENTS</b> .....                          | <b>3</b>  |
| <b>INTRODUCTION</b> .....                               | <b>4</b>  |
| <b>DATA EXCHANGE CRITERIA</b> .....                     | <b>5</b>  |
| <b>DEVELOPMENT OF A REMOTE DATA ACCESS SYSTEM</b> ..... | <b>6</b>  |
| DATA SYSTEM GENERAL STRUCTURE .....                     | 8         |
| <b>CURRENT STATE OF DEVELOPMENT</b> .....               | <b>18</b> |
| ACCESS TO DATASETS .....                                | 20        |



## Introduction

Information systems help us to manage what we know, by making it easy to organize and store, access and retrieve, manipulate and synthesize, and, finally, to apply data to the solution of problems. Ecomanager has the objective of making all data and tools relevant to the project available for every partner and to interested stakeholders in an efficient way. The success of this task is directly related to the definition and implementation of a sustainable data structure that is able to incorporate and relate all different types of data and also taking advantage of all recent WEB developments for dealing with the geographic dispersion associated to the project. This problem are not new in the framework of European marine research in fact structuring projects such as MERSEA (<http://www.mersea.eu.org/>) and SeaSearch ([www.sea-search.net](http://www.sea-search.net)) are committed in the development of efficient data management systems. These projects are serving and communicating with different thematic communities. So, the challenge is to find a balance and to merge the practices and experiences of these different thematic communities into a coherent set of data products that will take into account specific features of the products and provide simultaneously interoperability to these surrounding communities and networks. A common unified framework for product coherency and standardization as well as data transport and exchange procedures has been set to provide a coherent set of dissemination services with the ability to access the data in an interoperable manner from client applications, relying on a decentralized but compatible system architecture for distribution on Internet. EcoManager project doesn't aim to be a GMES service provider like MERSEA, neither have the data management complexity associated to the thematic and geographical coverage of that project, nevertheless there is the perception that, in order to growth in the future, EcoManager needs a well structured approach to the data management problem.



## Data Exchange Criteria

It is important to understand how information is being processed. To simplify our evaluation, we will distinguish three functional groups: (i) Stakeholders (ii) Data Owners (iii) Modelers

- (i) Stakeholders are individuals with an interest in the success of project in delivering intended results and maintaining the viability of the project's products and services. Stakeholders influence programs, products, and services;
- (ii) Data Owners are individuals that provide relevant data sets and information to the project;
- (iii) Modelers are individuals that use all available tools to study the necessary processes in order to provide answers to Stakeholders needs using all available data provided by Data Owners.

Each partner of the project, and also research groups, public institutes or private companies not directly funded but involved in the project, may be present in one or more of these functional groups and its involvement can change over time.

### 1st Stage

In the first stage, that coincided more or less with the first 6 months of the project, Stakeholders received general information concerning project objectives and deliverables then they presented their needs and expectations on the project and in some cases they delivered also specific information (mostly in situ data) relevant to the project, acting then as Data Owners. Modelers from different research groups exchanged information regarding the fact that some elements have greater knowledge on processes and others have more experience on the numerical tools available. Specific Data Owners were contacted in order to gain access to their data sets.

### 2nd Stage



On the second stage, which is more or finishing at present time modelers have overcome all initial difficulties and all numerical model applications are working and delivering results. A validation stage already began, supported by Data Owners. Stakeholders are being kept informed of all developments. In this stage the effort on the development and implementation of indexes is more intense and needs contribution from all 3 functional groups.

### **3rd Stage**

The third and final stage will be dedicated to present information to stakeholders and general public. This stage will be supported by the results of the previous stages and it is assumed that models are well developed and that all data and knowledge are documented, catalogued and available.

The purpose of this simple exercise is to focus on the development of the data structure that should support all these data fluxes between functional groups and the general public. The proposed structure is described on the next topic.

## **Development of a Remote Data Access System**

In general, data are available in digital form, but not always in possession of the organisation which needs these data and information and most of the time these data are not available in the required format. Moreover, 1) there is not always appropriate overview of the available data including of the source holders, and 2) these data are in general not (remote) accessible from a "central focus point". A direct and uniform access to external regional data (bases), based on Web technologies, is of great importance.

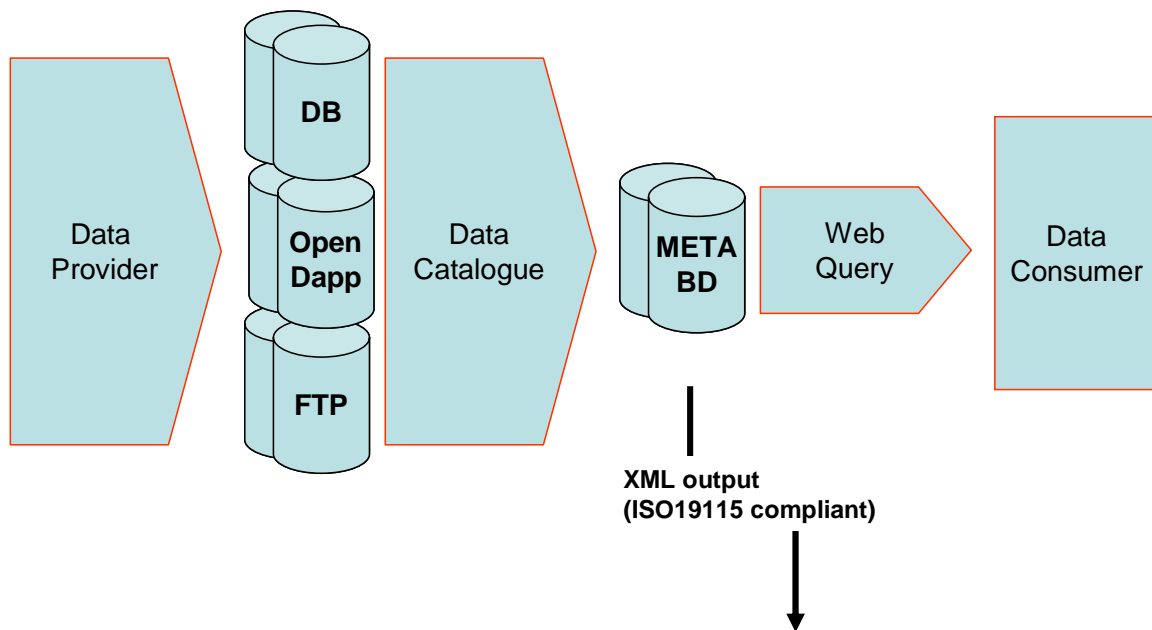
The aim is to develop a Remote Data Access System, providing uniform query facilities, to search on database catalogues within the EcoManage project where these databases are geographically distributed over different data providers.

A uniform query interface makes it possible for scientists, governmental decision makers, managers and administrators, industrial enterprises and the general public to search for data, based on the data type, geographical location and time together with other relevant parameters, as well as to retrieve data remotely. The Remote Access System is being developed to use the capabilities of the Internet (WWW Browser, HTTP Servers).



## Data system general structure

The data system can be structured in the following manner:



**Figure 1 – Scheme for data management system**

Data providers will produce datasets and stored them on FTP, typical relational databases (DB) or OpenDap (see description bellow), and then there is a need to create a metadata record characterizing the produced dataset. IST as developed a web application to help this task based on the CDI metadata index philosophy (see description bellow). A query interface will be available to data consumers via WEB with geospatial, feature and temporal query capabilities. It is important to enhance the fact that it will be the data providers decision to give full access to the dataset or not, thus accordingly to access policy defined by the data provider, the results of the queries will be links to ftp, opendap or relational databases containing the data providers datasets or the contact (phone, mail or fax) to the entity who controls those datasets.



## CDI

The Common Data Index (CDI) is a development, initiated by Sea-Search, a EU VIFP project (<http://www.sea-search.net/>) which will be adapted to fulfill Ecomanager's data management needs. Its purpose is to enable users to get highly detailed insight in the availability and geographical spreading of data across the different Ecomanager partners and other data providers for the Ecomanager project. It will cover all types of environmental datasets available under Ecomanager, but can easily be expanded to contain other datasets. The CDI thus provides an index (metadatabase) to individual datasets. These data sets can be physically located at partner databases and the CDI will provide a data link to them.

The CDI principle is that each participating data centre produces at regular intervals up-to-date CDI metarecords, giving an index overview of the content and coverage of its databases. These partner contributions are centrally collated in a central CDI metadatabase, located at IST Lisbon, which is equipped with a CDI user interface to serve users. For purposes of standardization and international exchange it was decided to adopt the ISO19115 metadata standard and to prepare the CDI metadata as a dedicated subset of this standard to make it ISO compliant. Therefore the CDI metadata format has been translated into a CDI XML format, because this supports the interoperability with other systems and networks. The ISO19115 schema provides the basis and is used as reference model. Within the framework of Ecomanager a pilot operational CDI system is developed, with participation of a selected number of Ecomanager data centres.

### Logical description of CDI metadata format

The definition of a data set is arbitrary for each partner, because of the differences in granularity that partners are applying in their archive systems for storing and accessing data sets. However a number of examples are given to illustrate the objectives and focus of the CDI.

#### Case: CTD measurements

CTD casts are collected at different geographical locations, e.g. during a scientific cruise. Each cast is represented by a data file / data set, which is stored and can be reproduced by a Data Centre. The CDI record reflects the metadata of a single CTD cast, covering multiple



parameters, and including the information for accessing this specific CTD data set or getting a copy of this specific CTD data set. The same approach can be applied for other in-situ and discrete measurements, such as sediment grabs, geological cores, water bottles, etc.

#### Case: Sea level / wave / current observations

Hydrodynamical observations are collected at different geographical locations, and might be part of a station, equipped with a number of instruments. Each instrument produces a timeseries of observation data, which is stored and can be reproduced by a Data Centre. The CDI record reflects the metadata of the resulting data set of a single instrument at a single station, covering multiple parameters, and including the information for accessing this specific data set or getting a copy of this specific data set. This can cover a long timeseries. However in case there are large gaps in time coverage, the Data Centre might have decided to split the data set into a sub series of data sets, each covering a consistent observation period. In that case each sub serie is represented by a separate CDI record.

#### Case: Hydrographic measurements

The bathymetry of the seabed is measured by hydrographic surveys, which cover an specific area. Each survey can comprise a consistent timeperiod in which the area is sampled by sailing a number of tracks, during which the seabed bathymetry is recorded in singular tracks or zones by specific instruments. Each instrument during a specific survey produces a hydrographic survey data set, which is stored and can be reproduced by a Data Centre. The CDI record reflects the metadata of the resulting data set of a single instrument during a single area survey, including the information for accessing this specific data set or getting a copy of this specific data set. The same approach can be applied for other area measurements, such as seismic surveys, satellite images, etc.

#### Case: Model Results

Results datasets are obtained by running different models, which cover a specific area and time period. Each dataset usually comprises a 3D or 2D spatial grid and several time instants within the specific time period. Each z spatial layer can contain data for several geophysical parameters. The CDI record reflects the metadata of the resulting dataset of a specific area and time period, including the information for accessing directly this specific data set (e.g. via OpenDap) or getting a copy of this specific data set (e.g asking by email). The same approach can be applied for other area type measurements, such as satellite images, etc.

The CDI is to give answers to the following basic questions:

- Where?
- When?
- What?
- How?
- Who?
- Where to find data?
- Other relevant information?

These basic questions are covered by specifying the following fields:

WHERE?

| • FIELD            | • COMMENTS   |
|--------------------|--|
| • <i>Latitude1</i> | Geographical coordinate (Mercator projection).   |
| <i>Longitude1</i>  | Geographical coordinate (Mercator projection):   |
| <i>Latitude2</i>   | Geographical coordinate (Mercator projection). Filled in case observation data were collected along a track or over an area. |
| <i>Longitude2</i>  | Geographical coordinate (Mercator projection). Filled in case observation data were collected along a track or over an area. |



|  |  |
|--|--|
|  |  |
| <ul style="list-style-type: none"> <li><i>Measuring area type</i></li> </ul> | <ul style="list-style-type: none"> <li>Point, Track or Area observation. Supported by controlled list (library)</li> </ul>   |
| <i>Datum system</i> <i>Coordinate</i>  | <ul style="list-style-type: none"> <li>CDI uses Geographical coordinates, preferably with Datum = WGS84. Other Datums might be used. Supported by controlled list (library)</li> </ul> |
| <i>Water depth</i>   | Water depth at location of observation.  |
| <i>Vertical datum</i>  | Vertical reference of water depth  |
| <i>Minimum instrument depth</i>  | Minimum depth of instruments collecting data   |
| <i>Maximum instrument depth</i>  | Maximum depth of instruments collecting data   |
| <i>Unit of Min and Max depth</i>   | Default specified as 'metre'   |

WHEN?

| • FIELD                | • COMMENTS  |
|------------------------|---|
| <i>Start date</i>      | Start date of the observation data                              |
| <i>Start time (UT)</i> | Start time of the observation data                              |
| <i>End date</i>        | End date. Filled in case of timeseries / repeated observations  |
| <i>End time (UT)</i>   | End time. Filled in case of timeseries / repeated observations. |



|                          |   |
|--------------------------|---|
| <i>Sampling Interval</i> | Temporal resolution in case of timeseries / repeated observations. Supported by controlled list (library) |
|--------------------------|---|

WHAT?

| • FIELD                    | • COMMENTS   |
|----------------------------|--|
| <i>Parameters measured</i> | For the purpose of the CDI a dedicated CDI Parameter Discovery Vocabulary (PDV) has been defined in a cooperation between BODC and IFREMER, which is now maintained by BODC. The CDI PDV contains a hierarchical structure to support the CDI User Interface: from Discipline => Agreed Parameter Groups => BODC Parameter Groups => BODC Parameter Dictionary. XML coding of parameters is done by including BODC Parameter Group codes. These are supported by a controlled list (library). Multiple codes can be entered to characterize the dataset. |
| <i>Abstract</i>            | Short description of dataset, if available. Default: Not specified   |

HOW?

| • FIELD   | • COMMENTS                               |
|---|--|
| <i>Instrument or gear type used to collect the data</i>             | Supported by controlled list (library)   |
| <i>Type of platform on which the sensors providing data for the</i> | • Supported by controlled list (library) |



|                            |  |
|----------------------------|--|
| <i>series were mounted</i> |  |
|----------------------------|--|

WHO?

| • FIELD             | • COMMENTS  |
|---------------------|---|
| • <i>Originator</i> | • This specifies the name, full address and profile of the organisation, that is originator of the data set. It is supported by a controlled list (library), which contains all the data holding centres, research institutes and monitoring organisations, that are active in Europe in the field of marine & ocean data and research. |



WHERE TO FIND DATA?

| • FIELD                                    | • COMMENTS   |
|--|--|
| <i>CDI Partner</i>                         | This specifies the name, full address and profile of the organisation, that is holding the data set and implicitly has produced this CDI record. It is supported by controlled list (library) of CDI partners, which is a subset of the overall controlled list of all the data holding centres, research institutes and monitoring organisations, that are active in Europe in the field of marine & ocean data and research. |
| <i>Data reference</i>                      | <ul style="list-style-type: none"> <li>• Unique reference code OR query string, serving a unique identification of the CDI dataset record by the CDI partner in its database.</li> </ul>   |
| <i>Database reference</i>                  | <ul style="list-style-type: none"> <li>• Identification of the database holding the dataset record at the CDI partner.</li> </ul>  |
| <i>Name by which the data set is known</i> | <ul style="list-style-type: none"> <li>• Data set name, if available. Default: Not specified.</li> </ul>   |
| <i>Short name of data set</i>              | <ul style="list-style-type: none"> <li>• Short name or acronym, if available. Default: Not specified.</li> </ul>   |
| <i>Access/ordering of data</i>             | <ul style="list-style-type: none"> <li>• The CDI provides the necessary information to enable a user to retrieve the identified data. Depending on the present state of the data web services of the CDI partner a number of situations might be applicable. This is supported by a controlled list (library).</li> </ul>  |
| <i>Internet access/ordering</i>            | The Direct 'Data URL' / registration URL / webshop URL, if applicable. See also above.   |
| <i>E-mail data contact</i>                 | The e-mail address for data requests, if applicable.   |
| <i>Restrictions on access to the data</i>  | Supported by controlled list (library).  |



OTHER RELEVANT INFORMATION?

| • FIELD                   | • COMMENTS   |
|---------------------------|--|
| • Creation Date           | Date that the data set was created.  |
| • CDI creation date       | Date that the CDI metadata was created (date stamp during generation)  |
| CDI Metadata originator   | This specifies the name, full address and profile of the organisation, that has produced this CDI record. It is supported by controlled list (library) of CDI partners, which is a subset of the overall controlled list of all the data holding centres, research institutes and monitoring organisations, that are active in Europe in the field of marine & ocean data and research. In CDI practice it is the same organisation as used for the field 'CDI partner' in 'WHERE TO FIND DATA'. |
| <i>Project name</i>       | If available and relevant  |
| <i>Station name</i>       | If available and relevant  |
| <i>Station ID</i>         | Alternative station name / number  |
| <i>Station Start date</i> | Start date of station observations   |
| <i>Cruise name</i>        | If available and relevant  |
| <i>Cruise ID</i>          | Alternative cruise name / number   |
| <i>Cruise Start Date</i>  | Start date of cruise   |
| • <i>Data size</i>        | Estimated size of the transferred data expressed in Megabytes. Can be null.  |
| <i>Data format</i>        | Format name of the transferred data.   |



## **OPeNDAP**

OpenDAP" stands for "Open-source Project for a Network Data Access Protocol". It is a middleware (XML native) that provides uniform access to scientific data on the Internet (via a URL). OpenDAP is a method of passing data over the Internet wrapped in self-describing metadata independent of file format. OpenDAP is not a data catalogue, it is associated with data stored as netCDF files. The protocol may be thought of as XML-tagged data descriptions (metadata) with associated non-sparsed (binary data) content. The data transport protocol relies on HTTP for requesting and providing data across the World Wide Web. OpenDAP also supports server side subsetting of data (and aggregation). It has been designed to minimize the barriers to sharing data over the Internet (it benefits for that of the netcdf structure and XML configuration information and greatly reduces the volumes of data that needs to be transferred across the Internet). The goal is to allow end users, whoever they may be, to access immediately whatever data they require in a form they can use, all while using applications they already possess and are familiar with.



## Current state of development

On the early stage of the project, Ecomana coordination team developed an ftp site to work as a central data archive available to all partners, where everyone could deliver and retrieve data. This ftp site was only accessible with username and password given to each partner. It had a backup system to avoid any problems regarding data lost. The ftp site was organized in different areas:

- Deliverables – where all the deliverables defined in the project work plan are archived
- Meetings – regarding all information related to project meetings (kick off, workshops)
- Modelling – where models applications, model software and documentation for using the modeling tools is stored.
- Scientific Information – where relevant bibliography concerning the project study areas is assembled.

The ftp site served its purposes and continues to serve regarding management of the documentation produced by the project. The Ftp site is not adequate for managing geophysical data produced and gathered in the framework of the project, thus the IST team develop a metadata web database inspired in the CDI philosophy (described above) for cataloguing all datasets.

A web application has been developed by the IST team (Figure 2) for helping data providers producing metadata records. This application also provides different ways, by data provider, by feature, geospatial or/and temporal, of querying the metadata and accessing the link to the physical location of the datasets.

The functional requirements of the web application are :

- The data provider can register its products.
- The products descriptions are managed in a standardized and consistent catalogue.



- The product description can be requested as XML ISO19115 standard metadata.

The user can search products matching one or several of the following criteria :

- Product short name.
- Type (among forcing fields, in-situ or remote sensing dataset or ocean forecast).
- Parameters (category and name).
- Geographical scale (among global, regional or local).
- Geographical extent.
- Time window.

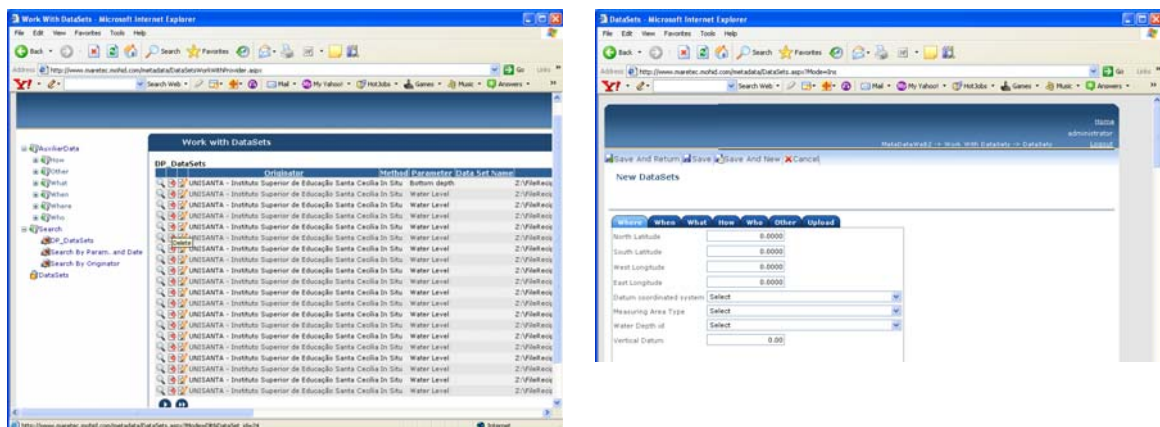


Figure 2 – Web interface for the metadata catalogue



## Access to DataSets

Currently when a metadata record is created, its correspondent dataset is archived on Ecomana**ge**'s FTP site. This means that any data user who wishes to view or work with available dataset will need to download it from server and then access it on the client machine.

### OpenDap Server

The IST team is currently implementing solutions to allow a more efficient access to data via web. Concerning grid data (models, remote sensing, etc), IST implemented an OpenDap server (described above) and a set of tools to convert MOHID HDF5 format to NETCDF OpenDap compatible. The OpenDap server is currently under test but already working. IST gathered all necessary information for implementation of this tool to help other partner to implement their one.

### Point and Line Data

Tabular data (in situ data and time series) is archived in a Post Gis + PostGreSQL data system. PostGIS adds support for geographic objects to the PostgreSQL object-relational database. In effect, PostGIS "spatially enables" the PostgreSQL server, allowing it to be used as a backend spatial database for geographic information systems (GIS), much like ESRI's SDE or Oracle's Spatial extension. PostgreSQL is an object-relational database management system (ORDBMS) based on POSTGRES, Version 4.2, developed at the University of California at Berkeley Computer Science Department. POSTGRES pioneered many concepts that only became available in some commercial database systems much later.. It supports SQL92 and SQL99 and offers many modern features:

- complex queries
- views
- transactional integrity
- multiversion concurrency control



Additionally, PostgreSQL can be extended by the user in many ways and because of the liberal license, PostgreSQL can be used, modified, and distributed by everyone free of charge for any purpose, be it private, commercial, or academic.

PostGis Class is the class where we program all functions that establish the interactions with data base to obtain, add, remove or change data from PostGis. The order for these actions is originated in the Web Gis client interface.

A Gis (Web) Client to allow mapping of data is supported by Map Server. The MapServer system includes MapScript that allows popular scripting languages such as PHP, Perl, Python, and Java to access the MapServer C API. MapScript provides a rich environment for developing applications that integrate disparate data. MapServer provides the core functionality to support a wide variety of web applications. Beyond browsing GIS data, MapServer is able to create "geographic image maps", that is, maps that can direct users to data tables with attributes that characterise the selected geographic feature. MapServer was originally developed at the University of Minnesota (UMN) through the NASA-sponsored ForNet project, a cooperative effort with the Minnesota Department of Natural Resources. Continued support has been provided through the NASA TerraSIP project, involving UMN and a consortium of land management interests. The software is developed and maintained by an increasing number of developers (nearing 20) from around the world and is supported by a diverse group of organizations funding enhancements.

## Conclusion

The main elements of Ecomanafe's data management system: metadata catalogue, OpenDap (for grid data), PostGis data system (for insitu and time series data) were developed and are currently under testing for consolidation. In the next months all gathered data will be archived in this system and connections will be made between the elements.