



Data e-Infrastructure Initiative for Fisheries management and Conservation of Marine Living Resources

## The iMarine Data Bonanza

Improving Data Discovery and Management through  
an **Hybrid Data Infrastructure**

Donatella Castelli, **Pasquale Pagano**, Leonardo Candela, Gianpaolo Coro

iMarine Technical Director

[pasquale.pagano@isti.cnr.it](mailto:pasquale.pagano@isti.cnr.it)

ISTI – CNR (Italy)

**i-marine.eu supports** the principles of the Ecosystem Approach (EA) to fishery management and conservation of marine living resources

- EA is a strategy for the integrated management of land, water and living resources that promotes conservation and sustainable use in an equitable way

**i-marine.eu operates** an Hybrid Data Infrastructure (HDI) offering access to a rich array of marine-related data and products via tailored environment

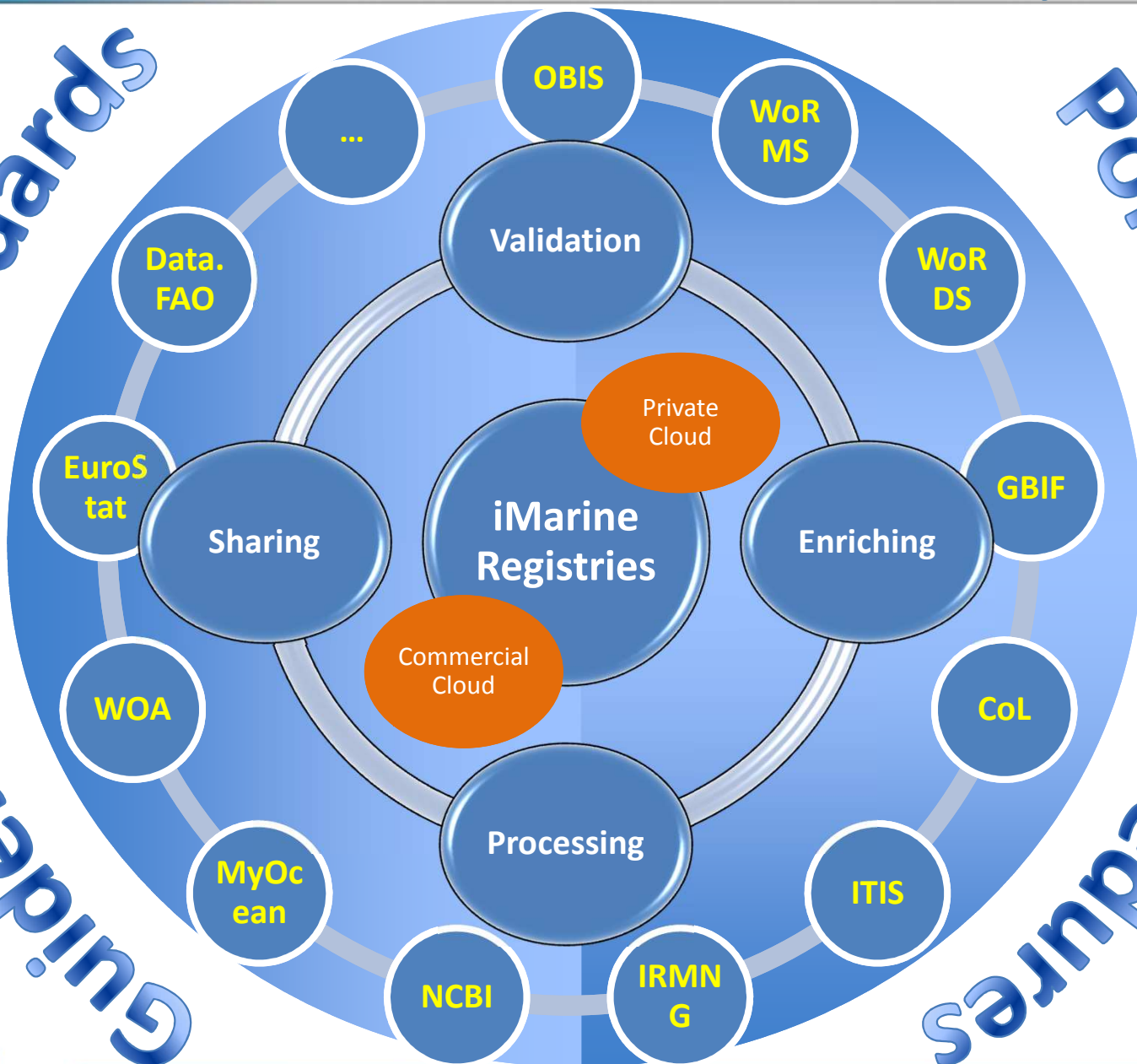
- HDI implements a data-management-capability delivery model in which computing, storage, data and software are made available as a utility (*as-a-Service*)

Standards

Guidelines

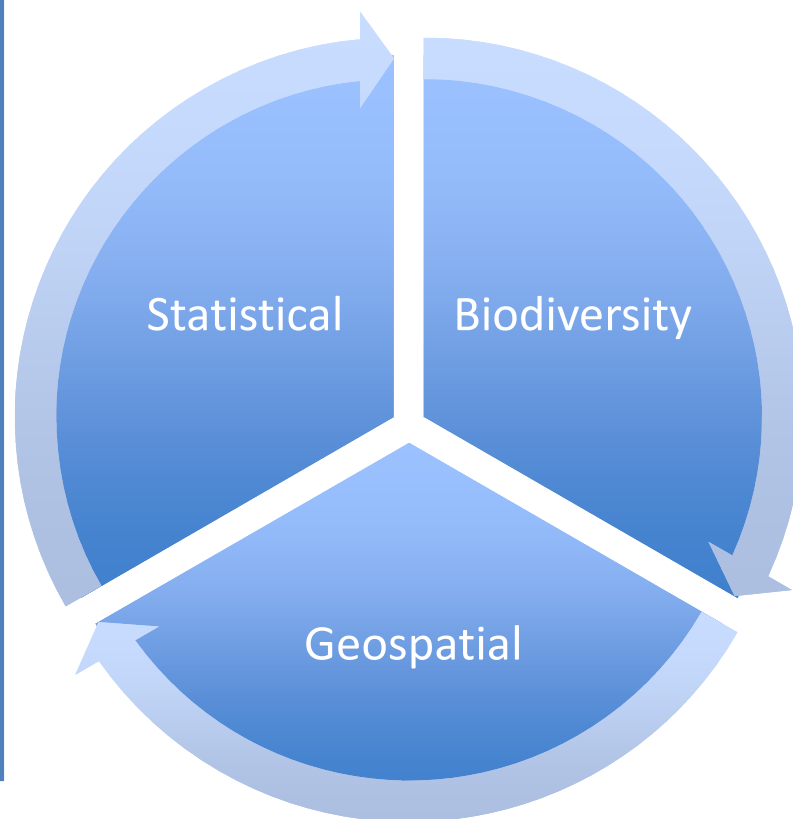
Policies

Procedures



## SDMX \*

- FAO CodeLists
- IRD CodeLists
- FAO Global Aquaculture Production
- FAO Global Capture Production
- FAO Global Production
- Eurostat
- ...



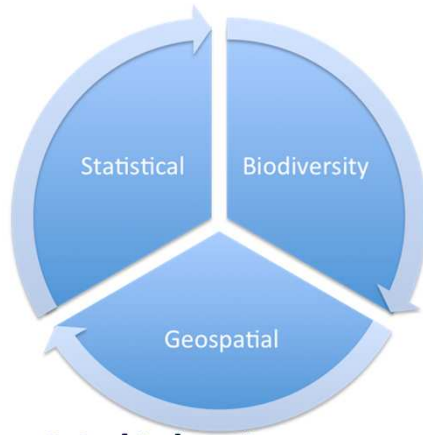
## DarwinCore / ISO19139

- >35 M Observations (OBIS)
- ≈ 120 K Observed Species (OBIS)
- ≈ 500 K Taxa (WoRMS)
- >600 K Scientific Names (ITIS)
- >12 K Species Distribution Maps (AquaMaps)
- ≈ 600 Species Extent (FAO)
- ... FishBase, SeaLifeBase
- ... CoL, GBIF

## ISO19139 (OGC W\*S)

> 300  
variables

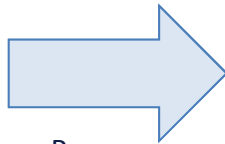
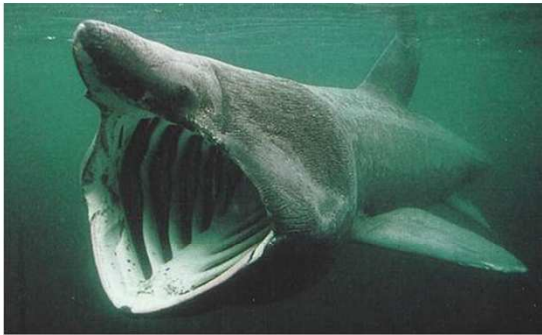
- 10 years Chemical and Physical variables in 2D space
  - Ice concentration and velocity, Chlorophyll, Oxygen, Nitrate, Phosphate, Phytoplankton as carbon, Salinity, Temperature, ...
- On-demand Chemical and Physical variables in 3D space
  - Apparent Oxygen Utilization, Dissolved Oxygen, Salinity, Temperature, ...



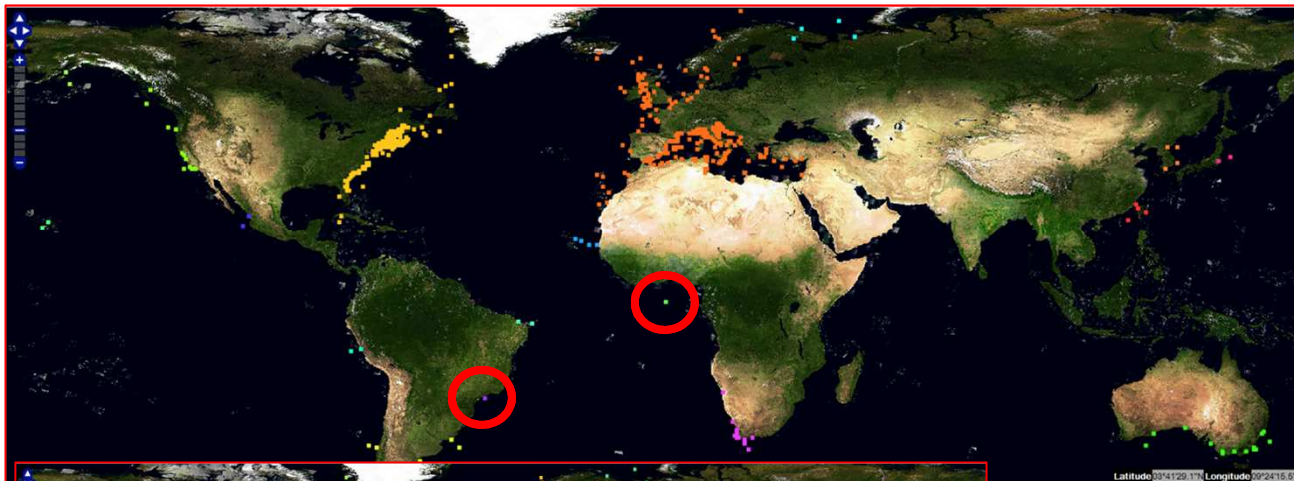
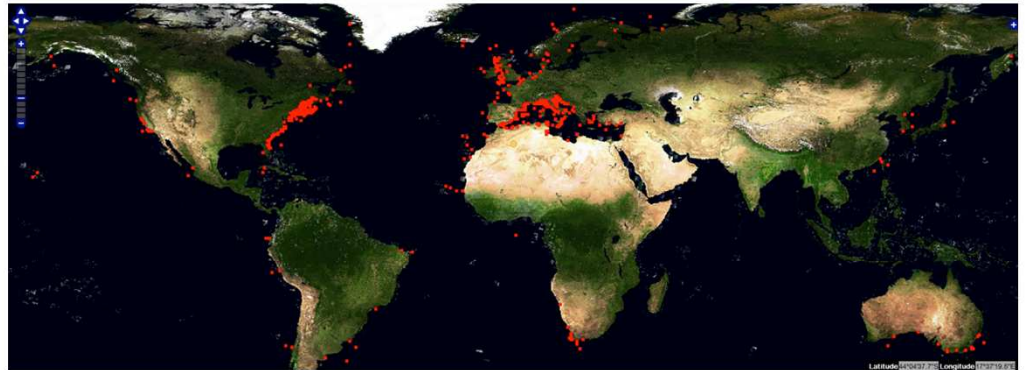
- Access
  - Retrieval of geospatial data as space/time-varying phenomena
  - Direct fine-grained access to feature and feature property level.
- Validation
  - User-defined quality and dissemination level
- Enriching
  - Generation metadata, exploitation of reference data, linking to environmental dataset
- Processing
  - Analysis and mining exploiting e.g. R, Weka and RapidMiner statistical frameworks
- Sharing
  - User-driven process to decide how other agents (human / machine) can access information



# Features Clustering with StatsCube

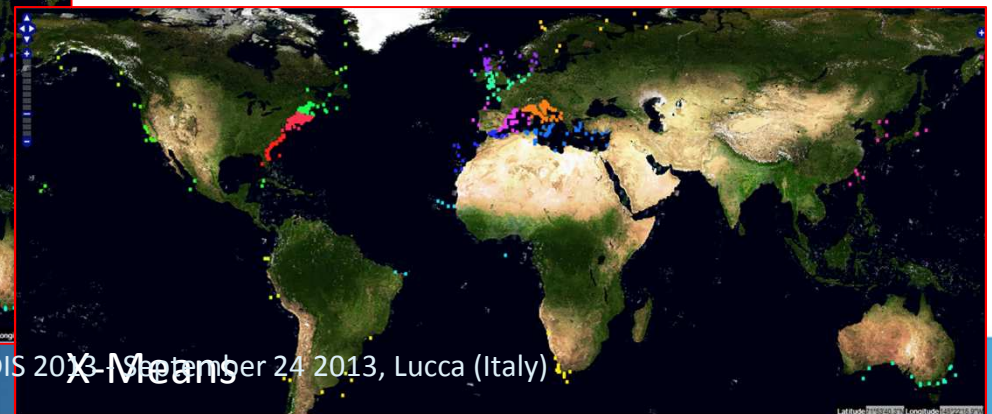
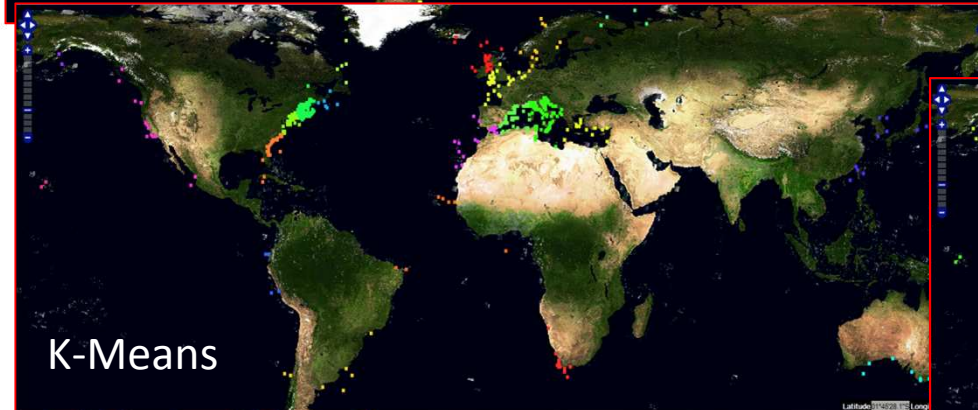


Presence Points  
(FishBase  
+  
Obis)



Density Based Clustering  
DBSCAN  
(with outliers)

Other methods are also  
available ...



The screenshot shows the Geo Explorer web interface. The main map displays a global temperature anomaly distribution with a color scale from blue (cold) to red (warm). Several layers are visible in the left sidebar, including 'Temperature in [12-15-09 01:00] (3D...)', 'Sarda sarda', and 'Exclusive Economic Zones Boundaries...'. The bottom panel shows a search results table for 'temperature' with columns for Layer Title, Abstract, Keywords, and Layer Name. A 'Summary layer info' panel on the right provides detailed metadata for the selected layer, including publication date, scope code, and abstract text.

This screenshot shows the iMarine search interface. At the top, there is a search bar and a 'Search' button. Below it, the 'Advanced Search Options' are visible, including a 'Switch grid view' section with icons for 'Images', 'Descriptive', 'Scientific', and 'Show related maps'. A grid of species images is shown, with 'Sarda sarda' selected. The 'Sarda sarda details' panel on the right displays a large image of the fish and a 'Meta Information' table. The table lists characteristics such as 'Deepwater', 'Stomach', 'Angling', 'Diving', 'Dangerous', 'Invertebrate', 'Algae', 'Seabird', and 'Freshwater', each with a 'N/A' value. It also lists species codes: 'Species ID: Fr-20018' and 'Species Code: 115'.

The screenshot shows the 'HSPec group generation settings' dialog box. It has several tabs: 'General details', 'Generation Settings', 'Select HCAFs', 'Select HSPENs', and 'Execution Environment'. The 'Generation Settings' tab is active, showing 'Algorithms' (Native Range, Native Range 2050, Suitable Range, Suitable Range 2050) and 'Data generation' options (Data only, Data and static images, Data, images and GIS layers). There are also options for 'Combine sources' (Maching only, All) and a 'Whether map' checkbox. A 'Send' button is located at the bottom right.



**MEAN=0.81**

VARIANCE=0.02

NUMBER\_OF\_ERRORS=6691

NUMBER\_OF\_COMPARISONS=259200

**ACCURACY=97.42**

MAXIMUM\_ERROR=1.0

MAXIMUM\_ERROR\_POINT=3005:363:1

COHENS\_KAPPA=0.218

COHENS\_KAPPA\_CLASSIFICATION\_LANDIS\_KOCH=Fair

COHENS\_KAPPA\_CLASSIFICATION\_FLEISS=Marginal

TREND=EXPANSION

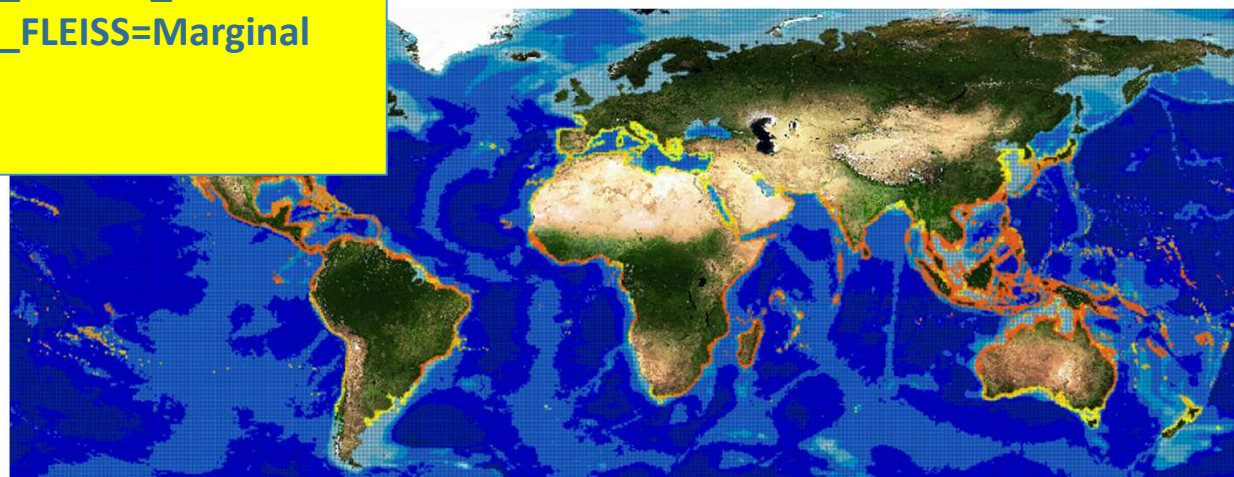
RESOLUTION=0.5



FAO Eleutheronema tetradactylum

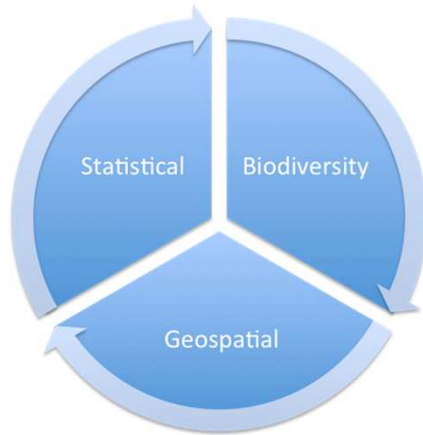
VS

AquaMaps Eleutheronema tetradactylum







# Not Only Access, Validation, Enriching, Processing, Sharing



- It is always possible to **save** the discovered data in various Standard formats
  - It is always possible to **collaborate** with co-workers through a dedicated workspace.
- **Mash-up** data across diversity
    - Accessing statistical datasets in SDMX, geo-referencing them, describing them in ISO19139, and making them available via OGC W\*S standard protocols
    - Accessing species observation datasets in DwC, analysing their distribution trend via R, and projecting them in geographical space
    - Accessing species taxonomies in DwCA and publishing them as reference data in SDMX

- The data and information are available under terms described in the product metadata
- Except where otherwise noted, this is the Creative Commons License 
- All derivative products are licensed under the Creative Commons License CC BY-SA 



- An **ecosystem** of participatory data e-Infrastructures
- Regulated by **policies**
- Enabled by **standards**
- Promoting not only access but **mash-up** of heterogeneous data

User centric





User-centric view of an **ecosystem** of participatory data e-Infrastructures to

- Cope with the overwhelming amount of data and capacities
- Promote re-use of data
- Encourage sharing of resulting products

User-centric and workflow-oriented

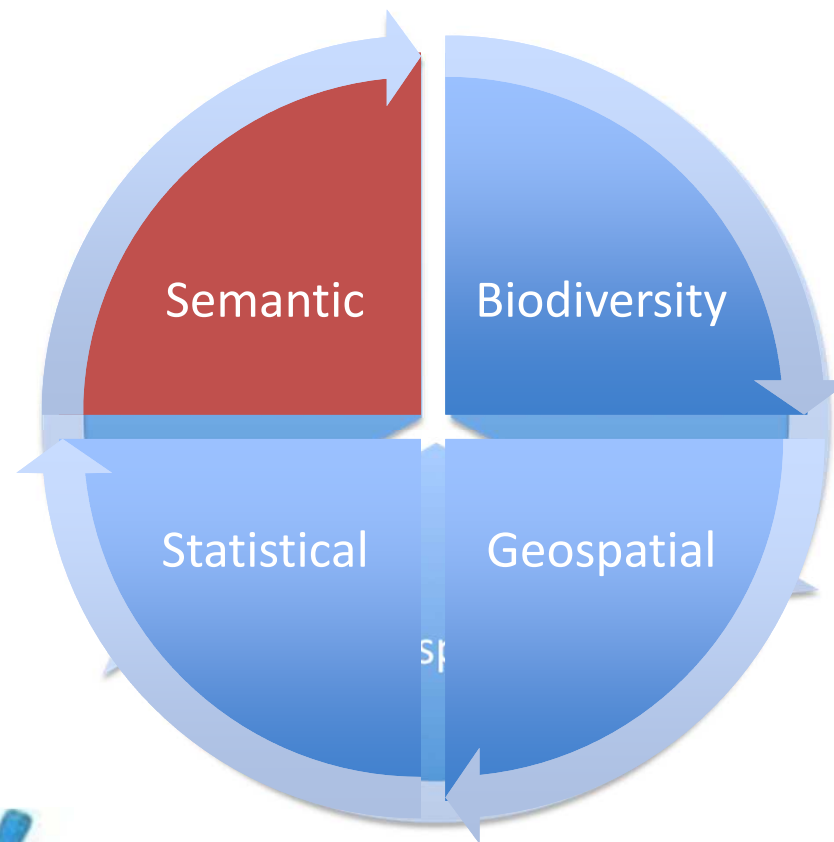


iMarine is user-centric and workflow-oriented thanks to the gCube VRE technology

**Virtual Research Environment (VRE) is**

- a **distributed and dynamically created** environment
- where **subset of** data, services, computational, and storage **resources**
- regulated by **tailored policies**
- are **assigned to a subset of users** via interfaces
- for a **limited timeframe**
- at **little or no cost** for the providers of the participatory data e-infrastructures

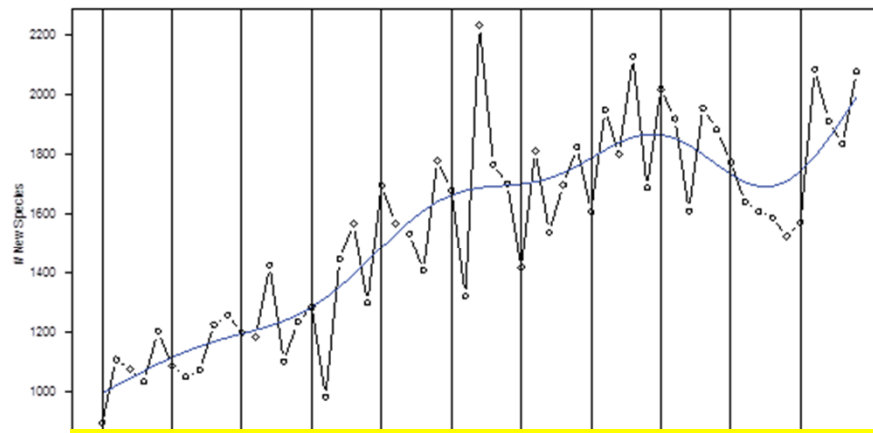






# A Service for Statistical Analysis of Marine Data in a Distributed e-Infrastructure

Gianpaolo Coro, Antonio Gioia, Pasquale Pagano, Leonardo Candela



Wednesday 25, September 2013 – 9:40

# Trendlyzer: a Long-Term Trend Analysis on Biogeographic Data

Ward Appeltans, Peter Pissierssens, IOC-UNESCO

Gianpaolo Coro, Angela Italiano, Pasquale Pagano, ISTI-CNR

Anton Ellenbroek, FAO

Tom Webb, University of Sheffield

**i marine**  
D4S e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources

A Service for Statistical Analysis of Marine Data in a Distributed e-Infrastructure  
Gianpaolo Coro, Antonio Gioia, Pasquale Pagano, Leonardo Candela  
Istituto di Scienza e Tecnologia dell'Informazione A. Faedo - CNR, Pisa, Italy

**Context**  
In the Computational Statistics domain, experts need environments in which they can easily manipulate data and run experiments by themselves. This is the case of Computational data mining procedures and datasets. In such domain, Distributed have been yet indicative layers for scientists who results with other remote experiments in easy way

**Proposal**  
We propose a workbench software, named **Statistical Manager (SM)** that offers a rich array of statistical computations and data mining applications for a variety of **biological and marine related problems**. Capabilities:

**Poster n° 80**

**Statistical Manager Workflow:**

- The requests coming from the clients are distributed to the available services.
- The first SM instance that has free computational resources accepts the request and notifies its status to the D4Science Information System.
- This SM executes the algorithm on local or distributed computational resources.
- The outputs are stored as tables or files.

**Features**  
The Service defines five base operations that can be requested by clients:

- **GetCapabilities:** to retrieve service metadata (Capabilities), which describe the algorithms supplied by the service.
- **GetProcessDescription:** to retrieve detailed information about the processes that can be executed through the service. Eg. inputs and outputs descriptions based on an xml schema.
- **Execute:** to run one of the processes listed among the Capabilities.
- **GetProcessStatus:** to check the completion status of a computation.
- **GetProcessOutput:** to retrieve the process output.

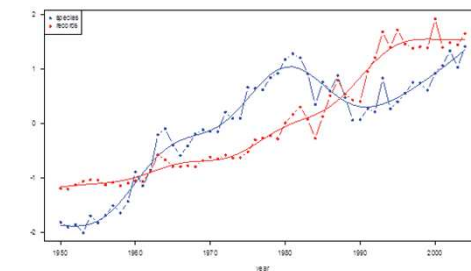
**Statistical Manager Web Interface:**  
On the left side: the list of Capabilities is grouped according to a user's oriented perspective. On the right side: the GetProcessDescription is called to get the types of the inputs and outputs for the selected algorithm (DBScan). The GUI generates the required fields "on the fly" and displays them with the proper format.

**Highlights**  
SM implements a simple approach to integrate algorithms, including R scripts. The Web GUI is automatically generated based on the algorithms inputs and outputs declarations. It allows scientists to execute and monitor the algorithms. The parallel executors through D4Science are transparent to the user. The features can be mapped onto OGC WPS specifications. SM is a component of the gCube open source software system, that supports the development of Distributed e-Infrastructures and the definition of Virtual Research Environments.

Contact Person: Gianpaolo Coro, gianpaolo.coro@isti.cnr.it

iMarine Consortium  
ERCIM  
FAO  
FiN  
GLOBIS  
CRIA  
IRD

www.imarine.org | www.d4science.org | www.gcube-system.org





# Thanks for your attention



Visit

<https://i-marine.d4science.org>

<http://www.i-marine.eu>



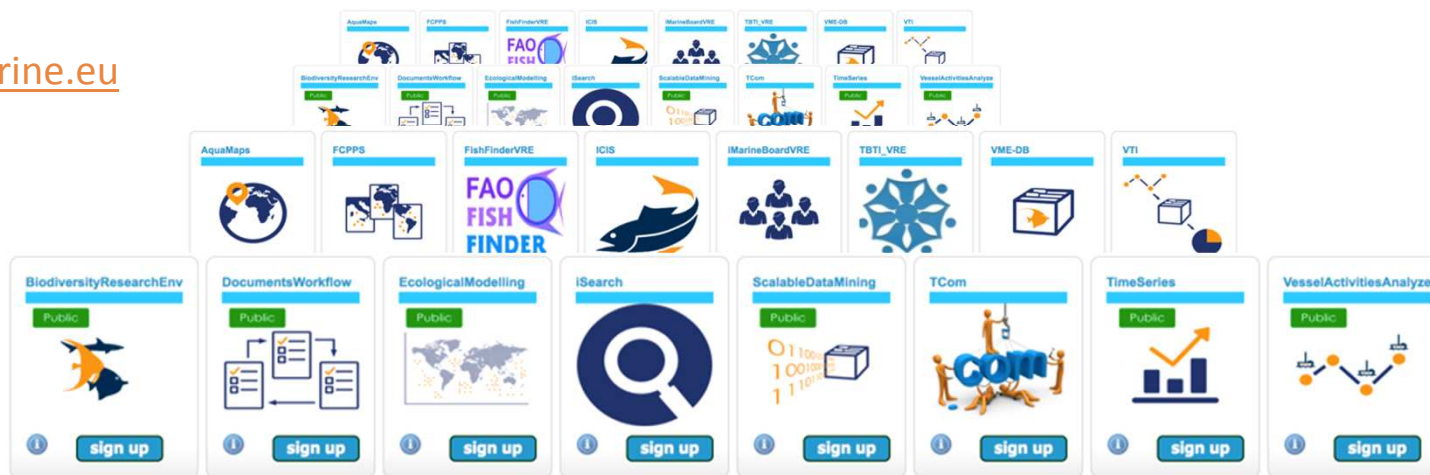
Join

gCube Apps



Enjoy

applications



## gCube is the iMarine empowering technology



Management and interpretation of biological and ecological data in the environment



Complete full life-cycle data framework, from observational data to aggregated data repositories enriched with validation and analytical tools



Storage and interpretation of geospatial explicit information, including WPS processing



Flexible sharing, storage, reporting, search and retrieval, aggregation and projection facilities

A BUNDLE is a set of services and technologies grouped according to a family of related tasks for achieving a common objective