The Technical Aspect of iMarine

Pasquale Pagano (CNR)

iMarine Technical Director
pasquale.pagano@isti.cnr.it
iMarine is exploiting a **Hybrid Data Infrastructure** combining over 500 software components into a coherent and centrally managed system of hardware, software, and data resources.
I need to host my applications in a secure and scalable environment.

I need to maintain my database.

I need to backup my data.

I need to deliver my data to a set of known people.

I want to offer a flexible sharing, storage, reporting, search and retrieval tool.
I need to manage and analyze biological and ecological data

I need to manage the full data life-cycle from import to validation, curation, harmonization and publication

I need to offer to my team a powerful tool to manage code-lists

I need to store and analyze geospatial explicit information

I need to analyse my big datasets
Born from the user needs

I need to access authoritative biological and ecological data

I need to simplify the access to my geospatial data

I need to mash-up statistical and biodiversity data

I need to reduce the costs of data maintenance of my dept.

I need to validate my datasets and provide a standard access to them
User Needs Analysis

• Needs
  – Not isolated
  – Not disconnected
  – Not trivial

• Solutions
  – Actual *but with an eye to the future*
  – Designed for individuals *but looking at the community*
Capacities: Storage as Service

to host and maintain data

**Database**
- High-availability
- Standard
- Ready-to-use

**Cloud Storage**
- Scalable
- Reliable
- Secure

**Geographical DB**
- Scalable
- OGC Standard
- Privacy and Attribution
to process and extract knowledge
Applications as a Service

to curate and manage data

Metadata Generation
Geospatial Data
Biodiversity Data
Statistical Data

Harmonization
Disambiguate
Validate
Integrate and Consistency Check

Data Exchange
OGC protocols
DarwinCore
SDMX
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.

iMarine data platform for collaborations
A BUNDLE is a set of services and technologies grouped according to a family of related tasks for achieving a common objective.
Presence Points (FishBase + Obis)

Density Based Clustering DBSCAN (with outliers)

Other methods are also available ...

K-Means

X-Means
Data Analysis with StatsCube

Import CodeLists

Validate Datasets

Analyse And Project
Ecological Modeling with BiolCube

iMarine data platform for collaborations
Maps Comparison with GeosCube

**FAO Eleutheronema tetratactylum**

**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**

**NUMBER_OF_ERRORS = 6691**

**NUMBER_OF_COMPARISONS = 259200**

**MEAN = 0.81**

**VARIANCE = 0.02**

**VS**

**AquaMaps Eleutheronema tetratactylum**

**VS**
iMarine data platform for collaborations
**OAI-PMH, OpenSearch**
- FAO Facksheets
- Aquatic Commons
- Bioline International
- Biodiversity Heritage
- OceanDocs
- Nature, PenSoft Journals
- ...

**SDMX ***
- FAO CodeLists
- IRD CodeLists
- FAO datasets
- Eurostat
- ...

**ISO19139 (OGC W*S)**
- 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, ...

**RDF, OWL**
- FAO FLOD
- Marine Top Level Ontology
- IRD Ecoscope
- FactForge, Yago2
- ...

**DarwinCore / ISO19139**
- >35 M Observations (OBIS)
- ≈ 120 K Observed Species (OBIS)
- ≈ 500 K Taxa (WoRMS)
- >600 K Scientific Names (ITIS)
- >12 K Species Maps (AquaMaps)
- >600 Species Extent (FAO)
- ... FishBase, SeaLifeBase
- ... CoL, GBIF
Is this enough?

- 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
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

to share and collaborate

Share
Database Tables
Workflow
Files

Communicate
Post
Favourite
Connection

Organize
Dynamic VRE Creation
Secure
Policy Control
iMarine Technology

- iMarine is powered by gCube

**Activity**

**30 Day Summary**
Feb 2 2014 — Mar 4 2014

- 1145 Commits
- 29 Contributors including 1 new contributor

**12 Month Summary**
Mar 4 2013 — Mar 4 2014

- 11102 Commits
  - Up +2564 (30%) from previous 12 months
- 43 Contributors
  - Up +8 (22%) from previous 12 months

... is mostly written in Java
with an average number of source code comments

... has a well established, mature codebase
maintained by a very large development team
with increasing Y-O-Y commits

https://www.ohloh.net/p/gCube
iMarine is exploiting D4Science.org

- Geographically Distributed Computing Infrastructure Across administrative boundaries Across private and commercial providers
- Service Allocations, Deployment, Monitoring, and Operation
- Uniform resource and data access

**Operation**
- Built on SLAs
- Support monitoring, auditing, reporting, and notification

**Trust**
- Privacy, governance, and attribution
- Security, trusted network
Multi-tenant Delivery Model

Infrastructure as a Service
- Dynamic deployment
- Hosting
- Resource Lifecycle
- Monitoring
- Accounting
- Security

Software as a Service
- BiolCube
- ConnectCube
- GeosCube
- StatsCube

Platform as a Service
- FeatherWeightStack
- SmartGears
- ApplicationSupportLayer
- SOA3

iMarine data platform for collaborations
www.i-marine.eu

i-marine.d4science.org
Google Analytics iMarine portal