



2nd International Conference
on Research Infrastructures
02-04 April 2014

Megaron Athens International Conference Centre

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

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



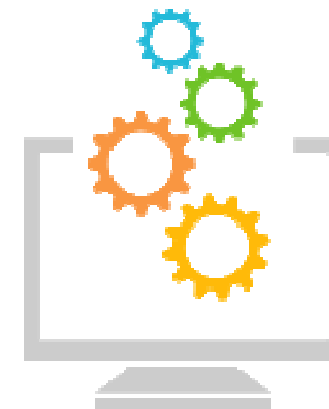
I need to backup my data



I need to delivery my data to
a set of known people



I want to offer a flexible sharing,
storage, reporting, search and
retrieval tool



Capacities



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

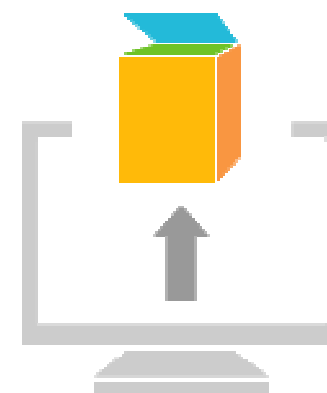


I need to offer to my team a powerful tool to manage code-lists

I need to store and analyze geospatial explicit information



I need to analyse my big datasets



Applications

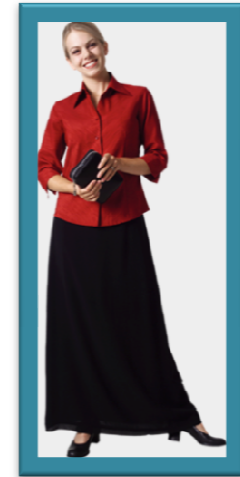
Born from the user needs



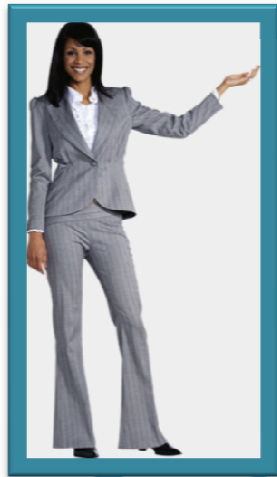
I need to access authoritative biological and ecological data



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



Data



- 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

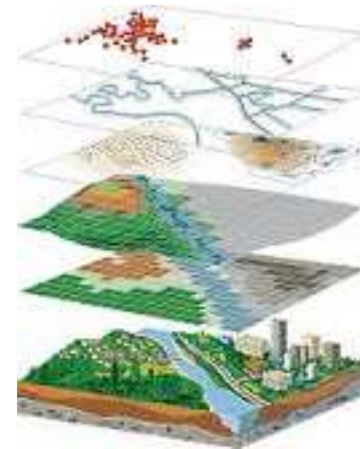
to host and maintain data



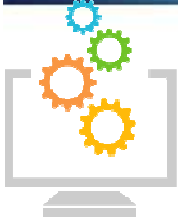
Database
High-availability
Standard
Ready-to-use



Cloud Storage
Scalable
Reliable
Secure



Geographical DB
Scalable
OGC Standard
Privacy and Attribution



to process and extract knowledge



Scalable

*Easy to Manage
Across Boundaries
Tailored*



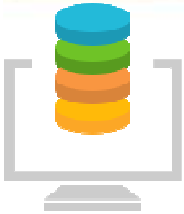
Elastic

*Assignment of Computing
Assignment of Processors
Virtual Research Environment*



Rich and Heterogeneous

*High Throughput
Map-Reduce
Parallel R*



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

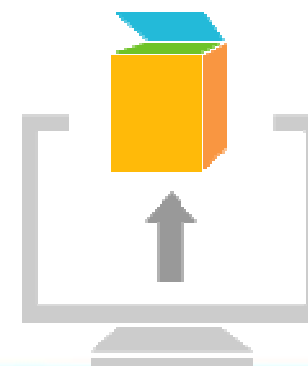


Storage and interpretation of geospatial explicit information, including WPS processing



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





Occurrence and Taxonomic Data Discovery
Occurrence Data Processing
Species Distribution Modeling
Species Distribution Maps Discovery
Taxonomic Data Comparison
Taxonomic Data Matching



Code List Discovery
Code List Management
Statistical Engine
Tabular Data Discovery
Tabular Data Enrichment
Tabular Data Management
Tabular Data Processing

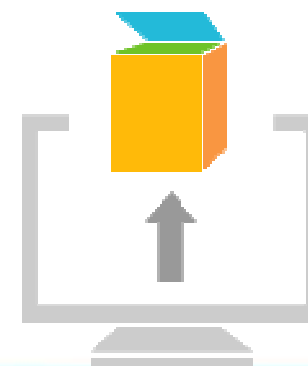


Geospatial Data Discovery
Geospatial Data Processing

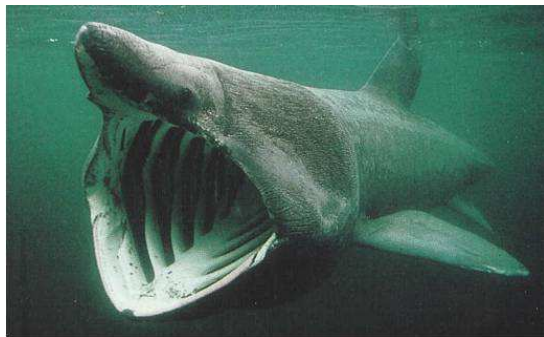


Enhanced Documents Management
Fact-sheets Management
Information Object Discovery
Messaging
Shared Workspace
Social Networking Facilities

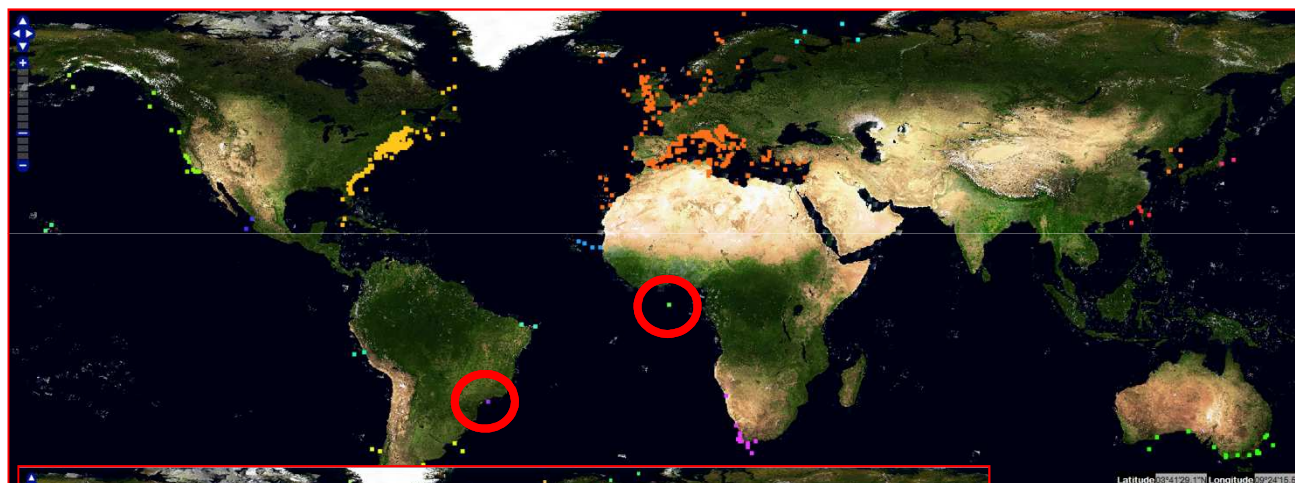
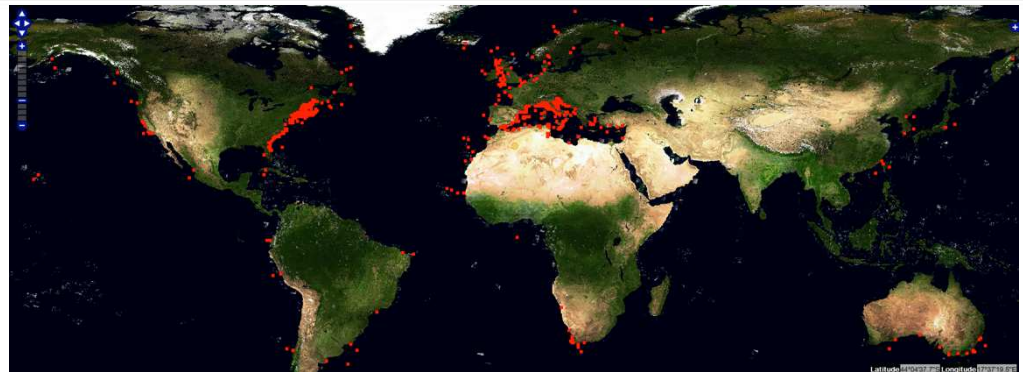
A BUNDLE is
a set of
services and
technologies
grouped
according to
a family of
related
tasks for achieving a
common
objective



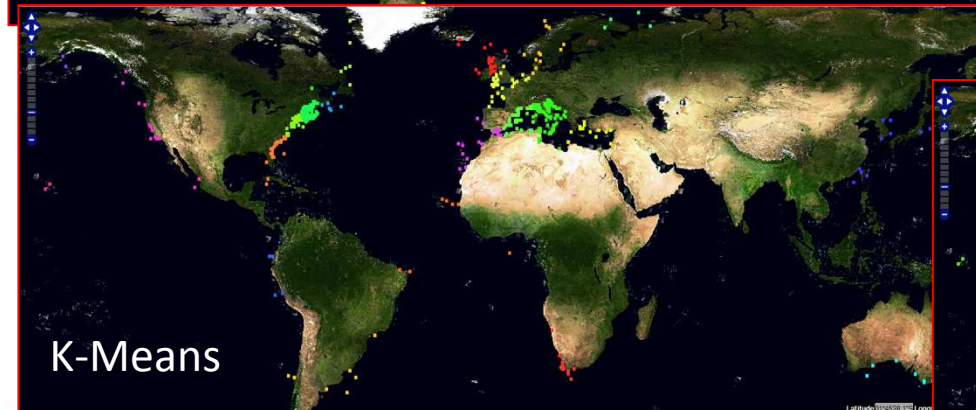
Features Clustering with StatsCube



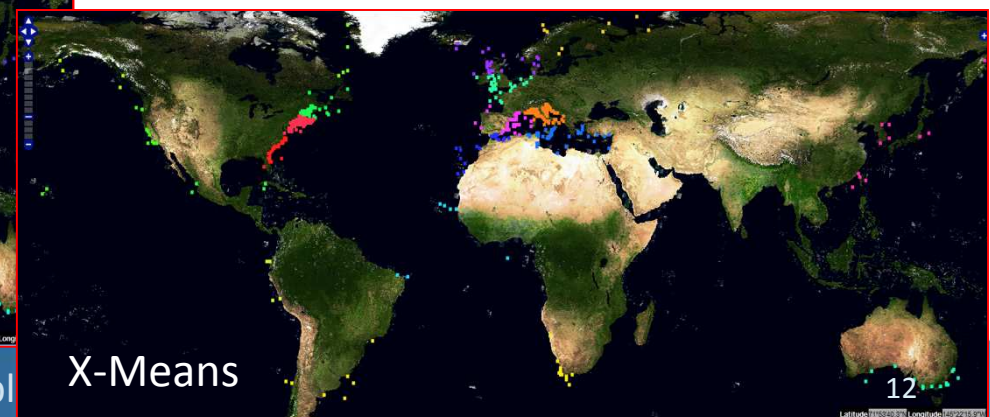
Presence
Points
(FishBase
+
Obis)



Density Based Clustering
DBSCAN
(with outliers)



K-Means



X-Means

Data Analysis with StatsCube

SDMX Import

SDMX Codelist selection

Search:

Name	Agency Id	Version
CL_ASFIS_TAX	FAO	0.1
CL_DIVISION	FAO	0.1
CL_FAO_MAJOR_AREA	FAO	0.1
CL_SPECIES	FAO	1.0
CL_UNIT	FAO	0.1
CL_UNIT_MULTIPLIER	FAO	0.1
CL_UN_COUNTRY	FAO	0.1
CL_DIVISION	SDMX	0.1
asfis-2012	SDMX	0

Import CodeLists

SDMX Import

Operation In Progress

Import Summary

Document: Codelist document
Source: SDMX Registry source
Url: Internal
Codelist Selected: Id: CL_FAO_MAJOR_AREA
Name: CL_FAO_MAJOR_AREA
Agency: FAO
Version: 0.1

30% importing...

CL_SPECIES

code	en	fr	es	la	AUTHOR
AAA	Adriatic sturgeon	Esturgeon de l'Adriatique	Esturión del Adriático	Acipenser naccarii	Bonaparte 1836
AAB	Twobar seabream	Pagre double bande	Sargo de dos bandas	Acanthopagrus bifasciatus	Forskåll 1775
AAC	Bowfin			Amia calva	Linnaeus 1766
AAD	Yangtze sturgeon			Acipenser dabryanus	Duméril 1869
	frogfish			Antennarius analis	(Gosline 1957)
	sturgeon			Acipenser fulvescens	Rafinesque 1817
	mottled eel			Anguilla bengalensis	(Gray 1831)
	sturgeon			Acipenser schrenckii	Brandt 1869
	se sturgeon			Acipenser sinensis	Gray 1835
	in longfin eel			Anguilla mossambica	(Peters 1852)
	in sturgeon			Acipenser mikadoi	Hilgendorf 1892
	mottled eel			Anguilla marmorata	Quoy & Gaimard 1824
	sturgeon	Esturgeon vert	Esturión verde	Acipenser medirostris	Ayres 1854
	barbel sturgeon	Esturgeon barbillons frangés	Esturión barba de flecos	Acipenser nudiventris	Lovetzky 1828
	ic sturgeon			Acipenser oxyrinchus	Mitchill 1815
	in sturgeon			Acipenser persicus	Borodin 1897
	Zealand longfin eel			Anguilla dieffenbachii	Gray 1842
	ed longfin eel			Anguilla reinhardtii	Steindachner 1867

Validate Datasets

ref_country

Column properties

Label: name_en Column Type: Description Value Type: Text (Lorem ipsum dicit ipse.)

id	fid	name_zh	name_ru	min_long	max_long	min_lat	max_lat	un_code	undp_code	iso_2_code	iso_3_code	name_en	name_fr	name_es	name
1	7			11	24	-18	-4	24	ANG	AO	AGO	Angola	Angola	Angola	
10	126			20	26	53	56	440	LIT	LT	LTU	Lithuania	Lituanie	Lituania	
100	217			0	1	6	11	768	TGO	TG	TGO	Togo	Togo	Togo	
101	250			12	31	-13	5	180	BRC	CD	COD	Congo, Dem. R	Rep. dém. du	Rep. Dem. del	
102	83			-174	176	-11	4	296	KIR	KI	KIR	Kiribati	Kiribati	Kiribati	
103	74			8	14	-3	2	266	GAB	GA	GAB	Gabon	Gabon	Gabon	
104	47			-165	-157	-21	-8	184	CKI	CK	COK	Cook Islands	Iles Cook	Islas Cook	
105	96			13	19	42	46	191	CRO	HR	HRV	Croatia	Croatie	Croacia	
106	49			-84	-74	19	23	192	CUB	CU	CUB	Cuba	Cuba	Cuba	
107	50			32	34	34	35	196	CYP	CY	CYP	Cyprus	Chypre	Chipre	
108	122			-165	-172	4	14	584	MAS	MI	MIB	Marshall Island	Iles Marshall	Islas Marshall	

Check Accept

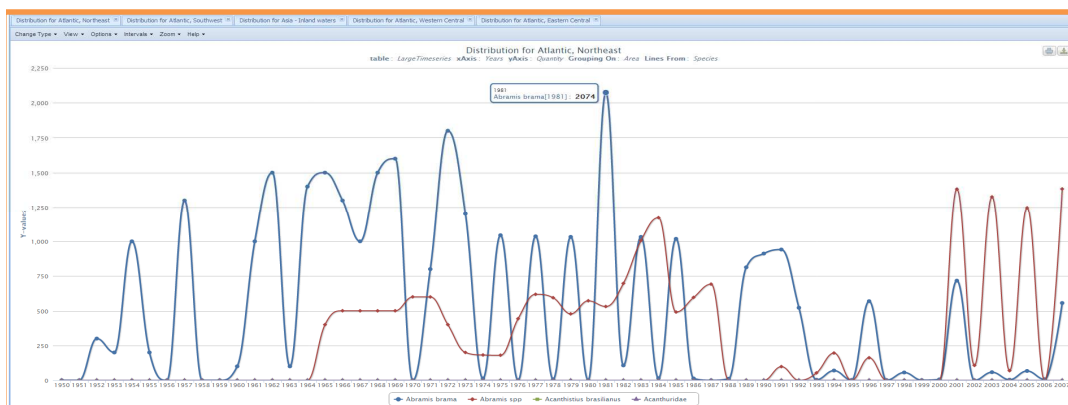
Country continent relations

Column properties

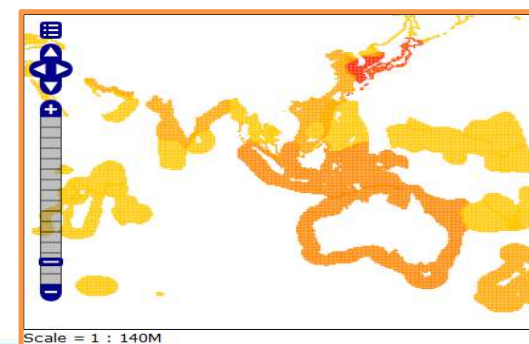
Label: child Column Type: Child Code Code List: COUNTRY_OLD

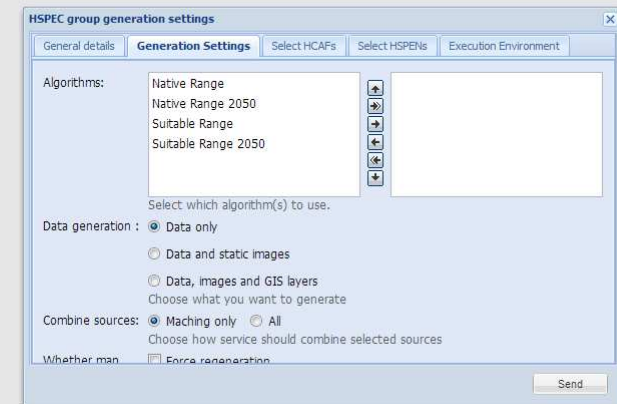
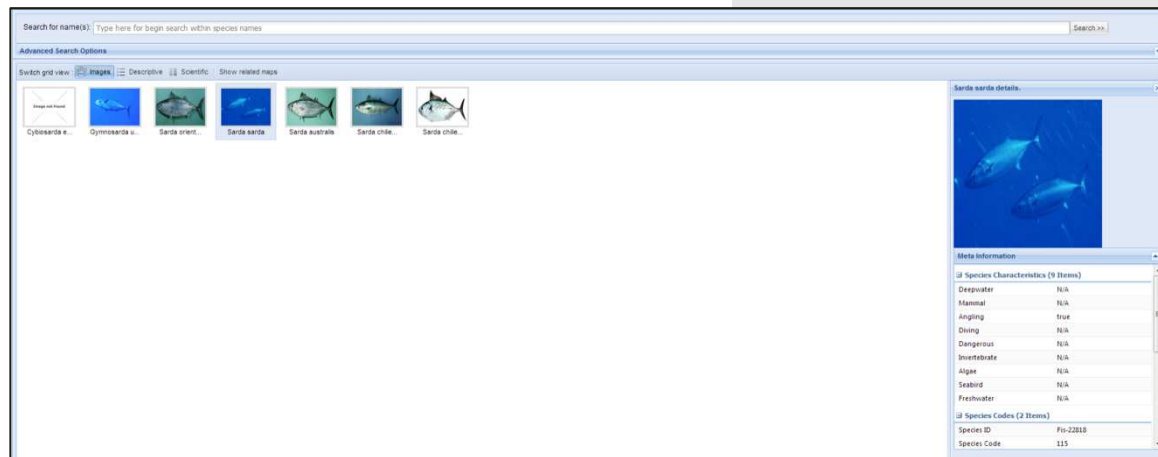
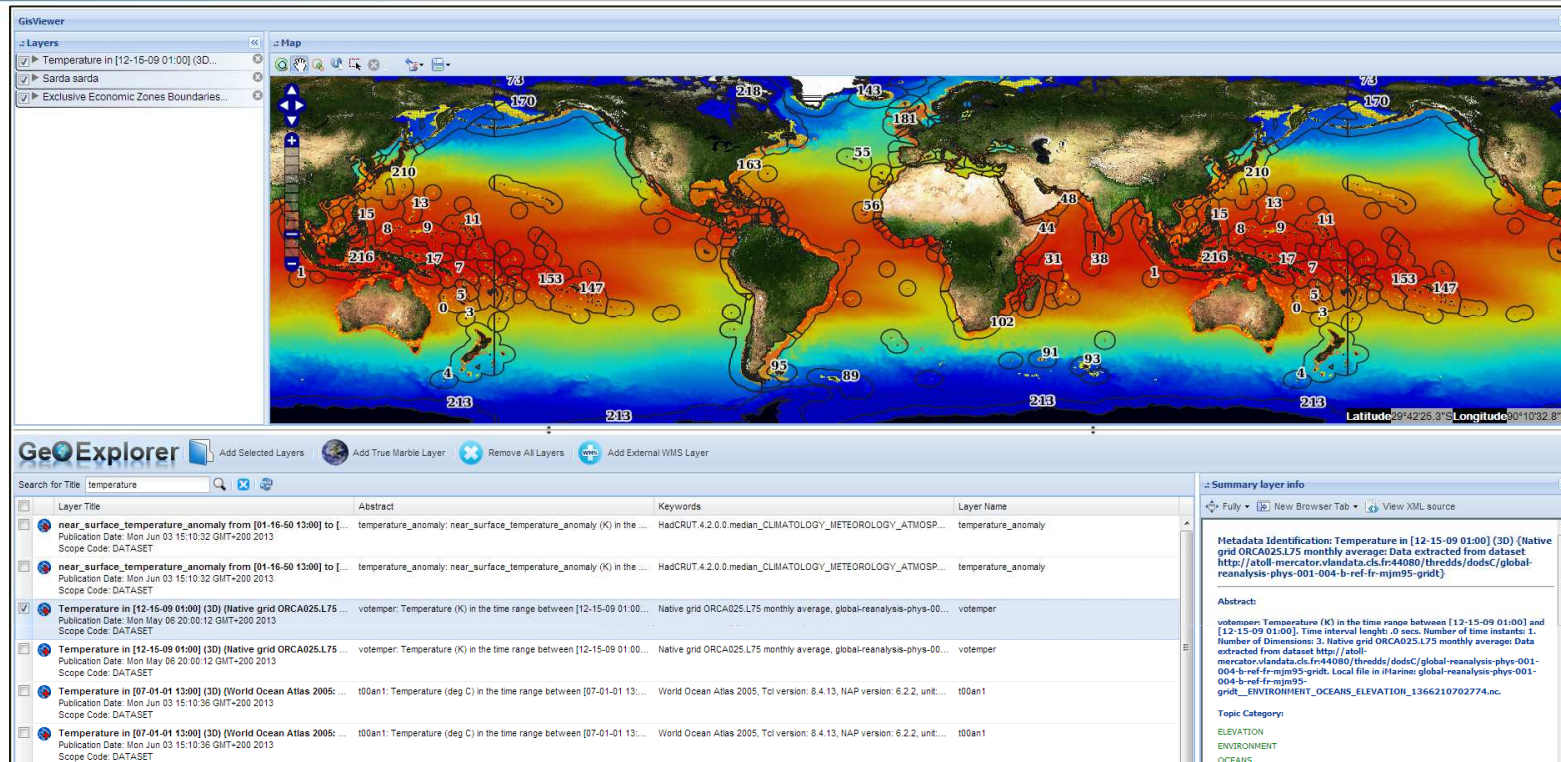
Name	Creation Date	Agency Id	Version
CL_SUB_UNIT	9/8/11 2:18 PM	FAO	1
CL_SUB_DIVISION	9/8/11 2:18 PM	FAO	1
CL_VESSEL_TYPE	9/12/11 10:03 AM	FAO	1
CL_VESSEL_TYPE	9/12/11 10:03 AM	FAO	1
CL_SUB_AREA	9/12/11 10:52 AM	FAO	1
CL_FAO_MAJOR_AREA	9/12/11 11:33 AM	FAO	1
CL_DIVISION	9/12/11 11:43 AM	FAO	1
COUNTRY_OLD	9/12/11 1:38 PM	undefined	1
SPECIES_OLD	9/12/11 1:45 PM	undefined	1
CONTINENT_OLD	9/12/11 2:21 PM	undefined	1
AREA_OLD	9/12/11 2:32 PM	undefined	1
CONTINENT_COUNTRY	9/12/11 4:06 PM	undefined	1

Select



Analyse And Project





Maps Comparison with GeosCube

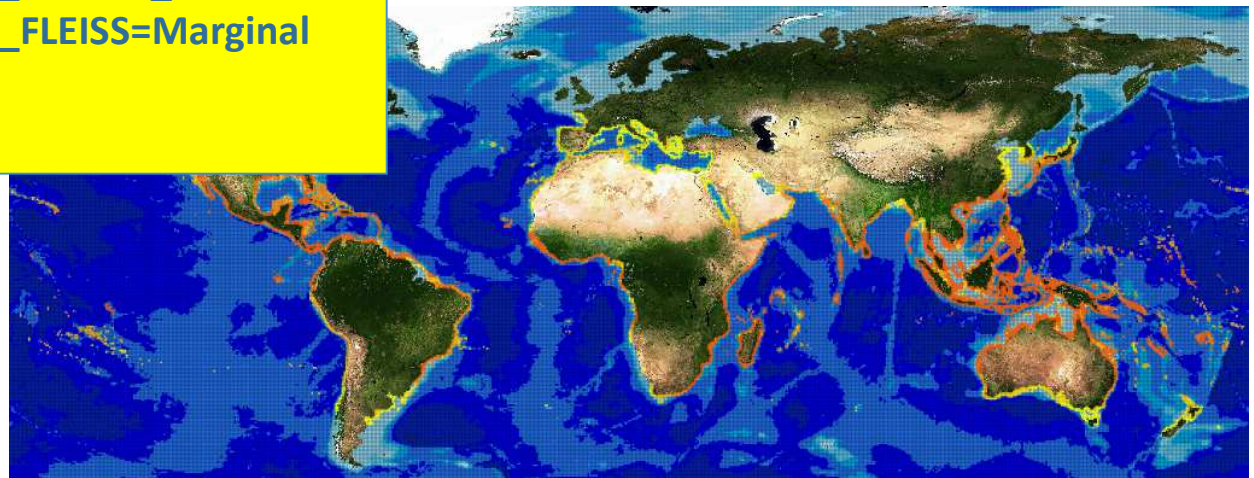
MEAN=0.81
VARIANCE=0.02
NUMBER_OF_ERRORS=6691
NUMBER_OF_COMPARISONS=259200
ACCURACY=97.42
MAXIMUM_ERROR=1.0
MAXIMUM_ERROR_POINT=3005:363:1
COHENS_KAPPA=0.218
COHENS_KAPPA_CLASSIFICATION_LANDIS_KOCH=Fair
COHENS_KAPPA_CLASSIFICATION_FLEISS=Marginal
TREND=EXPANSION
RESOLUTION=0.5

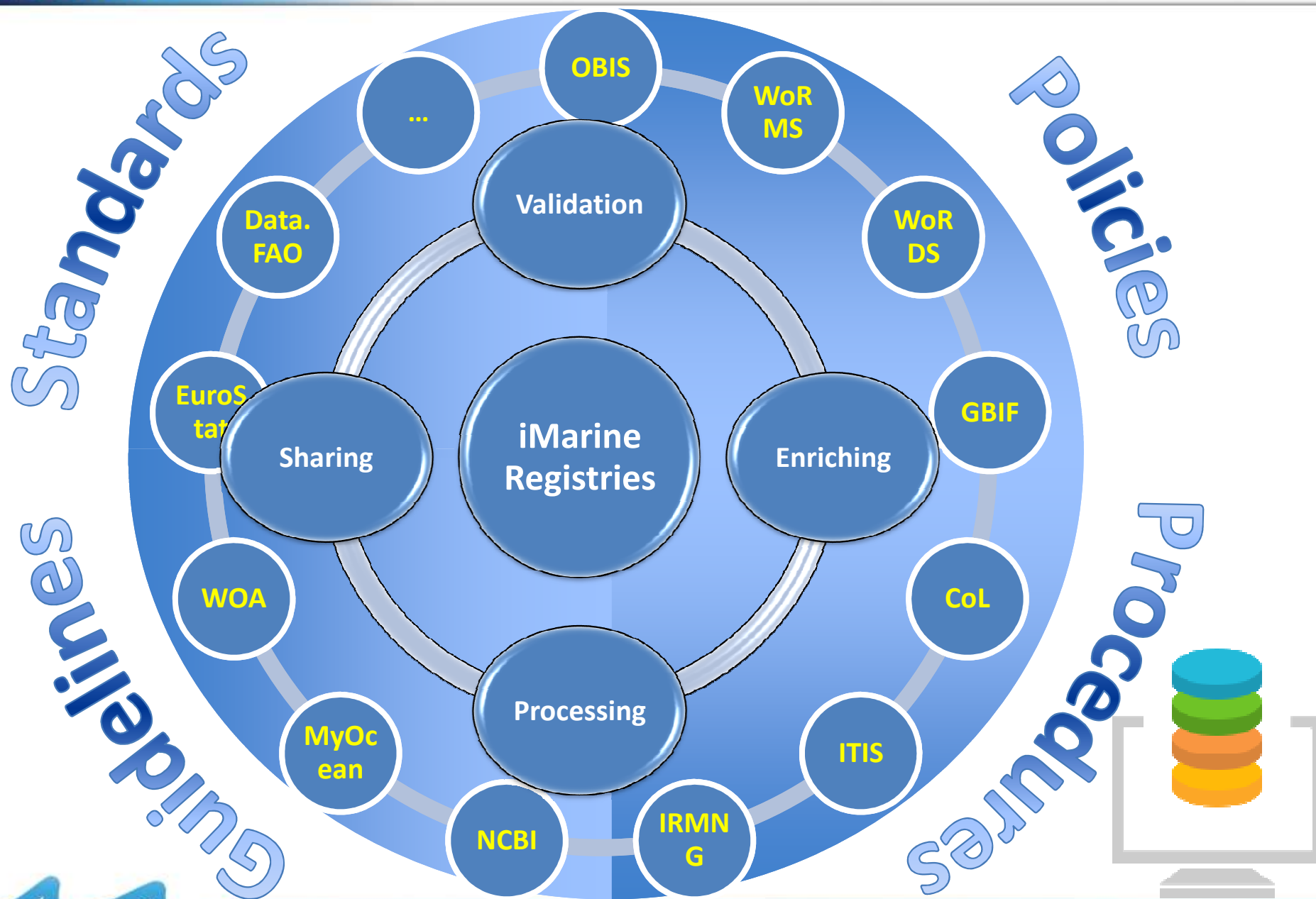


FAO Eleutheronema tetradactylum

VS

AquaMaps Eleutheronema tetradactylum



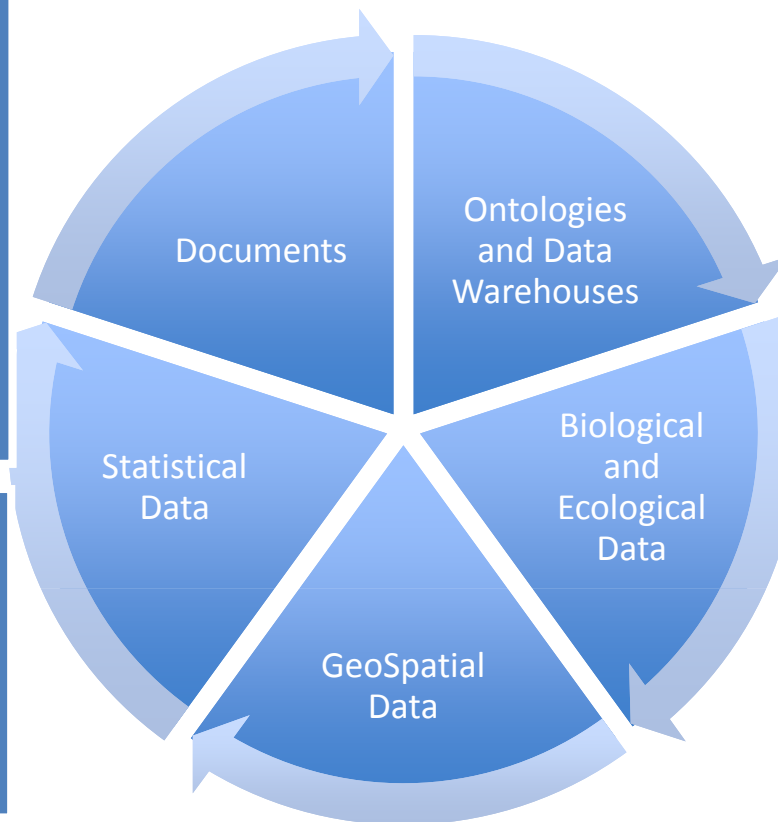


OAI-PMH, OpenSearch

- FAO Facksheets
- Aquatic Commons
- Bioline International
- Biodiversity Heritage
- OceanDocs
- Nature, PenSoft Journals
- ...

SDMX *

- FAO CodeLists
- IRD CodeLists
- FAO datasets
- Eurostat
- ...

**RDF, OWL**

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

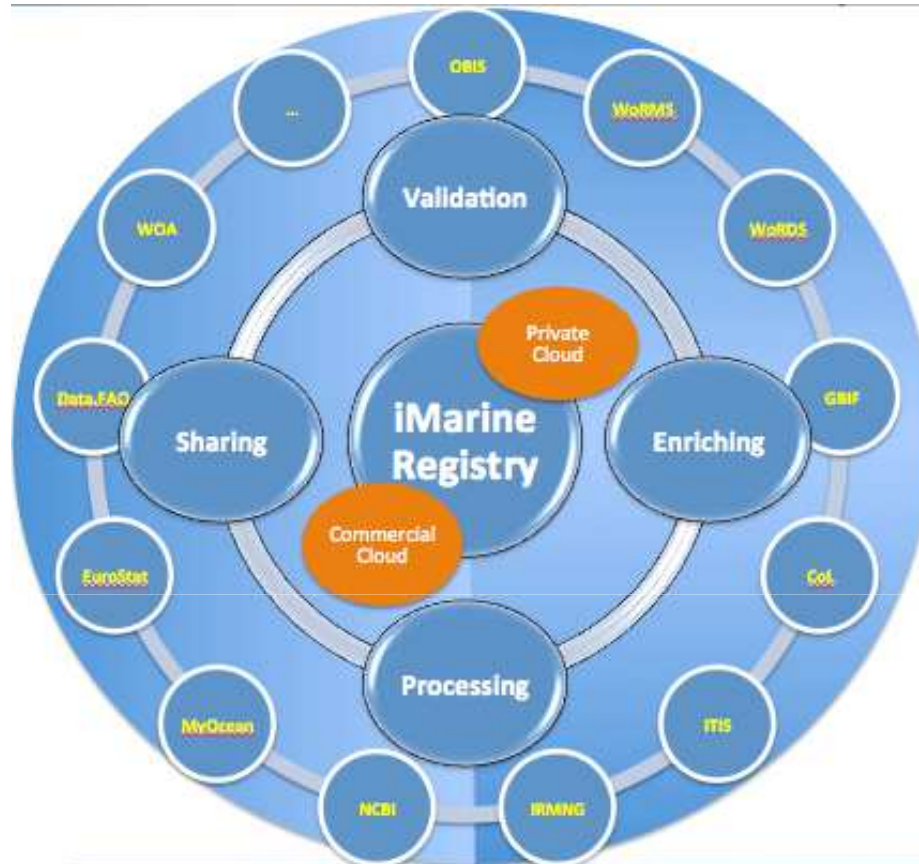
DarwinCore / ISO19139

- >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

ISO19139 (OGC W*S)

> 350
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, ...



- 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



L. Candela, D. Castelli, P. Pagano (2013) Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal, Vol. 12

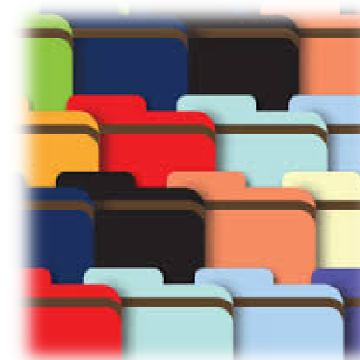
to share and collaborate



Share
Database Tables
Workflow
Files



Communicate
Post
Favourite
Connection



Organize
Dynamic VRE Creation
Secure
Policy Control

- iMarine is powered by gCube



Hot Projects on Ohloh

Filter by Languages: All Languages

Rank	Name	Claimed By	PAI	Hotness Score
1	Chromium (Google Chrome)			80.660
2	Chromium Blink			77.498
3	IntelliJ IDEA Community...			72.798
4	KDE	KDE		70.475
5	Mozilla Core	Mozilla Foundation		67.878
6	Mozilla Firefox	Mozilla Foundation		67.209
7	Kubuntu Packaging			63.997
8	bleeding_edge			61.607
9	gCube			60.910
10	commcare-hq			58.314

<https://www.ohloh.net/p/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

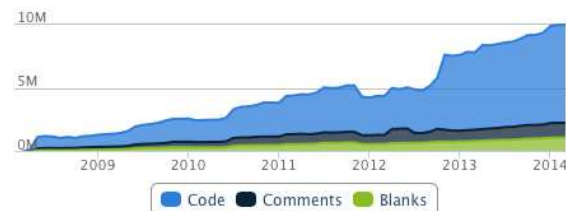
11102 Commits

Up +2564 (30%) from previous 12 months

43 Contributors

Up +8 (22%) from previous 12 months

Lines of Code



... 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

iMarine is exploiting D4Science.org



Geographically
Distributed
Computing
Infrastructure

Across
administrative
boundaries

Across private
and commercial
providers

Operation Built on SLAs

Support monitoring,
auditing, reporting, and
notification

Service
Allocations,
Deployment,
Monitoring, and
Operation

Uniform
resource and
data access

Trust Privacy, governance,
and attribution

Security, trusted
network



D4SCIENCE
INFRASTRUCTURE

Infrastructure
as a Service

- Dynamic deployment
- Hosting
- Resource Lifecycle
- Monitoring
- Accounting
- Security



Software as a
Service

- BioCube
- ConnectCube
- GeosCube
- StatsCube



Platform as a
Service

- FeatherWeightStack
- SmartGears
- ApplicationSupportLayer
- SOA3





www.i-marine.eu

i-marine.d4science.org

