

# Managing Virtual Research Environments in Hybrid Data Infrastructures

Pasquale Pagano (CNR, Italy)

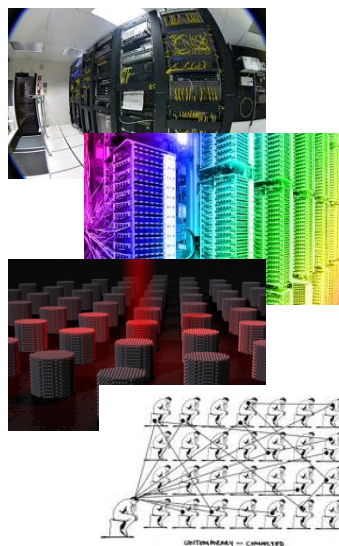
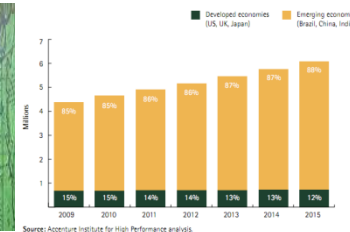
iMarine Technical Director

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

## The Context

Science is increasingly *global, multipolar, and networked*

Data continue to grow in *Volume, Variety, and collection, processing and consumption Velocity*



## The Needs

**Computational environments** dealing with the volume of the data

Efficient and tailored **storage and access technologies** dealing with the variety of the data types

**Elastic management** of the resources dealing with the innovative approaches for collection, processing and consumption of the data

**World-wide collaborative environment** between distributed scientific communities dealing with the federation of heterogeneous data sources

## The Solution

## Hybrid Data Infrastructures

*integrated technologies supporting efficient data management*

- Well suited for typical **biodiversity processes**
- Provides access to
  - **computational and storage resources** offered by commercial cloud providers
  - new storage technologies generally identified as **no-sql databases**
  - distributed computing platform supporting **MapReduce**
  - several algorithms for performing **data analysis and mining**
- Offers **scalable platforms for data interoperability** and efficient **data management**
- Offers a **scalable infrastructure for efficient spatial data access, processing, and visualization** (WCS, WPS, WMS, WFS)

D4Science HDI hosts biodiversity communities federated by the **iMarine** and the **EUBrazilOpenBio** initiatives

D4Science HDI will provide **ENVRI** RIs with seed resources



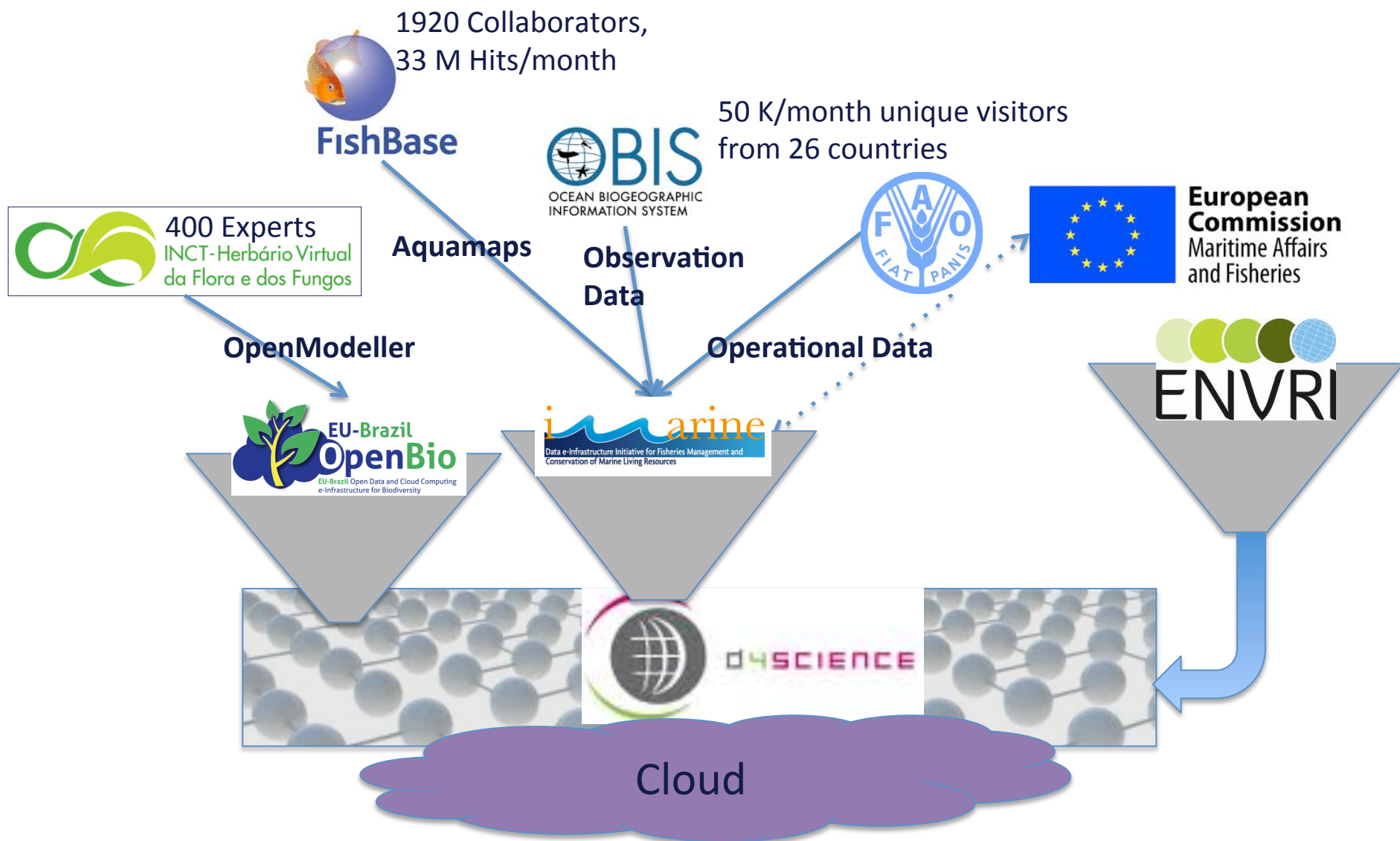
## D4Science Hybrid Data Infrastructure

Support to  
providers willing  
to share  
hardware, data,  
software  
resources

Transparent  
access to  
hardware, data,  
software  
resources of  
third-party  
providers

Harmonization,  
integration  
mining and  
analysis of  
particular types of  
data and support  
to process  
workflows

Cost effective  
creation,  
operation and  
maintenance of  
Virtual Research  
Environments



- gCube offers solutions to **abstract over differences in location, protocols, and models** by
  - scaling no less than the interfaced resources,
  - keeping failures partial and temporary,
  - reacting and recovering from a large number of potential issues.
- gCube **turns infrastructures and technologies into a utility** by offering a single registration, monitoring, and access facilities.





# gCube Enabling Layer

## Information System [1/2]

A scalable and reliable framework

- supporting an **extensible notion of resource**
- open to **modular extensions at runtime** by arbitrary third parties

- registration
- discovery
- Notification
- ...



### Hardware:

- Storage (RBDMS, blob, ColumnStore),
- Computing (gCube Container, Hadoop, EMI, Azure, ...)
- Cloud resources



### Services & Applications:

- gCube Apps
- Third party Software and Applications



### Data & Auxiliary Resources:

- Data sets, Metadata, Indexes, Annotations
- Schemas, Mappings, Transformation programs

A scalable and reliable framework

- supporting an **extensible notion of resource**
- open to **modular extensions at runtime** by arbitrary third parties

- ...
- Monitoring
- Inspection
- Assignment
- Accounting

The image displays two screenshots of the gCube Enabling Layer Information System interface. The top screenshot shows the 'Resource Management' tab, which includes a 'Resource Details (GHN)' section with a table of resources and a 'Node Accounting' section with a table of node usage. The bottom screenshot shows the 'VRE Management' tab, which includes a 'Select nodes to deploy VRE Services' section with a table of nodes and a 'VRE Deployer' section with a table of deployment options.

**Resource Details (GHN)**

Name	Status	Last Updated	gCore v.	ghn v.	V. Mem left	HD Spac...	V. Memory...
SubType: beaces (1 Item)							
bsgr006.bsc.es9090	certified	2012-04-26T01:01:47+02:00	1.2.3	3.2.5	46% (475 MB)	148,559 MB	1,017 MB

**Node Accounting**

Node	Service	ServiceName	CallerScope	StartTime	EndTime	Inv No	Avg Inv T...	Caller
node65.p.d4sc	IS-Collector	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	13	0.0670769	node65...
node66.p.d4sc	IS-Collector	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	3	0.018	146.48...
node66.p.d4sc	IS-Collector	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	3	0.0186667	146.48...
node66.p.d4sc	IS-Collector	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	3	0.0186667	146.48...
node66.p.d4sc	IS-Collector	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	3	0.0176667	146.48...
node66.p.d4sc	IS-Collector	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	3	0.0186667	146.48...
node66.p.d4sc	IS-Collector	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	8	0.01155	146.48...
node66.p.d4sc	IS-Registry	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	6	0.0176667	node62...
node65.p.d4sc	IS-Registry	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	12	0.06475	portal1...
node65.p.d4sc	IS-Registry	/d4science.research-infrastruct...	/d4science.research-infrastruct...	2012-04-...	2012-04-...	12	0.01725	node63...

**VRE Deployer**

Host name	mem avail.	up time	Select	Selectable	Use for GHN Ma...
studio.p.d4science.research-infrastructures.eu.8080	822 MB	195 days	<input type="checkbox"/>	no	<input type="checkbox"/>
portal1-marine.d4science.org.9000	512 MB	36 days	<input type="checkbox"/>	no	<input type="checkbox"/>
portal1.d4science.org.9000	585 MB	22 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node66.p.d4science.research-infrastructures.eu.8000	1217 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node63.p.d4science.research-infrastructures.eu.8080	639 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node62.p.d4science.research-infrastructures.eu.8080	693 MB	158 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node61.p.d4science.research-infrastructures.eu.8080	877 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node60.p.d4science.research-infrastructures.eu.8080	866 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node69.p.d4science.research-infrastructures.eu.8080	554 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node64.p.d4science.research-infrastructures.eu.8080	1190 MB	208 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node53.p.d4science.research-infrastructures.eu.8080	639 MB	68 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node48.p.d4science.research-infrastructures.eu.8080	1440 MB	29 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node36.p.d4science.research-infrastructures.eu.8080	729 MB	56 days	<input type="checkbox"/>	no	<input type="checkbox"/>
newportal1-marine.d4science.org.9000	612 MB	54 days	<input type="checkbox"/>	no	<input type="checkbox"/>



A distributed framework managing a trusted resource network

### *Dynamic Deployment*

- *remote deployment of resources across the infrastructure*

### *Resource lifetime management*

- *running of the lifetime of resources ranging from creation and publication to discovery, access and consumption*

### *Self-elastic management*

- *(re-)configuration of resources across the infrastructure*

### *Virtual Research Environment Management*

- *Cost effective creation, operation and maintenance of Virtual Research Environments*

### *Interoperability, openness and integration at software level*

- *third-parties software can be added to the Data e-Infrastructure at runtime - Web Applications (Running in Tomcat); Web Services (Running in service containers, e.g. JAX-WS, Axis); Executable (e.g. pojo, shell script, ...)*

## Workflow Engine

The following list of adaptors is currently provided:

- **WorkflowJDLAdaptor** - parses a Job Description Language (JDL) definition block and translates the described job or DAG of jobs into an Execution Plan which can be submitted to the ExecutionEngine for execution.
- **WorkflowGridAdaptor** - constructs an Execution Plan that can contact a EMI UI node, submit, monitor and retrieve the output of a grid job.
- **WorkflowCondorAdaptor** - constructs an Execution Plan that can contact a Condor gateway node, submit, monitor and retrieve the output of a condor job.
- **WorkflowHadoopAdaptor** - constructs an Execution Plan that can contact a Hadoop UI node, submit, monitor and retrieve the output of a Map Reduce job.

## Virtual Research Environment [1/4]

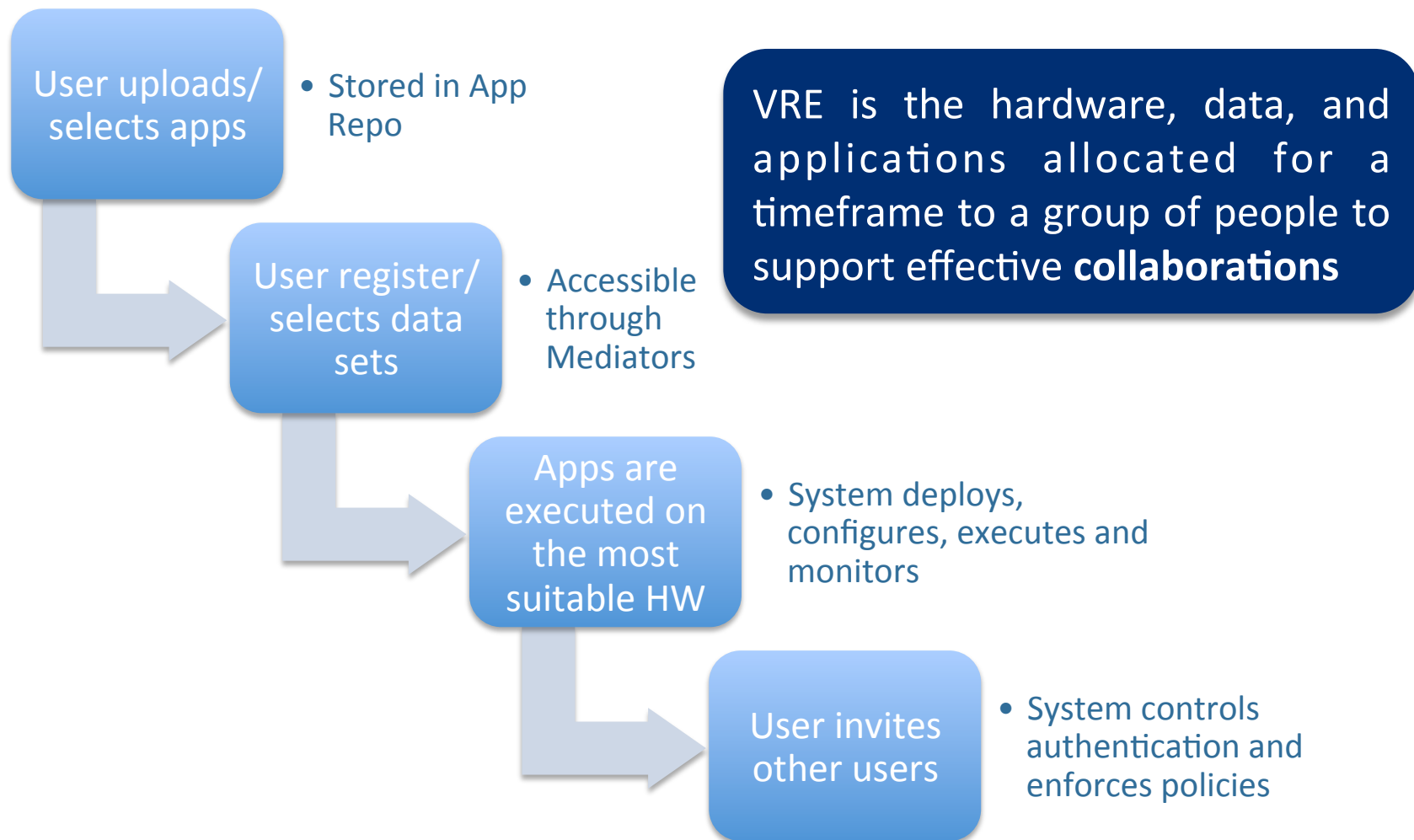
a distributed and dynamically created environment

where subset of resources are securely assigned and operated to a subset of users

for a limited timeframe

at little or no cost for the providers of the infrastructure

## Virtual Research Environment [2/4]



## Virtual Research Environment [3/4]

- Cost-effective creation and management

- Definition

- Creation

- Configuration

The screenshot displays the gCube Enabling Layer interface, which is used for managing Virtual Research Environments (VREs). The interface is divided into several sections:

- Defining the Virtual Research Environment:** This section allows users to define the VRE by selecting various services and functionalities. The 'Data Manipulation' section is currently selected, showing options like 'Time Series Management', 'Ecological n', 'Ontology Ma', 'Course Man', 'Annotation N', 'Report Man', 'Metadata Ed', and 'Data Type Adv'. The 'Time Series Management' section is also visible, explaining that selecting this function equips the VRE with an environment supporting the management of Time series objects.
- VRE Deployer:** This section shows a table of available nodes for deployment. The table includes columns for Host name, mem avail, up time, Select, Selectable, and Use for GHN Ma....
- Cloud available:** This section indicates that the cloud is available and shows the 'Resources Setup' section, which includes a 'Use Cloud' checkbox and a 'Virtual machines' dropdown menu.
- Users Management:** This section shows a list of registered users, including their username, email, and role. The 'Add New Users' button is also visible.

Host name	mem avail	up time	Select	Selectable	Use for GHN Ma...
restudio.p.d4science.research-infrastructures.eu/8080	822 MB	195 days	<input type="checkbox"/>	no	<input type="checkbox"/>
portal.i-marine.d4science.org/9000	512 MB	36 days	<input type="checkbox"/>	no	<input type="checkbox"/>
portal.d4science.org/9000	585 MB	22 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node66.p.d4science.research-infrastructures.eu/8000	1217 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node63.p.d4science.research-infrastructures.eu/8080	630 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node62.p.d4science.research-infrastructures.eu/8080	693 MB	168 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node61.p.d4science.research-infrastructures.eu/8080	677 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node60.p.d4science.research-infrastructures.eu/8080	866 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node59.p.d4science.research-infrastructures.eu/8080	5546 MB	67 days	<input type="checkbox"/>	no	<input type="checkbox"/>
node54.p.d4science.research-infrastructures.eu/8080	1190 MB	208 days	<input type="checkbox"/>	no	<input type="checkbox"/>

Username	Email	Role
daniele.izzi	daniele.izzi@isi.cnr.it	VO-Admin
daniele.strollo	daniele.strollo@isi.cnr.it	VRE-Manager
david.foster	david.foster@isi.cnr.it	Production-Support
david.fuegli	david.fuegli@isi.cnr.it	VO-Admin
david.landolf	david.landolf@isi.cnr.it	VRE-Manager
debbie.buckley	debbie.buckley@isi.cnr.it	Production-Support
defaveri	defaveri@isi.cnr.it	VO-Admin
dejan.kolundzija	dejan.kolundzija@isi.cnr.it	VRE-Manager
derek.law	derek.law@isi.cnr.it	Production-Support
detlev.balzer	detlev.balzer@isi.cnr.it	VO-Admin
diego.marcheggiani	diego.marcheggiani@isi.cnr.it	VRE-Manager
diego.milano	diego.milano@isi.cnr.it	Production-Support
dimtrakaramida	dimtrakaramida@isi.cnr.it	VO-Admin
dimtris.fotopoulos	dimtris.fotopoulos@isi.cnr.it	VRE-Manager
dimtris.katris	dimtris.katris@isi.cnr.it	Production-Support
dimtris.paparis	dimtris.paparis@isi.cnr.it	VO-Admin
dirk.noorda	dirk.noorda@isi.cnr.it	VRE-Manager
donatella.castelli	donatella.castelli@isi.cnr.it	Production-Support

## Virtual Research Environment [4/4]

### **addresses integration and presentation requirements**

when resources and researchers are widely apart

when research is computationally demanding

### **on-demand and interactive definition**

from resource pools allocated to communities

pools may overlap

### **self-deployed and self-monitored**

planned, based on match-making

with redeployment on detection of load and failures

### **value to e-Infrastructure**

lowers operational costs

encourages resource provision under federation





VREs

Exemplification

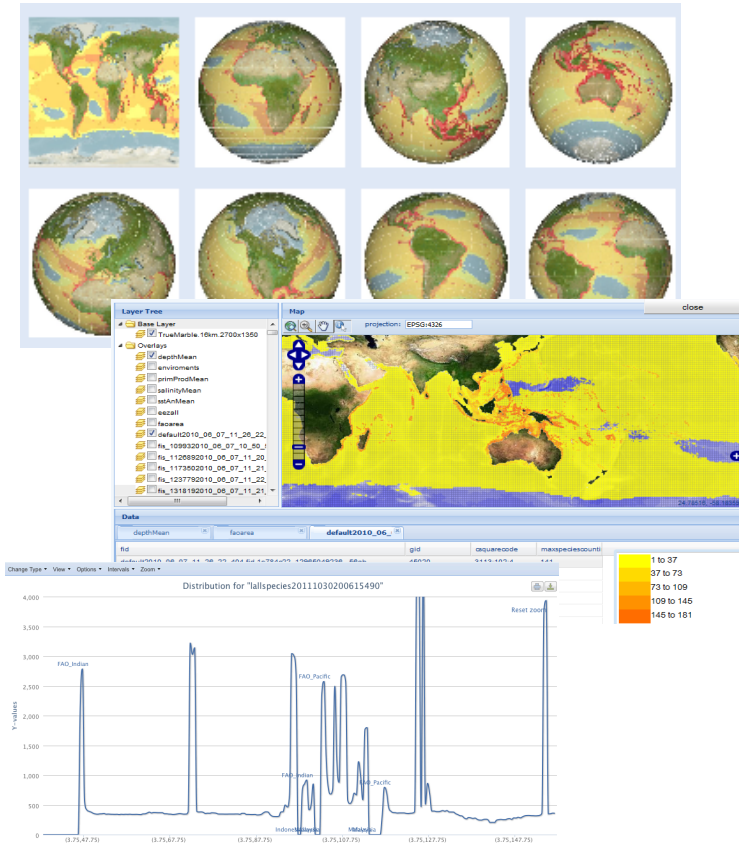
gCube Ecological Niche Modelling App is designed to

- work with **dataset versions**
- access to **external databases**
- **extensible** with predictive algorithms (aquamaps + feed-forward neural network algorithms)
- exploit **several computational back-ends** (multi-core server, distributed servers, and clouds)
- use **several storage technologies** (RDBMS, Column Store, Blob)
- publish distribution to **Geospatial Web** services
- support **evaluation** based on
  - CLASSIFICATION QUALITY ANALYSIS: given a probability distribution and a set of occurrences\absence points (True/False positives and negatives, accuracy, sensitivity, specificity)
  - DISCREPANCY ANALYSIS between two spatial distributions (variance, accuracy, mean error, ...)
  - HABITAT REPRESENTATIVENESS SCORE to assess the suitability of survey coverage for modeling the distribution of marine species

The **gCube Ecological Niche Modelling App** is instantiated with the four AquaMaps algorithms\*

Comparable with **AquaMaps Legacy** application but

- Data generation is 5-times faster on a single server, and up to 50-times faster on iMarine
- Adds generation and publication of GIS layers
- Supports generation of transect
- Supports data management facilities
- Solves scalability issues



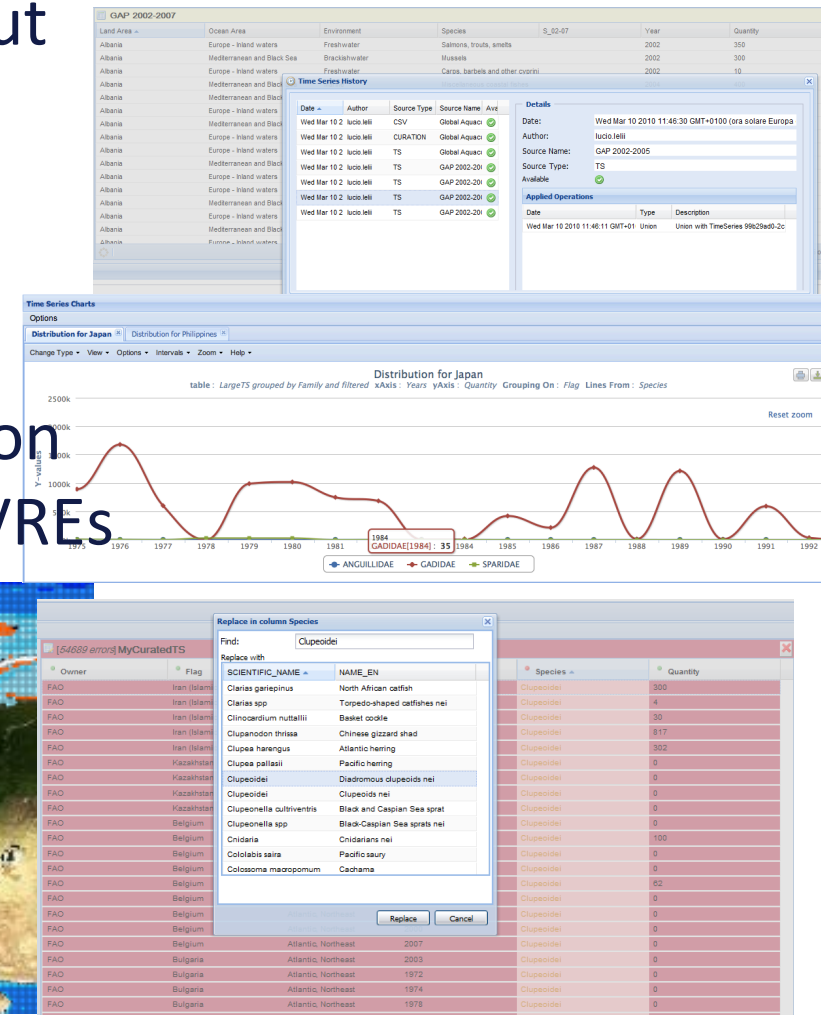
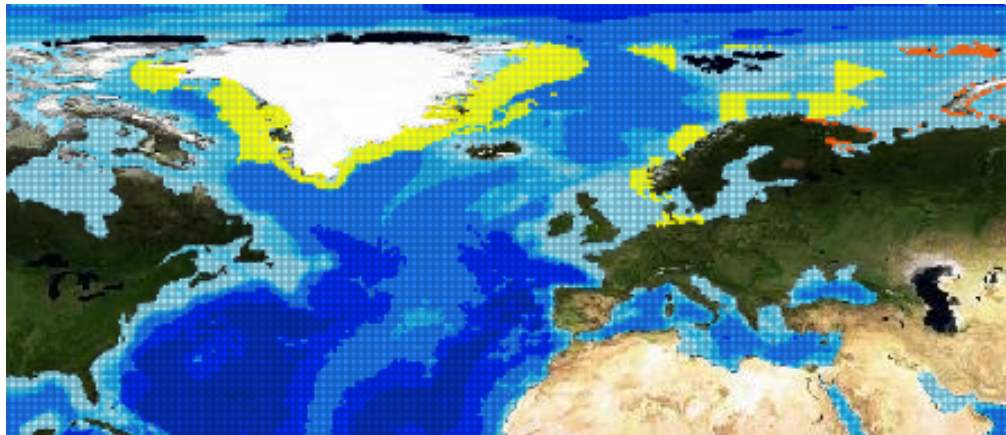
\* Algorithms by Kashner et al. 2006

Timeseries App is designed to

- support the **complete TS lifecycle**
- manage **multiple versions** enriched with **provenance** data
- support **validation, curation**, and **analysis** (filtering, grouping, and aggregation on multidimensional data)
- provide support for **data reallocation**
- supports **code list management** through SDMX
- statistical **data analysis** with R
- supports a rich set of **visualization**
  - Chart (histogram, bar, pie, line)
  - Map

## Comparable with Google Fusion but

- data import is 40-times faster
- supports code list management through SDMX
- supports data curation
- supports a rich set of visualization
- supports sharing in and across VRES



The **D4Science Infrastructure** implementing the HDI approach enables heterogeneous resource sharing between cross-domain infrastructures

Collects under a common environment resources coming from several e-infrastructures

Interacts with existing cloud infrastructures to deliver elasticity of resources

Is the result of a long experience managing distributed infrastructures for different communities and use cases





Thanks for your attention



Visit



Join



Enjoy

[www.d4science.org](http://www.d4science.org)

[www.i-marine.eu](http://www.i-marine.eu)

[www.eubrazilopenbio.eu](http://www.eubrazilopenbio.eu)

gCube Apps

applications

