Managing Virtual Research Environments in Hybrid Data Infrastructures

Pasquale Pagano (CNR, Italy)
iMarine Technical Director

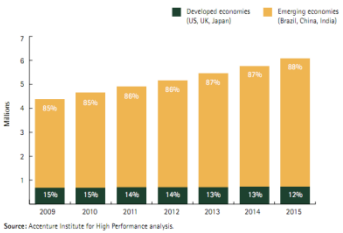
pasquale.pagano@isti.cnr.it

New Science Pattern

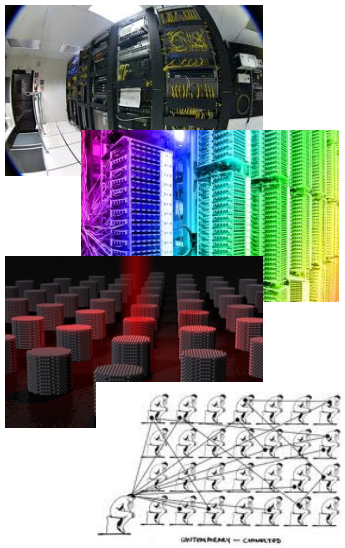
The Context

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

D4Science Hybrid Data Infrastructure

- 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

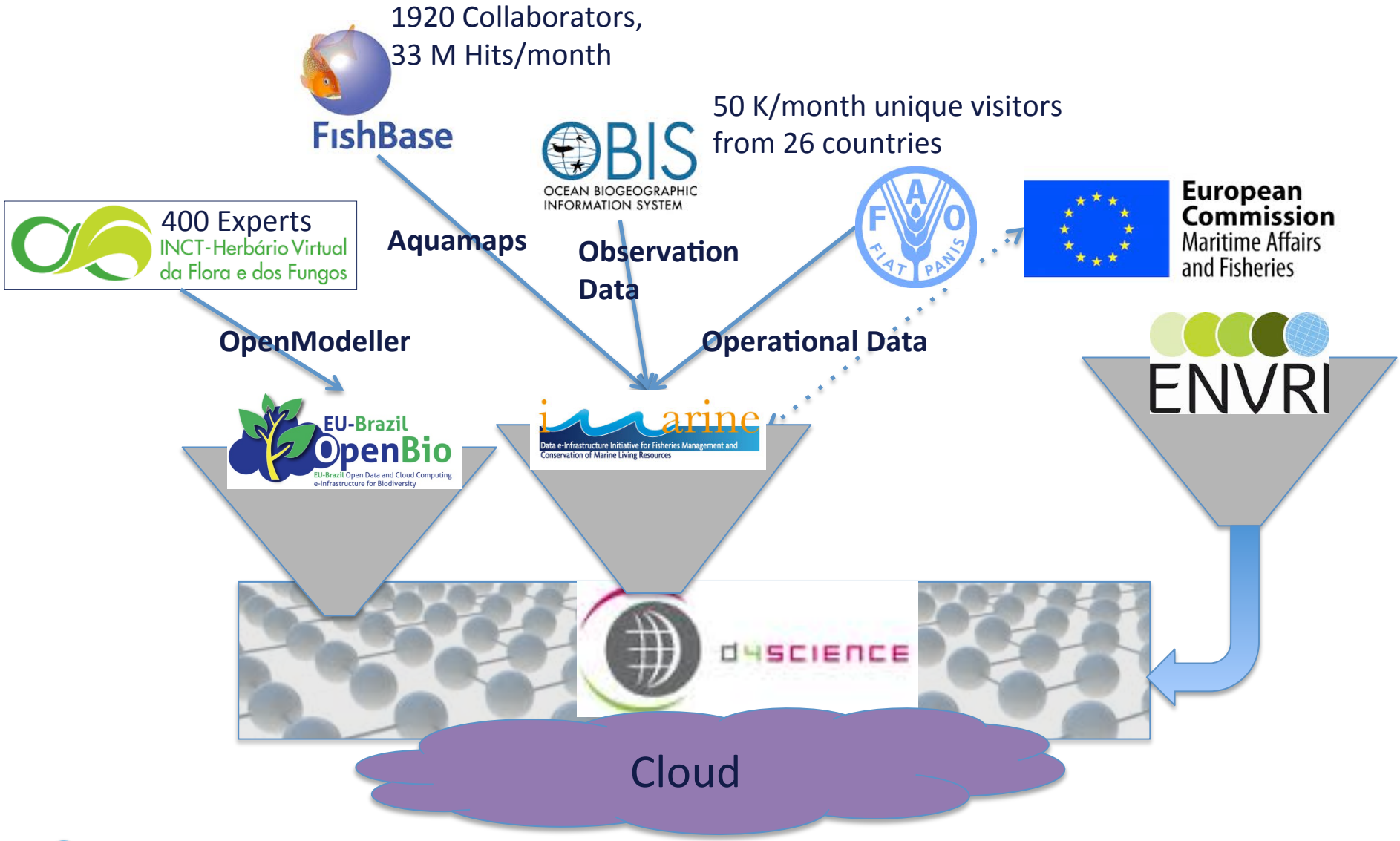
D4Science HDI will provide ENVRI RIs with seed resources

What type of e-Infrastructure?

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

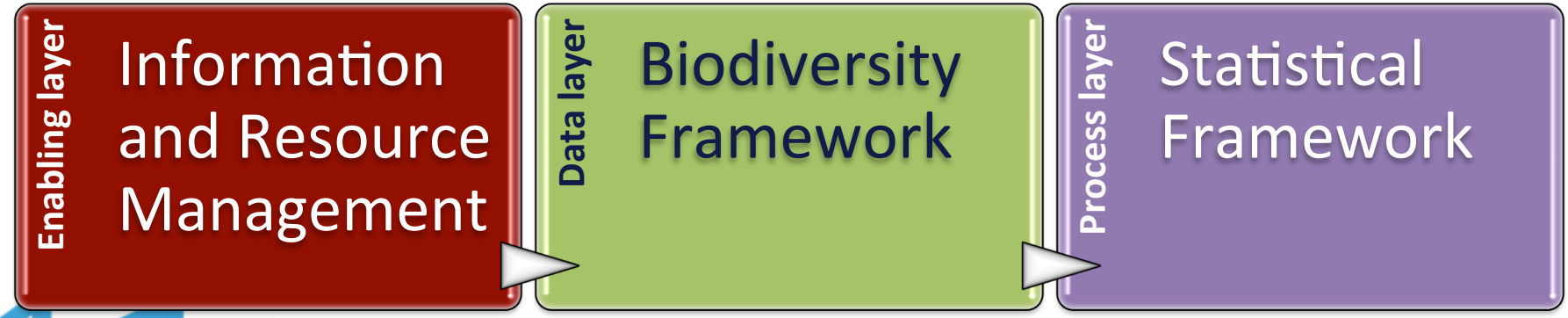
D4Science: example of communities



D4Science Technology: the gCube system

- 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.



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

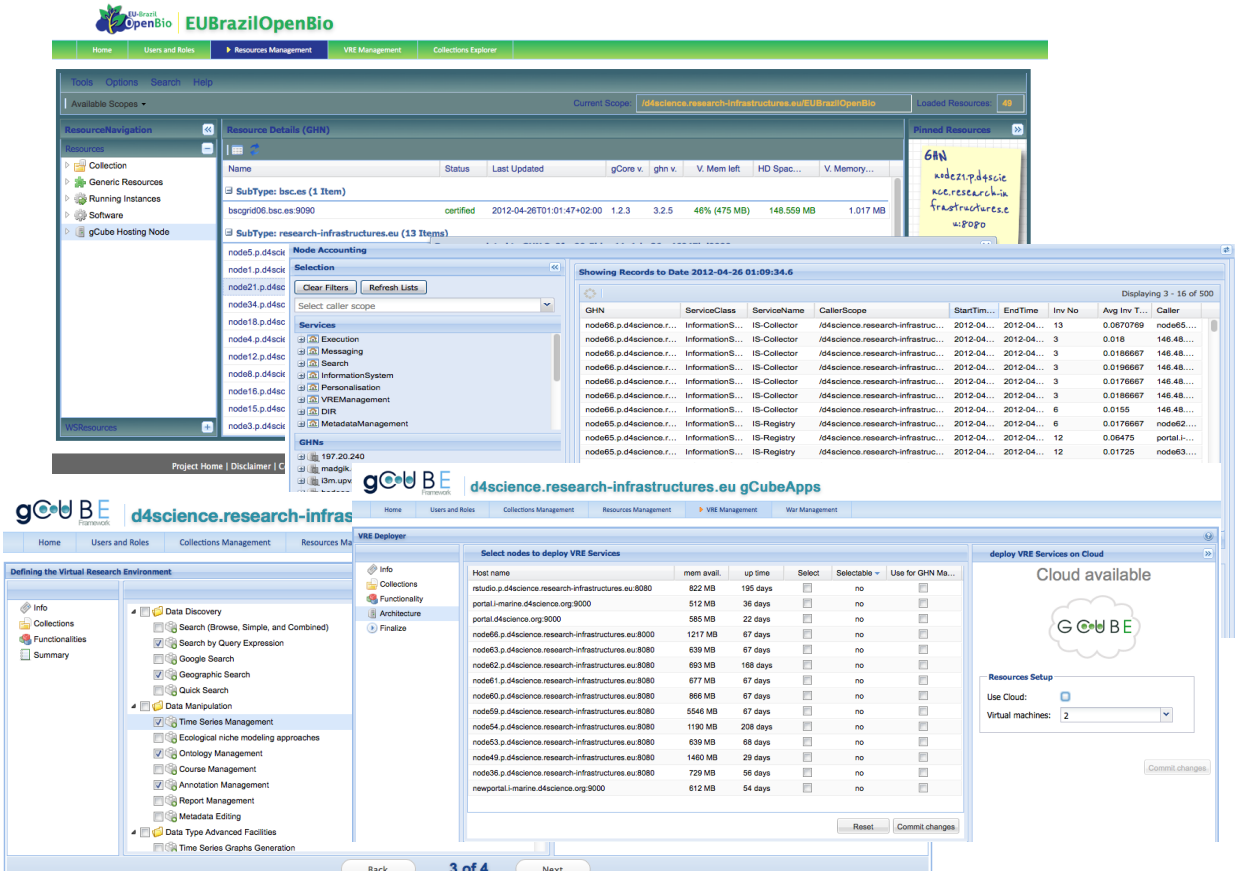
gCube Enabling Layer

Information System [2/2]

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



Resource Management

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]

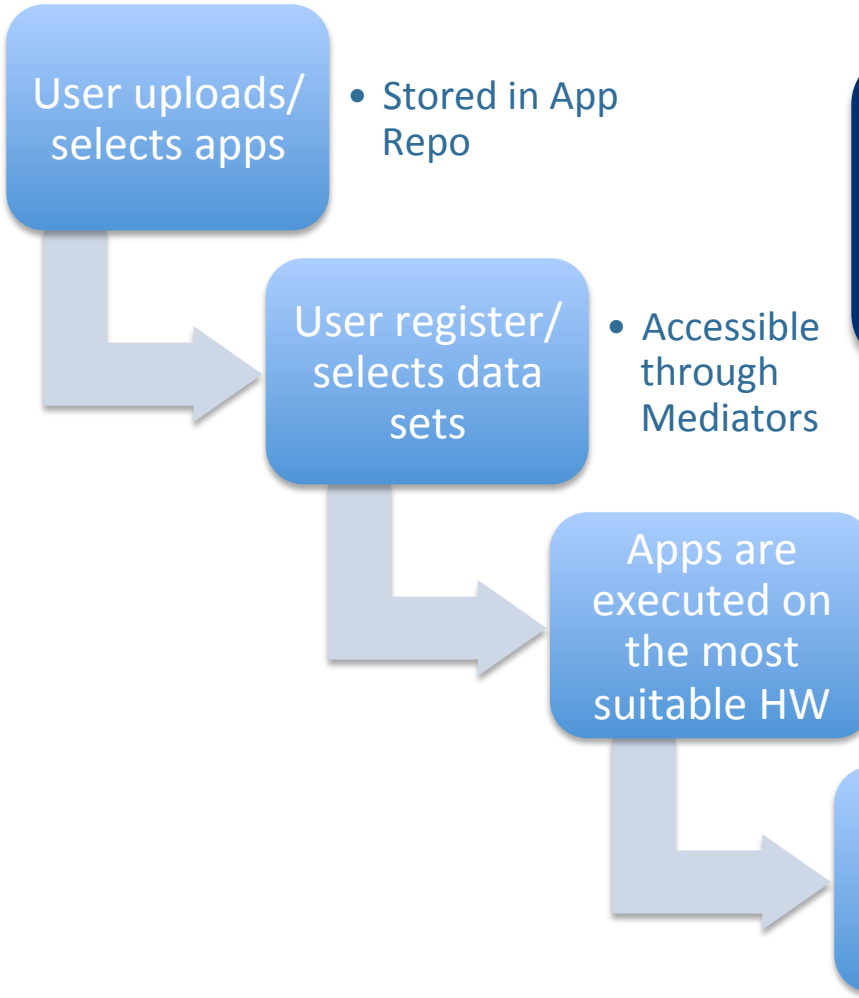
a distributed and dynamically created environment

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]

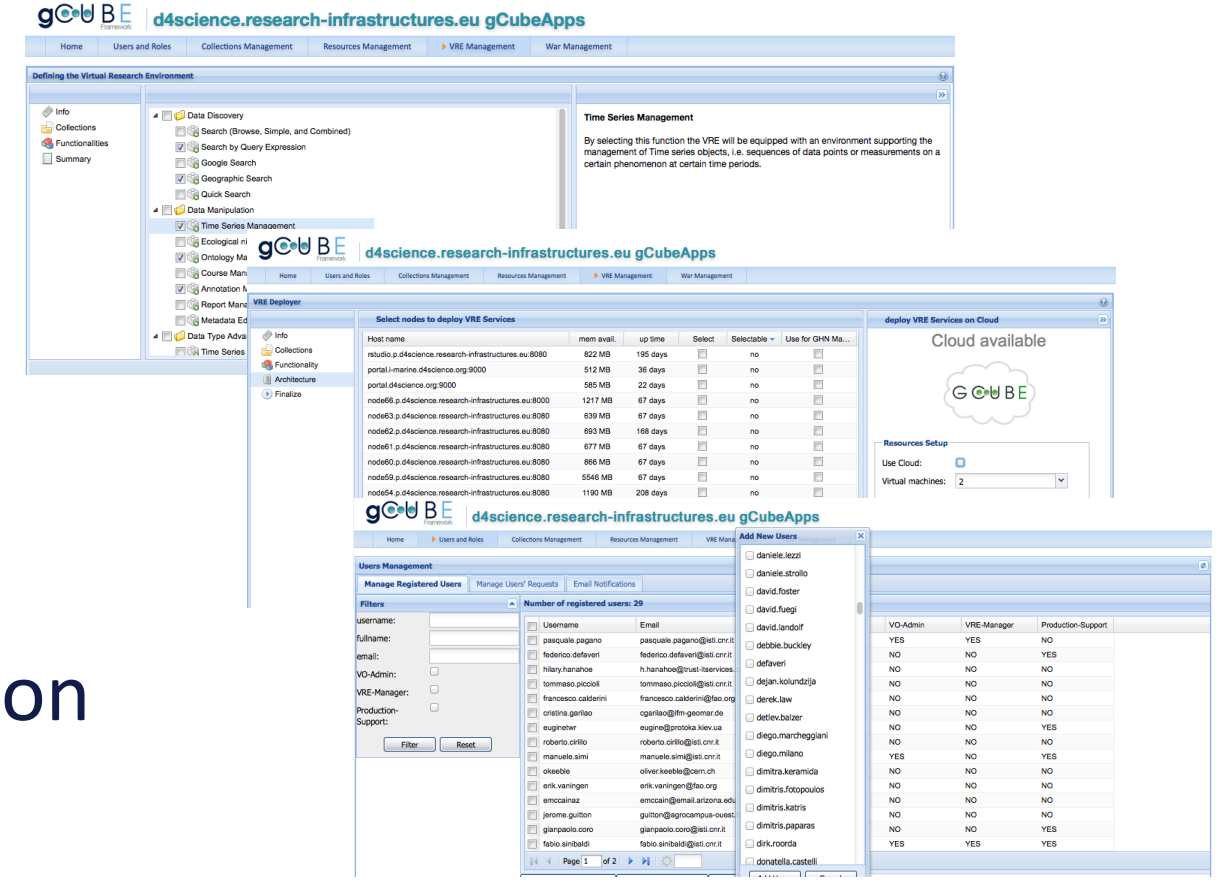


VRE is the hardware, data, and applications allocated for a timeframe to a group of people to support effective **collaborations**

gCube Enabling Layer

Virtual Research Environment [3/4]

- Cost-effective creation and management
- Definition
- Creation
- Configuration



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

Ecological Niche Modelling

gCube Ecological Niche Modelling App is designed to

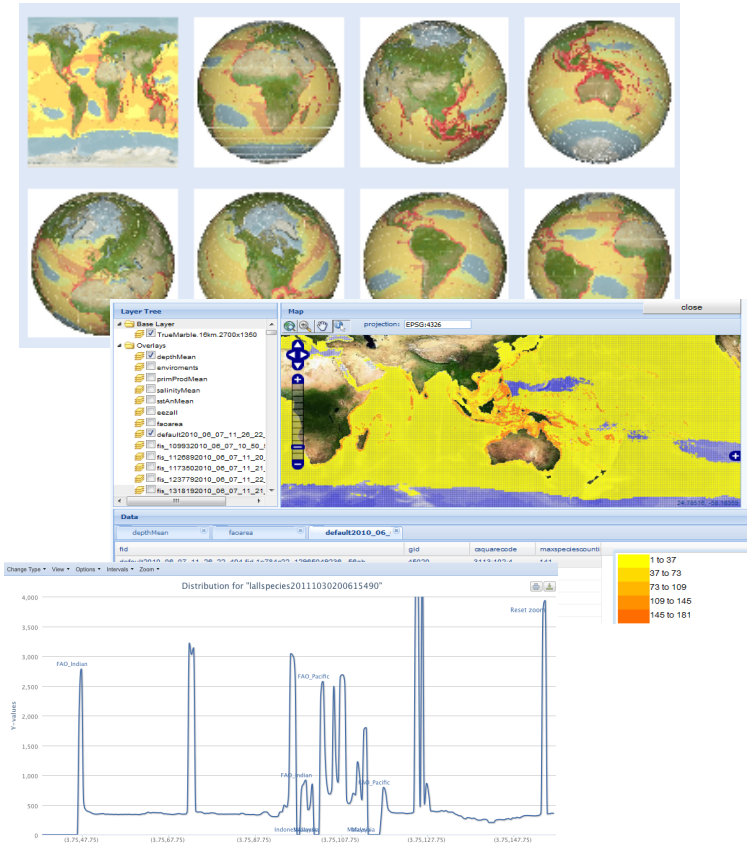
- work with **dataset versions**
- access to **external databases**
- **extensible** with predictive algorithms (aquamaps + feed-forward neural network algorithms)
- exploit **several computational back-ends** (multi-core server, distributed servers, and clouds)
- use **several storage technologies** (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

Ecological Niche Modelling

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 Harmonization

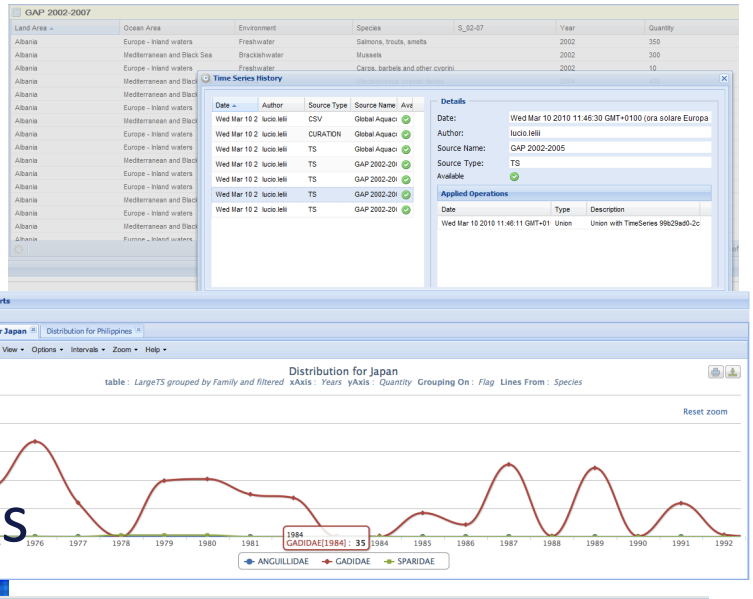
Timeseries App is designed to

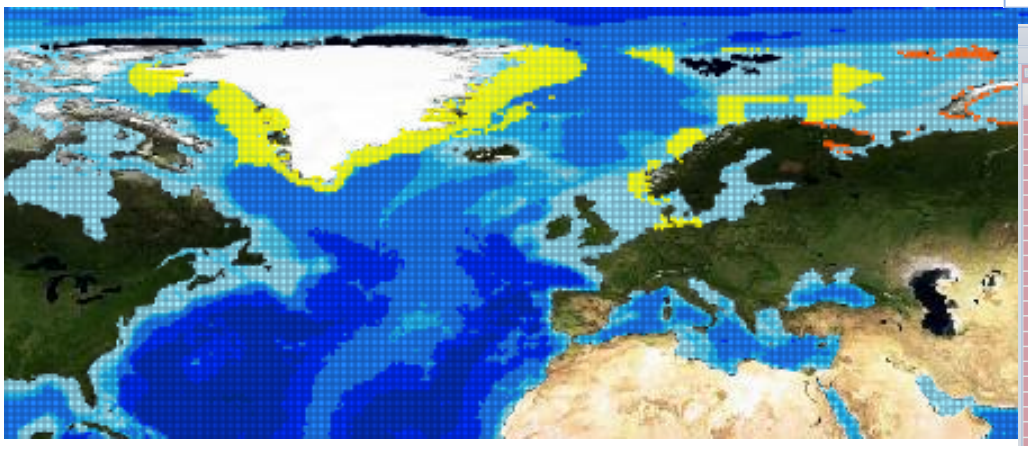
- support the **complete TS lifecycle**
- manage **multiple versions** enriched with **provenance** data
- support **validation**, **curation**, and **analysis** (filtering, grouping, and aggregation on multidimensional data)
- provide support for **data reallocation**
- supports **code list management**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

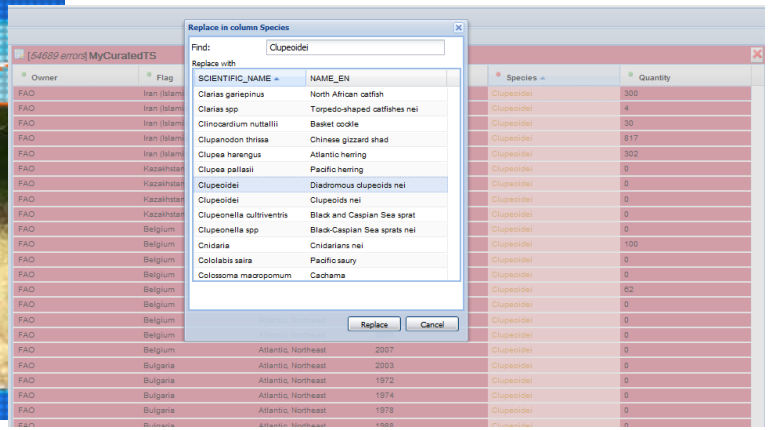
TimeSeries Harmonization

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







Summary

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

Discussion time

Thanks for your attention

Visit

www.d4science.org
www.i-marine.eu
www.eubrazilopenbio.eu

Join

gCube Apps

Enjoy

applications