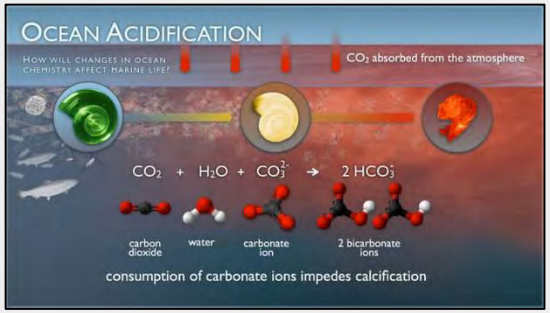
Bridging environmental **data providers** and **SeaDataNet** DIVA service within a collaborative and distributed **e-Infrastructure**

Gianpaolo Coro, Pasquale Pagano, Umberto Napolitano
ISTI-CNR, Italy

BlueBRIDGE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 675680

Context

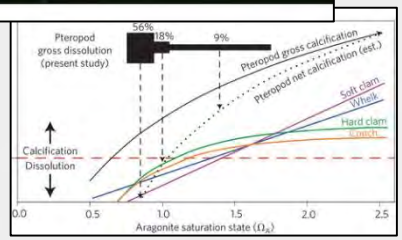
- Among its devastating effects, **ocean acidification harms** organisms whose life depends on **shells** and on **coral reefs**. It affects important marine ecosystems hosting high biodiversity and food availability



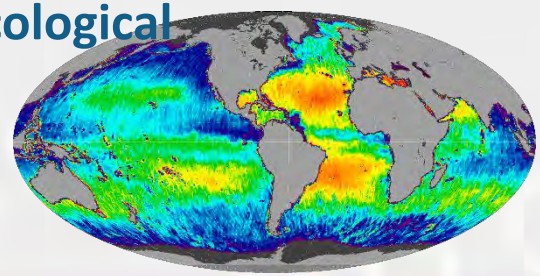
From Pmel-NOAA - http://www.pmel.noaa.gov/co2/story/Ocean+Acidification

- Ries, J. B. (2012). Oceanography: A sea butterfly flaps its wings. *Nature Geoscience*, *5*(12), 845-846.

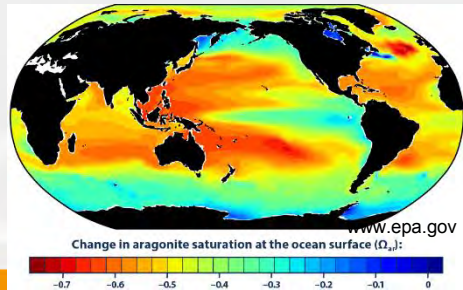
Highlighted the **decrease of calcification** of a planktonic mollusc, the sea butterfly (*Limacina helicina*). The shell-building capacity of this organism declines with **decreasing aragonite saturation**, which is due to the decrease of average pH in the seas.



- **Monitoring** and **containing ocean acidification** helps preventing coral reefs and shell-building organisms to dissolve, and thus helps **preventing ecological disasters**.



http://earthobservatory.nasa.gov/



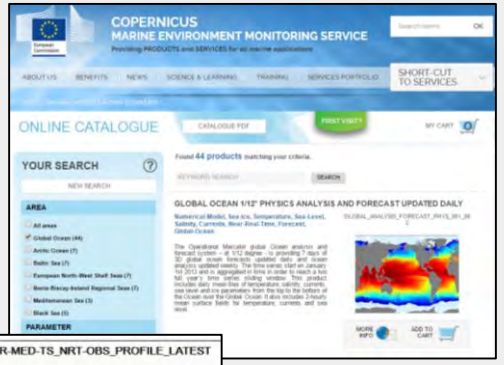
Issues

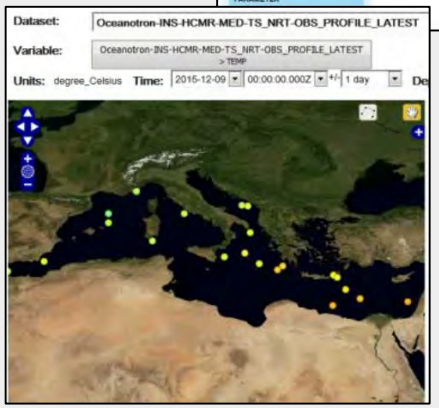
- **Environmental observations** of parameters like aragonite saturation and pH are usually available as **scattered *in situ* data**, published on restricted-access data e-Infrastructures (e.g. the Copernicus Marine Environment Monitoring Service)

- **Interpolation services exist** (e.g. the SeaDataNet Data-Interpolating Variational Analysis service, DIVA) to estimate global, uniform distributions of environmental parameters from scattered observations.

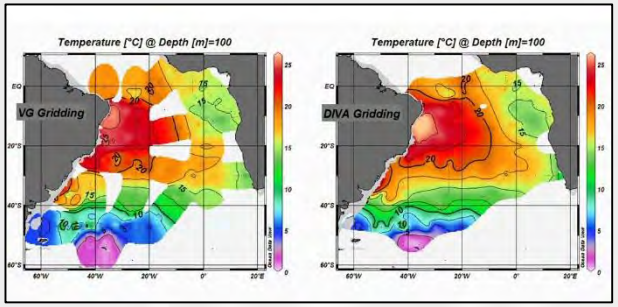
- However, these services usually **require data** to be **compliant** with a **non-standard format** and cannot accept *in situ* data formats directly.

- Furthermore, they may **benefit** from **facilities to communicate and** to **publish** their results for a large public.





DIVA



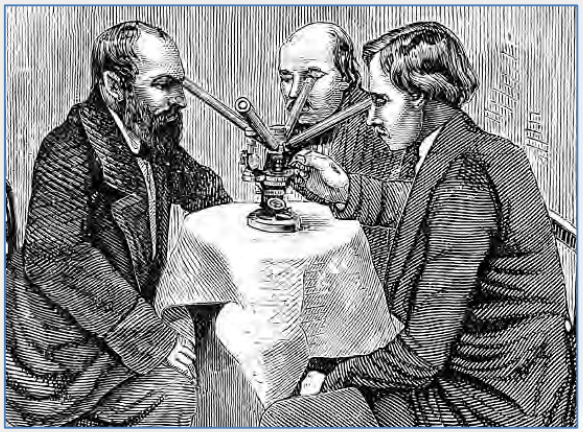
e-Infrastructures

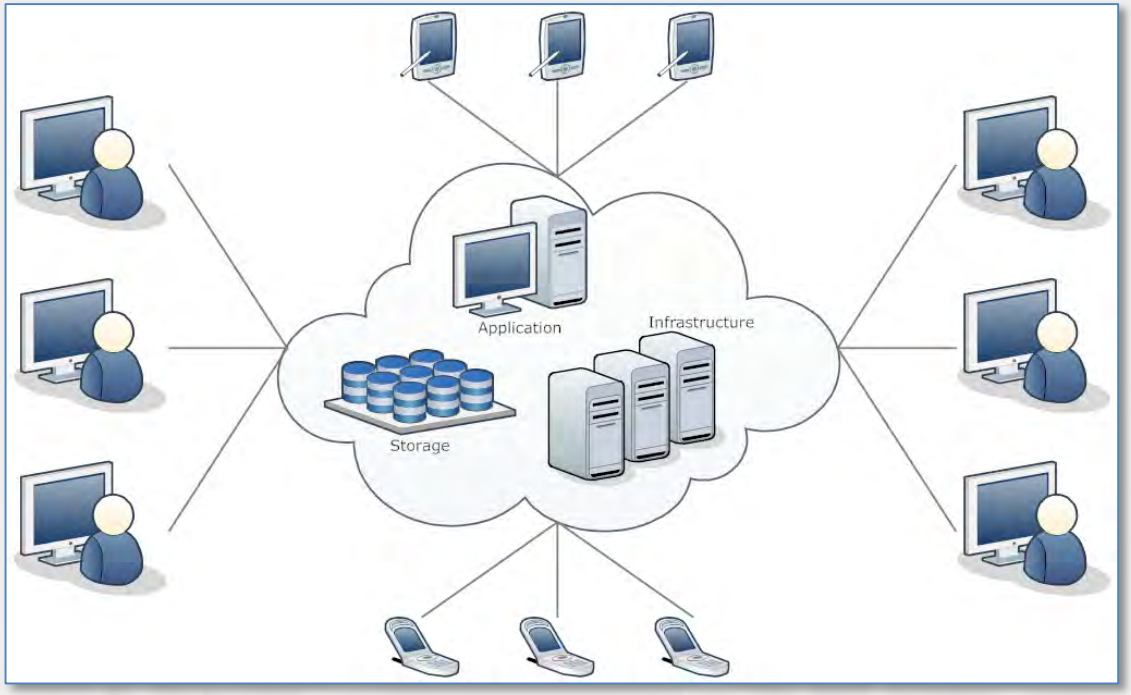
e-Infrastructures enable researchers at different locations across the world to **collaborate** in the context of their home institutions or in national or multinational scientific initiatives.

- People can work together having shared access to unique or distributed scientific facilities (including data, instruments, computing and communications).

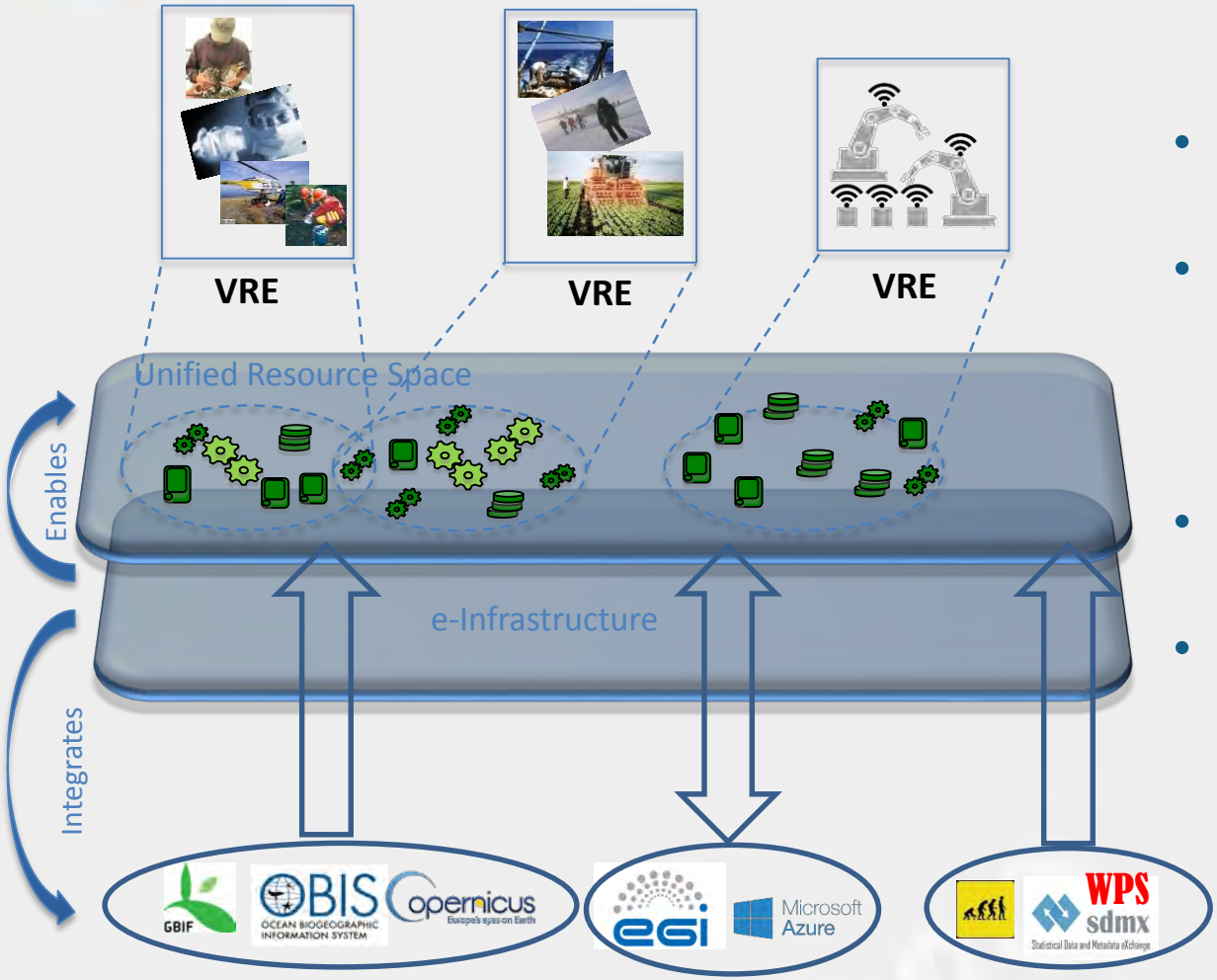
Examples:

Belief, http://www.beliefproject.org/
OpenAire, http://www.openaire.eu/
i-Marine, http://www.i-marine.eu/
EU-Brazil OpenBio,
http://www.eubrazilopenbio.eu/





Virtual Research Environments

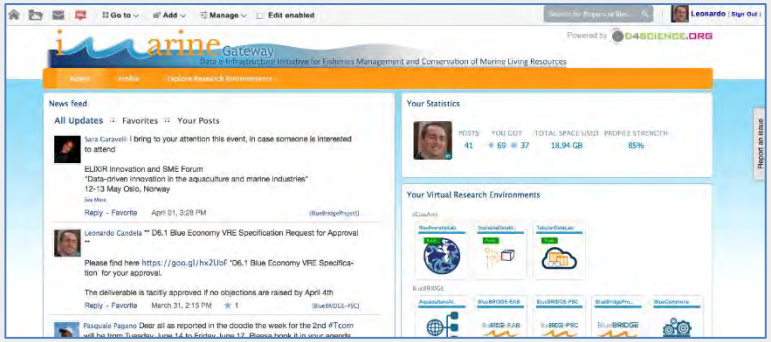


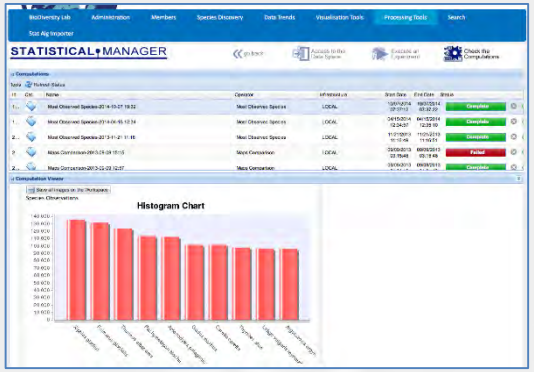
- Define sub-communities
- Allow temporary dedicated assignment of computational, storage, and data resources
- Manage policies
- Support data and information sharing

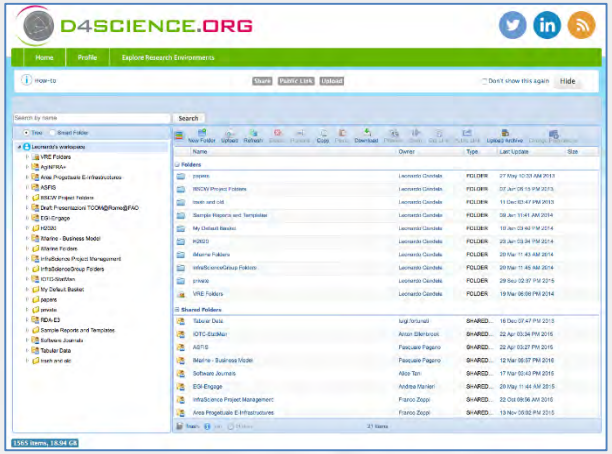
Virtual Research Environments

Innovative, web-based, community-oriented, comprehensive, flexible, and secure working environments.

- Communities are provided with applications to interact with the VRE services
- Client services are provided both with APIs (Java, R) and simple HTTP-REST interfaces







New Science Paradigms

- **Open Science**: make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.

  Keywords: Open Access, Open research, Open Notebook Science

- **E-Science**: computationally intensive science is carried out in highly distributed network environments that use large data sets and require distributed computing and collaborative tools.

  Keywords: Provenance of the scientific process, Scientific workflows

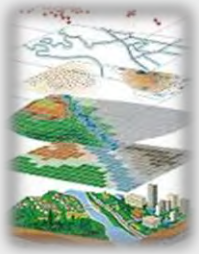
- **Science 2.0:** process and publish large data sets using a collaborative approach. Share from raw data to experimental results and processes. Support collaborative experiments and Reproducibility-Repeatability-Reusability (R-R-R) of Science.

  Keywords: collaborative and repeatable Science

BlueBRIDGE Facilities: Overview

Databases

Cloud storage



Geospatial data

Storage

Data management

Metadata generation and management

Harmonisation

Sharing

Cloud computing

Elastic resources assignment

Multi-platform: R, Java, Fortran

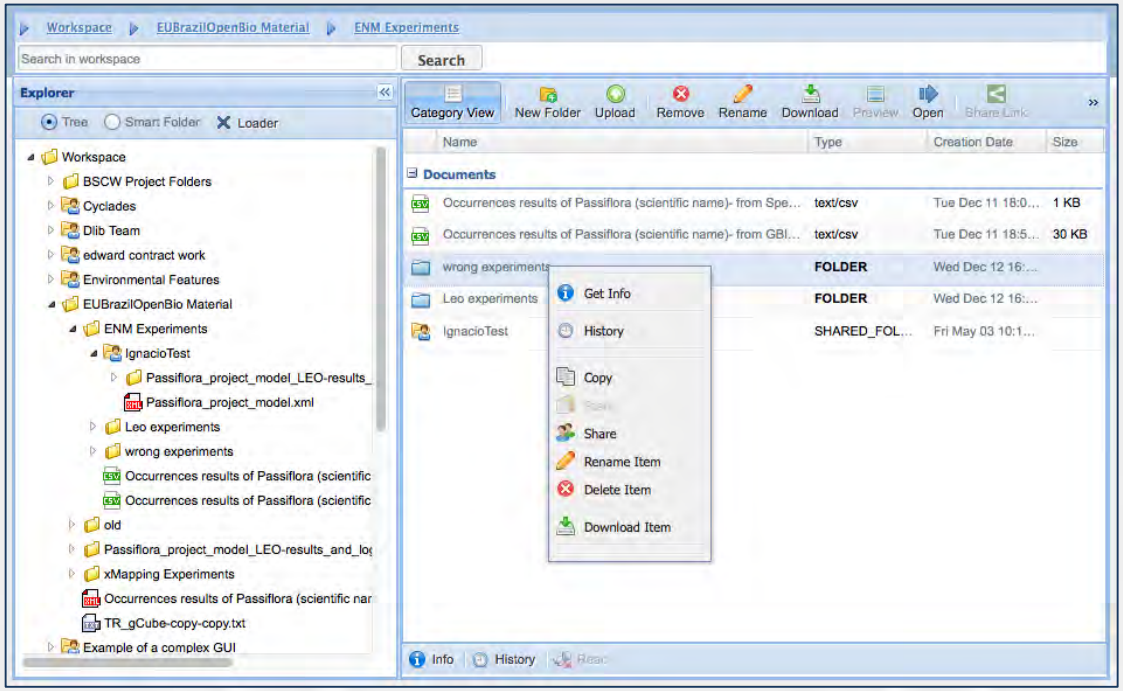
Processing

BlueBRIDGE VREs:
The Workspace – an online files storage system

A free-of-use folder-based file system allows managing and sharing information objects.

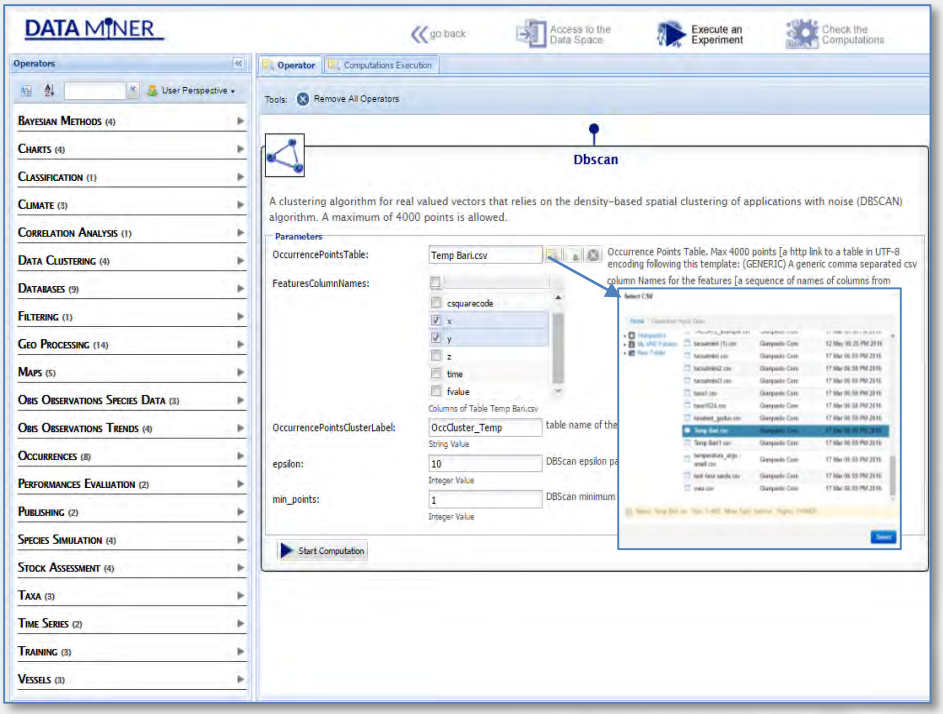
Information objects can be

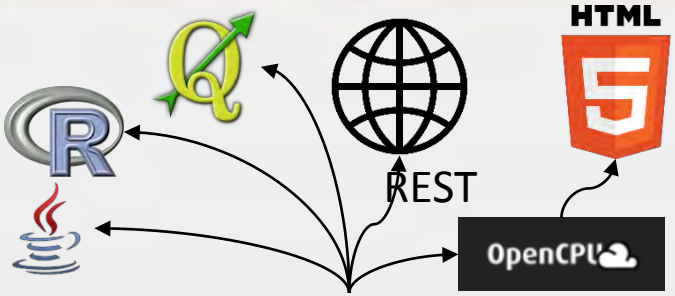
- files, dataset, workflows, experiments, etc.
- organized into folders
- shared
- disseminated via public URLs



Cloud Computing Platform

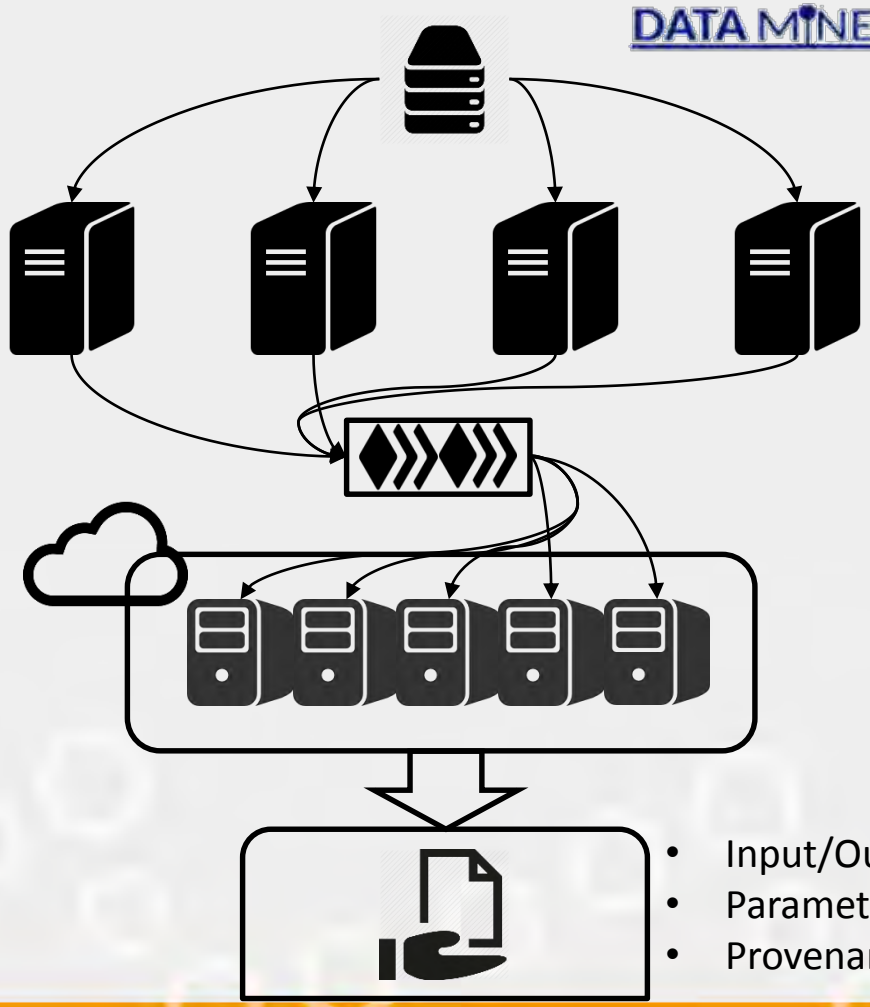
- Experiments on Big Data
- Sharing inputs and results
- Save the *provenance* of experiments
- Supports R-R-R of experiments





WPS

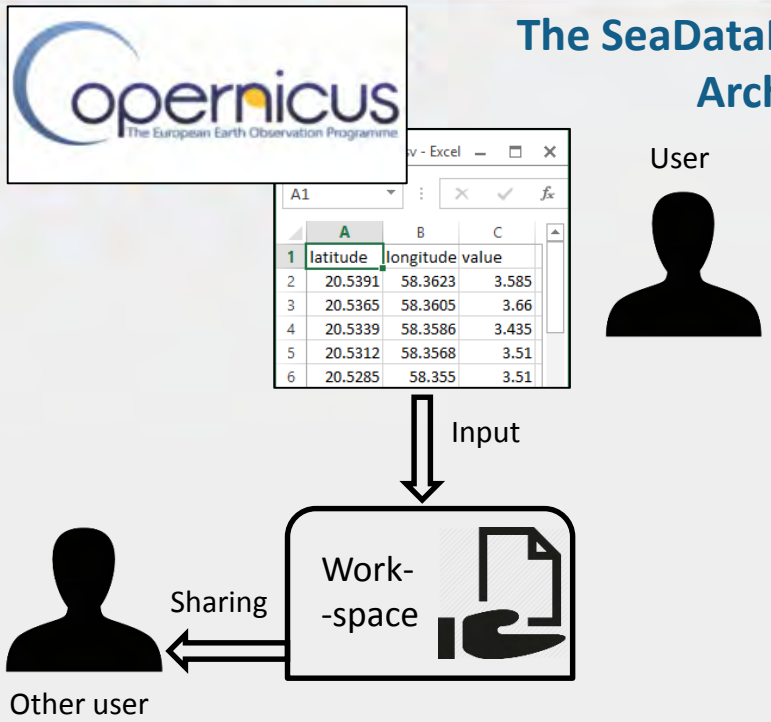
DATA MINER



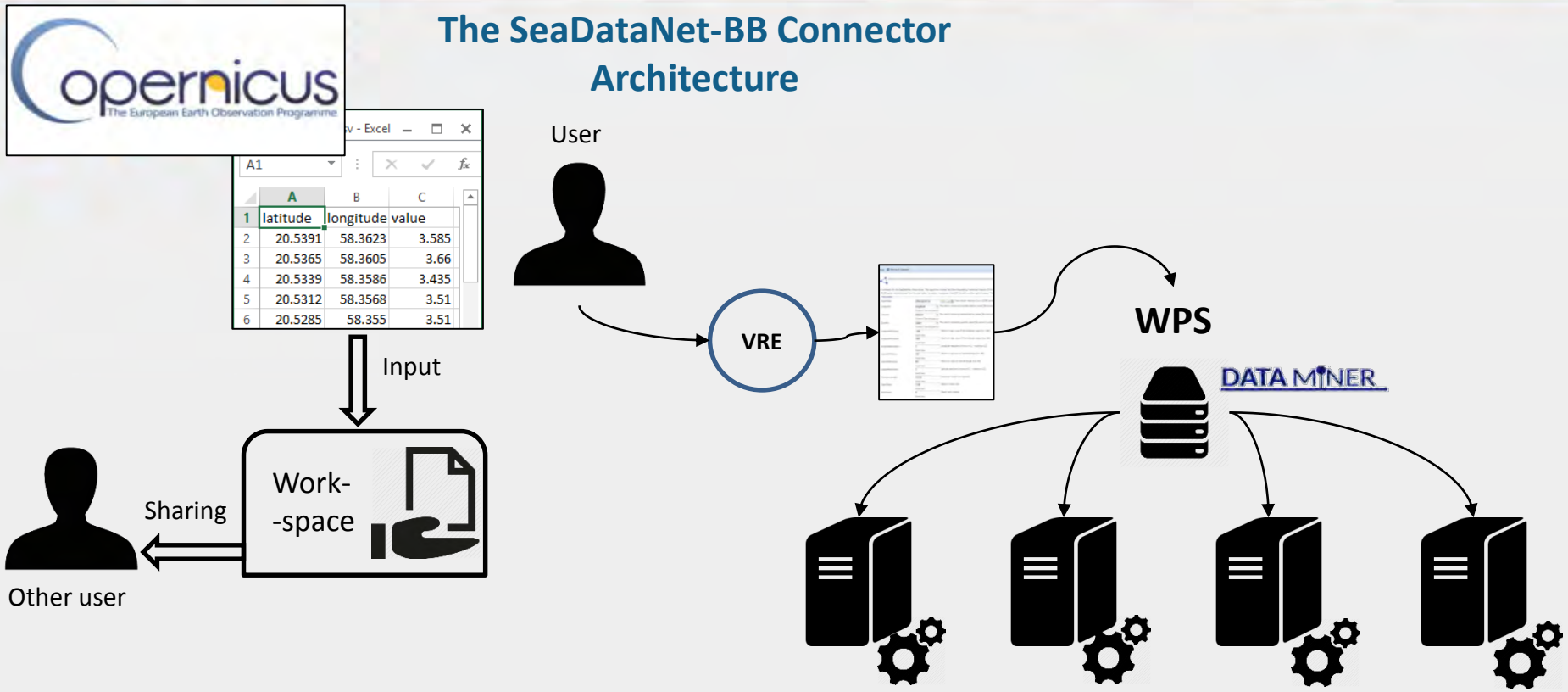
- Input/Out
- Parameters
- Provenance

Proposed solution

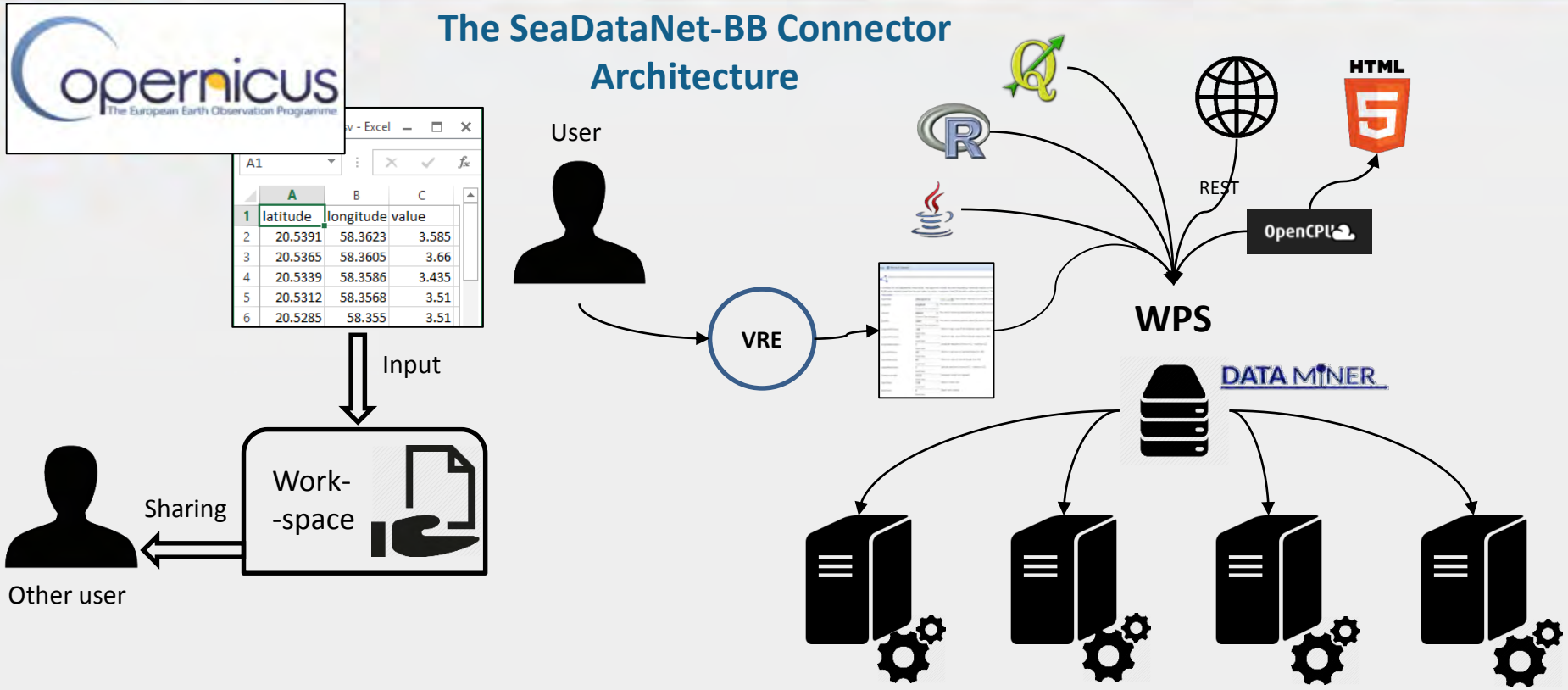
The SeaDataNet-BB Connector Architecture



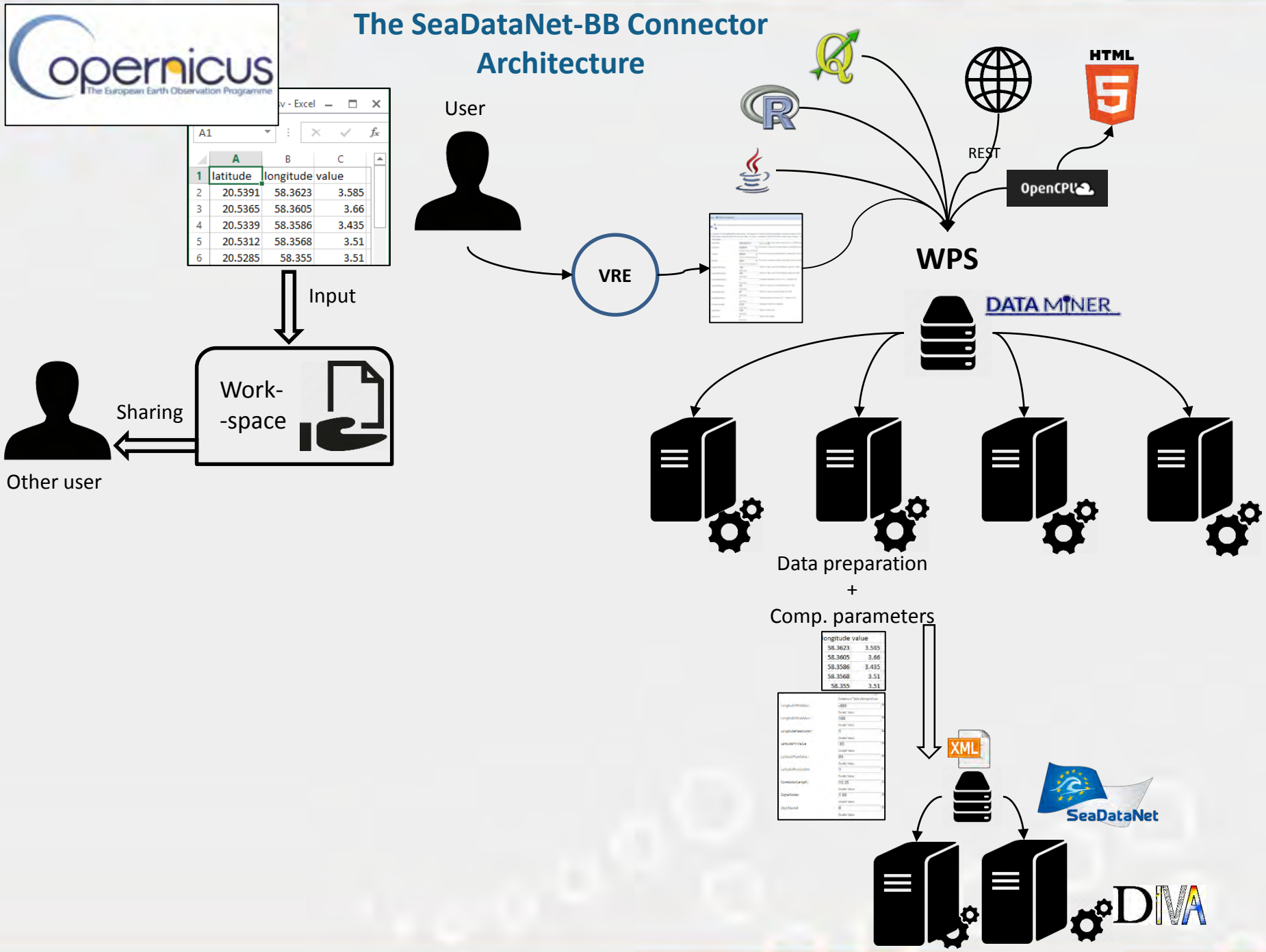
The SeaDataNet-BB Connector Architecture



The SeaDataNet-BB Connector Architecture

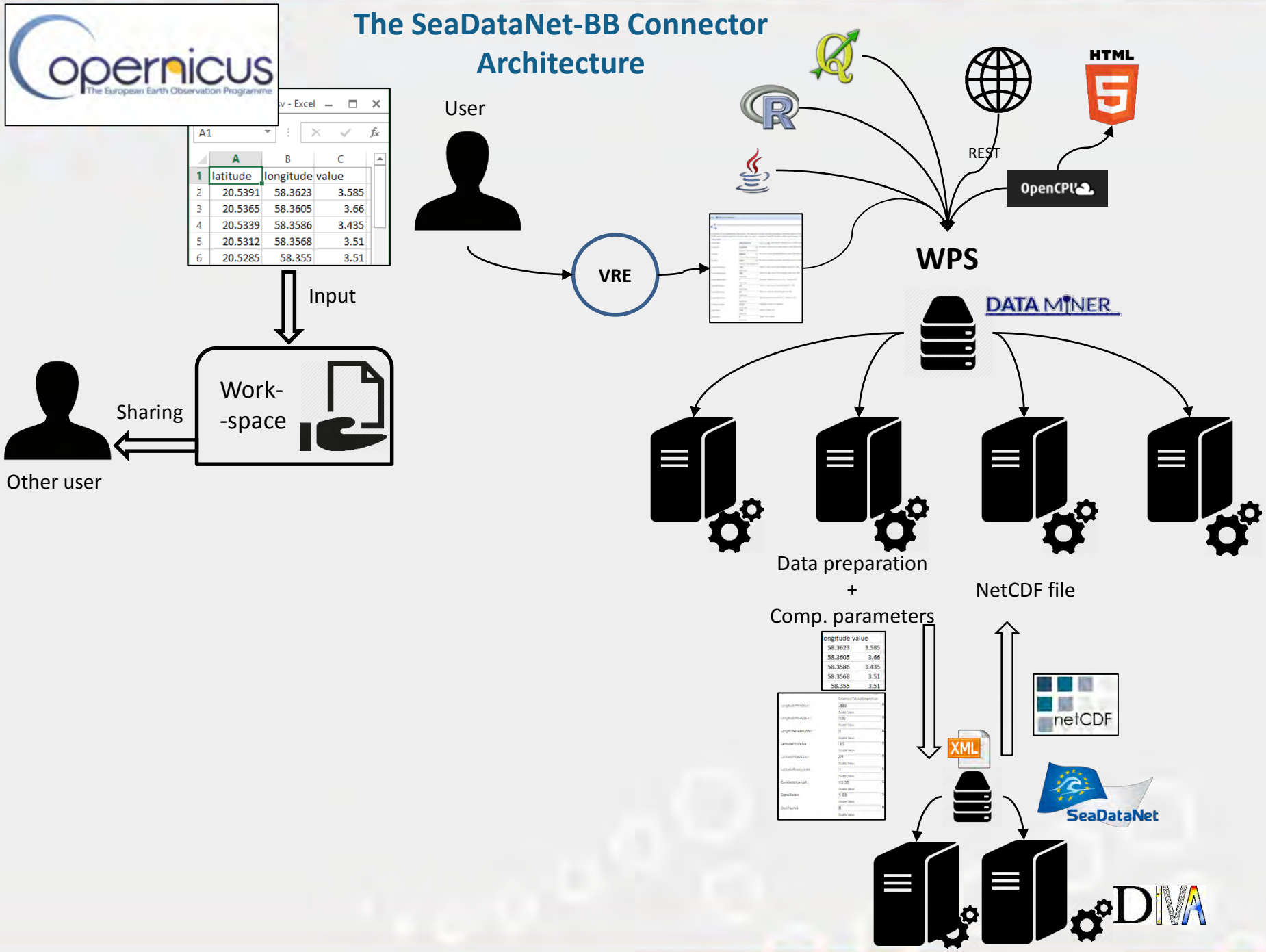


The SeaDataNet-BB Connector Architecture

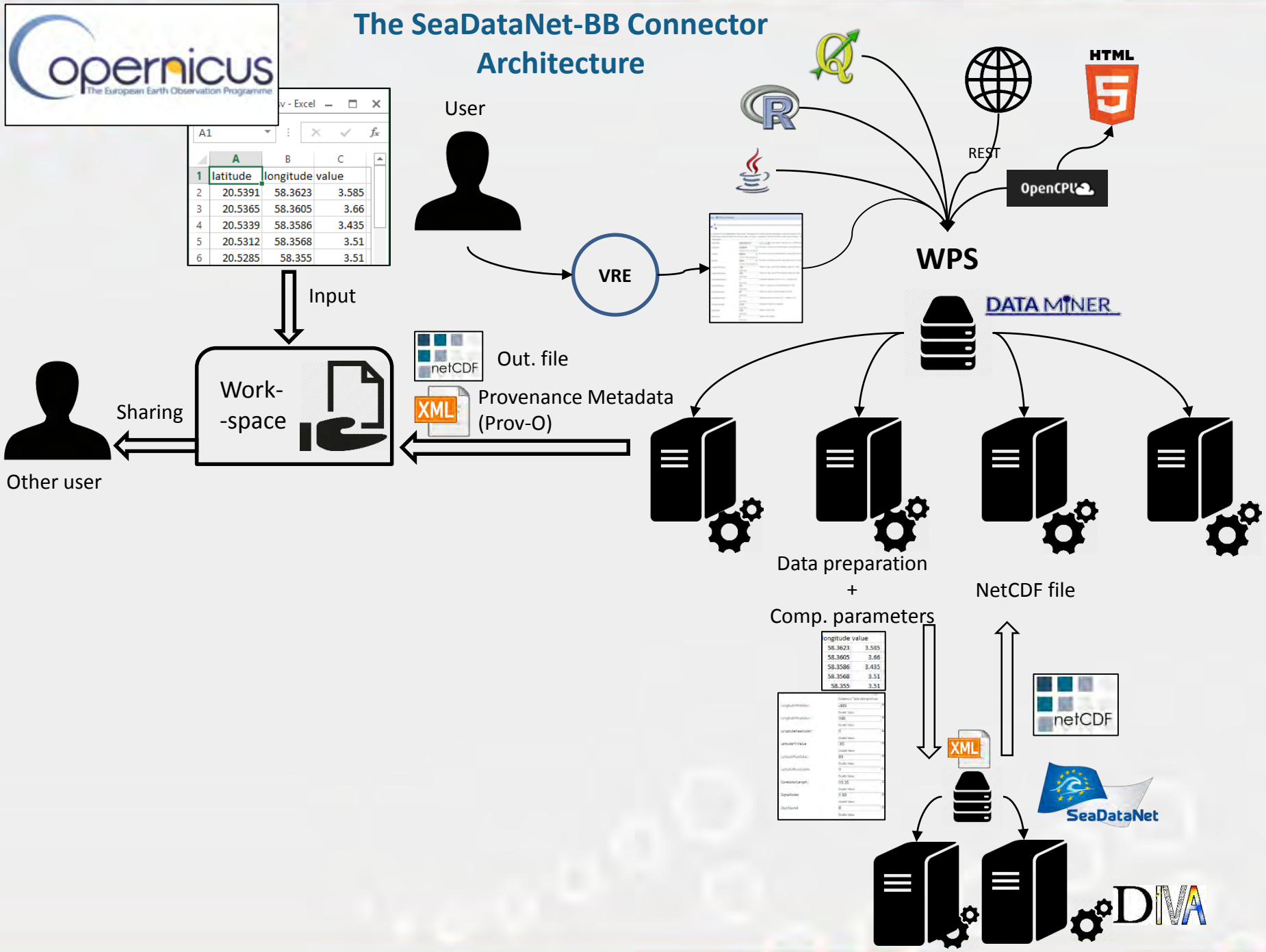


	A	B	C
1	latitude	longitude	value
2	20.5391	58.3623	3.585
3	20.5365	58.3605	3.66
4	20.5339	58.3586	3.435
5	20.5312	58.3568	3.51
6	20.5285	58.355	3.51

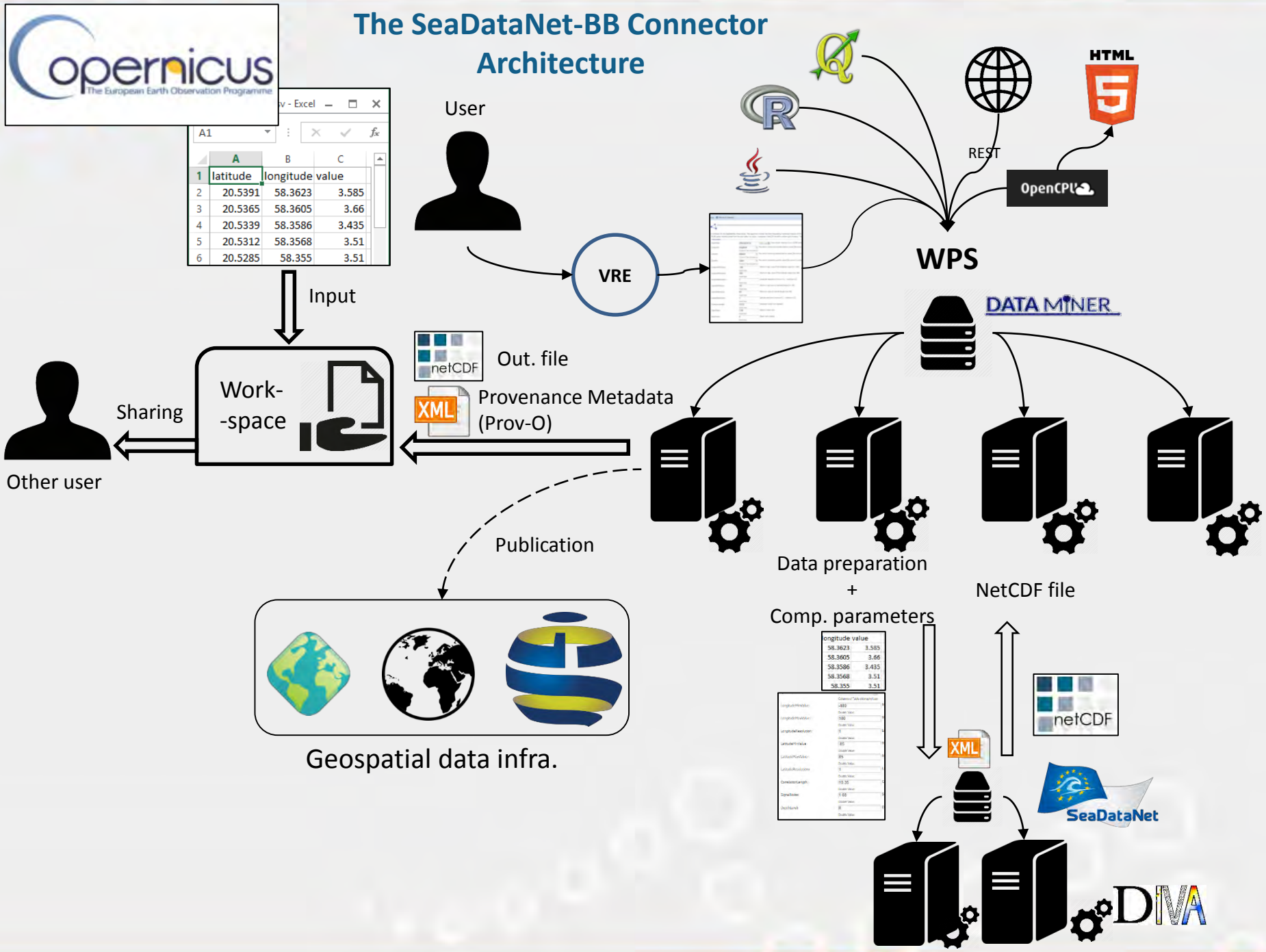
The SeaDataNet-BB Connector Architecture



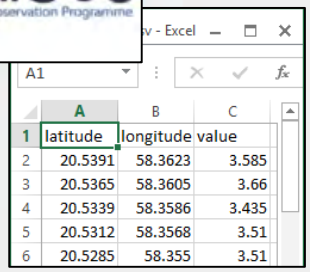
The SeaDataNet-BB Connector Architecture



The SeaDataNet-BB Connector Architecture



The SeaDataNet-BB Connector Architecture



User

VRE

REST

HTML

OpenCPU

WPS

DATA MINER

Input

Work-
-space

Sharing

Other user

netCDF

Out. file

XML

Provenance Metadata
(Prov-O)

Publication

Data preparation
+
Comp. parameters

NetCDF file

Geospatial data infra.

VRE

Visualisation

WMS
WCS
GeoTiff
NetCDF
OPeNDAP

OGC Standards

SeaDataNet

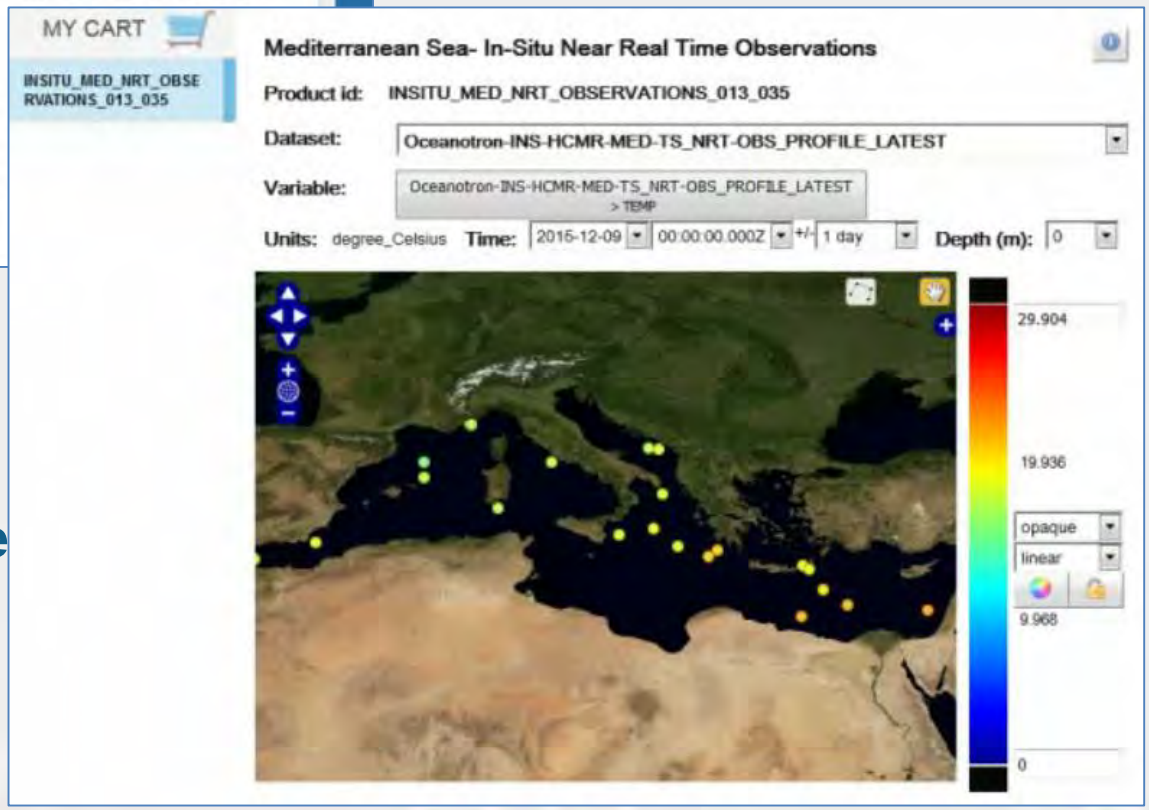
DIVA

BlueBRIDGE

Example

Demo Video : http://goo.gl/yX3kww

STEP 1: get *in situ* or satellite data



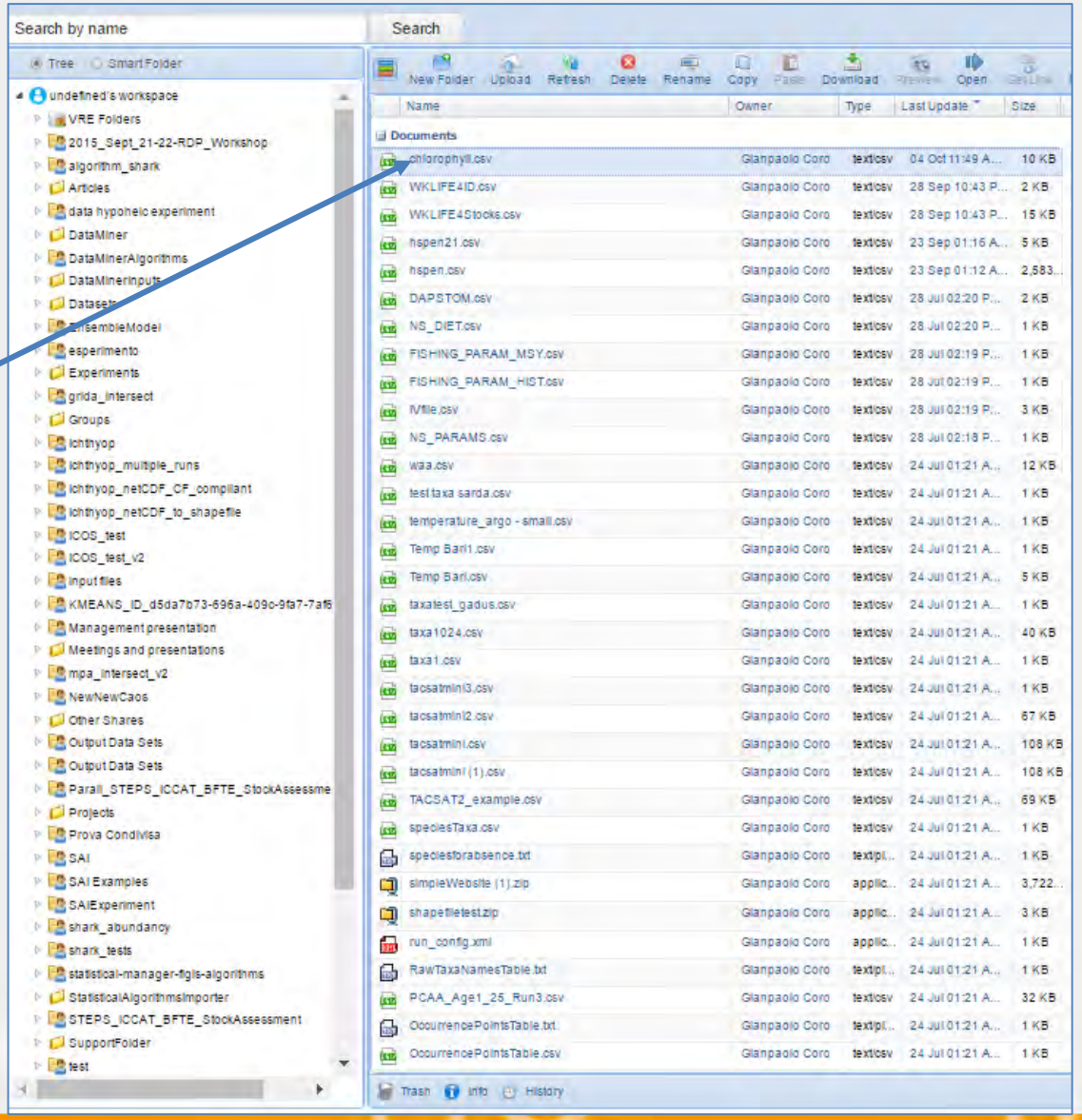
Example: **Copernicus Marine Environment Monitoring Service**

STEP 2: extract data in tabular format and upload them on the Workspace

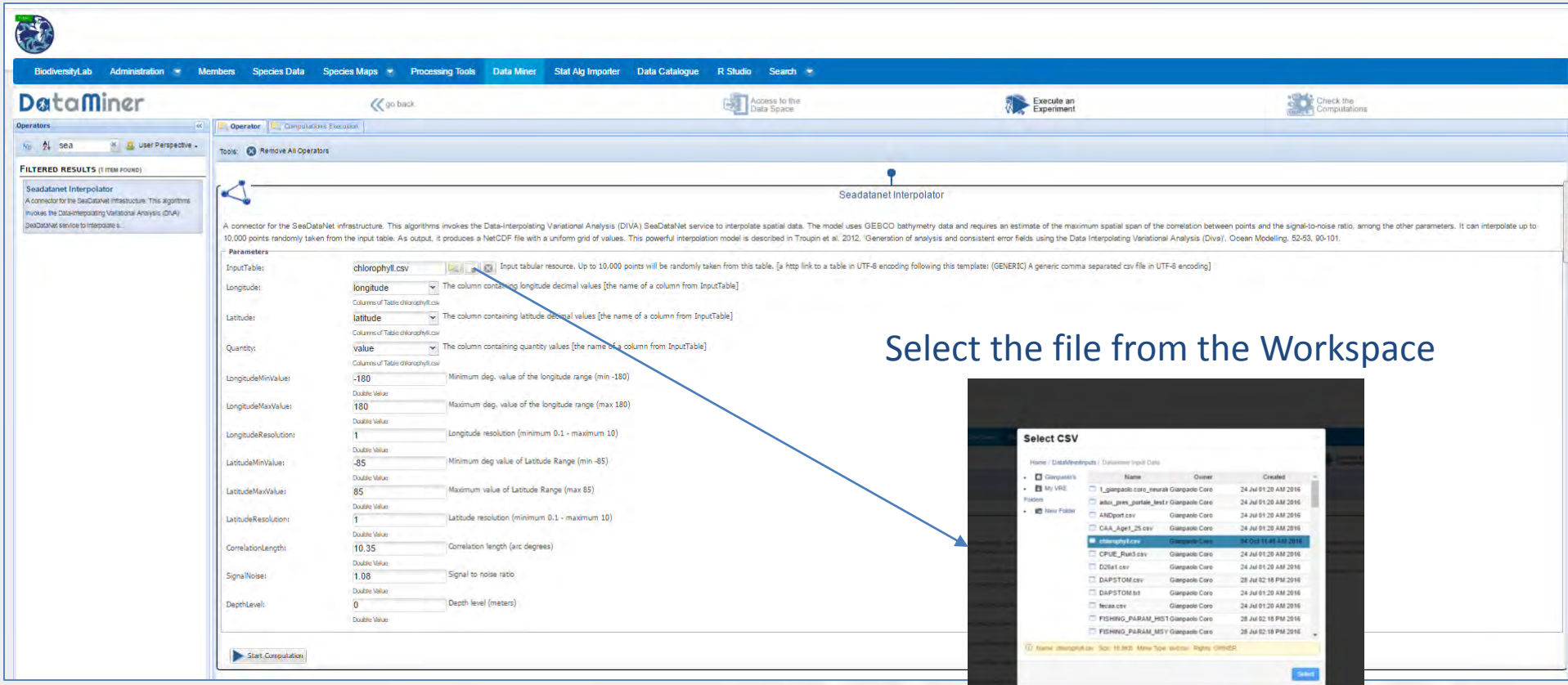
chlorophyll.csv - Excel

	A	B	C
1	latitude	longitude	value
2	20.5391	58.3623	3.585
3	20.5365	58.3605	3.66
4	20.5339	58.3586	3.435
5	20.5312	58.3568	3.51
6	20.5285	58.355	3.51
7	20.5258	58.3531	3.51
8	20.5232	58.3513	3.585
9	20.5204	58.3494	3.66
10	20.5177	58.3476	3.66
11	20.5151	58.3458	3.735
12	20.5124	58.344	3.81
13	20.5097	58.3421	3.585
14	20.5071	58.3403	3.585
15	20.5044	58.3385	3.36
16	20.5017	58.3367	3.36
17	20.499	58.3349	3.285
18	20.4964	58.333	3.285
19	20.4936	58.3312	3.36
20	20.491	58.3294	3.435
21	20.4883	58.3276	3.36
22	20.4856	58.3257	3.585
23	20.483	58.3239	3.51
24	20.4803	58.3221	3.435
25	20.4775	58.3203	3.585
26	20.4748	58.3184	3.51
27	20.4721	58.3166	3.585
28	20.4695	58.3148	3.585
29	20.4668	58.313	3.585
30	20.4641	58.3111	3.585
31	20.4614	58.3093	3.735
32	20.4587	58.3074	3.435
33	20.4561	58.3056	3.51
34	20.4534	58.3037	3.435
35	20.4508	58.3019	3.585
36	20.4481	58.3	3.435
37	20.4455	58.2982	3.51
38	20.4429	58.2964	3.585
39	20.4402	58.2945	3.435
40	20.4375	58.2927	3.66
41	20.4349	58.2908	3.435
42	20.4323	58.2889	3.51
43	20.4297	58.2871	3.435
44	20.4272	58.2852	3.435
45	20.4245	58.2834	3.585

Drag and drop



STEP 3: setup the experiment through the DataMiner interface



Select the file from the Workspace

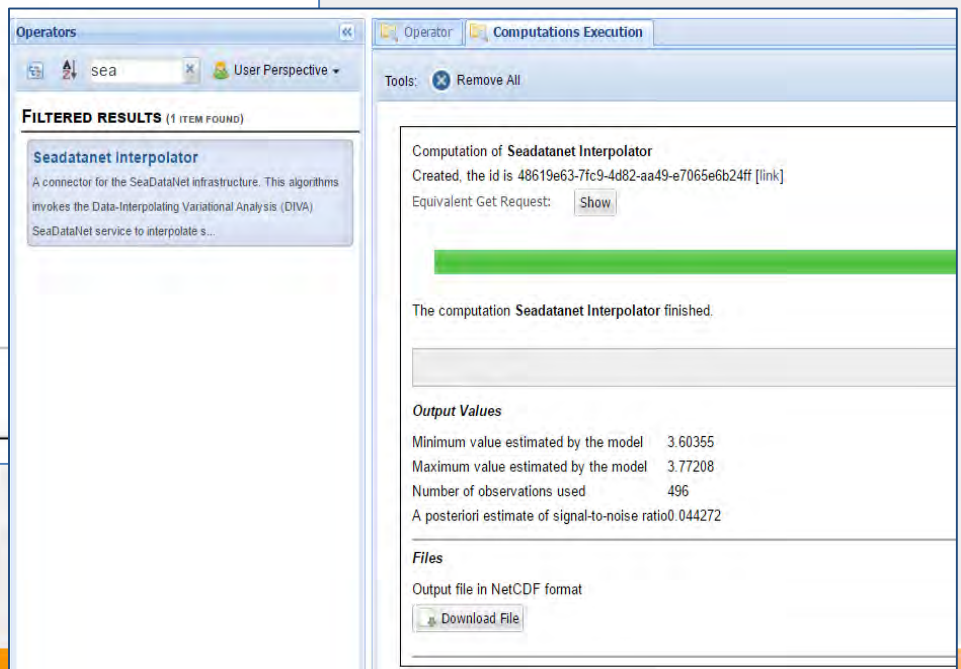
STEP 4: execute the experiment and get the results

A connector for the SeaDataNet infrastructure. This algorithms invokes the Data-Interpolating Variational Analysis (DIVA) SeaDataNet service to interpol
10,000 points randomly taken from the input table. As output, it produces a NetCDF file with a uniform grid of values. This powerful interpolation model is

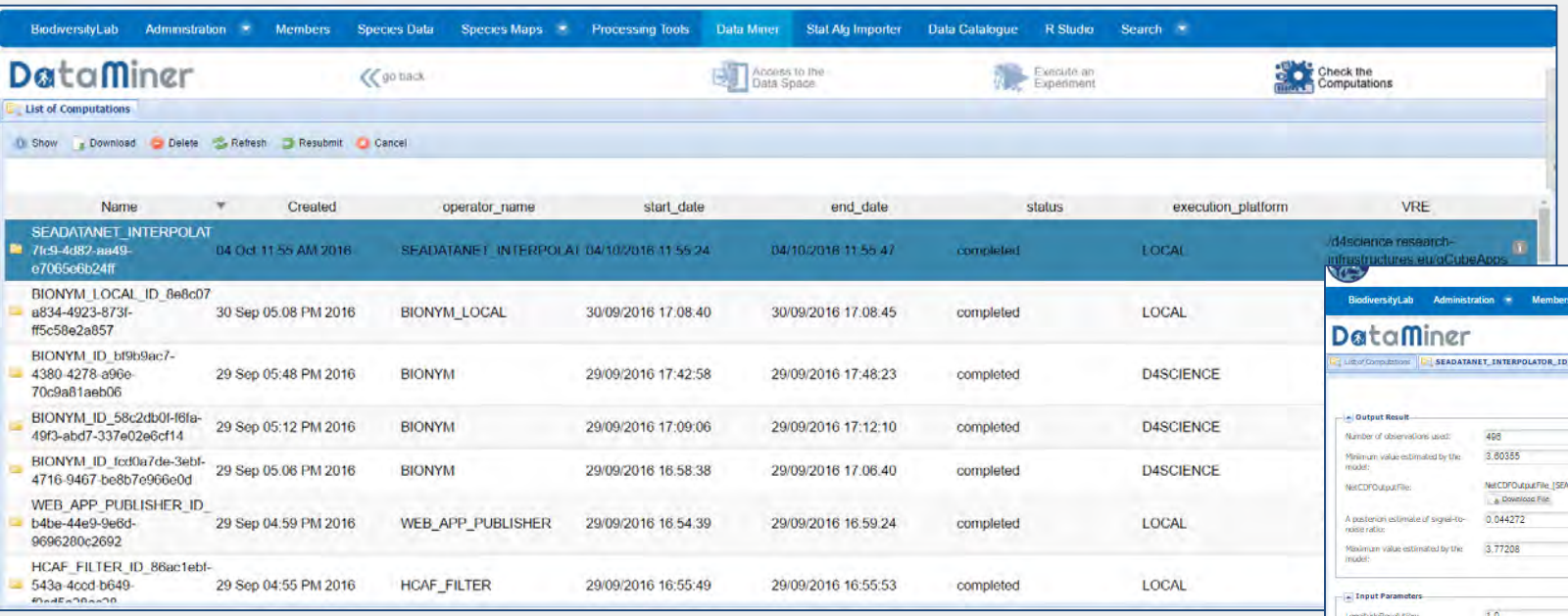
Parameters

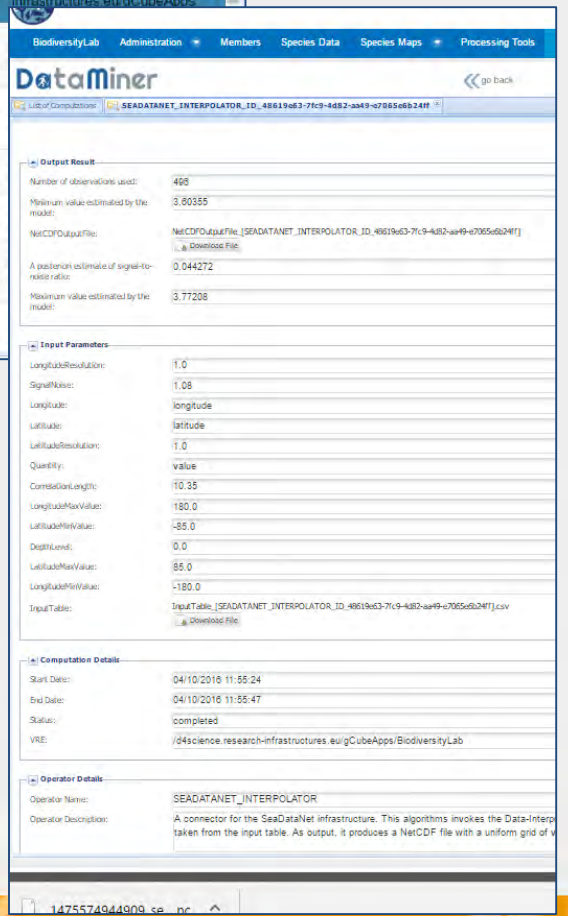
InputTable:	chlorophyll.csv	Input tabular resource. Up to 10,000 points will be randomly taken from thi
Longitude:	longitude Columns of Table chlorophyll.csv	The column containing longitude decimal values [the name of a column from InputTable]
Latitude:	latitude Columns of Table chlorophyll.csv	The column containing latitude decimal values [the name of a column from InputTable]
Quantity:	value Columns of Table chlorophyll.csv	The column containing quantity values [the name of a column from InputTable]
LongitudeMinValue:	-180 Double Value	Minimum deg. value of the longitude range (min -180)
LongitudeMaxValue:	180 Double Value	Maximum deg. value of the longitude range (max 180)
LongitudeResolution:	1 Double Value	Longitude resolution (minimum 0.1 - maximum 10)
LatitudeMinValue:	-85 Double Value	Minimum deg value of Latitude Range (min -85)
LatitudeMaxValue:	85 Double Value	Maximum value of Latitude Range (max 85)
LatitudeResolution:	1 Double Value	Latitude resolution (minimum 0.1 - maximum 10)
CorrelationLength:	10.35 Double Value	Correlation length (arc degrees)
SignalNoise:	1.08 Double Value	Signal to noise ratio
DepthLevel:	0 Double Value	Depth level (meters)

Start Computation



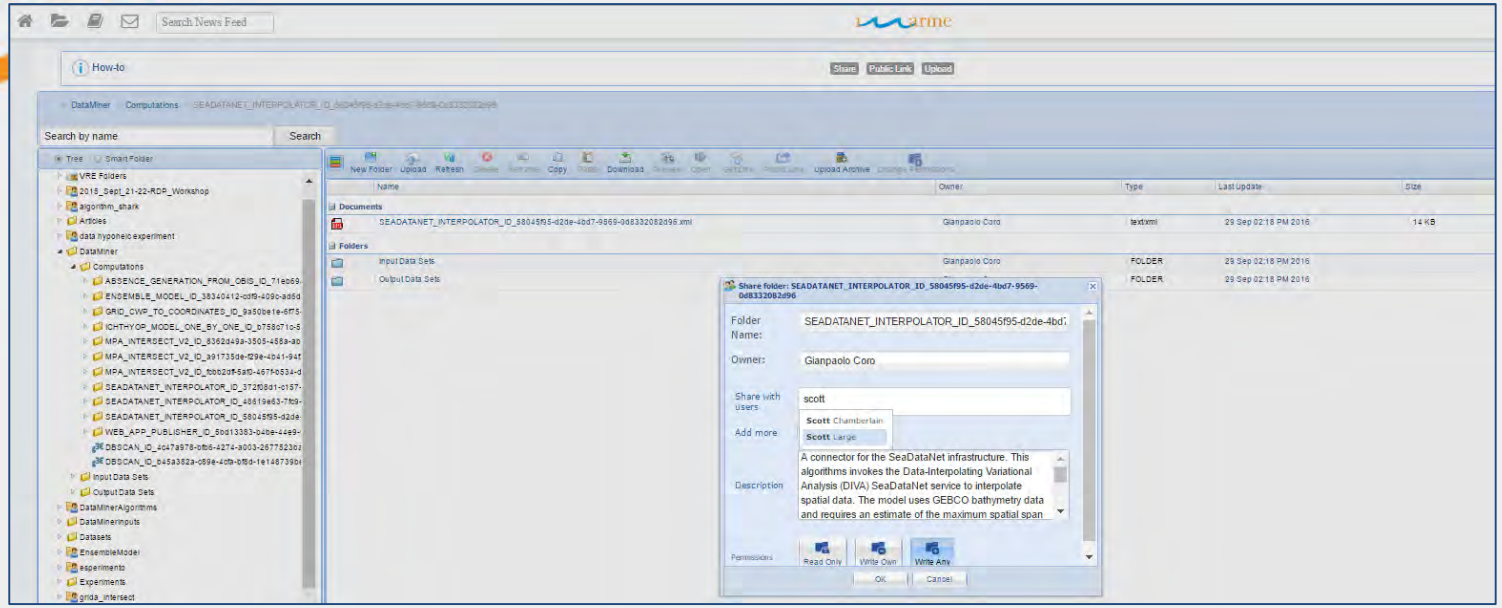
STEP 5: get the results and inspect the metadata



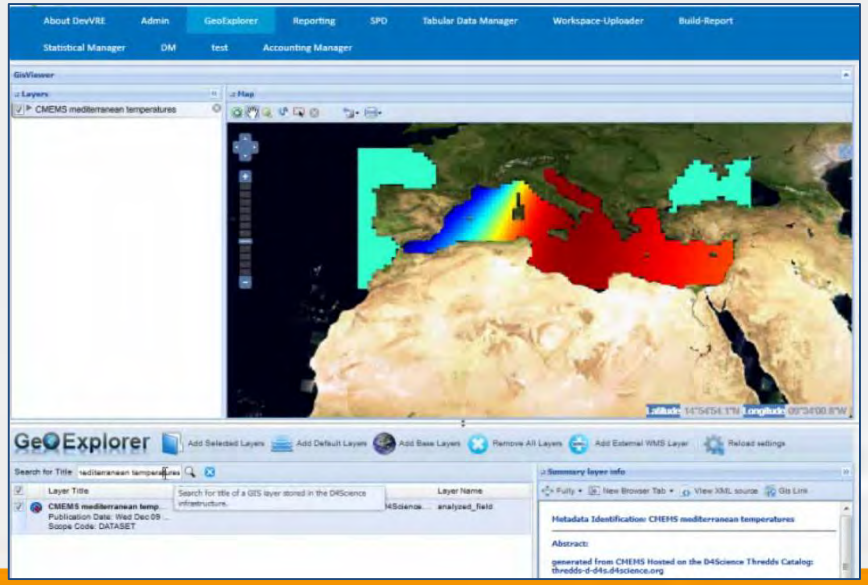


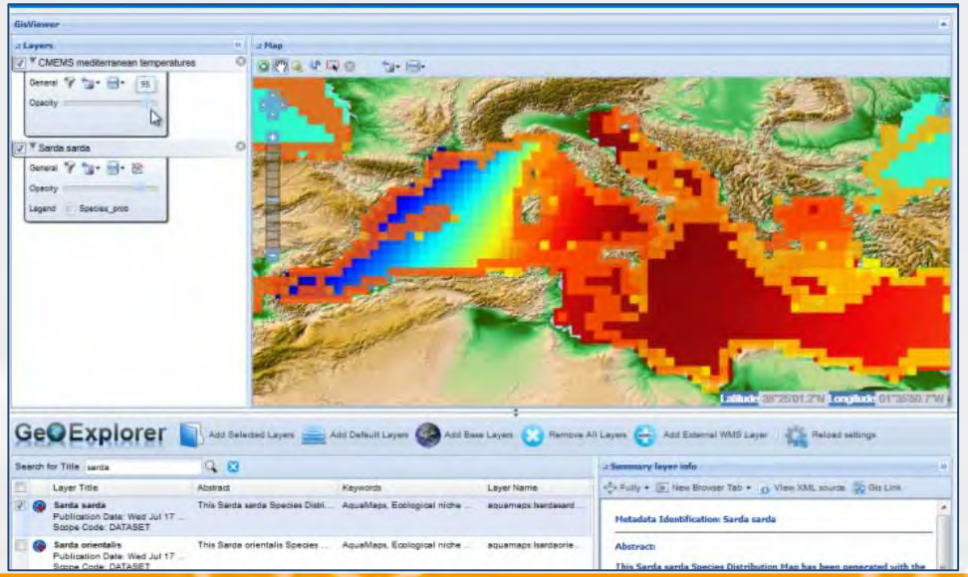
The PROV-O ontology is used to represent the metadata of the computation

STEP 6: share the results with other colleagues



STEP 7: publish the output as a map – overlay with other maps in the infrastructure



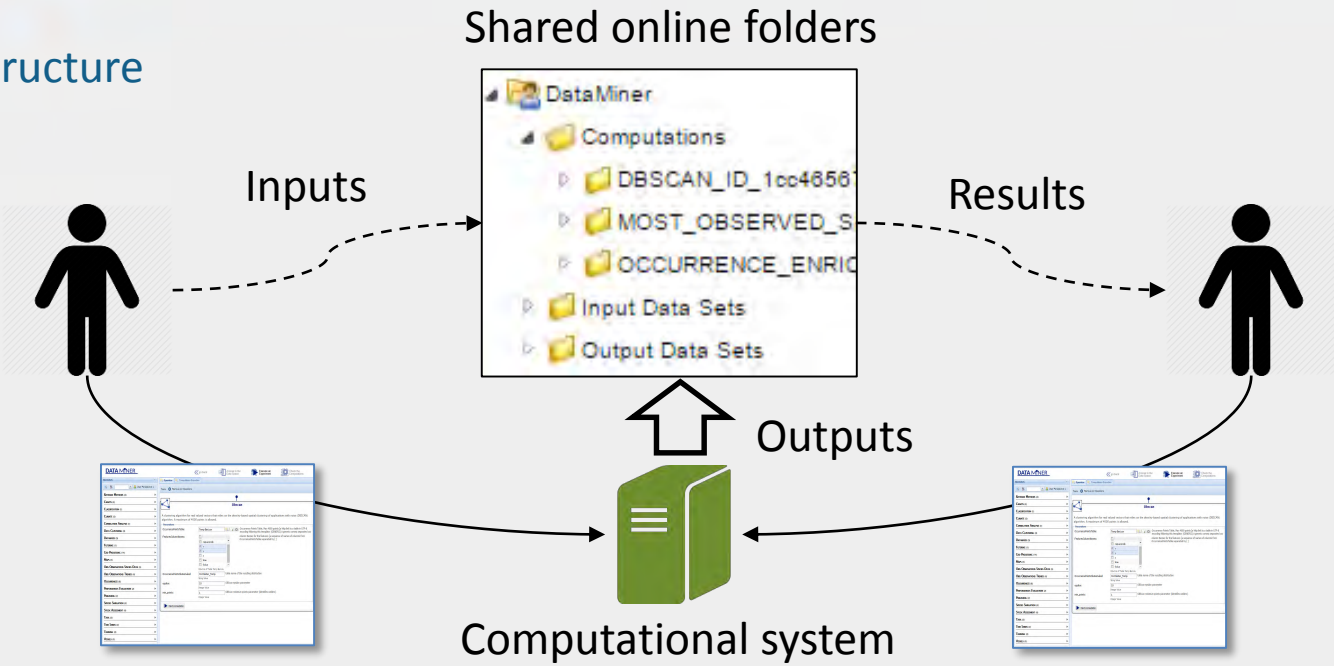


Benefits of the integration

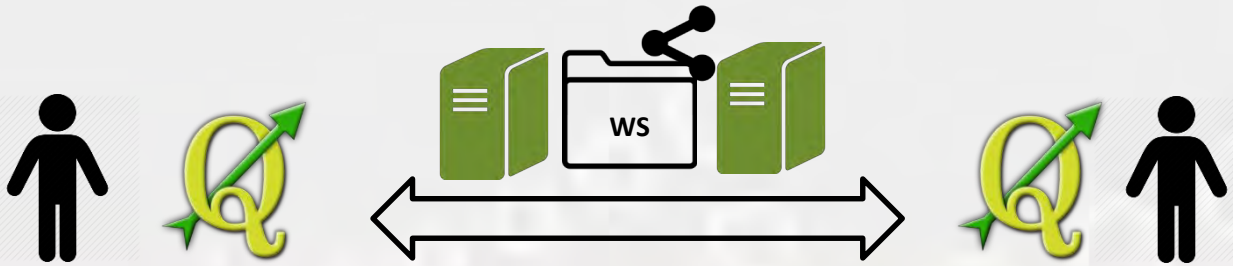
- ✓ The process can accessed through VREs
- ✓ Invoked via communication standards
- ✓ Available for the BlueBRIDGE community
  (stock assessment, MPAs, aquaculture)
- ✓ Automatic Web interface for the process
- ✓ Provenance management
- ✓ Storage of results on a high-availability system
- ✓ Collaboration and sharing
- ✓ R-R-R also from other software, e.g. QGIS

Collaborative experiments

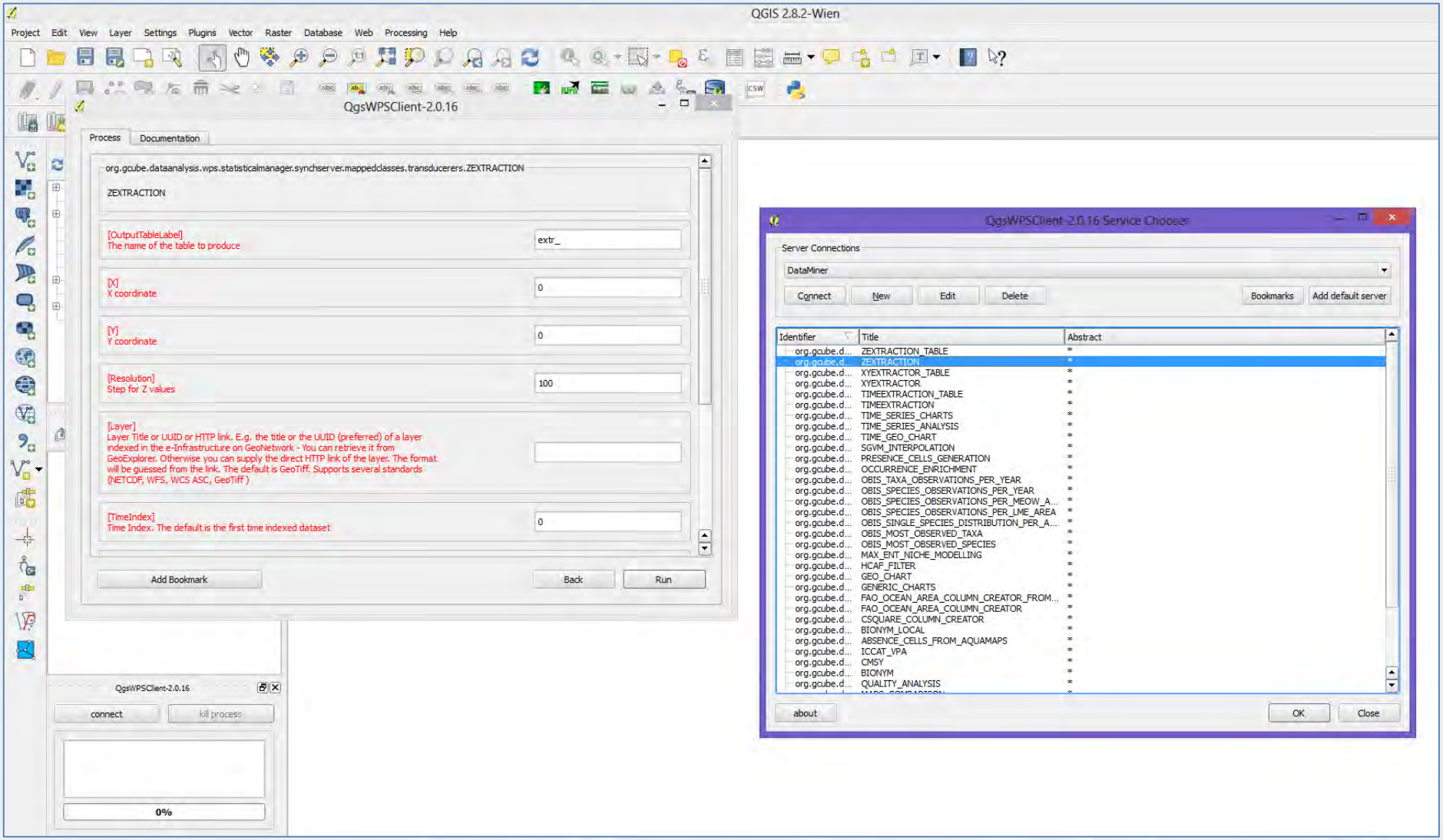
In the e-Infrastructure



Through third party software

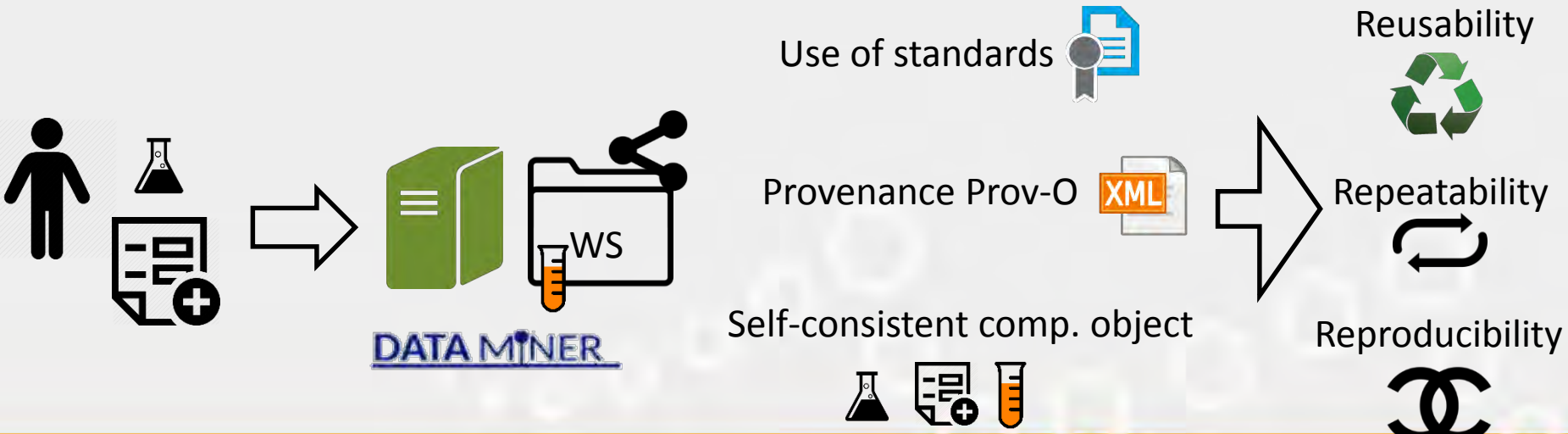


Usage from QGIS



Conclusions

- We have presented the benefits brought by a bridge between infrastructures;
- Our solution adds collaboration facilities and extends the usages of the connected service/data;
- This work goes in the direction of modern Science paradigms.



Useful links

General links:

D4Science e-Infrastructure website: https://www.d4science.org
Portal of the D4Science infrastructure: http://services.d4science.org
Portal of the BlueBRIDGE infrastructure: http://i-marine.d4science.org

About DataMiner:

DataMiner user's guide: https://wiki.gcube-system.org/gcube/DataMiner_Manager
DataMiner installation guide: https://wiki.gcube-system.org/gcube/DataMiner_Installation

Contributing with new algorithms:

Tutorial on the integration of new algorithms with the DataMiner
http://gcube.wiki.gcube-system.org/gcube/index.php/How-to_Implement_Algorithms_for_the_Statistical_Manager

Related work:

- Candela, L., Castelli, D., Coro, G., Pagano, P., & Sinibaldi, F. (2013). Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*.
- Coro, G., Candela, L., Pagano, P., Italiano, A., & Liccardo, L. (2015). Parallelizing the execution of native data mining algorithms for computational biology. *Concurrency and Computation: Practice and Experience*, *27*(17), 4630-4644.
- Candela, L., Castelli, D., Coro, G., Lelii, L., Mangiacrapa, F., Marioli, V., & Pagano, P. (2015). An infrastructure-oriented approach for supporting biodiversity research. *Ecological Informatics*, *26*, 162-172.
- Coro, G., Magliozzi, C., Ellenbroek, A., Kaschner, K., & Pagano, P. (2016). Automatic classification of climate change effects on marine species distributions in 2050 using the AquaMaps model. *Environmental and Ecological Statistics*, *23*(1), 155-180.
- Coro, G., Large, S., Magliozzi, C., & Pagano, P. (2016). Analysing and forecasting fisheries time series: purse seine in Indian Ocean as a case study.*ICES Journal of Marine Science: Journal du Conseil*, fsw131.
- Coro, G., Webb, T. J., Appeltans, W., Bailly, N., Cattrijsse, A., & Pagano, P. (2015). Classifying degrees of species commonness: North Sea fish as a case study. *Ecological Modelling*, *312*, 272-280.

BlueBRIDGE

Thank you!