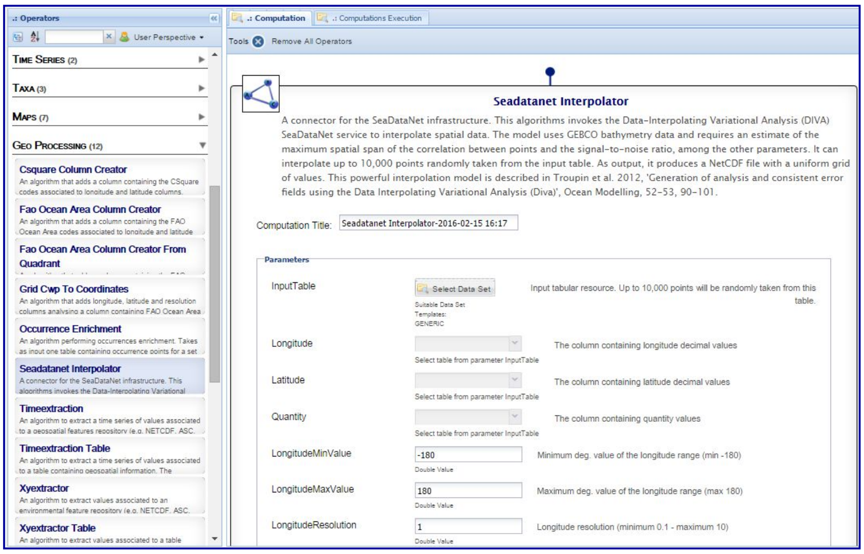
Bridging environmental data providers and SeaDataNet DIVA service within a collaborative and distributed e-Infrastructure

Gianpaolo Coro, ISTI-CNR (Italy), gianpaoloninited de.coro@isti.cnr.it
Pasquale Pagano, ISTI-CNR (Italy), pasquale.pagano@isti.cnr.it
Umberto Napolitano, ISTI-CNR (Italy), umberto.napolitano571@gmail.com

Among its devastating effects, ocean acidification harms organisms whose life depends on shells and on coral reefs. Therefore, it affects important marine ecosytems hosting high biodiversity and food availability. A recent paper published on Nature[1] has highlighted the decrease of calcification of a very important planktonic mollusc, the sea butterfly (*Limacina helicina*). The shell-building capacity of this organism declines with decreasing aragonite saturation, which is due to the increase 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. Unfortunately, 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). On the other hand, 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 do not support facilities neither to communicate nor to publish their results for a large public.



Csquare Column Creator
An algorithm that adds a column containing the CSquare codes associated to longitude and latitude columns.

Fao Ocean Area Column Creator
An algorithm that adds a column containing the FAO Ocean Area codes associated to longitude and latitude

Fao Ocean Area Column Creator From Quadrant

Grid Cwp To Coordinates
An algorithm that adds longitude, latitude and resolution columns analysing a column containing FAO Ocean Area

Occurrence Enrichment
An algorithm performing occurrences enrichment. Takes as input one table containing occurrence points for a set

Seadatanet Interpolator
A connector for the SeaDataNet infrastructure. This algorithms invokes the Data-Interpolating Variational

Timeextraction
An algorithm to extract a time series of values associated to a geospatial features repository (e.g. NETCDF, ASC,

Timeextraction Table
An algorithm to extract a time series of values associated to a table containing geospatial information. The

Xyextractor
An algorithm to extract values associated to an environmental feature repository (e.g. NETCDF, ASC,

Xyextractor Table
An algorithm to extract values associated to a table

Seadatanet Interpolator

A connector for the SeaDataNet infrastructure. This algorithms invokes the Data-Interpolating Variational Analysis (DIVA) SeaDataNet service to interpolate spatial data. The model uses GEBCO bathymetry data and requires an estimate of the maximum spatial span of the correlation between points and the signal-to-noise ratio, among the other parameters. It can interpolate up to 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 described in Troupin et al. 2012, 'Generation of analysis and consistent error fields using the Data Interpolating Variational Analysis (Diva)', Ocean Modelling, 52-53, 90-101.

Computation Title: Seadatanet Interpolator-2016-02-15 16:17

Parameters

InputTable — Suitable Data Set Templates: GENERIC — Input tabular resource. Up to 10,000 points will be randomly taken from this table.
Longitude — Select table from parameter InputTable — The column containing longitude decimal values
Latitude — Select table from parameter InputTable — The column containing latitude decimal values
Quantity — Select table from parameter InputTable — The column containing quantity values
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)

Fig. 1: Interface of the *D4Science-SeaDataNet Interpolator* service.

The D4Science e-Infrastructure is a distributed computer system that aims at supporting large-scale resource sharing (both hardware and software) via the definition of Virtual Research Environments (VREs) and allows data to be processed with distributed computing[2]. D4Science is able to create a bridge between several e-Infrastructures to fill the communication gap between them. This bridge is realised through a set of data storage and processing web services that are easy to extend. As for the ocean acidification monitoring case, we present a D4Science process (*D4Science-SeaDataNet Interpolator* service) that can transform a set of Copernicus *in situ* marine observations of an acidification parameter into a uniform global distribution map, published on a data catalogue under standard geographical representation formats (e.g. Web Map Service, Web Feature Service, Web Coverage Service). This process is hosted on the D4Science computing platform, which uses also

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

[2] 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.

European Grid Infrastructure FedCloud resources, and is able to (i) import Copernicus observations, (ii) organize them into a suitable format for the SeaDataNet interpolation service, (iii) execute the interpolation process and (iv) publish or share the output as a gradient map. The process invokes the SeaDataNet DIVA interpolation service using a distributed computing strategy to process both the input and output. Finally, the e-Infrastructure provides a geospatial services network, on which the computing system can store and publish data as GIS maps. The process is also endowed with a graphical user interface, available through a Web portal[3] (i-marine.d4science.org, Fig. 1). Through D4Science, users can establish the access policy for the input and output data, e.g. they can share data either with selected colleagues or with all the participants to a Virtual Research Environment they are involved in.

The integration realised by our service produces several benefits both to the data and model providers of SeaDataNet. We can summarise these benefits as follows:

Multi-users and concurrent load management: D4science services manage concurrent requests for computationally intensive processes and monitor the total allowed requests per user, using accounting facilities.
Virtual Research Environments: D4Science supports Virtual Research Environments that allow domain scientists to work together, exchange results, data etc. The *D4Science-SeaDataNet Interpolator* service can be published for selected users only and the D4Science social platform allows discussing about the results.
Storing output on a high-availability storage system: input and output are stored on a high-availability service offered by D4Science, based on a secure, fault-tolerant and fully replicated storage system.
Automatic generation of a web user interface: D4Science automatically endows integrated algorithms and services with a Web user interface, based on a declaration of the input and the output of the model. This is valid for the *Interpolation* service too.
Standard execution interface: the *Intepolation* service is made available through the Web Processing Service (WPS) standard of the Open Geospatial Consortium. This allows invoking models via REST communication and obtaining standard description of input and output. WPS increases the possibility to integrate the DIVA service with other software (e.g. QGIS) or workflows management systems (e.g. Taverna, Galaxy etc.).
Management of provenance information: D4Science tracks the experimental setup, the input and output of the DIVA service and allows other people to reproduce any experiment while getting the same results.
Data sharing and publishing facilities: D4Science users are endowed with a distributed file system accessible through a web interface (Workspace). This system allows sharing folders and files with (i) selected people, (ii) all the participants to a certain VRE, (iii) people outside of the e-Infrastructure (through the generation of persistent public links).
Applicability of the DIVA model to more data coming from direct observations: D4Science takes care of data harmonization and staging. Data from *in situ* observations are imported as generic CSV files and formatted for the SeaDataNet DIVA service.
Publication of the results as a GIS map: the NetCDF files produced by the DIVA service can be made publicly available under a number of standard representations (WMS, WCS, OPeNDAP, Esri GRID etc.). This increases the accessibility of the results by other systems.

In summary, our bridge expands the possibilities offered by SeaDataNet. In particular, it adds efficient data pre- and post- processing, sharing and publication facilities. Additionally, it increases the applicability of the DIVA service and expands the dissemination of the results. Therefore, it can facilitate producing global information about ocean acidification from scattered sea observations, with respect to the original scenario of independent infrastructures.

[3] The complete process is explained in a video available at http://goo.gl/yX3kww