

Leveraging Hybrid Data Infrastructure for Ecological Niche Modelling: The EUBrazilOpenBio Experience

Ignacio Blanquer^{**1}, Leonardo Candela², Pasquale Pagano², and Erik Torres¹

¹ Institute of Instrumentation for Molecular Imaging (I3M), Universitat Politècnica de València, Camino de Vera s/n, Valencia, Spain,
Email: {iblanque, ertorser}@i3m.upv.es

² Istituto di Scienza e Tecnologie dell'Informazione (ISTI), Consiglio Nazionale delle Ricerche (CNR), Via G. Moruzzi 1, Pisa, Italy
Email: {candela, pagano}@isti.cnr.it

Abstract. Ecological niche models are essential instruments in the development of strategies and policies in various application domains. The development of ever better models requires that scientists and practitioners are provided with *(i)* both the data and the processing capacities they need on demand and *(ii)* innovative mechanisms supporting the effective sharing of such products at scale. This paper describes the Ecological Niche Modelling service developed by the EUBrazilOpenBio project. This service is offered by relying on a underlying Hybrid Data Infrastructure (HDI) that *(i)* aggregates data, services and processing capacities from existing infrastructures and information systems and *(ii)* offers the aggregated resources in a seamless way. The paper presents the overall architecture of the service and details how its constituents have been developed by leveraging HDI offering on-demand data and processing facilities.

1 Introduction

Ecological niche models [17, 16, 31] aiming at estimating the presence of a species in a given area are essential instruments in the development of strategies and policies in various application domains. For example, ecological niche models have been used to predict the abundance and impact of invasive species [5, 21], to model distributions of disease vectors at continental scales [15], to study genetic diversity [29], and to study climate change consequences to biodiversity [11, 24]. These models rely on occurrence records to investigate the relationships between observed species presence and the underlying environmental parameters that – either directly or indirectly – determine a species distribution in a known area and use this information to predict the probability of a species occurrence in other areas [18].

In order to support the production of these models in a large scale, it is needed to develop innovative working environments for individual scientists potentially having at their disposal scarce resources. These environments should support complex and iterative processes which include at least the following key steps [16]: *(i)* identification of *relevant data*; *(ii)* *modelling*, i.e., deciding how to deal with the

^{**} e-mail of corresponding author: iblanque@i3m.upv.es

correlated prediction variables, selecting the appropriate algorithm, training the model, assessing the model; and (iii) *mapping* predictions to geographic space.

Scientists normally take models created by experts and work with them. The individual problem of creating a model, testing it and making a projection may take few minutes, depending on the number of occurrence points, the algorithm used and the resolution and number of layers. For instance, if the Brazilian Virtual Herbarium would like to create models for their 31,718 angiosperms species (18 Sept. 2012), and assuming that 30% of them will have enough points to generate models, the computational cost will go beyond 10 months (495k models, 540k tests and 90k projections). The problem will be even more complex if the spatial resolution is increased, or if the models consider not only Brazil and different climate conditions. Moreover, the models may be regenerated every time new data is available for each species.

There are some developments that makes the realisation of these working environments feasible including the availability of data and tools. Having a critical amount of species occurrence data is a pre-requisite to create a large repository of models, which have the potential to make significant impact in conservation biology and conservation planning. The consolidation of large biodiversity databases with high-quality data, such as GBIF [14], is a breakthrough in the way that ecological niche modelling is made. While many areas of research, such as life sciences, biomedicine or bioinformatics, are already benefiting from this approach, the exploitation of massive computing and data resources to accomplish ecological niche modelling presents an unprecedented opportunity to improve the quantity and quality of models available, also allowing the automatic computation of new models for geographical areas or species that has not been studied before. The availability of tools supporting the production of ecological niche models, such as openModeller [25], is another opportunity for scientists. openModeller is an open source software realising a single computing framework capable of handling different data formats and multiple algorithms that can be used in potential distribution modelling. The motivations leading to its development result from the willingness to provide scientists with a unified environment for species' distribution modelling. openModeller systematises the steps related to data management and to the algorithms exploited for modelling. It may exploit distributed computing resources through HTCondor [30] but it requires tailored worker nodes specifically equipped with openModeller libraries. It does not provide any facility for data discovery. It is expected that users acquire data by themselves and transform them to the supported formats. The current version supports a number of algorithms including ANN, Bioclim, GARP, SVM, and Envelope Scores. Its functionality is offered via a command line tool, a desktop tool as well as via a Web service.

However, there are some facts that makes the realisation of the envisaged working environments challenging, namely (i) relevant data are potentially huge, heterogeneous and scattered across many Information Systems, (ii) the production of models is a computational intensive task that potentially goes beyond the capacities of an average single scientist.

This paper presents an innovative facility for ecological niche models that benefits from the EUBrazilOpenBio Hybrid Data Infrastructure [9], i.e., an IT infrastructure offering *resource-as-a-service* where the resources range from computing

capacities to data and software. The HDI offers a rich array of data and data management services by leveraging other infrastructures (including Grid and Cloud) and Information Systems. Thus the HDI acts as a mediator for scientists by providing them with a unified and transparent way to access the data they need and the computing power by borrowing them on demand from heterogeneous providers.

The remainder of the paper is organised as follows. Section 2 gives a comprehensive description of the Ecological Niche Modelling facilities developed in the context of the EUBrazilOpenBio initiative. In particular, the section reports (i) the overall architecture of the services collaboratively realising the needed facilities (Sec. 2.1); (ii) the mechanisms and services supporting data management (Sec. 2.2); (iii) the mechanisms and services supporting effective data processing (Sec. 2.3); and, (iv) the web-based environment providing scientists with user friendly tools for niche modelling (Sec. 2.4). Concluding remarks are given in Section 3.

2 The EUBrazilOpenBio Ecological Niche Modelling Service

The EUBrazilOpenBio Ecological Niche Modelling service was devised to be offered *as-a-Service* in the context of a biodiversity-oriented Hybrid Data Infrastructure. It relies on a number of technologies including openModeller for niche modelling core facilities. In essence, it manifests in a service that can be used via a set of portlets or via a set of Web APIs (Application Programming Interfaces). Behind the scene, a number of Web services and software libraries contribute to realise the back-end needed to implement the resulting environment by integrating and interfacing with existing systems, infrastructures and information systems.

2.1 Overall Service Architecture

Ecological niche modelling requires data-intensive processing capabilities.

Figure 1 presents a schema of the Ecological Niche Modelling Service (ENM Service), which was specifically conceived to leverage computing and data resources delivered *as-a-Service* by the EUBrazilOpenBio HDI.

A portlet (ENM GUI) facilitates the use of the ENM Service. This portlet is integrated in the EUBrazilOpenBio Gateway³, which enables users with access to a variety of tools to support collaborative research. The organization of such tools in Virtual Research Environments (VREs)

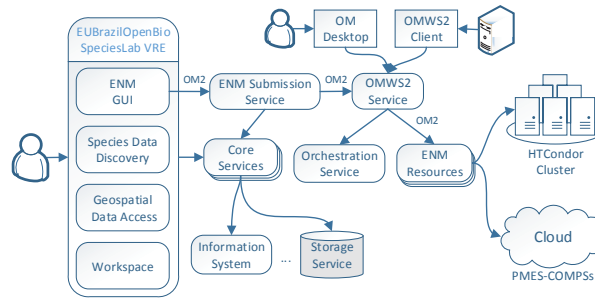


Fig. 1. Ecological Niche Modelling Architecture

³ EUBrazilOpenBio Gateway: <https://portal.eubrazilopenbio.d4science.org/>

[8, 6] provides a practical way in which users may effectively customize the tools to suit their specific needs. In addition to the ecological niche modelling portlet, others, such as the Species Data Discovery and Access Service (cf. Sec. 2.2) or the Workspace, are available to assist users in preparing ecological niche modelling experiments and analysing the results produced by these experiments.

The core services of the EUBrazilOpenBio infrastructure mainly rely on gCube [7, 8] facilities and D4Science.org⁴ support. Among these services, one of the most relevant for ecological niche modelling is the Information System, which is used to publish and discover information about the services available in the infrastructure. The Storage Service is also important for storing results and other information produced in the experiments, such as the execution traces.

Moreover, gCube provides a way to seamlessly integrate the different services and applications that supports the execution of ecological niche modelling experiments in the infrastructure. In particular, the services can be registered with the Information Service, facilitating their discovery and usage. Service registration has the additional benefit of not requiring endpoints with fixed paths. Instead of associating a service with a specific endpoint, clients can search the Information System for the services they need. Also, gCube provides system administrators with a way to automatically deploy their services to the gCube Hosting Nodes (gHN). These nodes are continuously monitored and, in some cases, corrective actions can be taken in a fully automatic way.

GUI operations are implemented through the ENM Submission Service, which manages the complete lifecycle of the experiments submitted from the GUI. However, the OMWS2 Service is the central entry point to the ENM Service, providing access to the computing back-ends that are available to execute ecological niche modelling experiments in EUBrazilOpenBio. The access to these computing facilities can occur in several ways: (i) through PMES-COMPSs [23] to access Cloud-based resources; or (ii) by using customized services to access computational clusters managed with HTCCondor. The interaction with the OMWS2 Service, no matter how the access to the computing back-end is carried out, is achieved through a common set of operations provided by the OMWS⁵ API.

By using this approach, a variety of OMWS clients, including the openModeller Desktop⁶, are supported directly in EUBrazilOpenBio. In this way, users have also the option to either use the openModeller Desktop or any OMWS-compatible client to access directly the OMWS2 Service offered by the infrastructure through a dedicated endpoint⁷.

The version 2 of the openModeller data format (OM2)⁸ is the language used to intercommunicate the different processes, components and services that make up the ENM Service.

⁴ D4science.org: <http://www.d4science.org>

⁵ OMWS v. 2 Definition: <http://openmodeller.cria.org.br/ws/2.0/openModeller.wsdl>

⁶ openModeller Desktop: <http://openmodeller.sourceforge.net/index.php>

⁷ EUBrazilOpenBio OMWS2 endpoint <http://enm.eubrazilopenbio.d4science.org/om/omws2>

⁸ openModeller v. 2 data format: <http://openmodeller.cria.org.br/xml/2.0/openModeller.xsd>

2.2 Data Management

In order to support the ecological niche modelling service, the EUBrazilOpenBio infrastructure should provide facilities for managing (i) biodiversity data, namely species occurrences, (ii) environmental data, and (iii) file-oriented content. The former two typologies of data are used as input data for the niche modelling activity, while the latter is used for storing data accompanying a modelling activity including experiment specification, execution logs, and static images resulting from model projection.

For biodiversity data, the EUBrazilOpenBio HDI has been equipped with the *Species Data Discovery and Access Service* (SDDA). This service provides users with seamless access to both nomenclature data and species occurrences from major information systems including GBIF [14] and speciesLink [2] for occurrence data, CoL [28] and List of Species of the Brazilian Flora [1] for nomenclature data. The SDDA service promotes a data discovery mechanism based on species name(s), either the scientific name or the common name of the target species. Moreover, to overcome the potential issues related to taxonomy heterogeneities across diverse data sources, the service supports an automatic query expansion mechanism, i.e., the query might be augmented with “similar” species names. Also, queries might be augmented with criterion aiming at explicitly selecting the databases to search and the spatial and temporal coverage of the data. Discovered data are presented in a homogenised form, e.g., in a typical Darwin Core [32] format. The service is designed to promote the openness with respect to the information systems it is capable to interface. In fact it is sufficient to implement (or reuse) a plug-in in order to enlarge the number of information systems and data sources integrated into SDDA. Each plug-in interacts with an information system or database by relying on a standard protocol, e.g., TAPIR [12], or by interfacing with its proprietary protocol. Every plug-in mediates queries and results from the query language and data model envisaged by SDDA to the requirements of a particular database.

For environmental data, the EUBrazilOpenBio HDI has been equipped with services realising the facilities of a *Spatial Data Infrastructure* (SDI) by relying on state-of-the-art technologies and standards. In particular, for the sake of the niche modelling, the infrastructure offers standard-based services for data discovery (a catalogue), storage and access (a federation of repositories), and visualisation (a map container). The standard-based catalogue service enables the discovery of geospatial data residing in dedicated repositories by relying on GeoNetwork [27] and its indexing facilities. GeoNetwork is the OGC recommended catalogue supporting ISO19115, ISO19119, and ISO19110 metadata standards represented and transported via the ISO19139 standard. For data storage and access, EUBrazilOpenBio relies on a federation of repositories built by relying on GeoServer [20] and THREDDS [26] technologies. In essence, the infrastructure hosts a number of repositories and a *GIS Publisher Service* that is conceived to enable the publication of geospatial data by relying on an open set of back-end technologies for the actual storage and retrieval of the data. Because of this, the GIS Publisher Service is designed with a plug-in-oriented approach where each plug-in interfaces with a given back-end technology. To enlarge the array of supported technologies it is sufficient to develop a dedicated plug-in. Metadata on available data are pub-

lished via the GeoNetwork catalogue. For data visualisation, the infrastructure offers *Geo Explorer* and *GIS Viewer*, two components dedicated to support the browsing and visualisation of geospatial data. In particular, the Geo Explorer is a web application that allows users to navigate, organize, search and discovery layers from a GeoNetwork instance via the CSW protocol. The GIS Viewer is a web application that allows users to interactively explore, manipulate and analyse geospatial data.

For file-oriented content, the EUBrazilOpenBio HDI has been equipped with the *gCube Storage Manager*, a Java based software library that presents a unique set of methods for services and applications running on the infrastructure and willing to store files. This software library relies on a network of distributed storage nodes managed via specialized open-source software for document-oriented databases. In its current implementation, two possible document store systems are used [10], MongoDB and Terrastore, while the implementation relaying on a third one, U.STORE [13], is ongoing. The software library was designed to promote the openness with respect to the storage back-end technology thus to reduce the time required to add a new storage system to the HDI and rely on it. In essence, it is sufficient to develop a specific mediator, deploy it and configure the Storage Manager accordingly.

2.3 Data Processing

Most of the ecological niche modelling experiments that are performed in EUBrazilOpenBio go through several stages of computation before a final result can be obtained. This is the case, for example, when a new model is computed with high-resolution climate layers (e.g., 30-arcsec WorldClim Version 1.4 Bioclimatic Variables [19]) for a large number of species occurrence points, and then validated using a statistical approach based on cross-validation –which is the common practice currently. This process may take several hours or days to complete, depending on the complexity of the modelling algorithm. The EUBrazilOpenBio infrastructure relies on Distributed Computing Infrastructures (DCI), such as Grid and Cloud computing systems, to provide users with access to massive computing and data resources, which contributes to reduce the average execution time. Through the use of these systems, users have found that in some cases it is more effective to run several experiments in parallel rather than sequentially. After the experiments are completed, the result can be analysed and compared by the expert to select manually the models with sufficient quality to be used in further studies.

The openModeller web service interface (OMWS)⁹ provides common programmatic access to the ecological niche modelling facilities that are made available in an open-access form to scientists and developers through the EUBrazilOpenBio infrastructure. OMWS was extended to meet the specific requirements for the execution of multi-staging and multi-parametric experiments [4]. These extensions were planned and undertaken in conjunction with the openModeller developers team to make them available to the openModeller community.

⁹ The openModeller Web Service: http://openmodeller.sourceforge.net/web_service.html

One of the main advantages of OMWS is that allows the complete specification of ecological niche modelling experiments, using the resources available in the infrastructure. This includes the occurrence datasets required to build and test models, the parameters of the modelling algorithms and the information to access the climate layers. Besides the expected operations to submit new experiments and to retrieve the results of already completed ones, OMWS defines operations that enable clients to discover which ecological niche modelling algorithms the service offers, as well as the climate collections that are available to the clients. Even though biodiversity databases can be directly queried for occurrence points, the SDDA service (cf. Sec. 2.2) provides an integrated, seamless interface for combining results from several databases in the same query.

However, OMWS does not cover all the aspects that are encountered in a multi-tenant application. A further, important consideration in any realistic application is the identification of the user who initiated the experiment request. This feature is not readily supported in the OMWS specification, but it is needed to: (i) manage (e.g., index or query) the experiments submitted by a specific user; and (ii) plan the use of the resources effectively, prioritizing critical requests (e.g., from a priority project) over non-critical ones, helping to ensure that the infrastructure would support the required performance.

This last aspect has been addressed in the ENM Submission Service (cf. Fig. 1), which is a separate service that manages the complete lifecycle of the experiment. In addition to keeping track of the experiments, the ENM Submission Service monitors the progress of the execution, automatically resubmitting failed jobs and managing the products of the experiments. This relieves applications of accomplishing such goals. Ecological niche models, as well as model projections and statistical results that are produced in the experiments, are initially stored in the computing back-end where the experiment is actually executed and then transferred to a shared storage in the infrastructure, before users can use them. This is also the case for the execution traces, which follow a similar pathway. The ENM Submission Service periodically checks the completion of the experiments and moves the results and the execution traces to their final location in the infrastructure. An interesting feature of the ENM Submission Service is that its public API, which is accessible from the GUI, uses the information that is locally available to the service. This means that when a user queries the ENM application for the status of his/her experiments, the ENM Submission Service will return the last status retrieved by the monitoring system, instead of querying recursively the computing back-end where the experiment is actually executed. In this way, the possible delays due to the network traffic are hidden from the user, contributing to provide a sense of real-time interactively in the ENM application.

The OMWS2 Service is a server implementation compatible with the OMWS Version 2.0 API. Contrary to the openModeller server, the OMWS2 Service does not connect directly with either the ecological niche modelling library or the job scheduler¹⁰. Instead, the OMWS2 Service uses the Orchestration Service of the infrastructure to select the most appropriate computing back-end for the execution of the experiment. The Orchestration Service is a meta-scheduler that dynami-

¹⁰ openModeller Server: http://openmodeller.sourceforge.net/om_server.html

cally discovers the available execution resources and performs a mapping between requests and resources, based on the requirements of computing power and data availability of the request (e.g., access to a specific modelling algorithm or climate layer set) and the workload of the resources.

Another major difference is that the OMWS2 Service can be extended with new plug-ins to new computing back-ends. EUBrazilOpenBio provides plug-ins to access PMES-COMPSs services [23] and OMWS2 instances. The VENUS-C Cloud infrastructure [3] and a dedicated HTCondor cluster [30] are the two major providers of computing resources in the infrastructure. While VENUS-C is accessible through a PMES-COMPSs service, HTCondor is not publicly available in the infrastructure and the OMWS2 API provides access to this resource.

The OMWS2 Service does not necessarily require synchronized ecological niche modelling facilities –same version of the ecological niche modelling algorithms are not needed– or data between the different computing resource providers because this information is readily accessible to the Orchestration Service, which allocates the resource according to it. In the case of the climate layers, missing layers are automatically downloaded to the local storage of the computing back-end from the GeoServer.

One of the most important details of the EUBrazilOpenBio ecological niche modelling service is that it connects users with experiments and computing resources. For example, when a user submits a new experiment to the infrastructure, the ENM Submission Service automatically registers the submission date and other details that may be important to the user. The request itself is stored as a document based on the openModeller data format that can be copied, edited or shared with other users of the infrastructure. Similarly, the OMWS2 Service registers where the request was sent to – the computing back-end where the experiment is actually executed –, as well as the identifiers assigned in the computing back-end to the jobs that are included in the experiment execution. In this manner, the information of the experiment can be retrieved and presented to the user in the most appropriate way. One of the most useful reports allows the user to list his/her experiments, filtering and sorting the list in various ways. This is addressed in detail in the Section 2.4.

The experiment information is stored in two different databases: the user information is stored in the accounting database, while the allocation database stores the information of the resource. These databases only share the identifier of the experiment, which is unique across the infrastructure. The accounting database is accessed mainly from the ENM Submission Service and is designed to provide fast access to the details of the experiments of the authenticated user. On the other hand, the allocation database is mainly accessed to consult the status of completeness of an experiment from the OMWS2 Service. Therefore, this database is optimized to support this kind of access.

2.4 The Graphical User Interface

Usability is one of the key aspects that drive the adoption of new tools and applications. In EUBrazilOpenBio, the use of well-established tools and protocols, such as openModeller and OMWS, allows the engagement of prospective users in the

design and validation of the ENM Service, enabling a broader user community the ability to evaluate the service and provide their feedback. Since the interface of the service was designed to be compatible with OMWS, any application that implements this API can easily access the computing and data capabilities provided by EUBrazilOpenBio.

However, there is a group of features that were developed in EUBrazilOpenBio to improve data utilization and are not available in OMWS. These features include the possibility of creating new data collections, for example, by querying simultaneously several databases for occurrences. Moreover, users may upload their own geospatial datasets, which is not possible with OMWS. Data collections and research findings can be shared with other researchers.

All these improvements in data management have lead to the design of a new GUI for ecological niche modelling. This new GUI was conceived as an experiment-centric environment that provides the necessary methods to retrieve, filter and format the input parameters for the experiment, and also to submit the experiment for execution on the computing back-ends, retrieve and analyse its results.

Fig. 2 shows a screenshot of the ecological niche modelling experiment editor, which is the main entry point of the GUI. Although the design of the interface was initially inspired by openModeller Desktop, the new GUI has introduced many features that are not present in this appli-

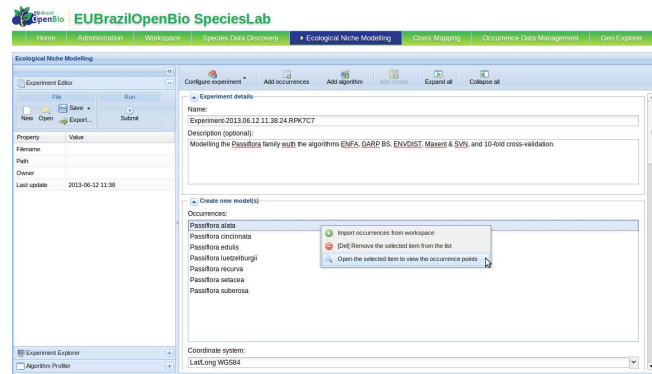


Fig. 2. Ecological Niche Modelling Experiment Editor

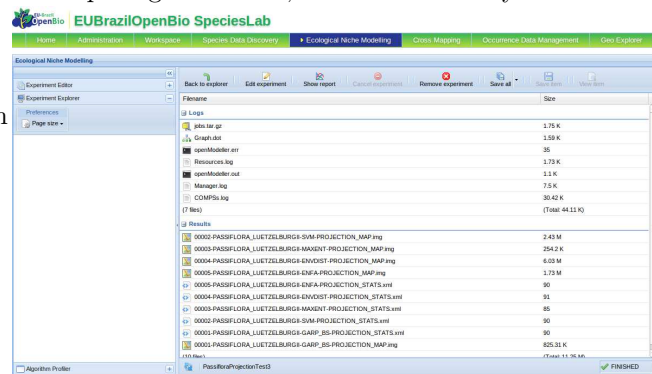


Fig. 3. Ecological Niche Modelling Experiment Explorer

cation. Among these features, the improvement of the sampling capabilities is an important characteristic of the GUI that allows users to set-up their occurrence datasets to be usable for cross-validation, bootstrapping or sub-sampling, depending on the desired validation strategy. Fig. 3 shows a screenshot of the ecological niche modelling explorer, which list all the experiments and provides access to the results and the report generation facilities. Report tools were also redesigned

to produce more expert-like interpretations of the results, including comparative charts, tabular views, error analysis reports, etc. Moreover, results can be exported in comma-separated values (CSV) file format, which facilitates their analysis with other tools as well as format them for papers and reports.

For models visualisation, the GUI relies on the fact that these are published in the SDI. Fig. 4 shows a screenshot of the geospatial information viewer.

Projected models can be selected to visualize them in the same view, thus facilitating the comparison of different predictions (e.g., different modelling algorithms or different climate layers).

However, the most significant difference between EUBrazilOpenBio and other applications based on

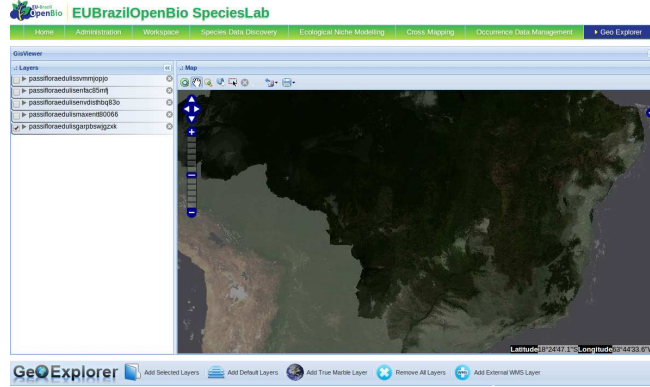


Fig. 4. Geospatial Information Viewer (GIS Viewer)

OMWS is that our service is the first implementation of the version 2.0 of this protocol. For this reason, the ENM GUI has a number of unique characteristics and advantages for the type of experiments that are addressed with this new release of OMWS. For example, the ENM GUI facilitates the realization of multi-parametric studies that involve variation in one or more parameters, such as different climate conditions or ecological niche modelling algorithms with customized parameters.

3 Concluding Remarks

The effective and large scale production of ecological niche models is a data and computation intensive task that calls for innovative working environments. This paper presented the innovative set of facilities developed by EUBrazilOpenBio infrastructure to leverage Hybrid Data Infrastructure offering where data, computing capacities and services are made available transparently and on demand. In particular, the set of services developed to realise the expected facility have been discussed ranging from those supporting the access to the diverse typologies of data to those oriented to support the execution of tasks by relying on diverse computational facilities. All of these services have been integrated in a Service-oriented application.

The potential benefits resulting from this environment are both quantitative and qualitative. Experimental results [22] demonstrate that the ENM service reaches good performance running on an on-demand provided environment (with an average performance loss around 9.6% with respect to a dedicated cluster), reaching a speed-up above 5 with the 10 machines. The potential benefits of such an integrated environment are evident from the final user perspective since they

are provided with a working environment offering the data and the tools they need, thus they can focus on the scientific investigation only.

EUBrazilOpenBio is starting a validation process to assess the impression of final users on the whole platform, considering performance and usability. Moreover, a sustainable plan is being defined which will use this validation input to define assets and future exploitation opportunities.

Acknowledgements

EUBrazilOpenBio - Open Data and Cloud Computing e-Infrastructure for Biodiversity (2011-2013) is a Small or medium-scale focused research project (STREP) funded by the European Commission under the Cooperation Programme, Framework Programme Seven (FP7) Objective FP7-ICT-2011- EU-Brazil Research and Development cooperation, and the National Council for Scientific and Technological Development of Brazil (CNPq) of the Brazilian Ministry of Science, Technology and Innovation (MCTI) under the corresponding matching Brazilian Call for proposals MCT/CNPq 066/2010.

References

1. List of Species of the Brazilian Flora. <http://floradobrasil.jbrj.gov.br/>, 2013.
2. speciesLink. <http://splink.cria.org.br>, 2013.
3. VENUS-C: Virtual multidisciplinary ENvironments USING Cloud infrastructures. <http://www.venus-c.eu>, 2013.
4. R. Amaral, R. Badia, I. Blanquer, L. Candela, D. Castelli, R. Giovanni, W. Gray, A. Jones, D. Lezzi, P. Pagano, V. Perez-Canhos, F. Quevedo, R. Rafanell, V. Rebello, and E. Torres. EU-Brazil Open Data and Cloud Computing e-Infrastructure for Biodiversity. In *Proceedings of International Workshop on Science Gateway (ISWG 2013)*, 2013.
5. N. Barve, V. Barve, A. Jiménez-Valverde, A. Lira-Noriega, S. Maher, A. T. Peterson, J. Soberón, and F. Villalobos. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11):1810 – 1819, 2011.
6. L. Candela, D. Castelli, P. Pagano. Virtual Research Environments: an overview and a research agenda. *CODATA Data Science Journal*, –, 2013.
7. L. Candela, D. Castelli, and P. Pagano. gCube: A Service-Oriented Application Framework on the Grid. *ERCIM News*, (72):48–49, January 2008.
8. L. Candela, D. Castelli, and P. Pagano. Making Virtual Research Environments in the Cloud a Reality: the gCube Approach. *ERCIM News*, (83):32–33, October 2010.
9. L. Candela, D. Castelli, and P. Pagano. Managing big data through hybrid data infrastructures. *ERCIM News*, (89):37–38, 2012.
10. R. Cattell. Scalable SQL and NoSQL data stores. *SIGMOD Rec.*, 39(4):12–27, May 2011.
11. G. C. Costa, C. Nogueira, R. B. Machado, and G. R. Colli. Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. *Biodiversity and Conservation*, 19(3):883–899, 2010.
12. R. De Giovanni, C. Copp, M. Döring, A. Güntscg, D. Vieglaiss, D. Hobern, J. Torre, J. Wiczorek, R. Gales, R. Hyam, S. Blum, and S. Perry. TAPIR - TDWG Access Protocol for Information Retrieval, 2010. Version 1.0.
13. F. A. Durão, R. E. Assad, A. F. Silva, J. F. Carvalho, V. C. Garcia, and F. A. M. Trinta. USTO.RE: A Private Cloud Storage System. In *13th International Conference on Web Engineering (ICWE 2013) - Industry track*, Aalborg, 2013.

14. J. L. Edwards, M. A. Lane, and E. S. Nielsen. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, 289(5488):2312–2314, 2000.
15. L. Eisen and R. Eisen. Using Geographic Information Systems and Decision Support Systems for the Prediction, Prevention, and Control of Vector-Borne Diseases. *Annual Review of Entomology*, 56:41–61, 2011.
16. J. Elith and J. R. Leathwick. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009.
17. A. Guisan and W. Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009, September 2005.
18. A. Guisan and N. E. Zimmermann. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147 – 186, 2000.
19. R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978, Dec. 2005.
20. S. Iacovella and B. Youngblood. *GeoServer Beginner’s Guide*. Packt Publishing, 2013.
21. A. Jiménez-Valverde, A. Peterson, J. Soberón, J. Overton, P. Aragn, and J. Lobo. Use of niche models in invasive species risk assessments. *Biological Invasions*, 13(12):2785–2797, 2011.
22. D. Lezzi, R. Rafanell, E. Torres, R. De Giovanni, I. Blanquer, and R. M. Badia. Programming ecological niche modeling workflows in the cloud. In *Proceed. of the 27th IEEE Int. Conf. on Advanced Information Networking and Applications, AINA-2013*, 2013.
23. F. Marozzo, F. Lordan, R. Rafanell, D. Lezzi, D. Talia, and R. Badia. Enabling Cloud Interoperability with COMPSs. In C. Kaklamanis, T. Papatheodorou, and P. Spirakis, editors, *Euro-Par 2012 Parallel Processing SE - 4*, volume 7484 of *Lecture Notes in Computer Science*, pages 16–27. Springer Berlin Heidelberg, 2012.
24. G. Mendoza-González, M. L. Martínez, O. R. Rojas-Soto, G. Vázquez, and J. B. Gallego-Fernández. Ecological niche modeling of coastal dune plants and future potential distribution in response to climate change and sea level rise. *Global Change Biology*, pages n/a–n/a, 2013.
25. M. Muñoz, R. De Giovanni, M. Siqueira, T. Sutton, P. Brewer, R. Pereira, D. Canhos, and V. Canhos. openModeller: a generic approach to species potential distribution modelling. *Geoinformatica*, 15:111–135, 2001.
26. S. Nativi, B. Domenico, J. Caron, E. Davis, and L. Bigagli. Extending THREDDS middleware to serve OGC community. *Advances in Geosciences*, 8:57–62, 2006.
27. Open Source Geospatial Foundation. GeoNetwork Opensource. <http://geonetwork-opensource.org/>.
28. Y. Roskov, T. Kunze, L. Paglinawan, T. Orrell, D. Nicolson, A. Culham, N. Bailly, P. Kirk, T. Bourgoin, G. Baillargeon, F. Hernandez, and A. De Wever. Species 2000 & ITIS Catalogue of Life, March 2013. Digital resource at www.catalogueoflife.org/col/. Species 2000: Reading, UK.
29. S. D. Schoville, A. Bonin, O. François, S. Lobreaux, C. Melodelima, and S. Manel. Adaptive genetic variation on the landscape: Methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):23–43, 2012.
30. D. Thain, T. Tannenbaum, and M. Livny. Distributed computing in practice: the Condor experience. *Concurrency and Computation: Practice and Experience*, 17(2-4):323–356, Feb. 2005.
31. A. Townsend Peterson, J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura, and M. B. Araújo. *Ecological Niches and Geographic Distributions*. Princeton University Press, 2011.
32. J. Wiczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. De Giovanni, T. Robertson, and D. Vieglais. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, 7(1), 2012.