EU-Brazil Open Data and Cloud Computing e-Infrastructure for Biodiversity

Rafael Amaral*, Rosa Badia†, ‡, Ignacio Blanquer§, Leonardo Candela¶, Donatella Castelli¶,
Renato de Giovanni‖, Willian A. Gray**,††, Andrew Jones**, Daniele Lezzi†, Pasquale Pagano¶,
Vanderlei Perez-Canhos‖, Francisco Quevedo**, Roger Rafanell†, Vinod Rebello* and Erik Torres§
* Instituto de Computação, Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil, Email: {rafaelbba, vinod}@ic.uff.br
† Department of Computer Sciences, Barcelona Supercomputing Center,
Barcelona, Spain, Email: {rosa.m.badia, daniele.lezzi, roger.rafanell}@bsc.es
‡Artificial Intelligence Research Institute (IIIA), Spanish National Research Council (CSIC)
§ Institute of Instrumentation for Molecular Imaging (I3M), Universitat Politècnica de València,
Camino de Vera s/n, Valencia, Spain, Email: {iblanque, ertorser}@i3m.upv.es
¶ Istituto di Scienza e Tecnologie dell'Informazione (ISTI), Consiglio Nazionale delle Ricerche (CNR),
Via G. Moruzzi 1, Pisa, Italy, Email: {candela, castelli, pagano}@isti.cnr.it
** School of Computer Sciences and Informatics, Cardiff University,
Cardiff, United Kingdom, Email: {F.Quevedo.Fernandez, W.A.Gray, Andrew.C.Jones}@cs.cardiff.ac.uk
†† Species 2000 Secreriat, Naturalis, Einsteinweg 2, 2333 CC Leiden, The Netherlands
‖ Centro de Referência em Informação Ambiental (CRIA), Campinas, SP, Brazil, Email: {renato, vcanhos}@cria.org.br

***Abstract*—EUBrazilOpenBio is a collaborative initiative addressing strategic barriers in biodiversity research by integrating open access data and user-friendly tools widely available in Brazil and Europe. The project deploys the EU-Brazil cloud-based e-infrastructure that allows the sharing of hardware, software and data on-demand. This e-Infrastructure provides access to several integrated services and resources to seamlessly aggregate taxonomic, biodiversity and climate data, used by processing services implementing checklist cross-mapping and ecological niche modelling. The concept of Virtual Research Environments is used to provide the users with a single entry point to processing and data resources. This article describes the architecture, demonstration use cases and initial experimental results.**

***Keywords*—Biodiversity, Data Infrastructure, Virtual Research Environments, Cloud, Taxonomy, Ecological niche modelling**

I. Introduction

The EUBrazilOpenBio project [1] aims to build an e-Infrastructure for research in biodiversity by leveraging primarily on resources (textual publications and datasets, maps, taxonomies, tools, services, computing and storage capabilities) provided by European and Brazilian e-Infrastructures available through existing projects and initiatives. Interoperation extends to all the infrastructure namely: hardware and computing facilities (Cloud and Grid computing, Internet), portals and platforms as well as the scientific data knowledge infrastructure.

A. State of the art

One of the most pressing needs in the biodiversity domain is the open and transparent access to data, tools and services. To support the worldwide sharing of various collections of biodiversity data [2], a number of large scale initiatives have occurred in recent years, either at global – e.g., *GBIF* [3], *OBIS* [4], *VertNet* [5], *Catalogue of Life* [6] – or at a regional level – e.g., *speciesLink* [7] and *List of Species of the Brazilian Flora* [8]. Moreover, standards for data sharing have been promoted by establishing appropriate interest groups, e.g., the Biodiversity Information Standards (TDWG - the Taxonomic Databases Working Group). Domain specific standards have been developed addressing different interoperability aspects, e.g., *Darwin Core* [9] and *TAPIR* [10] for distributed data discovery. In spite of these efforts and initiatives, the biodiversity domain is affected by a number of data sharing and reuse problems. [11].

New initiatives are creating global and web-based infrastructures to store, share, produce, serve, annotate and improve diverse types of species distribution information, such as the Map of Life [12]. Such initiatives highlight how the integration of disparate data types offers both new opportunities and new challenges for species distribution modelling.

The inherent complexity of using Distributed Computing Infrastructures (DCIs) to adapt, deploy and run applications and explore data sets have fostered the development of Science gateways, which facilitate the scientists' access to these tools, and simplify the organization of data repositories and the execution of experiments. There has been a concentrated effort to create portals and general-purpose services to address such issues. Portals and workflow engines such as Enginframe [13], eScienceCentral [14], Moteur [15], or P-Grade [16], address the problem of creating scientific workflows through individual modules and wrapping legacy code. However, these general approaches still require programming skills and background awareness of the features of the underlying infrastructure. Community portals, such as *WeNMR* [17], *GriF* [18], *Galaxy*[1], the Extreme Science and Engineering Discovery Environment

[1] http://galaxyproject.org/

(*XSEDE*)[2] or the gateway to nanotechnology online simulation tools *nanoHUB.org*[3] have developed customised solutions for their user communities. Current project efforts in Europe like *SCI-BUS*[4] are defining a flexible framework for developing science gateways based on the *gUSE*/*WS-PGRADE* portal family. In the area of data management, the *D4Science* [19] project has developed the gCube technology with a special focus on management of big data and the concept of Virtual Research Environments as its user interface. D4Science supports biodiversity [20] and other user communities.

B. Objective and motivation

This article describes the achievements of the EU-BrazilOpenBio project in creating an integrated infrastructure to assist research in biodiversity. The aim is to reduce the need for researchers to access data from multiple sources for local processing. Therefore, the project provides an access point to necessary data, services and computing capabilities to support research within the biodiversity community, demonstrated in two representative use cases.

The article is structured as follows. After this introduction, a description of the use cases is provided (cf. Sec. II). Section III describes the infrastructure of EUBrazilOpenBio and Section IV details the implementation of the use cases. Section V presents the conclusions.

II. The Use Cases and Requirements

In order to demonstrate the benefits an infrastructure might bring to the biodiversity informatics community, the project uses the infrastructure facilities to realise two representative use cases: the integration of taxonomies and production of ecological niche models which help in estimating species distributions. Although the requirements were elicited by analysing these two use cases, the infrastructure was designed and implemented with the aim of fulfilling the needs of a wider range of biodiversity applications.

In brief, the first use case aims at comparing two lists of species with the objective of identifying missing and incomplete entries. This process involves seamlessly accessing and comparing different taxonomical information, and it is the basis for enriching and improving existing regional and global taxonomies. The second use case is a computational-intensive problem consisting on constructing models that can be used to estimate the suitability of the environmental conditions on a certain region for a given species to survive there.

A. Integration of Regional and Global Taxonomies

The Catalogue of Life (CoL) [6], is a Global Taxonomy which covers most sectors of the world-wide taxonomic hierarchy. It aims to cover all known organisms. In contrast, a regional taxonomy (such as the List of Species of the Brazilian Flora [8]) only covers species known to occur in the region addressed by the taxonomy. However, regional taxonomies often contain richer information than global taxonomies about species. For example, a more extensive set of *synonyms* (scientific names other than the "accepted name" for a species), which either relate to the same or a similar concept, and descriptive information, may be given. They may also hold more up-to-date regional information, including information about some endemic species, which the compilers of global species lists may not yet be aware of.

Taxonomies may also vary in the names used for the same species, and may even vary in the associated concepts they represent. For example, a single concept in one taxonomy may correspond to the union of two distinct concepts in another. The codes of nomenclature (such as for plants [21] and animals [22]) specify how nomenclature is to be performed when the taxonomy is revised, perhaps merging or splitting concepts, or rearranging them. Such operations leave clues in the scientific names generated, which can help in detecting relationships between these names.

It is desirable to integrate regional and global taxonomies to attain: (*i*) more complete and richer information about individual species than is held in any contributing taxonomy, and (*ii*) coverage of a wider range of species than is held in any one contributing taxonomy.

An automated process is used in this project to identify the relationships between species concepts in taxonomies being integrated, when the accepted scientific names for the concepts are not the same. This *cross-mapping* between regional and global taxonomies is desirable, because: (*i*) When taxonomies differ, the concepts may differ (not just the names) making it impossible to simply integrate them all without losing information about observations attached to the individual concepts, and (*ii*) A user of a regional taxonomy may wish to see how the species-related data maps into another taxonomy.

A further complication is that in this paper's scenarios, the CoL data comes from a number of *Global Species Databases* (GSDs), each with its own specialist coverage of a particular section of the taxonomic hierarchy. The additional names and concepts discovered in the other taxonomy can be fed back to the custodians of the GSDs, for curation to enrich the CoL. This needs a cross-mapping and a piping tool where the latter feeds the discoveries of the former to the custodians. This project provides an opportunity to add knowledge from new sources to the CoL, and to discover new candidate GSDs for the CoL which may enrich the information of the CoL.

B. Ecological Niche Modelling

Ecological Niche Modelling (ENM) recently became one of the most popular techniques in macroecology and biogeography. There is an impressive growth in related published papers [23]. One of the reasons for this trend is the broad range of applications that arise when the ecological niche of a species can be approximated and projected in different environmental scenarios and geographical regions. An ecological niche can be defined as the set of ecological requirements for a species to survive and maintain viable populations over time [24]. ENMs are usually generated by relating locations where the species is known to occur with environmental variables that are expected to influence its distribution [25]. The resulting model is generated by an algorithm and can be seen as a representation of the environmental conditions that are suitable for the species. This makes it possible to predict the impact of

[2]https://www.xsede.org/
[3]https://nanohub.org/
[4]https://www.sci-bus.eu/

climate changes on biodiversity, prevent the spread of invasive species, help in conservation planning, identify geographical and ecological aspects of disease transmission, guide biodiversity field surveys, and many other uses [26].

This use case addresses computational issues when ENM is used with a large number of species in complex modelling strategies involving several algorithms and high-resolution environmental data. The use case is based on the requirements of the Brazilian Virtual Herbarium of Flora and Fungi (BVH) [27]. BVH has a specific system that uses a standard strategy to generate ecological niche models for plant species that are native to Brazil. All species that can be modelled by BVH come from the List of Species of the Brazilian Flora [8], which currently contains ∼40,000 entries. Occurrence points are retrieved from speciesLink [7] - a network that integrates data from distributed biological collections, currently serving almost 4 million plant specimen records. The modelling strategy used by BVH involves generating individual models using five different techniques in openModeller [28] when the species has at least 20 occurrence points: Ecological-Niche Factor Analysis [29], GARP Best Subsets [30], Mahalanobis distance [31], Maxent [32] and One-class Support Vector Machines [33]. The model quality is assessed by a 10-fold cross-validation, which leads to a final model created by merging the individual models into a single consensus model, which is then projected into the present environmental conditions for Brazil in high-resolution. The aim of this use case is to investigate and propose efficient ways of generating a large number of ecological niche models through a Web Service interface that can be used by applications such as BVH and by a new ENM Web application that can be integrated in the EUBrazilOpenBio Virtual Research Environment.

C. Requirements

Requirements were identified in an iterative process of analysis and refinement by users, systems analysts and application developers. First of all, there is a need to have seamless access to fundamental biodiversity data spread across multiple information systems, like CoL, GBIF, Brazilian Flora Checklist or SpeciesLink from an integrated access point.

Secondly, the analysis of such data requires using facilities to define and execute efficiently and effectively data and computational intensive workflows, including (*a*) the execution of pipelines to search and cross reference taxonomy item checklists with the objective of identifying missing entities and inconsistencies; and (*b*) the generation of multiple species distribution models by the use of different algorithms and settings. This includes the need to provide concurrent execution and to ensure a reasonable Quality of Service by providing scalability of the resource accessing or processing the available data in a timely manner.

Thirdly, the infrastructure should offer a user friendly, integrated environment, where scientists have innovative services supporting their data discovery and processing tasks as well as the sharing and consumption of research results, e.g., the storage and sharing of species distribution models with other users thus to avoid recalculation and feedback on results, and the display of results of the pipelines included in the platform.

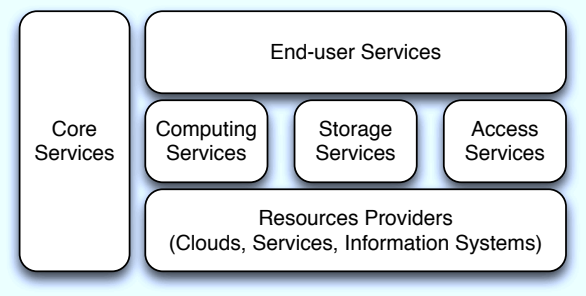

Fig. 1. EUBrazilOpenBio Infrastructure Conceptual Architecture

Fourthly, users must be able to upload their own data so that it can be readily processed using the facilities, e.g., to model ecological niches. Also, users need to download data stored in the "system" to be able to process such data with their own tools. This mitigates any infrastructure "lock-in" fear.

Finally, the infrastructure should cater for third-party service providers with web-based programming interfaces for using infrastructure facilities.

This project tackles these needs by providing an integrated infrastructure that has computing and storage resources, and integrates data and services through a user-friendly interface. These needs have led to the definition of specific requirements that are described in detail in the project's wiki[5], and they are fulfilled by the EUBrazilOpenBio platform.

III. The EUBrazilOpenBio Infrastructure

The EUBrazilOpenBio Infrastructure is an innovative *Hybrid Data Infrastructure* [34], conceived to enable a data-management-capability delivery model in which computing, storage, data and software are made available *as-a-Service*. In essence, it builds on the cloud paradigm offering "*computing as a utility*" and introduces *elasticity* of resources and infinite capacity as key features [35], [36] with the goal to make data and data management services available *on demand*.

The second distinguishing feature is its aggregative nature, i.e., the infrastructure is not built from scratch. Rather, it is a "system of systems" where the constituents include other infrastructures, services and Information Systems such as *GBIF* [3], *Catalogue of Life* [6], *speciesLink* [7], *List of Species of the Brazilian Flora* [8], *VENUS-C*[6]. EUBrazilOpenBio Infrastructure integrates these systems with the aim of exploiting the synergy amongst them, and thus offer biodiversity scientists a set of novel and enhanced services.

The third distinguishing feature is its capability to support the creation and operation of *Virtual Research Environments* (VREs) [37], [38], i.e., web based working environments where groups of scientists, perhaps geographically distant from each other, have transparent and seamless access to a shared set of remote resources (data, tools and computing capabilities) needed to perform their work.

Figure 1 is a schema of the infrastructure's underlying architecture. It consists of a number of services interacting according in a Service Oriented Infrastructure manner [39].

[5]wiki.eubrazilopenbio.eu

[6]VENUS-C, www.venus-c.eu (2012)

A. Core Services

Core services support the operation and management of the entire infrastructure. These services are from the gCube software framework [40], [37]. The Information Service has a key role here, as the infrastructure's Registry it caters for resource allocation, discovery, monitoring and accounting. Its role is to provide a continually updated picture of the infrastructure resources and their operational state where resources include service instances, hosting nodes, computing platforms, and databases. It also makes it possible for diverse services to cooperate by promoting a black-board mechanism, e.g., a service might publish a resource which is then consumed by another service. Overall, the service relies on a comprehensive yet extensible resource model.

Another core facility is the Resources Management service, which builds on the Information Service to realise resource allocation and deployment strategies. For resource allocation it enables the dynamic assignment of a number of selected resources to a given community (e.g., the creation of a VRE requires that a number of hosting nodes, service instances and data collections are allocated to a given application). For deployment, it enables the allocation and activation of both gCube software and external software on gCube Hosting Nodes (gHN), i.e., servers able to host running instances of services. By using this facility it is possible to dynamically create a number of service instances or enlarge the set of available computing nodes (by deploying a service on a gHN), to realise the expected elastic behaviour.

B. Biodiversity Data Access Services

Biodiversity data access services offer facilities enabling seamless data access, integration, analysis, visualisation and use of biodiversity data, namely nomenclature data and species occurrences. Such data represents a key resource for the target community that is spread across a number of Information Systems and databases making exploitation challenging [11]. EUBrazilOpenBio offers a species data discovery and access service (SDDA) which is a mediator over a number of data sources. SDDA is equipped with plug-ins interfacing with the major information systems: GBIF and speciesLink for occurrence data, CoL and List of Species of the Brazilian Flora for nomenclature data. In order to enlarge the number of information systems and data sources integrated into SDDA, it is sufficient to implement (or reuse) a plug-in. Each plug-in is able to interact with an information system or database by relying on a standard protocol, e.g., TAPIR [10], or by interfacing with its proprietary protocol. Every plug-in mediates queries and results from the language and model envisaged by SDDA to the requirements of a particular database.

SDDA promotes a data discovery mechanism based on queries containing 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 can specifically select the databases to search and other constraints on the spatial and temporal coverage of the data. Discovered data are presented in a homogenised form, e.g., in a typical Darwin Core [9] format.

A number of facilities for inspecting the retrieved data are available, e.g., a geospatial oriented one is available for occurrence data. Moreover, it is possible to simply "save" the discovered data in various formats – including CSV and Darwin Core [9] – and share them with co-workers through the *user workspace* (cf. Sec. III-E). This is a fundamental facility for the two use cases (cf. Sec. IV).

C. File-oriented Storage Services

The file-oriented storage facilities aims to offer a scalable high-performance storage service. In particular, this storage service relies on a network of distributed storage nodes managed via specialized open-source software for document-oriented databases. This facility is offered by the gCube Storage Manager, a Java based software that presents a unique set of methods for services and applications running on the e-Infrastructure. In its current implementation, two possible document store systems are used [41], MongoDB and Terrastore. The Storage Manager was designed to reduce the time required to add a new storage system to the e-Infrastructure. This promotes openness versus other document stores, e.g., CouchDB [42], while hiding the heterogeneous protocols of those systems to the services and applications exploiting the e-Infrastructure storage facility.

D. Computing Services

Computing services offer a rich array of computing platforms as-a-Service. This requires harnessing a wide range of computational resources (from individual computer servers, to clusters, grids and cloud infrastructures, probably distributed around the world) efficiently so as to have the potential capacity to handle the concurrent execution of significant numbers of experiments. This also implies the need to identify a set of technologies to allow scientific experiments and tools to exploit the synergy of the available aggregated processing capacity within the platform to the fullest extent.

Workflow and application management systems, such as the COMPSs programming framework and the EasyGrid AMS, benefit the infrastructure by acting as enabling technologies to leverage a range of distributed resource types, such as HPC clusters (with traditional workload management systems such as LSF, PBS, and SGE); HTCondor pools (also for opportunistic computing); and the VENUS-C cloud infrastructure (that can use both private and public providers including commercial ones such as Microsoft Windows Azure). The diversity of resources considered by EUBrazilOpenBio aims to reflect the most likely scenario of types of infrastructure resources that would be available to the biodiversity community.

The *VENUS-C middleware* has been adopted as one of the building blocks of the EUBrazilOpenBio computing services. In particular, the programming model layer, in conjunction with data access mechanisms, have proven to provide researchers with a suitable abstraction for scientific computing on top of virtualized resources. One of these resources is COMP Superscalar [43], leveraged in VENUS-C to enable the interoperable execution of use cases on the hybrid cloud platform. The COMPSs programming framework allows the development of scientific applications and their seamless execution on a wide number of distributed infrastructures. In

cloud environments, COMPSs provides scaling and elasticity features allowing the number of available resources to adapt to the execution [44].

HTCondor [45] is a workload management system for compute-intensive jobs on clusters and wide-area distributed systems of either dedicated or shared resources. Installed at over 3000 sites around the world, HTCondor provides a job queuing mechanism, scheduling policy, priority scheme, resource monitoring, and resource management that allows users to execute either serial or parallel jobs. With HTCondor's metascheduler, and Directed Acyclic Graph Manager (DAGMan) [46], HTCondor can manage task dependencies within a job, e.g., a Condor job may be configured to perform the modelling step first, and thereafter perform steps to test and project the model in parallel. Given the computational requirements of an experiment can be large, an additional feature in this deployment is pool elasticity. Node virtualization allows additional resources to be added on-demand to increase availability, and performance.

While systems like VENUS-C has COMPSs and HTCondor has DAGMan, others systems without a local workflow manager can use the *EasyGrid AMS* [47]. The EasyGrid middleware is a hierarchically distributed Application Management System (AMS) embedded into parallel MPI applications to facilitate efficient execution in distributed computational environments. By coupling legacy MPI applications with EasyGrid AMS, they can be transformed into autonomic versions which manage their own execution. The benefits of this approach include adopting (scheduling, communication, fault tolerance) policies tailored to the specific needs of each application thus leading to improved performance [48]. While the EasyGrid AMS is being used to accelerate phases of openModeller through parallelisation, given that workflows can be seen to be directed acyclic graphs, the AMS can also be used to encapsulate the entire workflow and manage their execution in distributed systems without workflow managers.

E. End-user Services

End-user services provide human users with facilities benefitting and building upon the resources the infrastructure aggregates. The majority of these services appear in a web-based user interface and all of them are conceived to be aggregated and made available via VREs hosted by a portal.

These services include infrastructure management facilities (e.g., VRE deployment facilities, user management, resource management) and user collaboration (e.g., shared workspace, data discovery facilities, data manipulation facilities). We now describe the facilities specifically exploited to support the two use cases.

The Workspace is a user interface implemented through portlets, that provide users with a collaborative area for storing, exchanging and organizing information objects according to any specific need. Every user of a VRE is provided with this area that resembles a classic folder-based file system, with seamlessly managed item types that range from binary files to compound information objects representing tabular data, species distribution maps, and time series. Every workspace item is equipped with rich metadata including bibliographic information like title and creator as well as lineage data. In addition to information object storage and organisation, the portlet allows easy exchange of objects among users as well as import/export of objects from/to the user file system to enable processing such objects using both the infrastructure and local users' computers.

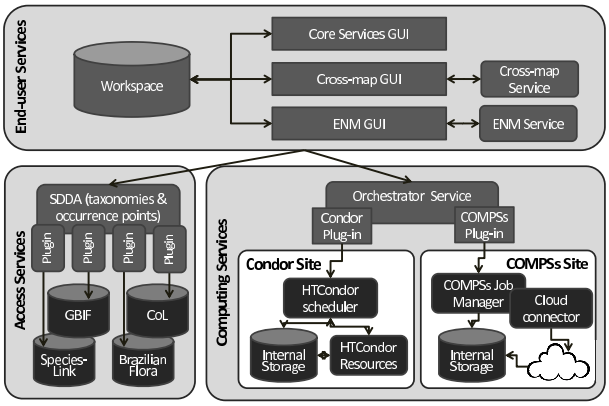

Fig. 2. Use Case Architecture Deployment

IV. Implementation of the Use Cases

The facilities of the EUBrazilOpenBio infrastructure are made available via a dedicated portal[7] which hosts the VRE resulting from the implementation of the use cases. The software artchitecture of both use cases is shown in Figure 2. End-users are provided with specific portlets, each realising one use case. These portlets are integrated with others portlets, namely SDDA and Workspace for data access. Each use cases' portlet interact with the processing services through specific services that implements the functionality of the use case. Different computing and data resources are accessed through a combination of services and plug-ins that hide the particularities of each source and back-end.

A. EUBrazilOpenBio Taxonomy Management Facilities

The basis of the software components developed in the first use case was the cross-mapping tool implemented in the i4Life[8] project. However, although the actual cross-mapping software is essentially the same in both systems, the environment and modes of interaction have been completely redesigned, moving from a simple web site, developed in PHP and Perl, in which the users had to interact with it manually, to a system in which all its functionality is accessible programmatically through a Web service interface, thereby making it more suitable for deployment as part of a distributed architecture. With its new portlet, the cross-mapping tool is now integrated with other software components provided by the EUBrazilOpenBio infrastructure (workspace, information system, SDDA, etc.), making it easier for the user to run cross-mapping experiments.

The migration has also achieved a reduction in the execution time of large cross-mapping tasks (e.g., by limiting the

[7] https://portal.eubrazilopenbio.d4science.org
[8] http://www.i4life.eu/

size of tables of names created during checklist import) and can display the results of the cross-map in a tree view perspective, not currently available in i4Life.

The software developed for this use case can be divided into two categories: a SOAP web service with MTOM [49] which exposes a set of methods that allow the clients to upload checklists, run cross-map experiments and export their results, and a portlet that interacts with the cross-map service alongside other services and tools provided by the infrastructure. Checklists are seamlessly obtained from the SDDA infrastructure service, and communication between the VRE and the processing services is done through the infrastructure storage services. This eases the development of applications and the sharing of data among users and services.

Internally, the cross-map service was developed using a layered approach. The service interface layer defines the public interface of the service (using a WSDL file). Another layer provides the logic of the application; this can also be called from a command line without using the web-service. These interfaces are specified and generated using the Tuscany SCA [50], a framework that allows declarative exposure of Java components into a plethora of different protocols, such as SOAP, Rest or JMS.

The portlet (cf. Figure 3) was developed jointly by CNR and Cardiff University; it is basically a GWT[9] project that uses the GXT[10] 3.0 library which provides rich web-based widgets. Also XML files have been added to deploy it as a Liferay portlet inside the project's portal. Internally the software interacts with the workspace to retrieve and store input and output data for the cross-mapping tool as well as querying the information system to obtain the instance of the cross-map web service to be used.

The web service and the portlet are deployed as Web Archive (war) files: (*i*) the portlet is deployed in the EUBrazilOpenBio portal; (*ii*) the cross-map web service is deployed in a web server container, registering its service endpoint in the Information System as an external resource.

At the moment, the cross-map service processes all the operations sent by the portlet locally. However, to make the application more scalable, the cross-map service will externalize the execution of some of those operations, depending on their complexity, to other resources provided by the infrastructure.

B. EUBrazilOpenBio Niche Modelling Facilities

The original openModeller Web Service (OMWS) API exposes a set of operations defined by an XML schema having all elements, attributes, structure and data types of the openModeller objects. Each operation defined by this scheme supports the execution of one simple action in openModeller, for example, to create, test or project a model. When a user wants to perform several actions on the same dataset, it is necessary to submit each operation to the service separately. For example, to create five models with the same species occurrence dataset using five different modelling algorithms, five different requests are needed (one per algorithm). This also occurs for experiments that create models for different species using the same modelling algorithm. When there are dependencies between the operations, such as, creating a model and then projecting it onto an environmental scenario, the client is responsible for sending the initial request, monitoring and retrieving the results of the operation that creates the model and also including the serialized model as an input parameter in the projection operation that follows.

[9]https://developers.google.com/web-toolkit/

[10]http://www.sencha.com/products/gxt

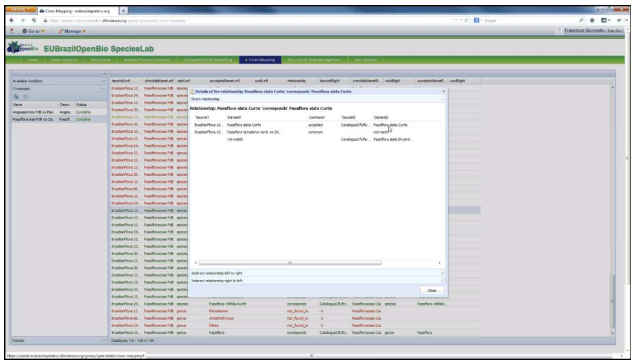

Fig. 3. Taxonomy Managament Facilities portlet snapshot

An extension of the openModeller Web Service API (namely OMWS+) provides a way to automatically convert multi-stage and multi-parameter experiments into a set of single legacy operations supported by the openModeller command-line suite. Support for user sessions is also included for enhanced monitoring and retrieving of results. This is currently implemented through COMPSs, which orchestrates the execution after automatically generating the execution graph (cf. Fig. 2).

This component exploits the advantages of the concurrent execution of the infrastructure and makes use of the storage services provided. The GUI implemented on the VRE integrates with the rest of the services so the user interacts with an editor application that enables the creation of experiments that are converted into multiple concurrent jobs with the results being gathered in a single view. It provides a comprehensive visualization of all the species and algorithms and provides a progress report that enables progress monitoring of the experiments and retrieving their results from any web browser.

The implementation of the COMPSs Job Manager on the VENUS-C PMES, receives the execution request from the Orchestrator dispatching users' requests received from the OMWS+ interface to support multi-staging and multi-parametric experiments through COMPSs and openModeller. These extensions are backward compatible with the original OMWS specification, which allows legacy clients to be fully supported in the new implementation and, therefore, still able to submit experiments to the execution resources without using the graphical user interface developed by the project. The infrastructure hides the complexity of accessing data sources and data repositories. Data sources are integrated through the SDDA and data storage provided by the infrastructure can be accessed by the VRE and the processing instances, facilitating the sharing of data and storing them permanently on the users' specific storage. A snapshot is shown in Figure 4.

This service was tested on 10 quad-core virtual instances with 2GB of memory and 1GB of disk space on the EU-

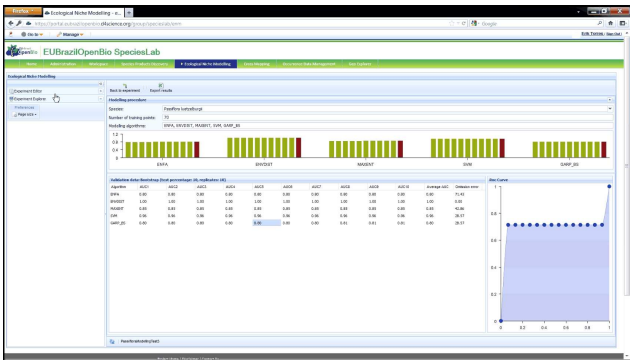

Fig. 4. Ecological Niche Modelling GUI snapshot

BrazilOpenBio infrastructure. The aim of these tests was to validate the workflow implementation and to evaluate the advantage of the elasticity features of these services. Eight species of the genus Passiflora, each one with more than 20 occurrence points, were used. Models were generated using 8 high resolution environmental layers from WorldClim. A simplified standard procedure consisting of model creation followed by an internal model test (confusion matrix and ROC curve calculation with the same input points) and a native model projection (with the same environmental layers) followed by a final image transformation was used for each species with a set of three algorithms used by BVH (SVM, ENVDIST and ENFA) called with a different set of parameters. The Brazilian territory served as a mask in all operations. This scenario composes a total of 46 simultaneous single operation requests. Experimental results [51] 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.

V. Conclusion

The EUBrazilOpenBio infrastructure is an integrated e-science platform for biodiversity researchers. It goes beyond integration of resources by providing seamless access to data, services and collaboration facilities. The concept of Virtual Research Environments requires a short learning curve, and integration of computing and visualization services reduces the need to transfer data from and to the infrastructure.

The integration of multiple technologies and services required development of intermediate services which orchestrate and virtualize different resources. The exploitation of commonly accepted protocols for data and services enables the use of the resources through web-based programming interfaces.

Requirements in these examples of biodiversity research also show the need to be able to use a user's own data, both lightweight data (taxonomies or occurrence points) and big data, such as environmental layers. Bandwidth usage minimization is a key issue in performance improvement.

The selection of two representative use cases has enabled the creation of demonstrators and the validation of specific requirements which are common to many other applications. Generic services provide building blocks for such applications.

The technologies used are open and extensible and interoperability is important to maximise the integration of different data sources and computing backends. Although the requirements elicited from the use cases focused on the cross-mapping of taxonomies and ecological niche models, the infrastructure has been designed and the services were implemented to fulfil the needs of a wide range of biodiversity applications. The main technical advance of EUBrazilOpenBio is the integration of diverse data sources, processing services, computing and data resources in a unique science gateway. Both applications available through the VRE have been tested by scientists as part of the project and are ready to be used.

EUBrazilOpenBio provides the scientists with a single access point to a wide range of Biodiversity resources. EUBrazilOpenBio storage enables a seamlessly and ubiquitous access to reference data and experiment results. Taxonomy checklists, occurrence points, ecologic niche models, projection maps, etc. can be exchanged and visualized from the VRE, without requiring local applications nor downloading output files. Users of this integrated framework also benefit from the high-performance computing back-ends of the platform.

Acknowledgment

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] EUBrazilOpenBio Consortium. (2013) EU-Brazil Open Data and Cloud Computing e-Infrastructure for Biodiversity. [Online]. Available: http://www.eubrazilopenbio.eu/

[2] D. Triebel, G. Hagedorn, and G. Rambold, "An appraisal of mega-science platforms for biodiversity information," *MycoKeys*, vol. 5, pp. 45–63, 2012.

[3] J. L. Edwards, M. A. Lane, and E. S. Nielsen, "Interoperability of biodiversity databases: Biodiversity information on every desktop," *Science*, vol. 289, no. 5488, pp. 2312–2314, 2000.

[4] J. Grassle, "The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context," *Oceanography*, vol. 13, no. 3, pp. 5–7, 2000.

[5] H. Constable, R. Guralnick, J. Wieczorek, C. Spencer, and e. a. Peterson, A. Townsend, "Vertnet: A new model for biodiversity data sharing," *PLoS Biol*, vol. 8, no. 2, p. e1000309, 02 2010.

[6] 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.

[7] (2013) speciesLink. [Online]. Available: http://splink.cria.org.br

[8] (2013) List of Species of the Brazilian Flora. [Online]. Available: http://floradobrasil.jbrj.gov.br/

[9] J. Wieczorek, 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*, vol. 7, no. 1, 2012.

[10] R. De Giovanni, C. Copp, M. Döring, A. Güntscg, D. Vieglais, D. Hobern, J. Torre, J. Wieczorek, R. Gales, R. Hyam, S. Blum, and S. Perry., "TAPIR - TDWG Access Protocol for Information Retrieval," Biodiversity Information Standards, 2010, version 1.0. [Online]. Available: http://www.tdwg.org/activities/abcd/

[11] A. Goddard, N. Wilson, P. Cryer, and G. Yamashita, "Data hosting infrastructure for primary biodiversity data," *BMC Bioinformatics*, vol. 12, no. Suppl 5, p. S5, 2011.

[12] W. Jetz, J. M. McPherson, and R. P. Guralnick, "Integrating biodiversity distribution knowledge: toward a global map of life," *Trends in Ecology & Evolution*, vol. 27, no. 3, pp. 151 – 159, 2012.

[13] NICE S.r.l. (2013) Enginframe. [Online]. Available: http://www.nice-software.com/products/enginframe

[14] H. Hiden, S. Woodman, P. Watson, and J. Cala, "Developing cloud applications using the e-science central platform," *Proceedings of Royal Society A*, 2012.

[15] T. Glatard, J. Montagnat, D. Lingrand, and X. Pennec, "Flexible and Efficient Workflow Deployment of Data-Intensive Applications On Grids With MOTEUR," *International Journal of High Performance Computing Applications*, vol. 22, no. 3, pp. 347–360, 2008.

[16] P. Kacsuk and G. Sipos, "Multi-grid, multi-user workflows in the p-grade grid portal," *Journal of Grid Computing*, vol. 3, no. 7-4, pp. 221–238, September 2005.

[17] T. Wassenaar, M. v. Dijk, N. Loureiro-Ferreira, G. v. d. Schot, S. d. Vries, C. Schmitz, J. v. d. Zwan, R. Boelens, and A. Bonvin, "WeNMR: structural biology on the grid," *CEUR Workshop Proceedings*, vol. 819, no. 4, pp. 1–8, 2011.

[18] C. Manuali, A. Lagan, and S. Rampino, "GriF: A Grid framework for a Web Service approach to reactive scattering," *Computer Physics Communications*, vol. 181, no. 7, p. 11791185, July 2012.

[19] L. Candela, D. Castelli, and P. Pagano, "D4science: an e–infrastructure for supporting virtual research environments," in *Post–proceedings of the 5th Italian Res. Conf. on Digital Libraries– IRCDL 2009*, 2009.

[20] L. Candela and P. Pagano, "The D4Science Approach toward Grid Resource Sharing: The Species Occurrence Maps Generation Case," in *Data Driven e-Science - Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010)*, S. C. Lin and E. Yen, Eds. Springer, 2011, pp. 225–238.

[21] J. McNeill *et al.*, *International Code of Nomenclature for algae, fungi and plants (Melbourne Code)*. Koeltz Scientific Books, 2012.

[22] W. Ride *et al.*, *International Code of Zoological Nomenclature*, 4th ed. The International Trust for Zoological Nomenclature, 1999.

[23] J. Lobo, A. Jiménez-Valverde, and J. Hortal, "The uncertain nature of absences and their importance in species distribution modelling," *Ecography*, vol. 33, pp. 103–114, 2010.

[24] J. Grinnell, "Field tests of theories concerning distributional control," *American Naturalist*, vol. 51, pp. 115–128, 1917.

[25] J. Sobern and A. Peterson, "Interpretation of models of fundamental ecological niches and species distributional areas," *Biodiversity Informatics*, vol. 2, pp. 1–10, 2005.

[26] A. Peterson, J. Sobern, R. Pearson, R. Anderson, E. Martinez-Meyer, M. Nakamura, and M. Arajo, *Ecological niches and geographic distributions*. Princeton University Press, 2011.

[27] (2013) Brazilian Virtual Herbarium. [Online]. Available: http://biogeo.inct.florabrasil.net/

[28] 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*, vol. 15, pp. 111–135, 2001.

[29] A. H. Hirzel, J. Hausser, D. Chessel, and N. Perrin, "Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data?" *Ecology*, vol. 83, no. 7, pp. 2027–2036, 2002.

[30] R. Anderson, D. Lew, and A. Peterson, "Evaluating predictive models of species distributions: criteria for selecting optimal models." *Ecological Modelling*, vol. 162, pp. 211–232, 2003.

[31] O. Farber and R. Kadmon, "Assessment of alternative approaches for bioclimatic modeling with special emphasis on the mahalanobis distance." *Ecological Modelling*, vol. 160, pp. 115–130, 2003.

[32] S. Phillips, R. Anderson, , and R. Schapire, "Maximum entropy modelling of species geographic distributions," *Ecological Modelling*, vol. 190, pp. 231–259, 2006.

[33] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[34] L. Candela, D. Castelli, and P. Pagano, "Managing big data through hybrid data infrastructures," *ERCIM News*, no. 89, pp. 37–38, 2012.

[35] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[36] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared," in *Grid Computing Environments Workshop, 2008. GCE '08*, 2008.

[37] L. Candela, D. Castelli, and P. Pagano, "Making Virtual Research Environments in the Cloud a Reality: the gCube Approach," *ERCIM News*, no. 83, pp. 32–33, October 2010.

[38] L. Candela, "Data Use - Virtual Research Environments," in *Technological & Organisational Aspects of a Global Research Data Infrastructure - A view from experts*, K. Ashley, C. Bizer, L. Candela, D. Fergusson, A. Gionis, M. Heikkurinen, E. Laure, D. Lopez, C. Meghini, P. Pagano, M. Parsons, S. Viglas, D. Vitlacil, and G. Weikum, Eds. GRDI2020, 2012, pp. 91–98.

[39] W. Tsai, "Service-oriented system engineering: a new paradigm," in *Service-Oriented System Engineering, 2005. SOSE 2005. IEEE International Workshop*, oct. 2005, pp. 3 – 6.

[40] L. Candela, D. Castelli, and P. Pagano, "gCube: A Service-Oriented Application Framework on the Grid," *ERCIM News*, no. 72, pp. 48–49, January 2008. [Online]. Available: http://ercim-news.ercim.eu/en72/rd/gcube-a-service-oriented-application-framework-on-the-grid

[41] R. Cattell, "Scalable SQL and NoSQL data stores," *SIGMOD Rec.*, vol. 39, no. 4, pp. 12–27, May 2011.

[42] J. C. Anderson, J. Lehnardt, and N. Slater, *CouchDB: The Definitive Guide*. O'Really, 2009.

[43] D. Lezzi, R. Rafanell, A. Carrión, I. Blanquer, V. Hernández, and R. M. Badia, "Enabling e-science applications on the cloud with compss," in *Proc. of the 2011 intl. conf. on Parallel Processing*, ser. Euro-Par'11. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 25–34.

[44] F. Marozzo, F. Lordan, R. Rafanell, D. Lezzi, D. Talia, and R. M. Badia, "Enabling cloud interoperability with compss." in *Euro-Par*, ser. Lecture Notes in Computer Science, C. Kaklamanis, T. S. Papatheodorou, and P. G. Spirakis, Eds., vol. 7484. Springer, 2012, pp. 16–27.

[45] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: the Condor experience," *Concurrency - Practice and Experience*, vol. 17, no. 2-4, pp. 323–356, 2005.

[46] P. Couvares, T. Kosar, A. Roy, J. Weber, and K. Wenger, "Workflow in Condor," in *Workflows for e-Science*, I. Taylor, E. Deelman, D. Gannon, and M. Shields, Eds. Springer Press, 2007.

[47] C. Boeres and V. E. F. Rebello, "EasyGrid: towards a framework for the automatic Grid enabling of legacy MPI applications: Research Articles," *Concurrency and Computation: Practice and Experience*, vol. 16, no. 5, pp. 425–432, Apr. 2004.

[48] A. Sena, A. Nascimento, C. Boeres, and V. Rebello, "EasyGrid Enabling of Iterative Tightly-Coupled Parallel MPI Applications," in *Proceedings of the 2008 IEEE International Symposium on Parallel and Distributed Processing with Applications*, ser. ISPA '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 199–206.

[49] N. Mendelsohn, M. Gudgin, H. Ruellan, and M. Nottingham, "SOAP message transmission optimization mechanism," W3C, W3C Recommendation, Jan. 2005, http://www.w3.org/TR/2005/REC-soap12-mtom-20050125/.

[50] S. Lawson, M. Combellack, R. Feng, H. Mahbod, and S. Nash, *Tuscany SCA in Action*. Manning Publications Company, 2011.

[51] 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*, ser. AINA-2013, 2013.