Supporting biodiversity studies with the EUBrazilOpenBio Hybrid Data Infrastructure

Rafael Amaral[1], Rosa M. Badia[2],[3], Ignacio Blanquer[4*], Ricardo Braga-Neto[5], Leonardo Candela[6], Donatella Castelli[6], Christina Flann[7], Renato De Giovanni[5], William A. Gray[8],[9], Andrew Jones[8], Daniele Lezzi[2], Pasquale Pagano[6], Vanderlei Perez-Canhos[5], Francisco Quevedo[8], Roger Rafanell[2], Vinod Rebello[1], Mariane S. Sousa-Baena[5] and Erik Torres[4]

[1] *Instituto de Computação, Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil*
[2] *Department of Computer Sciences, Barcelona Supercomputing Center, Barcelona, Spain*
[3] *Artificial Intelligence Research Institute (IIIA), Spanish National Research Council (CSIC), Catalonia, Spain*
[4] *Institute of Instrumentation for Molecular Imaging (I3M), Universitat Politècnica de València, Valencia, Spain*
[5] *Centro de Referência em Informação Ambiental (CRIA), Campinas, SP, Brazil*
[6] *Istituto di Scienza e Tecnologie dell'Informazione (ISTI), Consiglio Nazionale delle Ricerche (CNR), Pisa, Italy*
[7] *Netherlands Centre for Biodiversity Naturalis, Wageningen University, The Netherlands*
[8] *School of Computer Sciences and Informatics, Cardiff University, Cardiff, United Kingdom*
[9] *Species 2000 Secreriat, Naturalis, Einsteinweg 2, 2333 CC Leiden, The Netherlands*

Received . . .

KEY WORDS: Biodiversity; Data Infrastructure; Virtual Research Environments; Cloud; Taxonomy; Ecological Niche Modelling;

1. INTRODUCTION

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

1.1. 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 emerged 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, the biodiversity domain is still affected by a number of data sharing and reuse problems [11].

[*] Correspondence to: Institute of Instrumentation for Molecular Imaging (I3M), Universitat Politècnica de València, Camino de Vera s/n, Valencia, Spain E-mail: `iblanque@i3m.upv.es`

New initiatives continue to create 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.

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 major 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* [19], the Extreme Science and Engineering Discovery Environment (*XSEDE*) [20] or the gateway to nanotechnology online simulation tools *nanoHUB.org* [21] have developed customised solutions for their user communities. Current efforts in Europe like *SCI-BUS* [22] are defining a flexible framework for developing science gateways based on the *gUSE/WS-PGRADE* [23] portal family. In the area of data management, the *D4Science* [24] project has developed the gCube technology with special focus on management of big data and the concept of Virtual Research Environments (VRE) as its user interface. D4Science supports biodiversity [25] and other user communities.

1.2. Objective and motivation

This article describes the achievements of the EUBrazilOpenBio project in creating an integrated infrastructure to assist research in biodiversity. The aim has been to reduce the need for researchers to access data from multiple sources and process them locally. 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. 2). Section 3 describes the infrastructure of EUBrazilOpenBio and Section 4 details the implementation of the use cases. Section 5 covers the validation of the use cases from the users' point of view, and Section 6 presents the conclusions.

2. THE USE CASES AND REQUIREMENTS

To demonstrate the benefits an infrastructure having the characteristics of that developed by EUBrazilOpenBio might bring to the biodiversity informatics community [26], the project uses the infrastructure facilities to realise two representative use cases: the integration of taxonomies and the 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. This has been proven through the validation described in Section 5.

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 of constructing models that can be used to estimate the suitability of the environmental conditions of a certain region for a given species to survive.

2.1. 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. 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 compilers of global species lists may not be aware of due to their localised distribution.

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 [27] and animals [28]) 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 appropriate 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 CoL information. In particular, megadiverse countries such as Brazil, can contribute a lot of data from its regional checklists to the CoL, helping to create new GSDs, as well as improving coverage of existing ones.

2.2. *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 [29]. 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 [30]. ENMs are usually generated by relating locations where the species is known to occur with environmental variables that are expected to influence its distribution [31]. 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 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 [32].

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) [33]. 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 more than 4 million plant specimen records.

The modelling strategy used by BVH involves generating individual models using five different techniques in openModeller [34] when the species has at least 20 occurrence points: Ecological-Niche Factor Analysis [35], GARP Best Subsets [36], Mahalanobis distance [37], Maxent [38] and One-class Support Vector Machines [39]. Model quality is assessed by 10-fold cross-validation and, in the end, a final model is created by merging the individual models into a single consensus model, which is then projected into the present environmental conditions for Brazil at a high-resolution.

2.3. Requirements

Requirements were identified in an iterative process of analysis and refinement by users, systems analysts and application developers. Four main groups of requirements were identified. 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.

Users are asking for simple, generic, yet flexible data specification approaches allowing them to clearly define the data they are looking for without dealing with the heterogeneities of the data providers, such as Darwin Core Archive [9] and OpenModeller CSV [34]. Means to increase the quality of the data (e.g., by removing repeated occurrence points obtained in a search) are required and the performance should not deviate by more than 10% from the same time measured from the reference. The services should interact with a number of web services including those offered by Catalogue of Life, GBIF, and List of Species of the Brazilian Flora[†].

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 ecological niche models by using 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. The infrastructure needs to take into account also the support of different back-ends widespread over the global geography of the project. This requirement is of special interest to Use Case II (cf. Sec. 2.2).

Thirdly, the infrastructure should provide a user friendly, integrated environment, where scientists will have innovative services supporting their data discovery and processing tasks as well as the ability to share and consume research results, e.g., the storage and sharing of ecological niche models with other users avoids recalculation and feedback on results, and the display of results of the pipelines included in the platform. The visualization of the most important file formats must be supported from the portal, and all the actions that cannot be served interactively should be treated as batch jobs that can be consulted in future sessions. However, the infrastructure should allow programmatic access to services such as Ecological Niche Modelling, in order to make it accessible by other client applications.

Fourthly, users must be able to upload their own data so that it can be readily processed using the rest of infrastructure 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. This includes checklists of taxa, occurrence points and layer maps. The management of layers needs special attention as they may occupy up to one GB each.

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[‡].

[†] These web services are described via dedicated web pages: Catalogue of Life Web Services `http://www.catalogueoflife.org/content/web-services`, GBIF Web Services `http://data.gbif.org/ws/`, List of Species of the Brazilian Flora Web Service `http://checklist.florabrasil.net/service`.
[‡] EUBrazilOpenBio Wiki `http://wiki.eubrazilopenbio.eu`

DOI: 10.1002/cpe

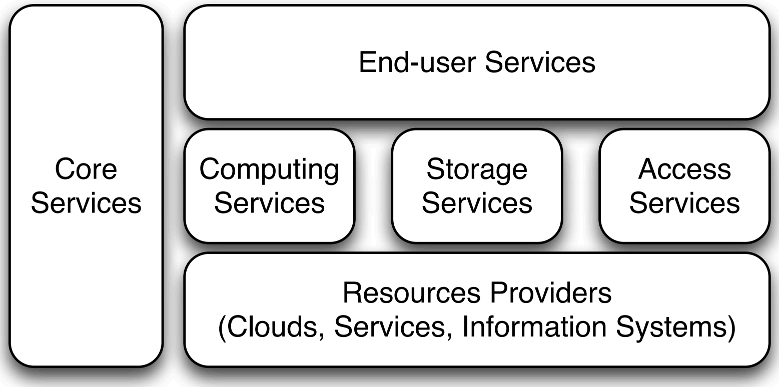


Figure 1. The Conceptual Architecture of the EUBrazilOpenBio Infrastructure

3. THE EUBRAZILOPENBIO INFRASTRUCTURE

The EUBrazilOpenBio Infrastructure is an innovative *Hybrid Data Infrastructure* [40], 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 [41, 42] 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 (including Grid and Cloud), services and Information Systems such as *GBIF* [3], *Catalogue of Life* [6], *speciesLink* [7], *List of Species of the Brazilian Flora* [8]. EUBrazilOpenBio Infrastructure integrates these systems with the aim of exploiting the synergy amongst them, and thus offer 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) [43, 44], 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 to Service Oriented Infrastructure patterns [45].

3.1. Core Services

Core services support the operation and management of the entire infrastructure. These services are provided by the gCube software framework [46, 43]. The Information Service has a key role here, as the infrastructure's registry, it supports resource discovery, monitoring, allocation 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 Resource 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.

3.2. Biodiversity Data Access Services

Biodiversity data access services offer facilities enabling seamless data access, integration, analysis, visualisation and use of biodiversity data, namely taxonomic, nomenclature, specimen and observational records. Such data represents a key resource for the target community but its extension across a number of Information Systems and databases makes 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 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 a 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 automatically augmented with "similar" species names by utilising synonyms resulting from the integrated data providers or species names belonging to a lower rank with respect to the specified species name. 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. 3.6). This is a fundamental facility for the two use cases (cf. Sec. 4).

3.3. Accessing Environmental Data

In addition to biodiversity data, biodiversity studies call for facilities for accessing environmental data, e.g., salinity, nitrate, precipitation, temperature. The EUBrazilOpenBio infrastructure hosts services forming a *Spatial Data Infrastructure*. In particular, it offers a catalogue service based on the GeoNetwork for the discovery and browsing of spatial datasets registered in the infrastructure. These datasets can be hosted on spatial data services of the infrastructure as well as harvested from existing catalogues and data services via standard protocols, e.g., CSW.

In fact, the infrastructure offers also spatial data services. In particular, it offers a GIS Publisher Service that enables 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 service is designed with a plug-in-oriented approach where each plug-in interfaces with a given back-end technology. The current implementation is capable of exploiting an array of THREDDS Data Server and GeoNetwork instances.

3.4. File-oriented Storage Services

The file-oriented storage facilities 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, three possible document store systems are used [47], MongoDB, Terrastore and USTO.RE [48]. 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 [49], while hiding the heterogeneous protocols of those systems from the services and applications exploiting the infrastructure storage facility.

3.5. 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, potentially 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 which 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 [50] programming framework and the EasyGrid AMS [51], 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* [52] 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 provided researchers with a suitable abstraction for scientific computing on top of virtualized resources. One of these resources is COMP Superscalar [50], 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 [53].

HTCondor [54] 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) [55], 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. Since 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* [51]. 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 [56]. 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.

For computing-intensive applications, an Experiment Orchestrator Service (EOS) has been developed. The EOS provides meta-scheduler functionalities to select the best resource to execute an application request. Such a decision depends on multiple factors such as the load of each computing infrastructure, the infrastructure availability, and application-specific metrics, such as the mean execution time, the availability of data in the local repository or the availability of a specific software configuration. A Job Resources Optimizer (JRO) tries to optimize the resources usage on the chosen resource provider by analysing the application topology and trying to guess which are the computational needs to obtain the best performance. The EOS service relies on the JRO for setting up the number of needed computing units, and its specifications in case of being virtual resources. The EOS retrieves the information from the Information Service, filled with data provided by each computational backend.

In order to support and to extend the access to a larger number of biodiversity scientists, the infrastructure has been enriched with the development of extensions to allow the access to federated cloud infrastructures. The main requirement to achieve this goal is interoperability at various levels as resources provision, authentication and authorization mechanisms, and deployment of the applications across multiple providers. The project followed the approach of the EGI Cloud Infrastructure Platform whose federation model provides an abstract cloud management stack to operate an infrastructure integrated with the components of the EGI Core Infrastructure Platform. This model defines the deployment of interoperable standard interfaces for VM management (OCCI [57]), data management (CDMI [58]) and information discovery (GLUE 2) [59]. The interaction with the core EGI components includes the integration with the AAI VOMS [60] system and with the monitoring and accounting services. An Appliance Repository and a VM Marketplace support the sharing and deployment of appliances across the providers. The EUBrazilOpenBio infrastructure supports the EGI Federated Cloud through the COMPSs-PMES components. A new connector has been developed in the COMPSs runtime to operate with the interfaces and protocols previously described. The VM management is implemented through a rOCCI [61] client connector that transparently interoperates with the rOCCI servers deployed in the testbed on top of different middlewares, such as OpenNebula and OpenStack. The connector supports different authentication methods, including X509 through VOMS proxy certificates. The connector is able to match the requirements of each task composing the application with the resource templates available on each provider.

3.6. End-user Services

End-user services provide human users with facilities benefitting and building upon the resources aggregated by the infrastructure. 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).

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, the portlet allows easy exchange of objects among users as well as import/export of objects from/to a user's file system to enable the processing of such objects using both the infrastructure and users' local computers.

4. IMPLEMENTATION OF THE USE CASES

The facilities of the EUBrazilOpenBio infrastructure are made available via a dedicated portal[§] 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

[§] `https://portal.eubrazilopenbio.d4science.org`

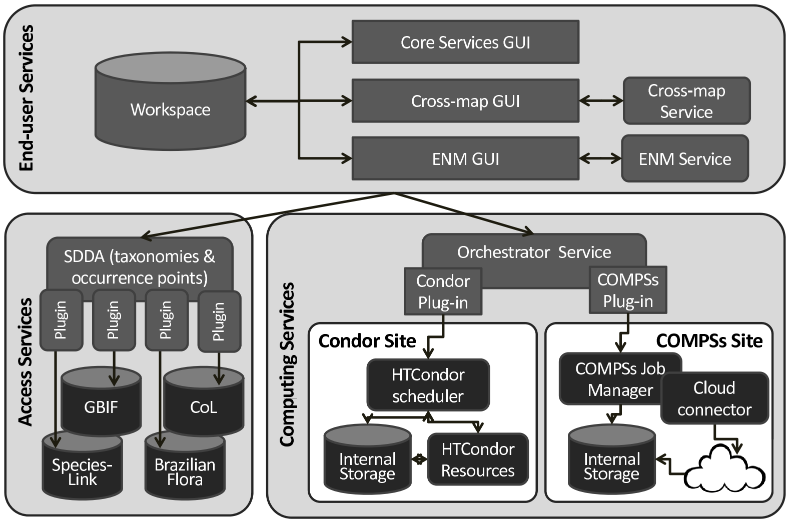


Figure 2. Use Case Architecture Deployment

one use case. These portlets are integrated with others portlets, namely SDDA and Workspace for data access. Each use case's portlet interacts with the processing services through specific services that implement 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.

4.1. 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[¶] 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 also achieved a reduction in the execution time of large cross-mapping tasks 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 message transmission optimisation mechanism [62] which exposes a set of methods that allows 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

[¶] `http://www.i4life.eu/`

DOI: 10.1002/cpe

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 [63], a framework that allows declarative exposure of Java components into a plethora of different protocols, such as SOAP, REST or Java Message Service.

The portlet (cf. Figure 5) was developed jointly by CNR and Cardiff University; it is basically a GWT [64] project that uses the GXT [65] 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.

4.2. EUBrazilOpenBio Niche Modelling Facilities

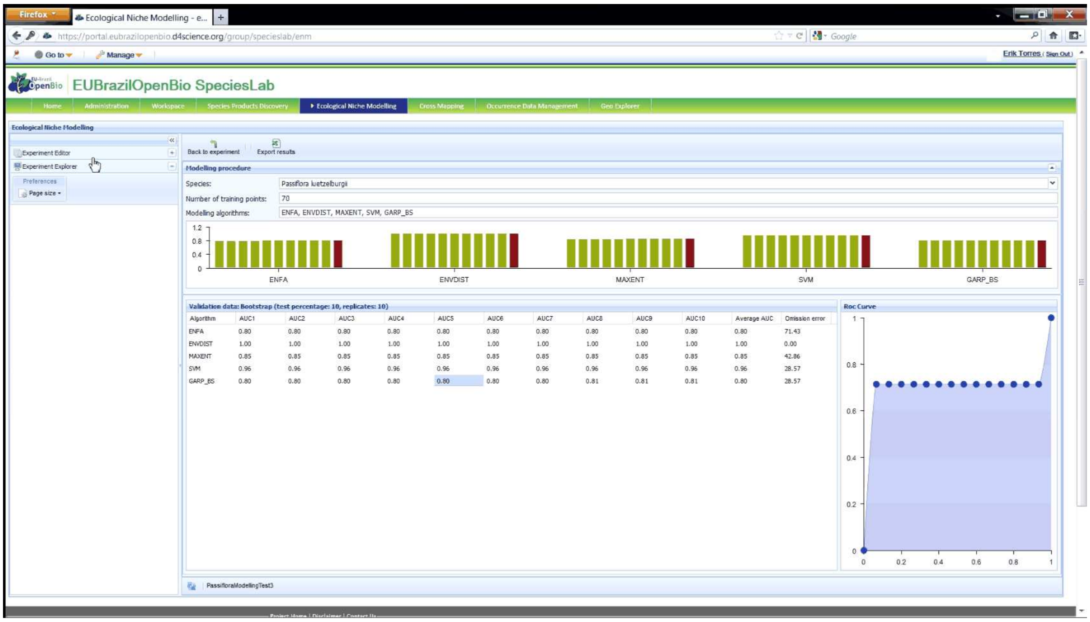
The implementation of Niche Modelling Facilities implied the development of (*i*) an ENM service offering the niche modelling facilities via a revised version of the openModeler protocol; (*ii*) a dedicated portlet interfacing with such a service; and (*iii*) a multi-parametric, multi-stage openModeller workflow implemented through COMPSs.

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 and several independent submissions are needed to perform several actions on the same dataset. A new version of the openModeller Web Service API (namely OMWS2) developed as part of this project, includes a new operation that allows multi-stage and multi-parameter experiments to be specified in a single request. This frees the clients from managing the workflow dependencies and allows back-ends such as COMPSs and HTCondor to orchestrate the execution after automatically generating an execution graph (cf. Figure 2).

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 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. Figure 3 shows the general appearance of the graphical interface of the ENM.

Programmatic access to the data is required by the services that support ecological niche modelling (OMWS and ENM submission services) and by the computing back-ends. However, final users are more likely to use graphical interfaces. The user workspace provides an abstraction of the storage service, which can be accessed from the portal. Workspace objects are presented in an interactive environment that can be used by biodiversity scientists with minimal understanding of the technological infrastructure.

The OMWS2 extension is 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 COMPSs-PMES service can be configured to boot a configurable number of VMs on the provider where the service is deployed; this solution allows the serving of requests that involve simple openModeller operations in a reasonable time avoiding the overhead of VM creation; if the number of requests exceeds the available resources, the service is still able to dynamically deploy new instances in order to cope with the burst of load. The ENM service is integrated with the Experiment Orchestrator Service (EOS).



Species: Passiflora luetzelburgii

Number of training points: 70

Modelling algorithms: ENFA, ENVDIST, MAXENT, SVM, GARP_BS

Validation data: Bootstrap (test percentage: 10, replicates: 10)

Algorithm	AUC1	AUC2	AUC3	AUC4	AUC5	AUC6	AUC7	AUC8	AUC9	AUC10	Average AUC	Omission error
ENFA	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	71.43
ENVDIST	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
MAXENT	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	42.86
SVM	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	28.57
GARP_BS	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.81	0.81	0.81	0.80	28.57

Figure 3. Appearance of the Ecological Niche Modelling GUI

This version of the ENM service has been provided to the BioVeL project [66] for the evaluation of the execution of workflows in the EGI Federated Cloud[||]. Leveraging the information provided by the Orchestrator Service, the ENM service is able to balance the load across different providers that expose a COMPSs-PMES endpoint. The EUBrazilOpenBio ENM endpoint is also officially available in the Biodiversity Catalogue[**], a Web Services registry publicly offered to scientists for the composition of their workflows.

To validate the workflow implementation and to evaluate the advantage of the elasticity features of these services, a test environment consisting of 10 quad-core virtual instances with 2GB of memory and 1GB of disk space was created on the EUBrazilOpenBio 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 [34]) 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 [67] 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.

Another test aimed at evaluating the performance of the PMES-COMPSs service in a scenario with multiple requests sent through the ENM service. In this experiment, each request requires low computation time, generating a dependency graph composed of only three OM tasks (*model*, *test* and *project*) where the last two are executed in parallel. The test used resources of the EGI Federated Cloud, configuring the PMES to deploy different types of virtual machines to serve different types of openModeller operations. Figure 4 depicts the response time together with the

[||] `https://wiki.egi.eu/wiki/FedCloudOPENMODELLER`
[**] `https://www.biodiversitycatalogue.org`

evolution of instances used in the execution. As represented in the *Resources* curve, two *XLarge* and one *Large* machines are pre-deployed to execute *Model* and *Test*, *Project* operations respectively. *XLarge* instances consist of 24-core servers, 96 GB of memory and 44 TB of shared filesystem. When multiple requests are issued at the same time the response time increases and the COMPSs runtime reacts to the rise of the load produced, adapting the number of resources needed to execute every kind of new task, according to the task constraints.

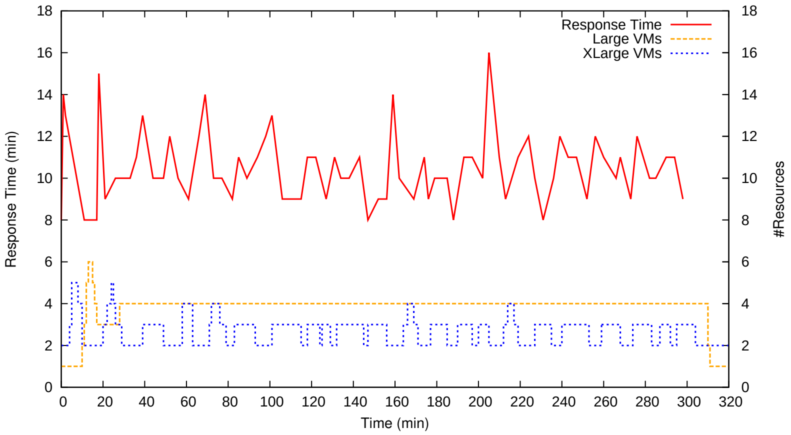


Figure 4. Response time *vs.* resource consumption.

5. VALIDATION OF THE USE CASES

Validation of the EUBrazilOpenBio infrastructure was performed through several pilot experiments defined by experts. These consisted of relevant, representative problems that could benefit from using the infrastructure and its VRE services.

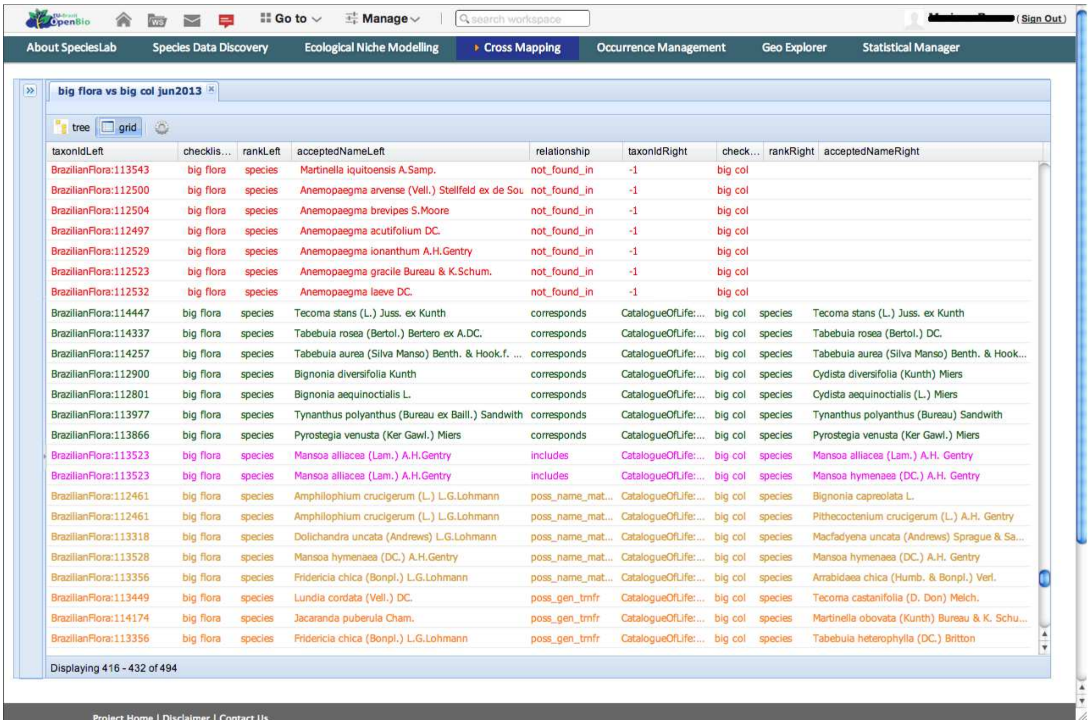
5.1. Validation of the EUBrazilOpenBio Taxonomy Management Facilities

Three different types of validation were carried out by cross-mapping taxa from the regional plant catalogue of Brazil – the List of Species of the Brazilian Flora (LSBF) – with the global index of plants within the the Catalogue of Life (CoL). These validations had different aims. The initial validation, using the *Angiosperms*, determined that the cross-mapping produced valid links between the lists. This was followed by validations aimed at two different uses of the cross-mapper that improve/enhance the CoL – plugging gaps in its coverage by creating a new protoGSD for the *Bignoniaceae*, and improving an existing GSD (the *Asteraceae* section of the Global Compositae Checklist (GCC)).

The assessment using *Angiosperms* revealed that 8909 species and 470 genera of flowering plants present in LSBF are not found in CoL. Also, 38,660 species names from LSBF do not entirely match with data in the CoL due to the possible transference of species from one genus to another in the LSBF. This clearly showed that cross-mapping produces useful, valid information and that the LSBF can be used to improve the CoL's content.

Bignoniaceae was selected for the second validation, because its greatest diversity is found in northern South America, and many specialists are currently working on various aspects of its biology in Brazil. Thus, it was expected that the LSBF would contain a large number of *Bignoniaceae* species not in the CoL. The family encompasses 85 genera and 860 species, of which 383 belong to a large tribe that represents the most diverse and abundant clade of lianas in the Neotropics [68]. The aim was to study the differences between the checklists to determine whether the results of this experiment could be used to create a proto-GSD of *Bignoniaceae* for future inclusion in the CoL. An analysis of the results showed that of the 393 species of *Bignoniaceae*

DOI: 10.1002/cpe



taxonIdLeft	checklis...	rankLeft	acceptedNameLeft	relationship	taxonIdRight	check...	rankRight	acceptedNameRight
BrazilianFlora:113543	big flora	species	Martinella iquitoensis A.Samp.	not_found_in	-1	big col		
BrazilianFlora:112500	big flora	species	Anemopaegma arvense (Vell.) Stellfeld ex de Sou	not_found_in	-1	big col		
BrazilianFlora:112504	big flora	species	Anemopaegma brevipes S.Moore	not_found_in	-1	big col		
BrazilianFlora:112497	big flora	species	Anemopaegma acutifolium DC.	not_found_in	-1	big col		
BrazilianFlora:112529	big flora	species	Anemopaegma ionanthum A.H.Gentry	not_found_in	-1	big col		
BrazilianFlora:112523	big flora	species	Anemopaegma gracile Bureau & K.Schum.	not_found_in	-1	big col		
BrazilianFlora:112532	big flora	species	Anemopaegma laeve DC.	not_found_in	-1	big col		
BrazilianFlora:114447	big flora	species	Tecoma stans (L.) Juss. ex Kunth	corresponds	CatalogueOfLife:...	big col	species	Tecoma stans (L.) Juss. ex Kunth
BrazilianFlora:114337	big flora	species	Tabebuia rosea (Bertol.) Bertero ex A.DC.	corresponds	CatalogueOfLife:...	big col	species	Tabebuia rosea (Bertol.) DC.
BrazilianFlora:114257	big flora	species	Tabebuia aurea (Silva Manso) Benth. & Hook.f. ...	corresponds	CatalogueOfLife:...	big col	species	Tabebuia aurea (Silva Manso) Benth. & Hook...
BrazilianFlora:112900	big flora	species	Bignonia diversifolia Kunth	corresponds	CatalogueOfLife:...	big col	species	Cydista diversifolia (Kunth) Miers
BrazilianFlora:112801	big flora	species	Bignonia aequinoctialis L.	corresponds	CatalogueOfLife:...	big col	species	Cydista aequinoctialis (L.) Miers
BrazilianFlora:113977	big flora	species	Tynanthus polyanthus (Bureau ex Baill.) Sandwith	corresponds	CatalogueOfLife:...	big col	species	Tynanthus polyanthus (Bureau) Sandwith
BrazilianFlora:113866	big flora	species	Pyrostegia venusta (Ker Gawl.) Miers	corresponds	CatalogueOfLife:...	big col	species	Pyrostegia venusta (Ker Gawl.) Miers
BrazilianFlora:113523	big flora	species	Mansoa alliacea (Lam.) A.H.Gentry	includes	CatalogueOfLife:...	big col	species	Mansoa alliacea (Lam.) A.H. Gentry
BrazilianFlora:113523	big flora	species	Mansoa alliacea (Lam.) A.H.Gentry	includes	CatalogueOfLife:...	big col	species	Mansoa hymenaea (DC.) A.H. Gentry
BrazilianFlora:112461	big flora	species	Amphilophium crucigerum (L.) L.G.Lohmann	poss_name_mat...	CatalogueOfLife:...	big col	species	Bignonia capreolata L.
BrazilianFlora:112461	big flora	species	Amphilophium crucigerum (L.) L.G.Lohmann	poss_name_mat...	CatalogueOfLife:...	big col	species	Pithecoctenium crucigerum (L.) A.H. Gentry
BrazilianFlora:113318	big flora	species	Dolichandra uncata (Andrews) L.G.Lohmann	poss_name_mat...	CatalogueOfLife:...	big col	species	Macfadyena uncata (Andrews) Sprague & Sa...
BrazilianFlora:113528	big flora	species	Mansoa hymenaea (DC.) A.H.Gentry	poss_name_mat...	CatalogueOfLife:...	big col	species	Mansoa hymenaea (DC.) A.H. Gentry
BrazilianFlora:113356	big flora	species	Fridericia chica (Bonpl.) L.G.Lohmann	poss_name_mat...	CatalogueOfLife:...	big col	species	Arrabidaea chica (Humb. & Bonpl.) Verl.
BrazilianFlora:113449	big flora	species	Lundia cordata (Vell.) DC.	poss_gen_trnfr	CatalogueOfLife:...	big col	species	Tecoma castanifolia (D. Don) Melch.
BrazilianFlora:114174	big flora	species	Jacaranda puberula Cham.	poss_gen_trnfr	CatalogueOfLife:...	big col	species	Martinella obovata (Kunth) Bureau & K. Schu...
BrazilianFlora:113356	big flora	species	Fridericia chica (Bonpl.) L.G.Lohmann	poss_gen_trnfr	CatalogueOfLife:...	big col	species	Tabebuia heterophylla (DC.) Britton

Displaying 416 - 432 of 494

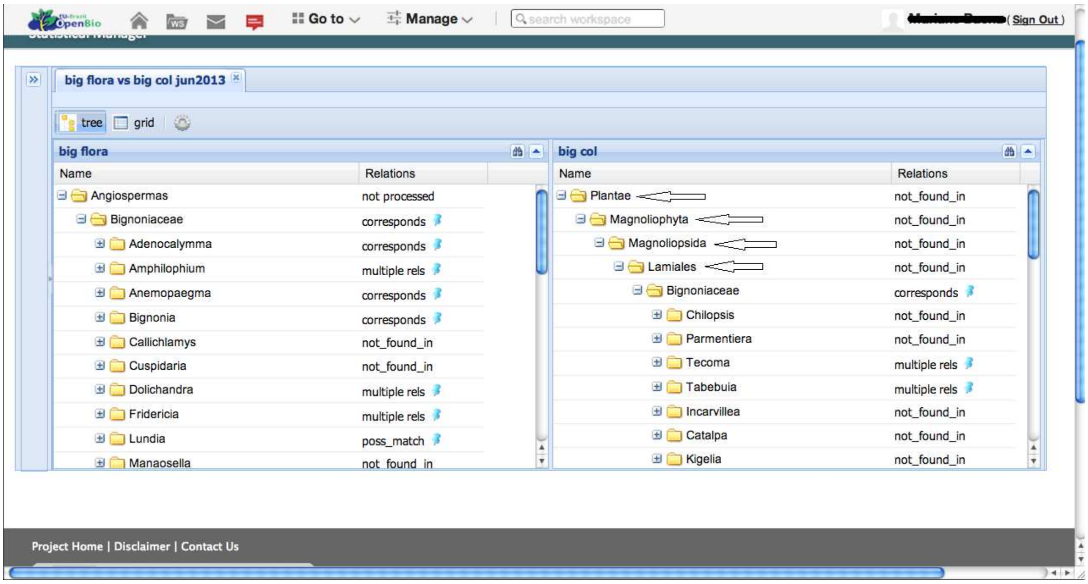
Figure 5. Snapshots illustrating the results of the cross-mapping experiment for *Bignoniaceae*. Top: Taxa from the checklist on the left (LSBF) that are not found in the checklist on the right (CoL) appear in red. A variety of types of relationships between species are pointed out, with each status in a different colour.

present in LSBF, 368 were not in the CoL. More specific information on taxonomic relationship is also provided by the cross-map. For instance, *Mansoa alliaceae* (accepted name) from the LSBF overlaps with *Mansoa hymenaea* (accepted name) from CoL as they share a synonym, *Adenocalymma obovatum* (Figure 5). Thus, the addition of a new GSD based on the LSBF's *Bignoniaceae* information will improve the CoL coverage of this plant family. The significant difference in the number of genera in these lists is partially due to the recent change in the *Bignoniaceae* taxonomy [69].

The *Asteraceae* family was chosen for the third validation, as it is already part of the GCC. Data from the LSBF was incorporated into the GCC in 2010, so any differences between the lists would be due to the fact that either there has been a more recent study, a different taxonomy has been used, or there were errors due to the tools and processes used in 2010. This validation was carried out by the GCC Custodian whose full report is available on the Species 2000 web site [70]. The GCC is a global list and it contains more than 30,000 names not in the LSBF (most of these are species not found in Brazil). While the LSBF contains 243 names not in CoL, 120 were missing due to name string mismatches, 12 were missing names and the remainder needing much fuller taxonomic investigation. This validation demonstrated the potential of the cross-mapper to enhance a GSD's contents as it led to 366 LSBF names being added to the GCC in the July 2013 edition of the CoL and further additions to the GCC in the September 2013 edition.

These comparisons between the checklists, if done manually, would have been very laborious and time consuming, and for *Angiosperms* and *Asteracea* probably not even possible. Although, using the cross-mapper tool, all analyses were done relatively quickly. In the process of analysing larger taxonomic groups, such as Divisions and Classes, we detected that the classification system used by the CoL and the LSBF are indeed different, for taxonomic ranks above family level (Figure 6); which does not hinder the comparisons by family, but makes any test above family level nearly impossible. These initial analyses revealed that the LSBF content and the CoL content is quite divergent, regarding species presence/absence and also classification systems employed.

DOI: 10.1002/cpe



big flora

Name	Relations
Angiospermas	not processed
Bignoniaceae	corresponds
Adenocalymma	corresponds
Amphilophium	multiple rels
Anemopaegma	corresponds
Bignonia	corresponds
Callichlamys	not_found_in
Cuspidaria	not_found_in
Dolichandra	multiple rels
Fridericia	multiple rels
Lundia	poss_match
Manaosella	not_found_in

big col

Name	Relations
Plantae	not_found_in
Magnoliophyta	not_found_in
Magnoliopsida	not_found_in
Lamiales	not_found_in
Bignoniaceae	corresponds
Chilopsis	not_found_in
Parmentiera	not_found_in
Tecoma	multiple rels
Tabebuia	multiple rels
Incarvillea	not_found_in
Catalpa	not_found_in
Kigelia	not_found_in

Figure 6. Snapshots illustrating the results of the more general cross-mapping experiment. Cross-mapping results can also be depicted in a tree format, in which relationships of higher taxonomic groups between the lists are promptly discernible. Arrows point groups that are considered by CoL, but not by LSBF

The results obtained in this validation indicate that national efforts should be carried out by countries to follow the same data format used by the CoL when producing their own lists, which is relatively simple. Another conclusion is that regional and local checklists are usually much more complete than the CoL regarding data on endemic species and can significantly contribute to its improvement. We have also shown that the cross-mapper has a significant role to play when new GSDs are being created and in improving existing GSDs. These results are a first step toward the alignment and standardisation of the taxonomies in regional and global catalogues. Having a taxonomically standardised global catalogue will allow the development of biodiversity studies in large scales, from which strategies for conservation may be defined.

5.2. Validation of the EUBrazilOpenBio Ecological Niche Modelling Facilities

The ENM service has been validated in two forms. First, the current service implementation successfully passed a set of tests for OMWS 1.0, which guarantees the compatibility with existing OMWS clients. Second, the interface developed for UCII was also used to generate ecological niche models to evaluate the potential distribution of selected plant species which are of conservation concern in Brazil. In 2008, the Brazilian Ministry of Environment published a Normative Instruction that listed possibly endangered species due to insufficient data (Annex II of [71]. However, the amount of information actually available through speciesLink may now be enough to allow a reassessment of their knowledge status regarding distribution. Hence, five *Bignoniaceae* species were selected to generate potential distribution models based on their ecological niches: *Jacaranda ulei*, *Godmania dardanoi*, *Tabebuia cassinoides*, *Handroanthus spongiosus* and *Adenocalymma dichilum*. Occurrence data was retrieved for each species and the corresponding records were filtered to check taxonomic and georeferencing inconsistencies. To generate the species' potential distributions maps, the ENMs were configured with the occurrence points, selected environmental variables, coordinate system, spatial resolution and algorithms. These models were evaluated and projected to depict the potential area of distribution for each species in Brazil.

Some models generated consistent results, while others can be considered preliminary models requiring more data to be improved (cf. Figure 7). The model for *Godmania dardanoi* was based on 24 points and indicates that its potential distribution is restricted to the semiarid region of Brazil. Models for *Adenocalymma dichilum* and *Handroanthus spongiosus*, based on 12 and 13 points respectively, predicted that the potential distribution of these species is also restricted to the semiarid domain. These patterns seem consistent with their biological traits. On the other hand, despite the model for *Jacaranda ulei* being based on 63 points, it indicates that this species potentially has a broad distribution area in central Brazil, far beyond the distribution of the available occurrence points. More data is required to determine its distribution limits. Similarly, the model for *Tabebuia cassinoides* shows that the potential area of distribution is composed of disjunct regions, including coastal, central, and western Brazil. The model was based on 13 points and the pattern indicates that more sampling is required to determine if the accessibility to environmentally similar but geographically distant areas is indeed constrained by dispersal limitation or is an artifact.

Results indicate that the e-infrastructure can be used to generate ecological niche models which in turn can be used in further research and conservation actions. Moreover, as data quality and coverage increase, there are countless examples where the ENM services could be employed, contributing to safeguard biodiversity and environmental services. As the ENM services are provided through friendly and comprehensive interfaces, the e-infrastructure is prone to be widely used by the scientific community and decision makers. Several benefits do contribute to attract the interest of users, most of which are related to the easiness of access to data integrated with up-to-date modelling facilities that produce fast and reliable results.

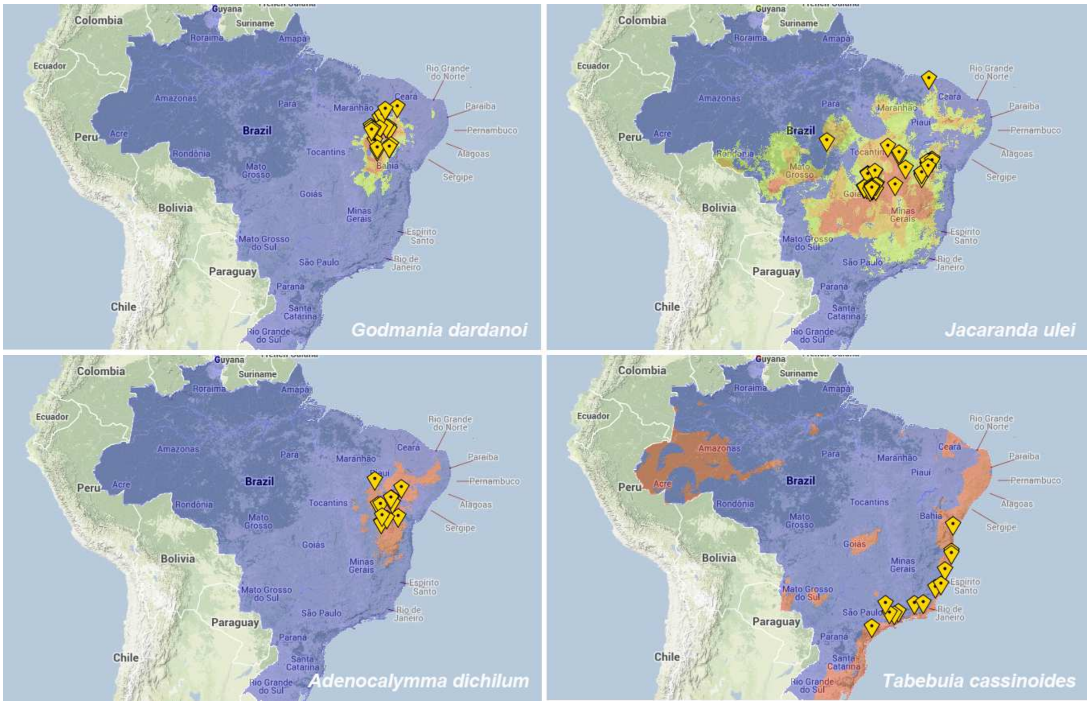


Figure 7. Snapshots of the projected models over the area of Brazil for selected Bignoniaceae species. Models were generated with different algorithms depending on the number of occurrence points available. For *Godmania dardanoi* and *Jacaranda ulei*, with more than 20 occurrence points, five algorithms were used: Maxent, GARP-BS, Mahalanobis Distance, ENFA and one-class SVM. Then, the models were transformed into binary models through a cut based on the lowest presence threshold (LPT) and aggregated into a single consensus model that displays the places where at least three of the algorithms agree (in red all algorithms agree, in orange four and in yellow three). For *Adenocalymma dichilum* and *Tabebuia cassinoides* with less than 20 points, two models were generated (Maxent and GARP-BS). Both models were transformed into binary models based on the LPT and then aggregated into a single consensus model which displays only those places where there is agreement between the two algorithms (in red). Yellow symbols depict the occurrence points used to generate the models.

DOI: 10.1002/cpe

6. 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 to and from 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.

One of the many advantages of the EUBrazilOpenBio data infrastructure is that each actor in the system (users, services, computing back-ends) has the possibility to decide which of the available methods to access the data is more suitable for its purposes. Moreover, independently of the method used to access the data, its security, integrity and consistency is ensured across the infrastructure. In this way, for example, new results obtained and copied to the storage in a remote computing back-end are immediately available to users, who only need a standard web browser to display and analyse the results in the portal.

This ubiquitous access to the information also allows synchronisation of the data between the different computing back-ends where the experiment requests are executed. This is especially important for environmental datasets, which often are large and cause difficulties because of their size. By relying on the storage service to access the data, new providers can enter the infrastructure easily.

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. A clear example of how interoperability is fostered in the infrastructure is implementation of the ENM Service that allows to access resources external to the project testbed, such as the EGI Federated Cloud.

The main 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. This has been demonstrated through the use case applications, which integrated a complete workflow including data retrieval, processing, visualization, storage and data sharing. EUBrazilOpenBio provides a complete research environment to users.

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 or downloading output files. Users of this integrated framework also benefit from the high-performance computing back-ends of the platform. Services such as the ENM facility have been made available to the user community of BioVeL to increase the variety of experiments in the validation.

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. BSC authors also acknowledge the support of the grant SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government and the Spanish Ministry of Science and Innovation under contract TIN2012-34557 and the Generalitat de Catalunya (contract 2009-SGR-980).

References

1. EUBrazilOpenBio Consortium. EU-Brazil Open Data and Cloud Computing e-Infrastructure for Biodiversity. `http://www.eubrazilopenbio.eu/` 2013.
2. Triebel D, Hagedorn G, Rambold G. An appraisal of megascience platforms for biodiversity information. *MycoKeys* 2012; **5**:45–63.
3. Edwards JL, Lane MA, Nielsen ES. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* 2000; **289**(5488):2312–2314.
4. Grassle J. 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* 2000; **13**(3):5–7.
5. Constable H, Guralnick R, Wieczorek J, Spencer C, Peterson ea A Townsend. Vertnet: A new model for biodiversity data sharing. *PLoS Biol* 02 2010; **8**(2):e1000 309, doi:10.1371/journal.pbio.1000309.
6. Roskov Y, Kunze T, Paglinawan L, Orrell T, Nicolson D, Culham A, Bailly N, Kirk P, Bourgoin T, Baillargeon G, *et al.*. Species 2000 & ITIS Catalogue of Life March 2013. Digital resource at www.catalogueoflife.org/col/. Species 2000: Reading, UK.
7. speciesLink Consortium. speciesLink. `http://splink.cria.org.br` 2013. URL `http://splink.cria.org.br`.
8. List of Species of the Brazilian Flora Consortium. List of Species of the Brazilian Flora. `http://floradobrasil.jbrj.gov.br/` 2013. URL `http://floradobrasil.jbrj.gov.br/`, rio de Janeiro Botanical Garden.
9. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, De Giovanni R, Robertson T, Vieglais D. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 2012; **7**(1).
10. De Giovanni R, Copp C, Döring M, Güntscg A, Vieglais D, Hobern D, Torre J, Wieczorek J, Gales R, Hyam R, *et al.*. TAPIR - TDWG Access Protocol for Information Retrieval. `http://www.tdwg.org/activities/abcd/` 2010. Version 1.0.
11. Goddard A, Wilson N, Cryer P, Yamashita G. Data hosting infrastructure for primary biodiversity data. *BMC Bioinformatics* 2011; **12**(Suppl 5):S5.
12. Jetz W, McPherson JM, Guralnick RP. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution* 2012; **27**(3):151 – 159.
13. NICE Srl. Enginframe. `http://www.nice-software.com/products/enginframe` 2013.
14. Hiden H, Woodman S, Watson P, Cala J. Developing cloud applications using the e-science central platform. *Proceedings of Royal Society A* 2012; .
15. Glatard T, Montagnat J, Lingrand D, Pennec X. Flexible and Efficient Workflow Deployment of Data-Intensive Applications On Grids With MOTEUR. *International Journal of High Performance Computing Applications* 2008; **22**(3):347–360.
16. Kacsuk P, Sipos G. Multi-grid, multi-user workflows in the p-grade grid portal. *Journal of Grid Computing* September 2005; **3**(7-4):221–238.
17. Wassenaar T, Dijk Mv, Loureiro-Ferreira N, Schot Gvd, Vries Sd, Schmitz C, Zwan Jvd, Boelens R, Bonvin A. WeNMR: structural biology on the grid. *CEUR Workshop Proceedings* 2011; **819**(4):1–8.
18. Manuali C, Lagan A, Rampino S. GriF: A Grid framework for a Web Service approach to reactive scattering. *Computer Physics Communications* July 2012; **181**(7):11791185.
19. Goecks J, Nekrutenko A, Taylor J, The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 2010; **11**(8):R86, doi: 10.1186/gb-2010-11-8-r86.
20. XSEDE consortium. Extreme science and engineering discovery environment. `https://www.xsede.org/` 2013.
21. NanoGUBorg. Online simulation and more for nanotechnology. `https://www.xsede.org/` 2013.
22. SCI-BUS consortium. Scientific gateway based user support. `https://www.sci-bus.eu/` 2011.
23. Kacsuk P, Farkas Z, Kozlovszky M, Hermann G, Balasko A, Karoczkai K, Marton I. Ws-pgrade/guse generic dci gateway framework for a large variety of user communities. *Journal of Grid Computing* Jan-12-2012 2012; **10**:601 – 630, doi:10.1007/s10723-012-9240-5. URL `http://link.springer.com/article/10.1007\%2Fs10723-012-9240-5`.
24. Candela L, Castelli D, Pagano P. D4science: an e–infrastructure for supporting virtual research environments. *Post–proceedings of the 5th Italian Res. Conf. on Digital Libraries– IRCDL 2009*, 2009.
25. Candela L, Pagano P. The D4Science Approach toward Grid Resource Sharing: The Species Occurrence Maps Generation Case. *Data Driven e-Science - Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010)*, Lin SC, Yen E (eds.), Springer, 2011; 225–238, doi:10.1007/978-1-4419-8014-4_18.

DOI: 10.1002/cpe

26. Hardisty A, Roberts D, The Biodiversity Informatics Community. A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology* 2013; **13**(16):–, doi:10.1186/1472-6785-13-16.
27. McNeill J, *et al.*. *International Code of Nomenclature for algae, fungi and plants (Melbourne Code)*. Koeltz Scientific Books, 2012.
28. Ride W, *et al.*. *International Code of Zoological Nomenclature*. Fourth edn., The International Trust for Zoological Nomenclature, 1999.
29. Lobo J, Jiménez-Valverde A, Hortal J. The uncertain nature of absences and their importance in species distribution modelling. *Ecography* 2010; **33**:103–114, doi:http://dx.doi.org/10.111/j.1600-0587.2009.06039.x.
30. Grinnell J. Field tests of theories concerning distributional control. *American Naturalist* 1917; **51**:115–128.
31. Sobern J, Peterson A. Interpretation of models of fundamental ecological niches and species distributional areas. *Biodiversity Informatics* 2005; **2**:1–10.
32. Peterson A, Sobern J, Pearson R, Anderson R, Martinez-Meyer E, Nakamura M, Arajo M. *Ecological niches and geographic distributions*. Princeton University Press, 2011.
33. Brazilian Virtual Herbarium Consortium. Brazilian Virtual Herbarium. `http://biogeo.inct.florabrasil.net/` 2013.
34. Muñoz M, De Giovanni R, Siqueira M, Sutton T, Brewer P, Pereira R, Canhos D, Canhos V. openModeller: a generic approach to species potential distribution modelling. *Geoinformatica* 2001; **15**:111–135.
35. Hirzel AH, Hausser J, Chessel D, Perrin N. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology* 2002; **83**(7):2027–2036.
36. Anderson R, Lew D, Peterson A. Evaluating predictive models of species distributions: criteria for selecting optimal models. *Ecological Modelling* 2003; **162**:211–232.
37. Farber O, Kadmon R. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the mahalanobis distance. *Ecological Modelling* 2003; **160**:115–130.
38. Phillips S, Anderson R, , Schapire R. Maximum entropy modelling of species geographic distributions. *Ecological Modelling* 2006; **190**:231–259.
39. Schölkopf B, Platt J, Shawe-Taylor J, Smola A, Williamson R. Estimating the support of a high-dimensional distribution. *Neural Computation* 2001; **13**(7):1443–1471, doi:doi:10.1162/089976601750264965.
40. Candela L, Castelli D, Pagano P. Managing big data through hybrid data infrastructures. *ERCIM News* 2012; (89):37–38.
41. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, *et al.*. A view of cloud computing. *Communications of the ACM* Apr 2010; **53**(4):50–58, doi:10.1145/1721654.1721672.
42. Foster I, Zhao Y, Raicu I, Lu S. Cloud Computing and Grid Computing 360-Degree Compared. *Grid Computing Environments Workshop, 2008. GCE '08*, 2008.
43. Candela L, Castelli D, Pagano P. Making Virtual Research Environments in the Cloud a Reality: the gCube Approach. *ERCIM News* October 2010; (83):32–33.
44. Candela L, Castelli D, Pagano P. Virtual research environments: an overview and a research agenda. *Data Science Journal* 2013; **12**:GRDI75–GRDI81, doi:10.2481/dsj.GRDI-013.
45. Tsai W. Service-oriented system engineering: a new paradigm. *Service-Oriented System Engineering, 2005. SOSE 2005. IEEE International Workshop*, 2005; 3 – 6, doi:10.1109/SOSE.2005.34.
46. Candela L, Castelli D, Pagano P. gCube: A Service-Oriented Application Framework on the Grid. *ERCIM News* January 2008; (72):48–49. URL `http://ercim-news.ercim.eu/en72/rd/gcube-a-service-oriented-application-framework-on-the-grid`.
47. Cattell R. Scalable SQL and NoSQL data stores. *SIGMOD Rec.* May 2011; **39**(4):12–27, doi:10.1145/1978915.1978919.
48. Durão FA, Assad RE, Silva AF, Carvalho JF, Garcia VC, Trinta FAM. USTO.RE: A Private Cloud Storage System. *13th International Conference on Web Engineering (ICWE 2013) - Industry track*, Aalborg, 2013.
49. Anderson JC, Lehnardt J, Slater N. *CouchDB: The Definitive Guide*. O'Really, 2009.
50. Lezzi D, Rafanell R, Carrión A, Blanquer I, Hernández V, Badia RM. Enabling e-science applications on the cloud with compss. *Proc. of the 2011 intl. conf. on Parallel Processing*, Euro-Par'11, Springer-Verlag: Berlin, Heidelberg, 2012; 25–34.
51. Boeres C, Rebello VEF. EasyGrid: towards a framework for the automatic Grid enabling of legacy MPI applications: Research Articles. *Concurrency and Computation: Practice and Experience* Apr 2004; **16**(5):425–432.
52. VENUS-C consortium. Deliverable 6.1 - report on architecture 2012. URL `http://www.venus-c.eu/Content/Publications.aspx?id=bfac02a9-9bc0-4c8f-80e0-7ceddc5c893b`.
53. Marozzo F, Lordan F, Rafanell R, Lezzi D, Talia D, Badia RM. Enabling cloud interoperability with compss. *Euro-Par*, *Lecture Notes in Computer Science*, vol. 7484, Kaklamanis C, Papatheodorou TS, Spirakis PG (eds.), Springer, 2012; 16–27.
54. Thain D, Tannenbaum T, Livny M. Distributed computing in practice: the Condor experience. *Concurrency - Practice and Experience* 2005; **17**(2-4):323–356.
55. Couvares P, Kosar T, Roy A, Weber J, Wenger K. Workflow in Condor. *Workflows for e-Science*, Taylor I, Deelman E, Gannon D, Shields M (eds.). Springer Press, 2007.
56. Sena A, Nascimento A, Boeres C, Rebello V. EasyGrid Enabling of Iterative Tightly-Coupled Parallel MPI Applications. *Proceedings of the 2008 IEEE International Symposium on Parallel and Distributed Processing with Applications*, ISPA '08, IEEE Computer Society: Washington, DC, USA, 2008; 199–206.
57. Edmonds A, Metsch T, Papaspyrou A, Edmonds A, Metsch T, Papaspyrou A. Open Cloud Computing Interface. *Grid and Cloud Database Management*, Fiore S, Aloisio G (eds.). Springer Berlin Heidelberg, 2011; 23–48, doi: 10.1007/978-3-642-20045-8_2. URL `http://dx.doi.org/10.1007/978-3-642-20045-8_2`.
58. Livenson I, Laure E. Towards transparent integration of heterogeneous cloud storage platforms. *Proceedings of the fourth international workshop on Data-intensive distributed computing*, DIDC '11, ACM: New York, NY, USA, 2011; 27–34.

59. Andreozzi S, Burke S, Field L, Knya B. Towards GLUE2: evolution of the computing element information model. *J. Phys.: Conf. Ser.* 2008; **119**:062 009.
60. Alfieri R, Cecchini R, Ciaschini V, dell'Agnello L, Frohner A, Gianoli A, Lõrentey K, Spataro F. VOMS, an authorization system for virtual organizations. *Grid Computing*, *Lecture Notes in Computer Science*, vol. 2970, Fernandez Rivera F, Bubak M, Gomez Tato A, Doallo R (eds.). Springer Berlin Heidelberg, 2004; 33–40, doi: 10.1007/978-3-540-24689-3_5. URL http://dx.doi.org/10.1007/978-3-540-24689-3_5.
61. The rOCCI framework. http://dev.opennebula.org/projects/ogf-occi/wiki. Last visited on 10/01/2013.
62. Mendelsohn N, Gudgin M, Ruellan H, Nottingham M. SOAP message transmission optimization mechanism. *W3C recommendation*, W3C Jan 2005. Http://www.w3.org/TR/2005/REC-soap12-mtom-20050125/.
63. Lawson S, Combellack M, Feng R, Mahbod H, Nash S. *Tuscany SCA in Action*. Manning Publications Company, 2011.
64. Tacy A, Hanson R, Essington J, T"okke A. *GWT in Action*. Manning Publications Company, 2013.
65. Sencha. Sencha GXT application framework for Google web toolkit. http://www.sencha.com/products/gxt/ 2013.
66. Vicario S, Hardisty A, Haitas N. BioVeL: Biodiversity Virtual e-Laboratory. *EMBnet.journal* 2011; **17**(2):5–6.
67. Lezzi D, Rafanell R, Torres E, De Giovanni R, Blanquer I, Badia RM. Programming ecological niche modeling workflows in the cloud. *Proceed. of the 27th IEEE Int. Conf. on Advanced Information Networking and Applications*, AINA-2013, 2013.
68. Lohmann LG. Untangling the phylogeny of neotropical lianas (Bignonieae, Bignoniaceae). *American Journal of Botany* 2006; **93**:304–318.
69. Lohmann L. A new generic classification of Bignonieae (Bignoniaceae) based on molecular phylogenetic data and morphological synapomorphies). *Annals of the Missouri Botanical Garden* 2013; .
70. Flann C. Use Case Study EUBrazilOpenBio Cross-mapping tool Assessment of usability for regional-GSD comparisons. http://www.eubrazilopenbio.eu/Content/Factfile.aspx?id=0750dcd8-23f2-4bf1-bad4-52aa3277d002 June 2013.
71. Brazilian Ministry of Environment. Instrução normativa no. 6, 23 de setembro de 2008 2008. URL http://www.mma.gov.br/estruturas/179/_arquivos/179_05122008033615.pdf.