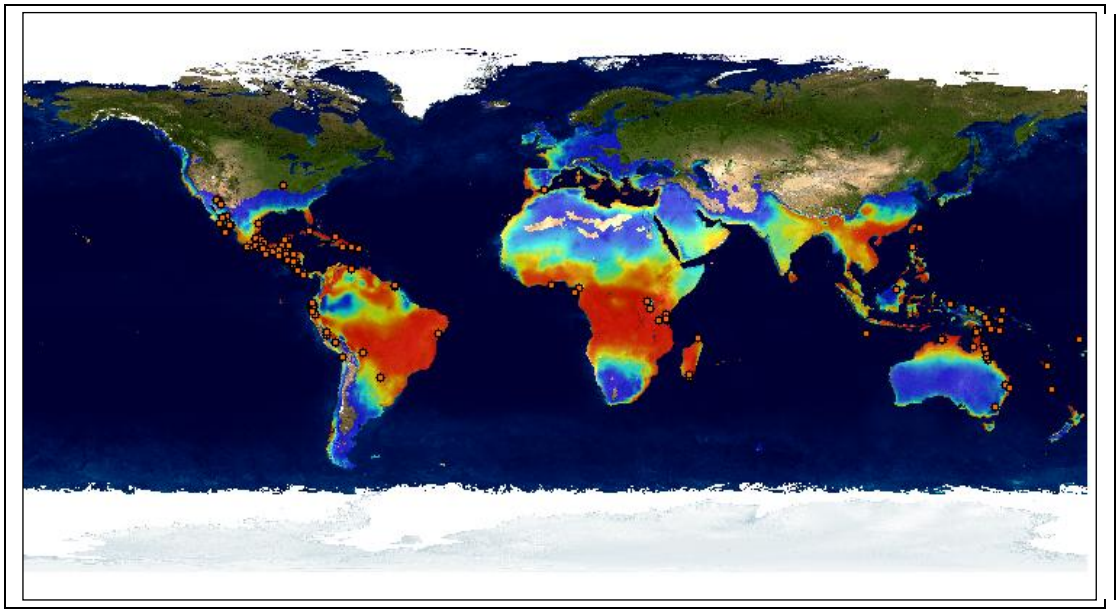
openModeller

A framework for species distribution modeling



Fapesp process: 04/11012-0

Final Report (April 2005 – October 2009)

Vanderlei Perez Canhos
coordenador do projeto

Index

Introduction

This report summarizes the activities carried out during the openModeller project, from April 2005 to October 2009. The main goal of this project was to develop a flexible and robust modeling framework to facilitate the work of scientists to predict species distributions under different scenarios. The project was funded by Fapesp and involved three institutions: CRIA (Centro de Referência em Informação Ambiental), Poli (Escola Politécnica da USP), and INPE (Instituto Nacional de Pesquisas Espaciais).

Objectives

The main objectives of the project were to:

- Develop a component-based modeling framework with reusable modules compliant with web services technology.
- Enable use and comparison of different modeling algorithms through the same framework.
- Enable use of pre-analysis and pos-analysis techniques using specific components.
- Develop multiple interfaces (web, desktop, command-line, web service).
- Facilitate access to distributed biological and environmental data networks.
- Allow usage of high performance computing in the modeling process.
- Carry out use cases to test and validate the framework.

Summary of main activities and results

During the project, there were 9 doctoral students, 5 master students, 7 undergraduate students and 6 fellowships for technical training involved, besides the direct involvement of staff members from each institution (CRIA, Poli, and INPE).

The number of publications includes 12 journal papers, 49 conference papers among other means of outreach. The number of international publications that mention openModeller is continuously increasing[1].

Fifteen releases of the openModeller framework were made – each release including bug fixes and new features – totalizing almost 6,000 downloads. The framework currently contains thirteen algorithms: AquaMaps, Artificial Neural Networks, Bioclim, Climate Space Model, Envelope Score, Environmental Distance, two versions of GARP (for both single run and Best Subsets), Maximum Entropy, Niche Mosaic and Support Vector Machines.

Occurrence data can be read from local TAB-delimited files, TerraLib[2] databases or remote services such as the *species*Link[3] and GBIF[4] portals or

[1] http://openmodeller.sourceforge.net/index.php?option=com_content&task=blogcategory&id=10&Itemid=6
[2] http://www.terralib.org
[3] http://splink.cria.org.br
[4] http://www.gbif.org

any TAPIR[5]/DarwinCore[6] service. More than 100 raster formats are supported for environmental layers, including remote services such as WCS[7]. The framework has different interfaces on top of it, including command-line/console, graphical user interface, web service, and web interface.

The framework and its Desktop interface have packages for all three major platforms: GNU/Linux, Mac OSX and Windows. More than 13,000 downloads of openModeller Desktop – the graphical user interface – were computed since the beginning of the project, reaching peaks of 600 downloads per month.

A computer cluster was purchased, installed and is available to researchers. A modeling service is running on the main cluster node, from where each individual request can be distributed across the other ten nodes. Parts of the framework code were parallelized to take advantage of cluster environments.

A service oriented infrastructure was developed to evaluate the potential of using service oriented architecture for ecological niche modelling. The implementation integrates niche modelling and GIS services.

The *species*Link network[8], which constitutes the main source of species occurrence data in the Brazilian territory and whose project originated openModeller, increased the number of records freely and openly available, from 550.000 records to approximately 3.2 million records during the project.

The publicity gained from regular software releases and interactions with other individuals and institutions contributed to the use of openModeller by other interfaces and tools, such as the GBIF portal and the LifeMapper[9] project.

Sixteen study cases were carried out during the project with the objective of testing the framework and offering feedback to developers, training graduate students, disseminating the tool through the involvement of users from other institutions, and producing papers to disseminate openModeller and its applications worldwide.

General framework activities and studies

Evaluation of parallelization techniques and analysis of parallel algorithms for biodiversity data analysis

Since openModeller may require intensive use of computational resources depending on the experiment, a detailed study was carried out to identify which tasks from the typical execution flow would take advantage of parallelization techniques. This analysis was performed using AOP (Aspect Oriented Programming) with asynchronous method invocation to monitor performance in various framework components. Two components were initially selected for optimization: the map projection module and the GARP

[5] http://www.tdwg.org/activities/tapir
[6] http://www.tdwg.org/activities/darwincore
[7] http://www.opengeospatial.org/standards/wcs
[8] http://splink.cria.org.br
[9] http://www.lifemapper.org

algorithm. Both were parallelized using LAM-MPI (Message Passing Interface) based on the master-slave model. Four different versions of the original projection code were developed and tested. Depending on input parameters (map resolution and extent), the final version achieved a 5.7 fold reduction in execution time when running with ten nodes (figure 1). To avoid file access latency, data partitioning is done dynamically by means of MPI messages.

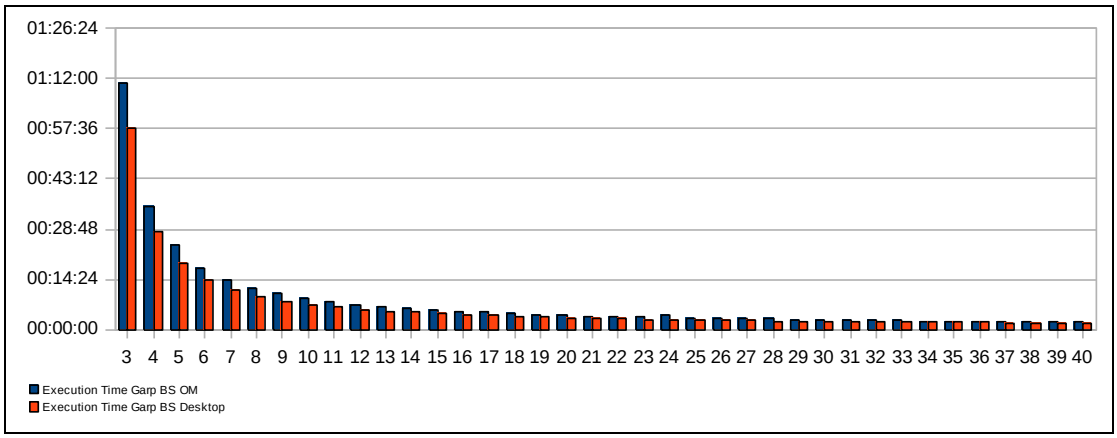


Figure 1: Performance (time) of the map projection procedure after parallelization. The same input data was used with two implementations of GARP Best Subsets: DesktopGarp implementation (red) and openModeller implementation (blue) – both are available in openModeller.

Parallelized versions known as P-GARP and P-GARP Best Subsets were created for GARP. During the last year, other parts of the code that were developed as part of this project were also parallelized: the jackknife algorithm used for pre-analysis (Rodrigues, F.A. et al. 2008) and the maximum entropy modeling algorithm (Rodrigues, E.S.C. et al. 2008).

A computer cluster was purchased as part of the project after analyzing different possibilities. The machine is an SGI Altix xE 1300 with 11 nodes, or 88 Cores of Intel Xeon 5335, with 2.0GHz/8MB cache, 1333MHz FSB and 88 GB RAM DDR 667MHz. An instance of the modeling service (also developed as part of this project) was installed on the main cluster node. The cluster was configured to distribute modeling tasks across all nodes using Condor as the management system. This procedure takes into account input data size and cluster load balancing. A web portal was also implemented to facilitate job submissions to the cluster. The parallelized version of the framework, including P-GARP and P-GARP Best Subsets, is successfully running on the cluster.

Identification of general requirements to access data and services

In the beginning of the project, a series of interviews and workshops were organized to understand all steps involved in species distribution modeling. Use-cases, limitations and opportunities for improvement were identified during the process. As a result, a reference business process for ecological niche modeling was published (Santana et al., 2008), serving as a starting point to plan a new architecture and helping to analyze software usability.

Being a multi-disciplinary project, these activities were also important to build a common understanding of all issues involved.

Definition of a service oriented architecture

Based on the reference business process for ecological niche modeling and a comparative study between precision agriculture and biodiversity modeling information systems (Santana et al., 2007) a strategy based on SOA (Service Oriented Architecture) was defined for openModeller.

During the project we also realized that some of the components, such as the locality and environmental components would be better implemented with a focus on supplying data to the modeling environment rather than to end-users. For example, TAPIR, WFS[10] and DiGIR[11] services all provide a robust and well-established protocol for obtaining occurrence data, so implementing another service with a similar goal would not make much sense. Similar logic applies to the environmental component, where protocols such as WCS are already well-established. The idea was to be able to retrieve data from these kinds of services and concentrate on building a cohesive and simplified web service API for modeling to ease integration with third parties and facilitate maintenance. The new architecture is represented in figure 2.

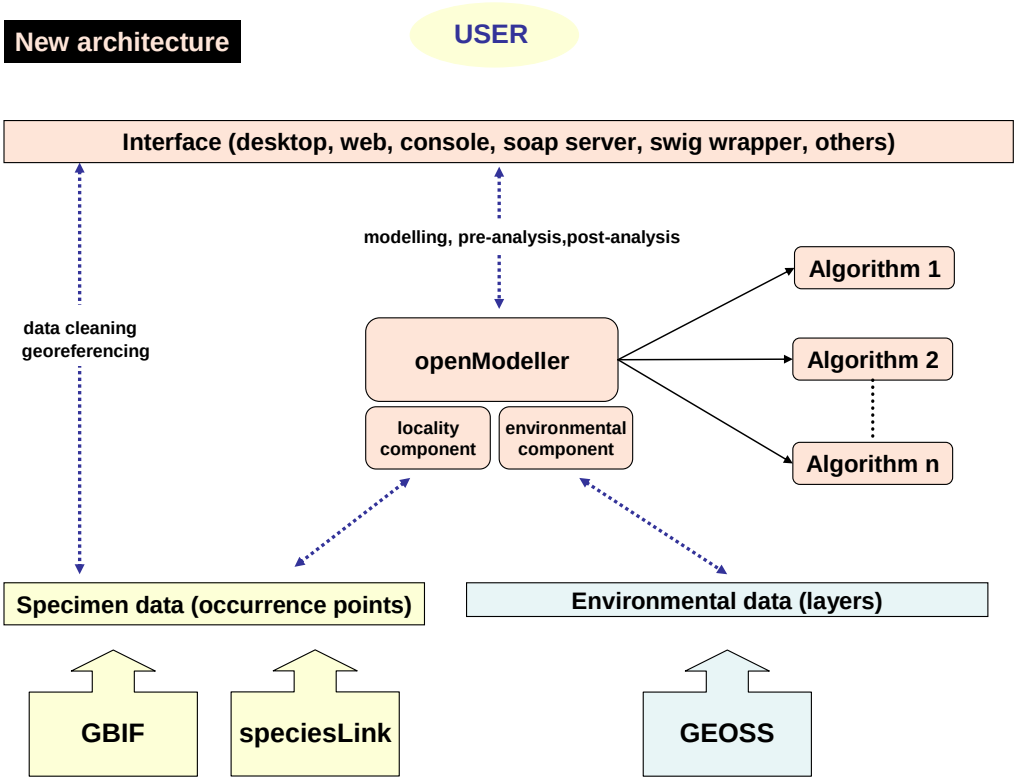


Figure 2: New architecture for openModeller

Therefore, part of the work included the definition and implementation of a core modeling service. The openModeller framework includes a full server implementation for the web services modeling API based on SOAP (Simple

[10] http://www.opengeospatial.org/standards/wfs
[11] http://digir.sourceforge.net

Object Access Protocol) developed in C++, and a simple command-line client developed in Perl. This API was developed in partnership with the BiodiversityWorld project. It enables the use of remote modeling servers that can be set up on top of a cluster. A formal definition of the web services API is available[12], including the following operations: ping, getAlgorithms, getLayers, createModel, getModel, projectModel, getProjectionMetadata, getLayerAsAttachment, getLayerAsUrl, testModel, getTestResult, getProgress and getLog. This API uses the Document/Literal SOAP style and references external elements and types from an XML Schema[13] that also validates openModeller serialized objects. Two instances of the service were made available: one in a modeling server at CRIA and the other in the Cluster. The first one is also being used by the GBIF portal.

An initial implementation following the SOA architecture was developed for openModeller using Jboss, Apache Tomcat, Apache Axis and Apache Ant following the ESB (Enterprise Service Bus) architecture. This SOA-based implementation does not replace the framework or any of its interfaces. It just provides an additional way to make use of the framework in an environment where modeling services can potentially interact with other types of web services. An improved solution was developed using J2EE, Glassfish V2, Open ESB (SUN) and BPEL (Business Process Execution Language), for service integration. A web interface for ecological niche modelling was developed to evaluate the new infrastructure. The new architecture also allows integration with GIS services through OGC (Open Geospatial Consortium) standards.

Locality data component

The first step to define a locality data component was to standardize the basic input of locality data. Data structures were changed, forcing each locality provided to openModeller to contain a unique identifier, so that serialized models and system messages could explicitly reference back to it. Adjustments were also made allowing absence data to be provided to the framework and used by the algorithms by adding a new “abundance” attribute. An interface (API) was created for the locality data component. Different implementations of the interface (i.e., drivers) allow locality data to be read from simple TAB-delimited text files, TerraLib tables and remote sources such as GBIF and any TAPIR/ DarwinCore provider.

The *species*Link network is the main source of on-line species occurrence data in Brazil today, also representing an important source of input data for modeling experiments. As part of this project, *species*Link changed its architecture from using an on-line dynamic distributed search system to a centralized search approach (Figure 3). A new data harvesting mechanism periodically searches the network retrieving new and modified records to update the central database. This has improved performance, making it easier to serve all *species*Link records to other applications. All regional servers providing data to *species*Link originally used a DiGIR/DarwinCore provider

[12] http://openmodeller.cria.org.br/ws/1.0/openModeller.wsdl
[13] http://openmodeller.cria.org.br/xml/1.0/openModeller.xsd

software. As DiGIR is becoming obsolete since TAPIR was created as a unifying solution for the DiGIR and BioCASe protocols, all regional servers were migrated to the TAPIR/DarwinCore standard. A web service also based on TAPIR/DarwinCore was installed and configured on top of the *species*Link central database, allowing the corresponding openModeller locality driver to read data from the entire network through a single endpoint. These new developments in *species*Link were co-funded by the JRS Biodiversity Foundation.

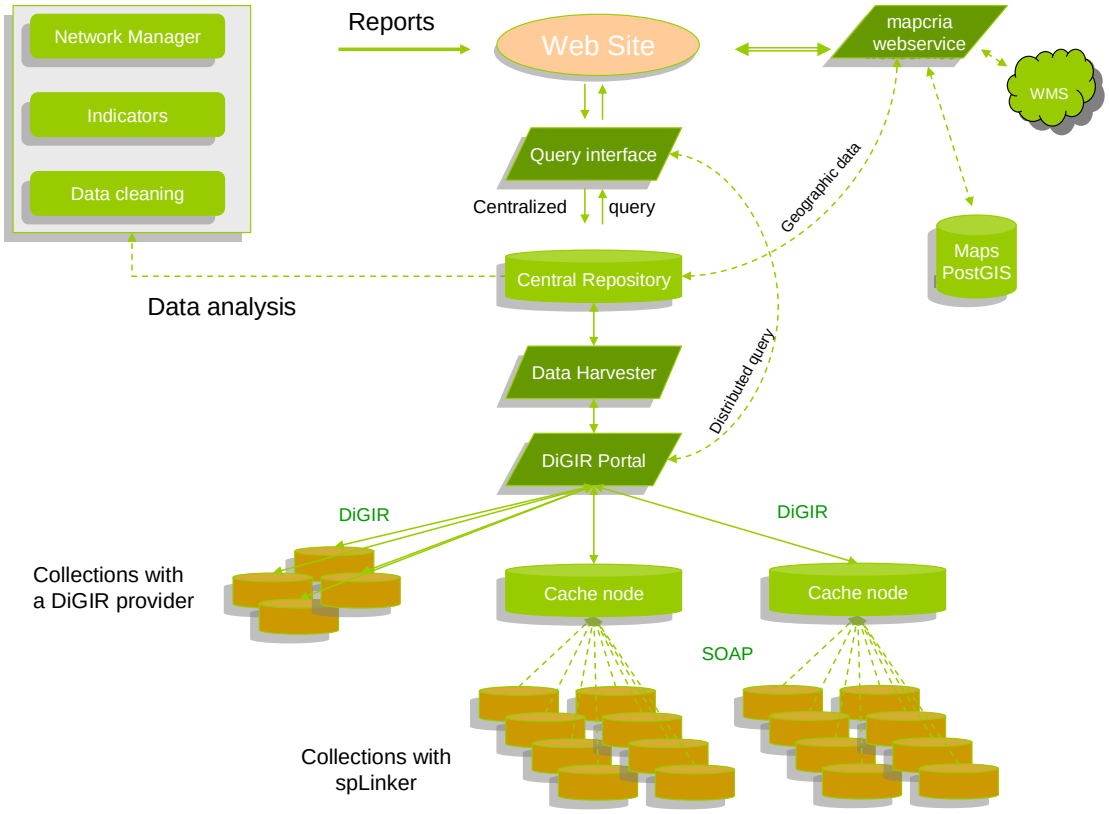


Figure 3: *species*Link network architecture.

Data cleaning

Several data-cleaning techniques were studied and implemented as part of the *species*Link project. In the beginning of the openModeller project the idea was to create a web service for data cleaning, but during the project it was perceived that more benefits would be gained by improving the existing data cleaning infrastructure already available at *species*Link. Such improvements now serve end users retrieving data from *species*Link through openModeller. Currently, data tests directly related to locality attributes include:

- Geographic error detection:
    - Check that coordinates are consistent with the administrative regions provided by the original records.
    - Check that coordinates are consistent with the species habitat (marine/terrestrial).
    - Check that coordinates do not fall outside the world boundaries.

- Elevation error detection:
    - Check that coordinates correspond to an elevation consistent with data provided by the original records.
- Itinerary error detection:
    - When records are associated to an individual (such as a collector) check that all points from the same individual are geographically consistent with the original collecting dates.
- Geographic outlier detection:
    - Detect outliers on latitude and longitude using a statistical method based on reverse-jackknifing procedure.

Records that do not pass all data cleaning tests are marked as “suspect”, helping providers to fix possible errors and helping users to decide whether the record is suitable or not.

Additionally, a new module in the *species*Link network enabled all records without coordinates to be georeferenced based on municipality data when data is present at this level (figure 4). An associated error (distance between the point and the most distant border) is calculated and stored. Original data remains unchanged, so that users can clearly distinguish between original coordinates and the ones suggested by the georeferencing module. Records that are georeferenced this way can still be useful for modeling experiments involving macro analysis, for example when the region of interest has a continental scale.

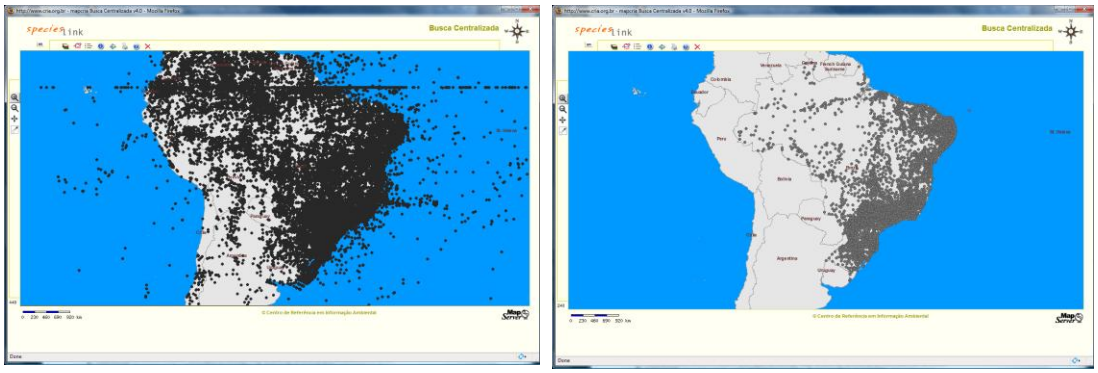


Figure 4: All points from speciesLink with original coordinates (left-hand image) contrasted with all points that were automatically georeferenced based on municipality (right-side image).

Environmental data component

Modeling experiments use environmental layers as parameters for model creation and model projection. In the beginning of the project, the environmental data component was restructured and a new interface (API) was created to allow the implementation of different drivers. Two drivers were implemented: a GDAL adapter and a TerraLib adapter. GDAL[14] is a library that handles raster geospatial data in different formats. More than 100 formats[15] are supported by GDAL. TerraLib is a library of GIS classes and

[14] http://www.remotesensing.org/gdal/

[15] http://www.remotesensing.org/gdal/formats_list.html

functions that can also store rasters on relational databases. TerraLib has tools to manage data in geographical databases using a set of spatial and spatio-temporal data structures. It also provides a set of spatial statistics algorithms, image processing algorithms and functions to execute map algebra. The new API makes it possible for openModeller to read and write raster data using both adapters. The integration between openModeller and TerraLib also enables close coupling with other TerraLib based programs, such as TerraView GIS or TerraWeb web application.

During the last years, more than 150000 environmental layers have been used by CRIA in modeling experiments. Of these, approximately 4000 are in the public domain and were organized during this project in specific directories to be made available through the modeling service. These layers basically consist of climate data from IPCC (past, present and future), Worldclim data (temperature, precipitation and altitude), topographic data from Hidro1K and vegetation index measurements derived from the Advanced Very High Resolution Radiometer[16] (AVHRR).

There are two ways in which openModeller can access remote rasters: 1) Through the TerraLib driver, that can open connections to remote TerraLib databases; 2) Through recent versions of the GDAL library, that include support to rasters accessible through the Web Coverage Service (WCS). This enables openModeller to interoperate with other systems that can provide environmental data, such as the GEOSS[17] initiative.

As part of this project, there were several improvements and new developments involving TerraLib and TerraView to help sharing raster data in a distributed environment through OGC standards. These include:

- **TerraOGC**: Library of classes which allows the development of clients and servers compatible with the OGC standards. The library already supported the Web Feature Service (WFS) and Web Mapping Service (WMS) standards, and now it also supports WCS.
- **SisAmBio**: TerraView GIS plugin that facilitates the creation of a database with environmental data and the sharing of it through WCS. The plugin improves the capacities of TerraView by allowing to:
    - Import multiple GeoTIFF files in a single operation;
    - Associate metadata (such as creation date, provenance, keywords, spatial reference systems) to multiple layers of environmental data;
    - Select layers based on metadata or region of interest;
    - Export multiple layers to files in different formats in a single operation;
    - Select multiple files to be exported as coverages in a WCS server.

[16] http://edc.usgs.gov/products/satellite/avhrr.html
[17] see GEO and GEOSS at http://www.earthobservations.org/

- **Collaboration server**: Package based on TerraOGC and TerraLib that reads a TerraView database and exports selected environmental layers as coverages, so that they can be accessed by WCS clients.

Pre-analysis component

In order to increase model accuracy and improve performance during model creation, a number of pre-processing techniques can be used. Studies that were carried out during the project showed that different sets of input layers may give completely different results in modeling experiments. This clearly suggests the implementation of pre-processing techniques to cross reference occurrence points with environmental values and determine which layers can better explain the distribution of a particular species.

A generic API to run different pre-analysis techniques was developed to help identifying the most relevant environment layers that can be used in each modeling experiment. Two techniques were implemented. One based on the jackknife procedure which generates a model with a specific algorithm for different subsets of layers (each time excluding one of the layers from the complete set). Results are then tested with an independent set of points, and the contribution of each variable (environment layer) in model accuracy is measured.

The other technique uses the chi Square method described at Li, L., et al. (2006) "An Integrated Bayesian Modelling Approach for Predicting Mosquito Larval Habitats". It builds a contingence matrix for each pair of layers and then applies the chi-square test to identify correlations at a significance level of 0.05.

Other pre-processing techniques like sub-sampling locality data into test and training datasets to allow extrinsic model validation were implemented.

Additionally, openModeller provides ways to visualize occurrence points from both the geographical and environmental viewpoints to check if the current samples are widely or poorly distributed in the region of interest from the users' perspective. Two command-line tools, "om_viewer" and "om_niche", can be used for these purposes, respectively.

Modeling component

During the project, 15 versions of the framework were released. The C++ API is now stable and includes methods to retrieve algorithm metadata, set parameters (input and output layers, input and output masks, localities data, output format, algorithm, and algorithm parameters), create models, project models generating distribution maps, and serialize/ deserialize models. Integration with GSL[18] (GNU Scientific Library) enables algorithm writers to make use of a broad range of mathematical functions. Documentation about the framework API is available on-line[19]. The framework is currently used by a console/command-line interface, a graphical user interface, and a web

[18] http://www.gnu.org/software/gsl/
[19] http://openmodeller.cria.org.br/doxygen/1.0.0/

services interface. The main modeling functionalities can also be used in the Python programming language.

The framework makes use of a multi-platform compilation system (CMake[20]) that allows openModeller to be more easily deployed in the three major platforms: GNU/Linux, Windows and Mac OSX. The adoption of a single build system for all platforms provided a significant reduction in maintenance, complexity of the code base, and duplication of effort. Additionally, for GNU/Linux two packages are being generated in each release using the two main formats: Red Hat Package Manager (rpm) and Debian (deb).

Considering the frequency in which changes are made to the library by different developers, another effort was to develop a new infrastructure for unit testing. Unit tests provide a way to check that specific parts of the source code are working as expected. Several C++ frameworks for unit tests were investigated before deciding to use CxxTest[21]. A set of unit tests were developed with the chosen framework, including documentation about how to write, compile and run unit tests.

Ten types of modeling algorithms are now available in the framework. In the beginning of the project, GARP (with the Best Subsets procedure), Bioclim, a couple of distance-based algorithms and Climate Space Model were already available. The following algorithms were implemented during the project:

- AquaMaps: Algorithm specifically created to model the distribution of marine organisms. Developed in partnership with the Incofish project[22]. AquaMaps is based on the idea of environmental envelopes, where each variable has an associated preferred range and a broader accepted range. This algorithm differs from other traditional ones since it requires a fixed set of layers to work and makes use of expert information for the lower and upper limits of each variable. This information is stored in local database provided by FishBase[23] which contains data for approximately 30 thousand species.
- Environmental Distance: Replaced the previous distance-based algorithms in openModeller (Minimum Distance and Distance to Average) with a more general approach. This new implementation allows users to specify the distance metric to be used (Euclidean, Gower, Mahalanobis or Chebyshev), the number of nearest points to be taken as reference, and the maximum distance to the centroid of the nearest points. When used with Gower and maximum distance of 1, this algorithm is equivalent to the algorithm known as DOMAIN[24].
- Envelope Score: Lax bioclimatic envelope algorithm where the probability of presence in a specific point is proportional to the number of environment layers whose envelope (min-max range calculated in model creation) contains the corresponding environment value for the point. The primary motivation for implementing Envelope Score was to

[20] Cross Platform Make: http://www.cmake.org

[21] http://cxxtest.sourceforge.net/

[22] http://www.incofish.org/

[23] http://www.fishbase.net/

[24] Carpenter, G., Gillison, A. N., and Winter, J. (1993). DOMAIN: A flexible modeling procedure for mapping potential distributions of animals and plants. Biodiversity and Conservation, 2, 667-680.

generate more meaningful models in paleo-climate scenarios, where the traditional Bioclim algorithm produces overly constrained models. This work was done in collaboration with Chris Yesson from the University of Reading. The algorithm is being used by the niche modeling interface offered by the GBIF web portal.

- Support Vector Machines (SVM): Machine learning technique based on concepts from the statistical learning theory. SVMs have been used in many different applications that involve pattern recognition. Recently, SVM was also applied to the problem of creating species' distribution models, but there is still considerable scope for further research. SVMs are known for their good generalization ability and robustness to high dimensional datasets. The openModeller implementation of SVM was carried out in a partnership with the "Instituto de Ciências Matemáticas e de Computação"[25] of the University of São Paulo.
- Artificial Neural Networks (ANN): Non-linear statistical modeling technique based on the idea of an interconnected group of artificial neurons that can be used to find data patterns. There are many variations of ANNs. The implementation in openModeller uses Multilayer Perceptron with Backpropagation. Two training methods are available: by epoch (maximum number of iterations) or by minimum error.
- Maximum Entropy: Algorithm that has been successfully applied in many different areas including species' distribution models, mainly due to a software known as MaxEnt[26]. An initial version was implemented in openModeller using the Maximum Entropy Modeling Toolkit developed by Zhang Le, which offered two traditional methods to estimate the maximum entropy parameters: GIS (Generalized Iterative Scaling) and L-BFGS (Limited-Memory Variable Metric). Unfortunately the results seemed to be always far behind all other algorithms, so this version was recently replaced by a new one which tries to follow the same training method and specific procedures used by the MaxEnt software.
- Niche Mosaic: This algorithm makes use of tabu search to find an optimal solution for a set of bioclimatic envelopes around each presence point provided as input.

An adaptive variation of the GARP algorithm, called AdaptGARP, was also implemented. AdaptGARP makes use of adaptive decision tables, including an adaptive crossover operator and an adaptive mutation operator. The two algorithms, GARP and AdaptGARP, were tested and compared using two different data sets, producing almost identical results. This work also showed that Adaptive Decision Tables can simulate any genetic algorithm without performance loss, and that previous developments on the subject are a special case for a particular genetic algorithm, e.g. the GARP algorithm. Furthermore, a significant conceptual gain is obtained since the simulation of genetic algorithms using adaptive devices allows both genetic representation and operators to be unified. The adaptive technique adopted here can be

[25] http://www.icmc.usp.br/

[26] http://www.cs.princeton.edu/~schapire/maxent/

applied to other algorithms. Currently, a similar approach is being used to create an adaptative version of the maximum entropy algorithm.

New algorithms can be included in the framework at any time. For this purpose, a document was created to help algorithm developers[27]. The text contains detailed explanation of the main data structures and methods, including a step-by-step algorithm creation example.

Post-analysis component

A wide range of post-processing techniques can be used after model creation and map projection. Model validation was considered the most important post-processing functionality for the framework. Besides calculating confusion matrix statistics such as accuracy, omission and commission errors, the framework can also perform ROC (Receiver Operating Characteristic) analysis. ROC analysis provides a global quality measure based on the "area under the curve" (AUC), an approach being used by several experiments since it does not depend on a cutoff limit such as confusion matrix does. ROC curve data and AUC are automatically calculated for training data after model creation using command-line tools, Web Service calls, and the Desktop interface (which includes a graphical display of the curve). When no absence points are used, openModeller generates background points to calculate the curve.

Besides automatic calculation for training data, the same functionality is also available for extrinsic tests. A new command-line tool (om_test) was developed for this purpose. It loads a serialized model, tests each point from a specified file and generates the confusion matrix and/or ROC statistics. Partial AUC can also be calculated when a maximum omission error if specified.

Distribution map statistics are also calculated, including the total number of cells and the number of cells where presence was predicted.

Hotspot analysis and consensus maps were developed as part of the Desktop interface. Hotspot maps aggregate multiple distribution maps from different species, and can be used to represent species richness. Consensus maps aggregate multiple distribution maps generated by different algorithms for the same species, therefore producing a single distribution map highlighting areas where most algorithms agree about the potential distribution.

Desktop interface

During the first years of the project, releases 0.3.1 and 0.3.2 (Figure5) of the existing openModeller graphical user interface were made, including many improvements and bug fixes. Although the 0.x version series has been a successful step toward making the library widely accessible to the general public, it became clear that it only covered the basic needs of researchers through a wizard-like interface – a step-by-step process for running simple modeling experiments.

[27] http://openmodeller.sourceforge.net/alg_manual.html

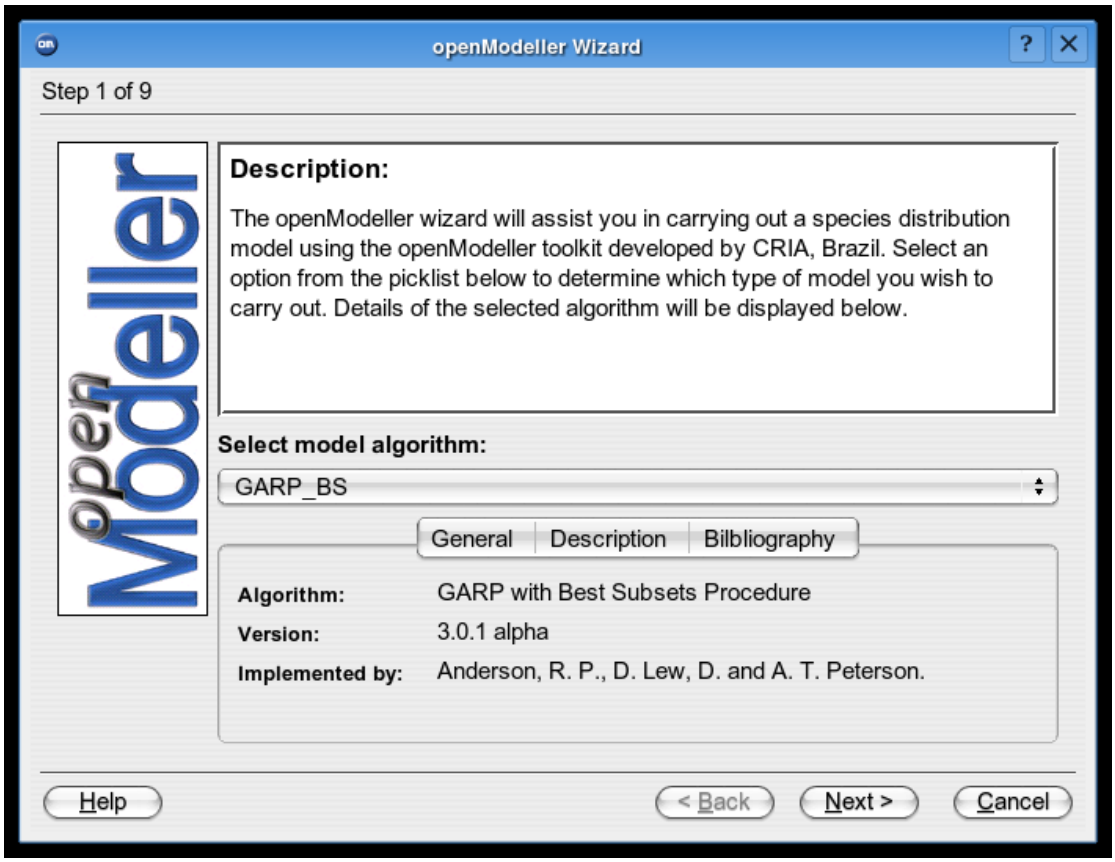


Figure 5: Omgui 0.3.2, the first graphical interface for openModeller, with its wizard style.

A new strategy was then conceived to create a more advanced and comprehensive graphical user interface and enable researchers to carry out complex experiments. As part of that strategy, openModeller Desktop (version 1.x series) was a complete rewrite of the previous “omgui” application. The new version employs a modular and extensible architecture. Experiments involving multiple species, multiple algorithms and multiple climate scenarios can be easily carried out (Figure 6). Plugins were written, allowing modeling tasks to be run on the local machine using the local openModeller library or on a remote machine via the openModeller Web Service API. Plugins were also written that facilitate automated occurrence data retrieval from the *species*Link and GBIF on-line databases. Multiple extrinsic model tests can be performed at any time. The new version has a new embedded GIS component that allows users to interactively explore resulting distribution maps. Post processing tools allow users to create hotspot maps and consensus maps. A special interface was also developed to run pre-analysis on layer sets. The latest release includes installation packages for Windows, GNU/Linux and Mac OSX.

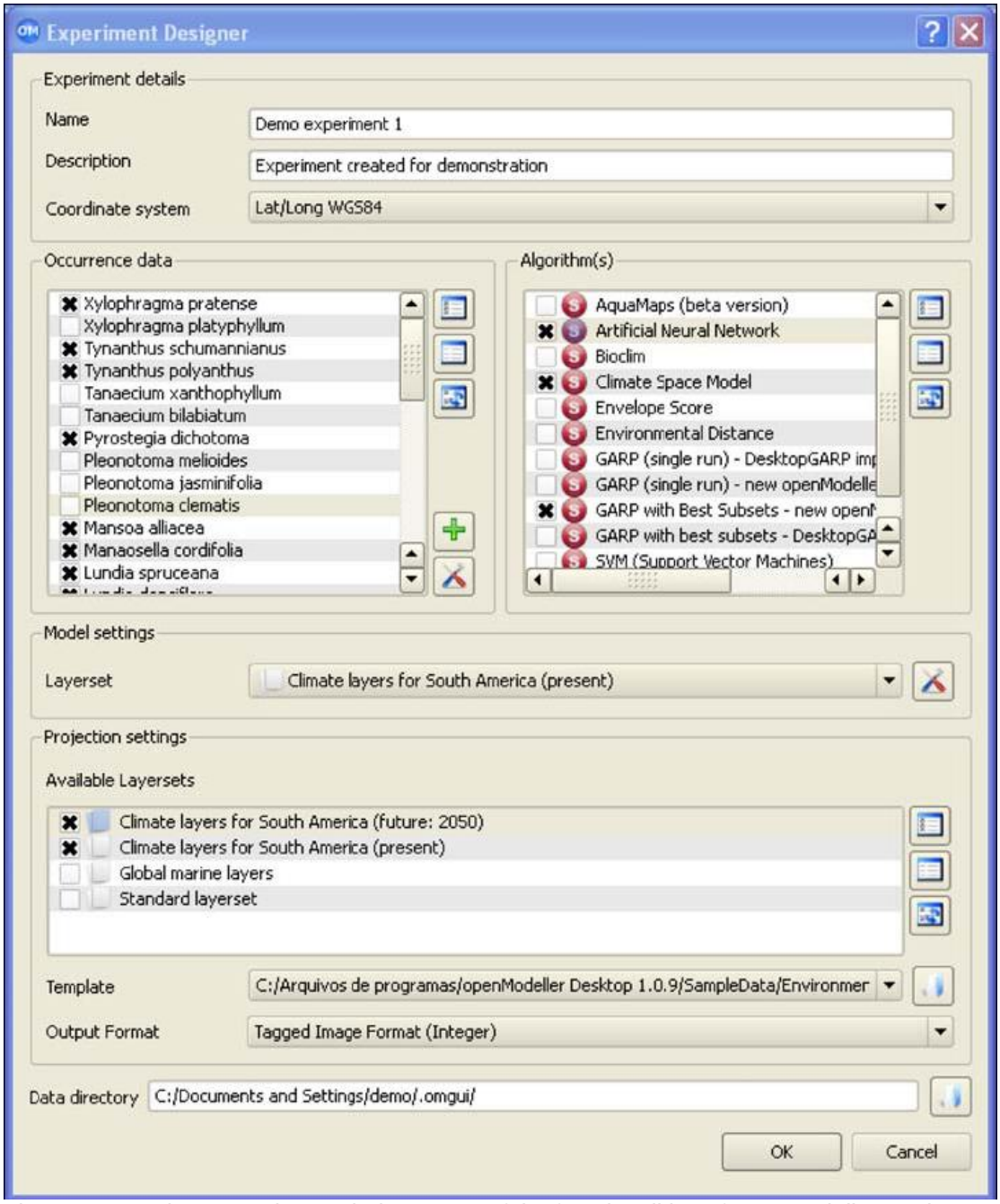


Figure 6: Experiment Designer window – one of the functionalities of openModeller Desktop.

A new TerraView[28] modeling plugin (Figure 7) was developed as a separate application. TerraView is a GIS that can be used to build a TerraLib database, as well as to query and visualize its contents in several types of output (such as graphs, maps or charts). It also includes a set of spatial analysis tools for knowledge discovering. This plugin is an independent interface that can be plugged to the currently distribution of TerraView for openModeller users, providing a bridge between openModeller components and the TerraView database.

[28] http://www.dpi.inpe.br/terraview

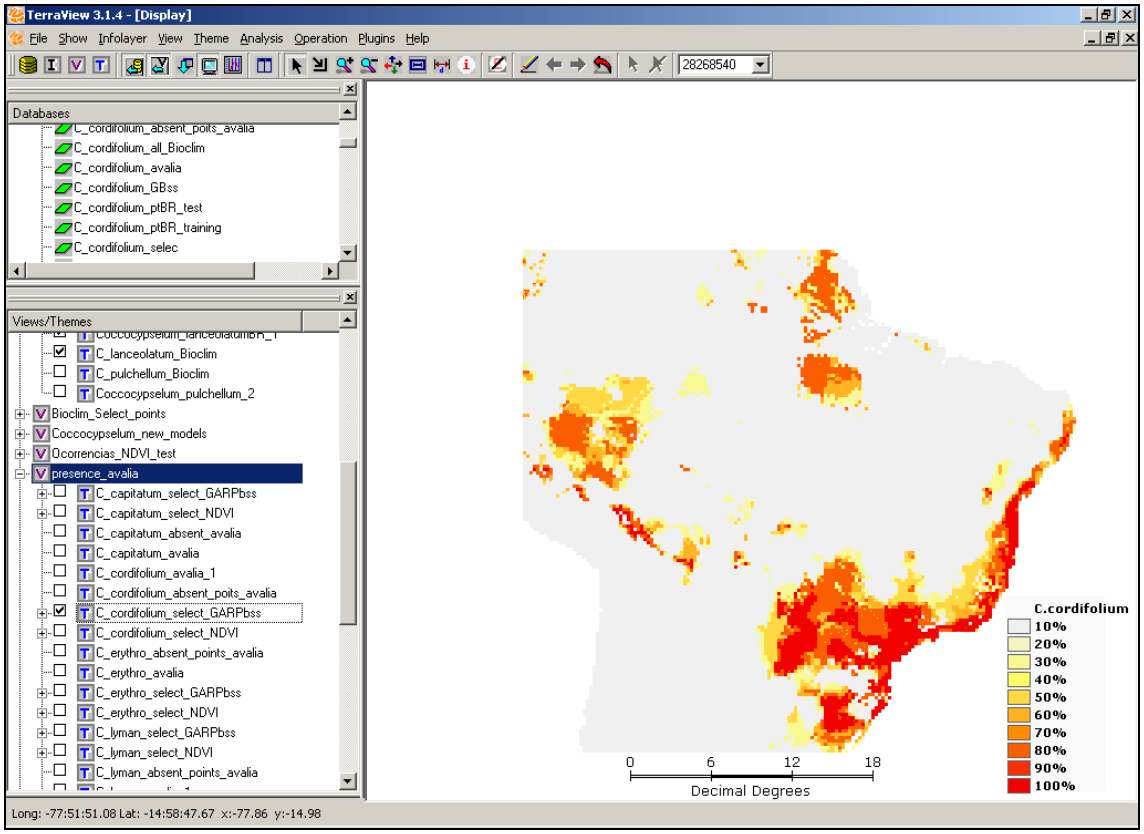


Figure 7: TerraView displaying a distribution map created with the modeling plugin.

Web interface

In the beginning of the project, the first prototype of an openModeller web interface (Figure 8) was developed in Perl making shell calls to command-line tools to generate the distribution maps. This initial effort was just a proof of concept, although it already had features such as job status management and map visualization.

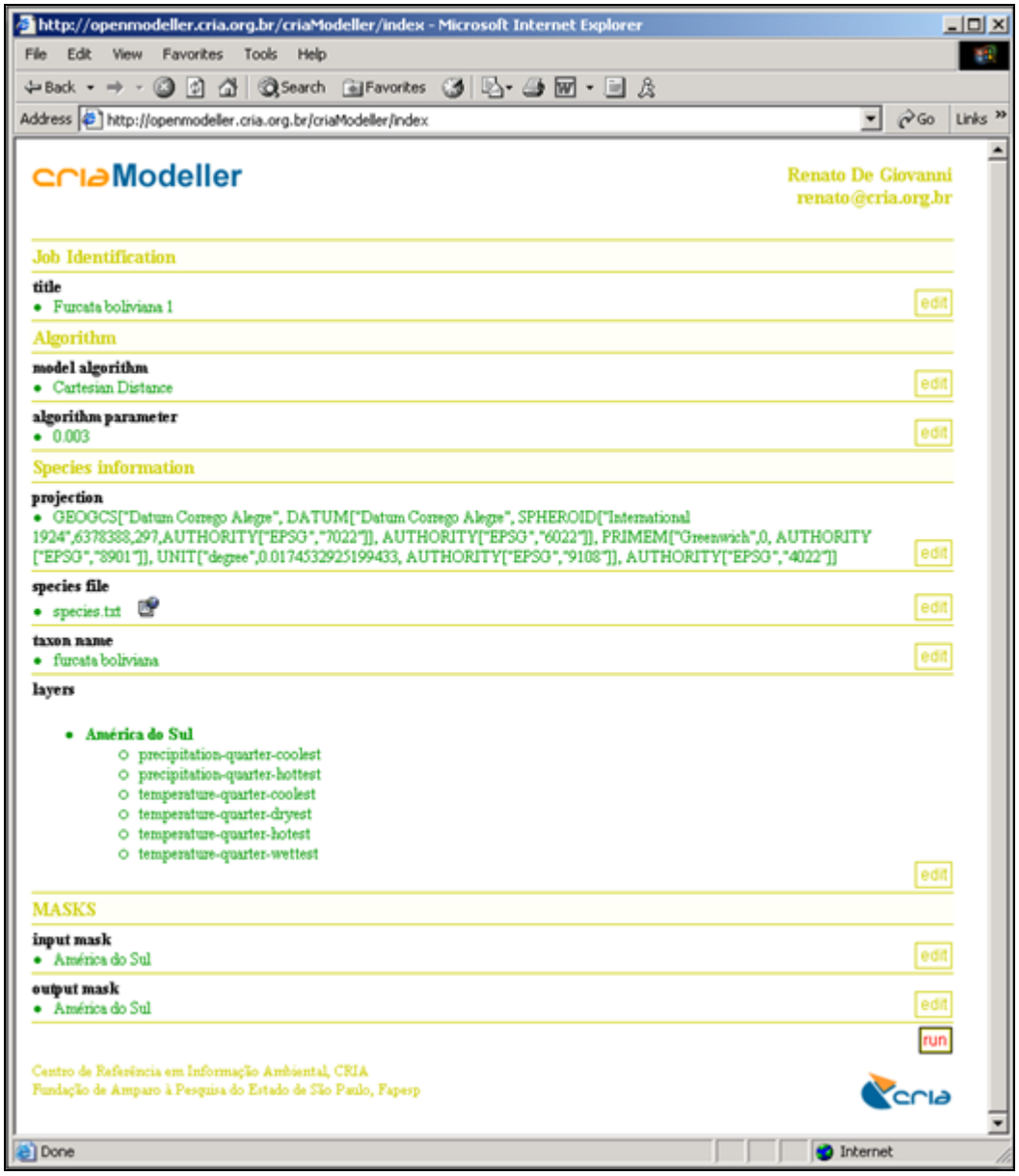


Figure 8: criaModeller – the first web interface developed for openModeller.

Other web interfaces for openModeller were developed. Two of them – one in PHP and the other in Flex (Figure 9) – were outcomes of the BioGeo Interoperability Workshop[29]. This workshop was organized by a specific task group from the TDWG Geospatial Interest Group[30] and was hosted by CRIA. The aim of this workshop was actually to test biodiversity and geospatial protocols and data standards and to investigate how they could interoperate. This was done by developing two demo applications which basically consisted of web interfaces making use of specific web services. One of the services to be tested was based on the openModeller Web Service API developed as part of this project to facilitate remote modeling. Both applications developed during the workshop are web interfaces that can generate niche models.

[29] http://www.tdwg.org/homepage-news-item/article/geointeroperability-workshop-outcomes/
[30] http://www.tdwg.org/activities/geospatial/

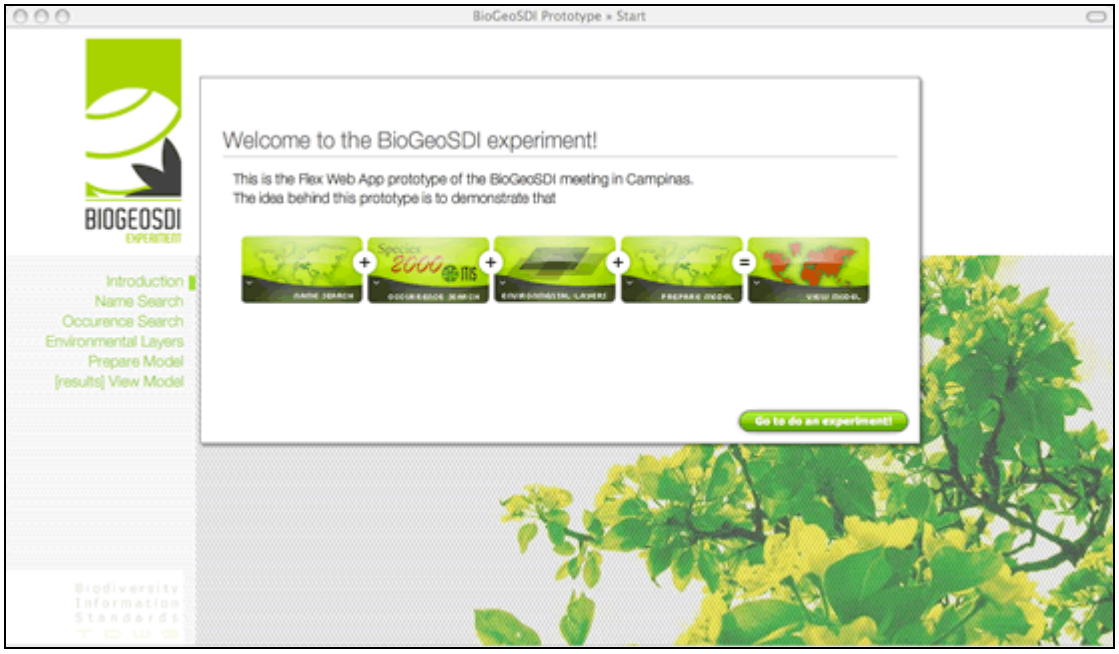


Figure 9: Front page from one of the BioGeoSDI demo interfaces, which included niche modeling functionalities by interacting with a modeling service.

The next interface to be implemented was the result of a partnership between CRIA, University of Colorado and GBIF. Besides developing a web interface for niche modeling[31], another objective was to develop a Java library to access the modeling service so that GBIF could use it. After that, GBIF incorporated the functionality into their data portal (Figure 10), allowing users to generate niche models when visualizing species occurrence data. The GBIF niche modeling interface interacts with a modeling server hosted at CRIA to generate distribution maps.

[31] http://dbmuseblade.colorado.edu/gbiftestbed/

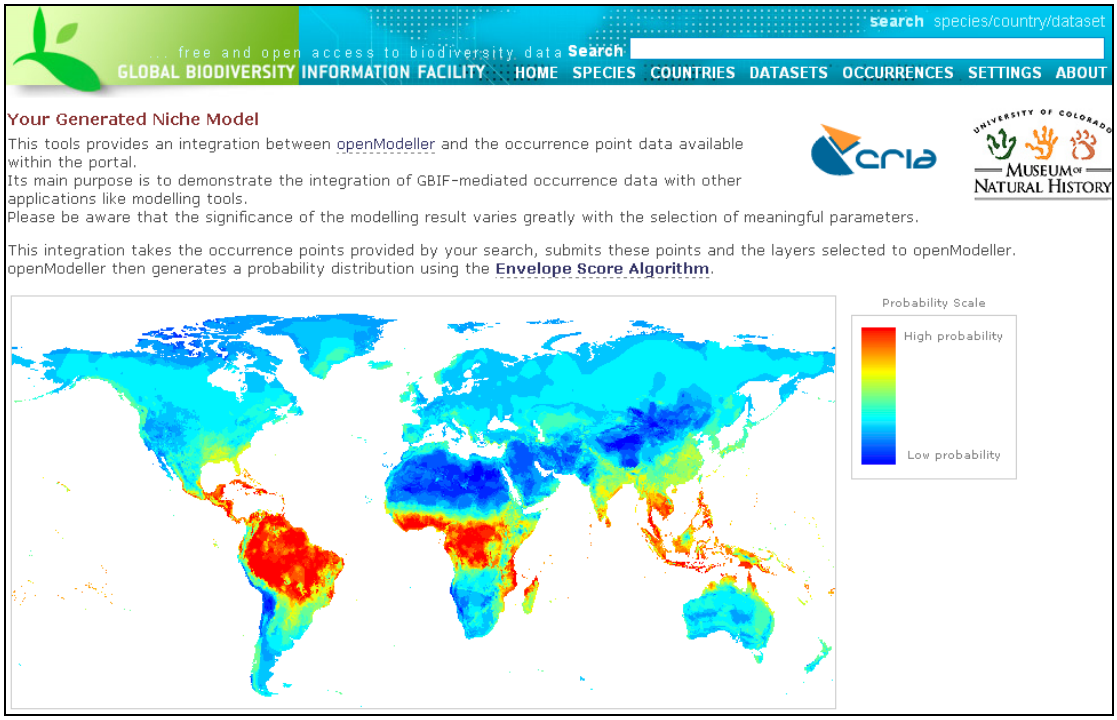


Figure 10: Distribution map generated by the GBIF portal with occurrence data for *Anhima cornuta.* The interface communicates with the modeling server at CRIA to generate the models.

Two other web interfaces involving openModeller were developed during the last year of the project. One of them was part of a Doctoral thesis (Fook, 2009) to build a niche model repository. The other one was part of the previously mentioned developments to build the new SOA-based architecture for openModeller (Figures 11 and 12).

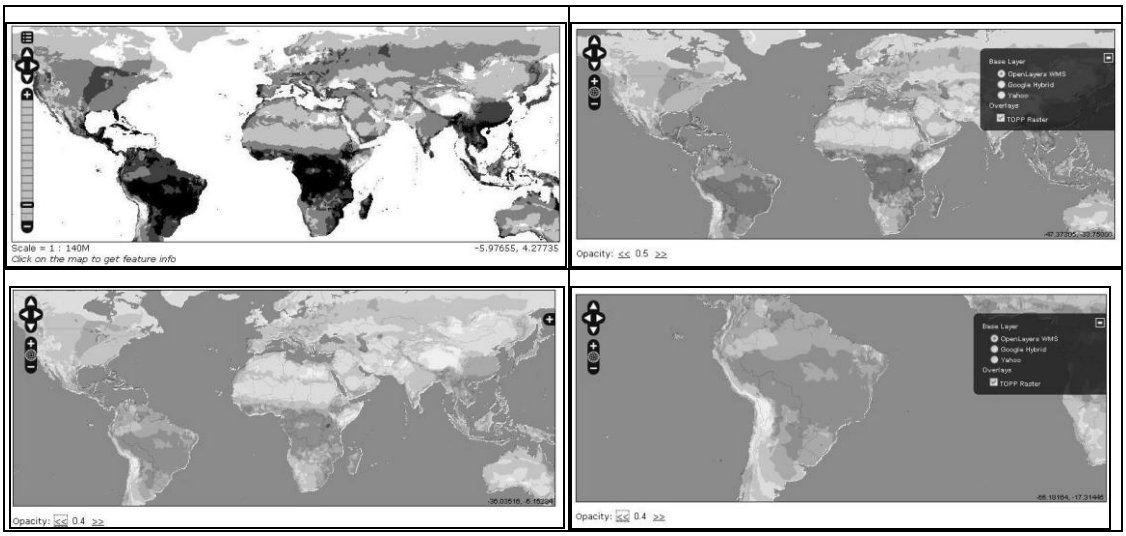


Figure 11: Distribution map displayed with openLayers (part of the web interface developed on top of the SOA.

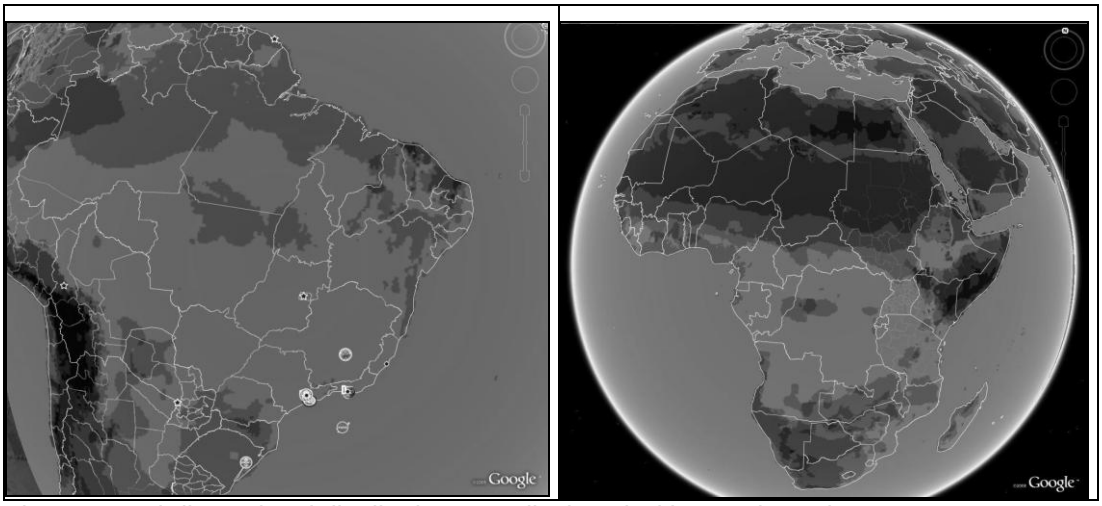


Figure 12: Tri-dimensional distribution map displayed with GoogleEarth.

Other relevant activities and results

Development of the aRT package

Statistical spatial data analysis and Geographical Information Systems (GIS) can act together to understand and model spatially distributed data. Geoprocessing operations can enhance statistical models with relevant information which can be used to better understand the main features of usually noisy and multidimensional data. Therefore integration between GIS and statistical software can be highly beneficial for both sides. The package aRT[32] enables the access to TerraLib from the statistical software called R[33]. aRT encapsulates C++ classes into S4, therefore the user can manipulate TerraLib objects directly in memory using the implemented wrappers. aRT can manipulate spatial data using the data structures of the sp package, reading and writing Spatial data in the database. Some spatial operations already implemented in the package as part of this project are:

- Manipulation of points, lines, polygons and raster data;
- Spatial predicates, such as “touches”, “within”, “contains”, “crosses” and “overlaps”;
- Polygons operations, as “union”, “intersection”, “difference” and “simplification”.

aRT is available as source code and also as a cross-compiled Windows binary. Along with the package, there are files documenting the implemented functions and also examples of scripts showing how to use aRT.

Model repository

A new web services architecture to support collaboration in a species distribution modeling network was also proposed during the project (Fook, 2009). In this scenario, users can share model instances (definition of

[32] http://www.est.ufpr.br/aRT
[33] http://www.r-project.org/

parameters, data used, chosen algorithms) and potential distribution maps. This allows them to access other models and compare results. The proposed architecture was implemented and is now being tested at INPE.

Study Cases

The following studies were carried out during the project with the objective of testing the framework, training and involving users from other institutions, and producing papers to disseminate openModeller and its applications:

- Assessment of the sensibility of species distribution models according to the precision of locality data. Two algorithms implemented in openModeller (GARP and BIOCLIM) as well as Maxent were evaluated (Iwashita, 2007).
- Comparison of two different algorithms (GARP and Maxent) in modeling the potential habitat of the manned wolf (*Chrysocyon brachyurus*). The main objective was to know the consequences of actual habitat fragmentation for this species (Kawashima et al., 2007). In collaboration with Marinez F. Siqueira (CRIA).
- Performance test of two different algorithms (GARP and SVM – Support Vector Machine) in modeling Cerrado tree species (*Stryphnodendron obovatum*). The main objective was to compare the accuracy of these algorithms and test the effect of using a high number of environmental layers in the process (Lorena et al., 2008).
- Application of species potential distribution modeling for *Hennecartia omphalandra* (Monimiaceae) (Souza, 2007). In collaboration with Marinez F. Siqueira (CRIA).
- Potential distribution modeling of species threatened of extinction from State of Minas Gerais, Brazil. Doctorate thesis (on going) of Luciana H. Yoshino Kamino (Universidade Federal de Minas Gerais – UFMG). In collaboration with Marinez F. Siqueira (CRIA).
- Potential distribution modeling and model validation using openModeller (Barreto, 2008). In collaboration with Marinez F. Siqueira (CRIA).
- Ecological niche modeling in the Brazilian Atlantic Forest: a comparative evaluation of presence-only methods for modeling the geographic distribution of anurans (Giovanelli, 2009).
- For the state of São Paulo, species of genus Croton (Euphorbiaceae) and species from the Tribo Cynodonteae (Poaceae – Chloridoideae) where studied in collaboration with the Instituto Botânico de São Paulo (IBt). The first analysis of the spatial distribution of the Tribo Cynodonteae indicated that the sampling effort has to be intensified to enable a better understanding of the biogeography and conservation status of the group (Santos et al., 2007). For the Croton genus, the importance of the soil variable in species distribution modeling for this group was analyzed (Caruzo et al. 2007). For both taxa, experiments considering species distribution models with different algorithms and variables can be further tested to better discuss the biological processes related to the species distribution

- A database for studying Arecaceae (Palmae) distribution in Brazil is under development at INPE. Occurrence data from the most important Brazilian and international herbaria were incorporated into the database. At the moment, a data cleaning process to correct the geographical coordinates and taxonomic nomenclature is being performed. The database with Arecaceae occurrence data will enable a study about the spatial distribution of Brazilian palms and also help testing openModeller algorithm implementations.
- Different environmental data and species distribution algorithms were tested at INPE to achieve better results and optimize modeling procedures. In particular, Normalized Difference Vegetation Index (NDVI) was used to model the genus Coccocypselum (Rubiacea) using GARP Best Subsets and Maxent (Amaral et al., 2007).
- A study case is under preparation at INPE to analyze the impact of climate change over the life forms in the boundaries between savanna and tropical rain forest in the Brazilian Amazon. A database containing occurrence data from savanna and forest vegetation is under construction and will be used to model the current and the predicted distribution of dominant life forms (Box model algorithm) as well as for a number of selected species using algorithms implemented in openModeller.
- Comparison among three modelling algorithms (GARP, AdaptGARP and P-GARP) in modelling species of *Peponapis timberlakei* and *Cucurbita palmata*. *Peponapis timberlakei* is the main pollinator of native *Cucurbita palmata*, therefore the spatial distribution of both species must be related. (Santana et al., 2009). Congresso de Ecologia do Brasil.
- Experimental integration of the SOA-based solution with GIS services provided by GeoServer, with solutions implemented in 2D and 3D according to the standards defined by the Open Geospatial Consortium. Ecological niche modelling experiments were performed with *Stryphnodendron obovatum*. (Santana et al., 2009) Evaluation of European Federation of IT in Agriculture.
- Evaluation of three GIS geospatial service providers: *Deegree*, *GeoServer* and *MapServer*. The purpose was to verify which of the solutions was adequate to implement a professional solution integrating GIS services and applying the standards defined by the Open Geospatial Consortium. Ecological niche modelling experiments were performed with *Stryphnodendron obovatum* and *Peponapis* and *Cucurbita* genera. (Santana et al., 2009) *SBIAgro 2009.*
- Evaluation of the geographical distribution of pollinator and plants, using *Peponapis fervens* and *Cucurbita* species. The purpose was to compare the species distributions so as to establish a relationship among the pollinators and the non-domesticated *Cucurbita* species. (Gianinni et al., 2009)
- Evaluation of the potential of ontologies in ecological niche modelling using species of *Peponapis* and *Cucurbita* genera. The software solution to perform this evaluation was developed using OWL-based

concepts and the Protégée ontology editor. openModeller was sligthly modified to generate output files presenting the species distribution in a georreferenced specific text format, so as to be used as the input for the ontology developed at Protégée. (Santana et al., 2008) *6th International Conference on Ecological Informatics.*

Seminars

To promote greater interaction between project participants, 11 seminars were held during the project to present new developments and to advance other discussions.

Additionally, a study group on Biodiversity Modeling was structured at INPE. Several research initiatives, as papers and thesis, are focused in questions related to biodiversity modeling theory and computational tools. Students and researchers involved promote monthly meetings, called "Referata Biodiversa"[34] where ecological and computational aspects of the biodiversity modeling process are presented and discussed. The main objective is to promote the debate about biodiversity and modeling, integrating this multidisciplinary team. The study group also offers opportunity for interaction with other groups such as the Instituto de Biociências (USP) and the Instituto de Botânica de São Paulo (IBt-SP).

On October 2008, a special seminar was organized in the *Instituto de Estudos Avançados (IEA-USP)* to present openModeller and to discuss issues, trends and opportunities involving ecological niche modeling[35]. Dr. Townsend Peterson from the University of Kansas was invited to participate.

Workshops

On February 2006 all developers involved in the project were invited to participate in a "code fest" at CRIA to explore the existing common interests between openModeller and the BiodiversityWorld project[36]. Two representatives from BDWorld attended: Tim Sutton and Peter Brewer. The main goals were to:

- Review requirements and desirable features for an advanced openModeller graphical user interface.
- Produce an API specification for remote invocation of openModeller jobs.
- Start implementing the next generation of a graphical user interface for openModeller.
- Document the release process for both openModeller and its GUI, in particular under the Windows platform, and release omgui 0.3.4.
- Familiarize new developers with the openModeller development environment.

[34] http://www.dpi.inpe.br/referata/index.html
[35] http://www.iea.usp.br/iea/boletim/contato123.html
[36] http://www.bdworld.org/

Significant advances were achieved during the meeting, when it was decided to use the SOAP protocol with “Document” mode and “Literal” encoding combined with openModeller serialization/deserialization capabilities. The first prototypes for remote invocation methods were implemented and a full featured version was planned for 2006.

On October 2008 a niche modeling workshop[37] was organized at INPE, this time not only involving developers, but also users and researchers. Dr. Townsend Peterson was a special guest, sharing his experience with the group. Each participant had the opportunity to show the results of his/her work and interact with the others.

Training

Dr. Marinez F. Siqueira from CRIA offered training courses, lectures, and acted as advisor to graduate students working with openModeller. Last year activities include:

- Lecture: Uso de modelos de nicho ecológico para avaliar a distribuição geográfica de espécies de plantas. Programa de Seminários do Curso de Pós-Graduação em Computação Aplicada – INPE. São José dos Campos/SP. 11th May, 2006.
- Presentation: Uso de modelos de nicho ecológico para avaliar a distribuição geográfica de espécies de plantas. Symposyum: Distribuciones geográficas y patrones de diversidad. Santo Domingo, República Dominicana. 19th June, 2006.
- Presentation: Conseqüências das mudanças climáticas globais nas espécies arbóreas de Cerrado. Mesa Redonda: Mudanças Climáticas Globais”. XVI Congresso da Sociedade Botânica de São Paulo. Piracicaba/SP. 19th September, 2006.
- Instructor: Advanced course “Potential distribution of species”. CENARGEM/EMBRAPA. Feb/2008.
- Co-adviser of Luciana H. Yoshino Kamino. PhD degree: Modelagem de espécies de plantas ameaçadas de extinção de Minas Gerais. Pós-graduação em Biologia Vegetal. Laboratório de Sistemática Vegetal. Depto de Botânica /ICB /UFMG. Jan/2008.
- Co-adviser of Francisco Candido Cardoso Barreto. PhD Degree. Potential distribution modeling and models validation in openModeller. Programa de Pós-Graduação em Entomologia. Universidade Federal de Viçosa - UFV, Brazil. Feb/2008.
- Instructor: Course “Species potential distribution modeling”. Instituto de Pesquisas Ecológicas – IPÊ. Nazaré Paulista, SP, Brazil. Mar/2007.
- Instructor: Course “Species potential distribution modeling”. Instituto de Pesquisas Ecológicas – IPÊ. Nazaré Paulista, SP, Brazil. Sep/2007.
- Instructor: Course “Species potential distribution modeling”. Instituto de Pesquisas Ecológicas – IPÊ. Nazaré Paulista, SP, Brazil. Apr/2008.

[37] http://www.dpi.inpe.br/referata/Oficina_2008.html

- Instructor: Course “Species potential distribution modeling”. Instituto de Pesquisas Ecológicas – IPÊ. Nazaré Paulista, SP, Brazil. Oct/2008.
- Lecture: Modelagem de distribuição geográfica de espécies. In: XVIII Semana de Estudos da Ecologia. Instituto de Biociências, UNESP, Campus Rio Claro. 10 – 14 September, 2007.
- Lecture: Modelagem de distribuição potencial de espécies. In: Faculdades Integradas Metropolitanas de Campinas – METROCAMP. 6th October, 2007.
- Lecture: Acesso a dados de coleções biológicas. In Faculdades Integradas Metropolitanas de Campinas – METROCAMP. October 6, 2007.
- Lecture: Mudanças ambientais: possíveis impactos na biodiversidade. In: Programa de Extensão da Escola Nacional de Botânica Tropical. Seminários em Ciência e Tecnologia. Jardim Botânico do Rio de Janeiro. 20th April, 2007.
- Lecture: Environmental satellite data: applications in studies of biodiversity. “Strategies for Open and Permanent Access to Scientific Information in Latin America: Focus on Health and Environmental Information for Sustainable Development”. Atibaia, SP. 8-10 May 2007.

Dra. Silvana Amaral presented the openModeller project as part of INPE´s activity in the following events:

- Visit of the Ministry of Forestry of Indonesia at INPE, June/2007, presentation entitled “Species Distribution Modeling in the Amazônia”;
- Lecture in the Post-Graduation in Remote Sensing at INPE (24/10/2007), in the course “Tópicos Especiais em Floresta” (SER 455-3), presentation entitled “Modelos de Distribuição de Espécies”;
- Rede GEOMA Symposium, Petrópolis-RJ (29-31/10/2007), presenting the paper “Estudos de Modelagem de Distribuição de Espécies no Componente Biodiversidade na Rede GEOMA”.

Additional presentations:

Cavalcante, P.V.L. (2009) “Course of Programming with Qt”. March 16-20, 2009 – EPUSP – São Paulo.

Corrêa, P.L.P. (2009) Lecture about Biodiversity Modeling. July 27 – 2009. University of Tennesse. Graduate Programme of School of Information Sciences. College of Communication & Information. Course: Problems in Information Sciences: Environmental Informatics. Coordinator: Prof. Mike Frame (University of Tennesse).

Giovanni, R. (2005) “openModeller: A new tool for fundamental niche modelling”. BDWorld Workshop, National e-Science Centre, Edinburgh, UK, June 2005. Oral presentation.

Giovanni, R. (2008). “speciesLink & openModeller: Network and Tool Enabling Biodiversity Research”. Taxonomic Database Working Group Annual Meeting, Fremantle, Australia, October 2008. Oral presentation.

Sutton, T. & Giovanni, R. (2006). “A Web Services API for Fundamental Niche Modelling”. Taxonomic Database Working Group Annual Meeting, Saint Louis, Missouri, USA, October 2006. Oral presentation.

Sutton, T. (2008) “Workshop of Collaborative development of Free Software”. May 10, 2008 – EPUSP – São Paulo.

The following people were involved with openModeller through scholarships and training:

Doctoral students:

- Cristina Giannini, Instituto de Biociências da Universidade de São Paulo (IB/USP) ;
- Elisângela Silva da Cunha Rodrigues, EPUSP (CAPES scholarship);
- Fabiana Soares Santana, EPUSP;
- Fabrício Rodrigues, EPUSP (CAPES scholarship);
- Francisco Candido Cardoso Barreto, UFV;
- Karla Donato Fook, INPE;
- Luciana H. Yoshino Kamino, UFMG;
- Nilton Cézar de Paula, EPUSP;
- Jeferson Martin de Araújo, EPUSP.

Master students:

- Fabio Iwashita, INPE Remote Sensing Program;
- João Gabriel R. Giovanelli, UNESP, Rio Claro;
- Marcos Gonzales, ENBT/JBRJ;
- Renata Luiza Stange, EPUSP;
- Silvio Luiz Stanzani, EPUSP.

Undergraduate students:

- Albert Massayuki Kuniyoshi, EPUSP;
- Alex Oshika, Computer Engineering, EPUSP;
- Danilo de Jesus da Silva Bellini, Electric Engineering, EPUSP (CNPq scholarship);
- Luciano Bergantini Lippi, Computer Engineering, EPUSP;
- Marcos Cabral Santos, Computer Engineering, EPUSP (Fapesp scholarship);
- Mariana Ramos Franco, Computer Engineering, EPUSP (Fapesp scholarship);
- Nelson Mimura Gonzalez, Computer Engineering, EPUSP;
- Pedro Victor Losada Cavalcante, EPUSP;
- Rafael Pelegrini Domingues, Computer Engineering, EPUSP.

FAPESP Technical Training scholarships:

- Alexandre Copertino Jardim, INPE, scholarship type TT4;
- Dr. César Alberto Bravo Pariente, EPUSP, scholarship type TT5;

- Luciana Satiko Arasato, INPE, scholarship type TT3;
- Missae Yamamoto, INPE, scholarship type TT5;
- Renata Luiza Stange, EPUSP, scholarship type TT4a;
- Tim Sutton, CRIA, scholarship type TT5.

Publications

The following journal papers and conference papers were published during the project:

Amaral, S., Costa, C.B. & Rennó, C.D. (2007). "Normalized Difference Vegetation Index (NDVI) improving species distribution models: an example with the neotropical genus *Coccocypselum* (Rubiaceae)". Anais do XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, INPE, p. 2275-22282 (http://marte.dpi.inpe.br/col/dpi.inpe.br/sbsr@80/2006/11.15.14.30/doc/2275-2282.pdf).

Andrade Neto, P.R. & Justiniano Jr., P.R., (2005) "A Process and Environment for Embedding the R Software into TerraLib". VII Brazilian Symposium on GeoInformatics, GeoInfo2005. Campos do Jordão, SP, Brazil.

Andrade Neto, P. R., Justiniano Jr., P. R. & Fook, K. D. (2008) "Integration of Statistics and Geographic Information Systems: the R/TerraLib Case". VII Brazilian Symposium on GeoInformatics, GeoInfo2005. Campos do Jordão, SP, Brazil.

Arasato, L.S., Amaral, S. & Costa, C.B. (2008). "Banco de dados de palmeiras para modelagem de distribuição de espécies." Conferência Científica Internacional LBA/GEOMA/PPBio. Amazônia em Perspectiva: Ciência Integrada para um Futuro Sustentável Manaus, LBA/GEOMA/PPBio, Brazil.

Arasato, L.S., Amaral, S. & Ximenes, A.D.C. (2009). "Densidade de Drenagem e HAND (Height Above the Nearest Drainage) do SRTM para modelagem de distribuição de espécie de palmeiras no Brasil ". 14º SBSR - Simpósio Brasileiro de Sensoriamento Remoto, Natal, RN, Brazil.

Araújo, J.M., Corrêa, P.L.P. & Saraiva, A.M. (2007) "A Framework for Species Distribution Modeling: a performance evaluation approach", I2TS'2007 Proceedings of the 6th International Information and Telecommunication Technologies Symposium, Brasília: IEEE R9. Editors: Fundação Bardall de Educação e Cultura, Boukerche, A., Loureiro, A.A.F., Melo, A.C.M.A. and Gondim, P.R.L. p. 111-118. Oral presentation.

Araújo, J.M., Corrêa, P.L.P., Saraiva, A. M., Sato, L., Sutton, T.P. & Franco, M. A. (2006) "Framework for Species Distribution Modeling - A performance evaluation approach". In: Ecological Informatics ISEI2006, Santa Barbara, California. Proceedings of Ecological Informatics, ISEI2006.

Araújo, J.M., Trevelin, A.L.C., Corrêa, P.L.P., Saraiva, A. M. & Sato, L. (2008) "A high performance computing environment for hosting openModeller Framework". In: 6th International Conference on Ecological Informatics ISEI2008, Cancún, Mexico.

Barreto, F.C.C. (2008) "Modelagem de distribuição potencial de espécies como ferramenta para a conservação: seleção e avaliação de algoritmos e aplicação com *Heliconius nattereri* Felder, 1865 (Lepidoptera: Nymphalidae)". Doctoral Thesis in Entomology (UFV) – Universidade Federal de Viçosa, MG, Brasil.

Bonaccorso, E., Koch, I. & Peterson, A.T. (2006) "Pleistocene fragmentation of Amazon species' ranges". Diversity and Distributions, 12:157-164. (http://www.specifysoftware.org/Informatics/bios/biostownpeterson/Betal_DAD_2006.pdf)

Bravo, C., Neto, J.J, & Santana, F.S. (2007) "Unifyinig Genetic Representation and Operators in an Adaptive Framework". Analysis of Genetic Representations and Operators, AGRO 2007.

Bravo, C., Neto, J.J, Santana, F.S. & Saraiva, A.M. (2007) "Towards an adaptive implementation of genetic algorithms". Anais da XXXIII Conferência Latinoamericana de Informática – CLEI 2007, Taller Latinoamericano de Informática para la Biodiversidad – INBI 2007, San José, Costa Rica. Proceedings of the CLEI – Centro Latinoamericano para Estudios en Informatica, 2007. v.1 p. 1-5.

Canhos, V.P., Siqueira, M.F., Marino, A. & Canhos, D.A.L. (2008) "Análise da vulnerabilidade da biodiversidade brasileira frente às mudanças climáticas globais". Parcerias Estratégicas. Centro de Gestão e Estudos Estratégicos. (http://www.cgee.org.br/prospeccao/doc_arq/prod/registro/pdf/regdoc5033.pdf)

Caruzo, M.B., Costa, C. B., Amaral, S. & Cordeiro, I. (2007). "Aplicação de classes de solo em modelos de distribuição de espécies: um exemplo com Croton L. (Euphorbiaceae)". Paper presented at Congresso Nacional de Botânica, São Paulo.

Chapman, A.D., Muñoz, M.E.S. & Koch, I. (2005). "Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context". Biodiversity Informatics, 2, pp. 24-41. (https://journals.ku.edu/index.php/jbi/article/viewFile/5/3)

Costa, C. B., Amaral, S. & Valeriano, D. M. (2006) "Presence-only modeling method for predicting species distribution and species richness: an example with the widespread Rubiaceae genus Coccocypselum". In: XVI Congresso da Sociedade Botânica de São Paulo, UNIMEP, Piracicaba, São Paulo. 18 - 21 de setembro de 2006.

De Marco Jr, P. & Siqueira, M.F. (2007) "Como determinar a distribuição potencial de espécies sob uma abordagem conservacionista?". Megadiversidade. (accepted)

Fonseca, R.R., Corrêa, P.L.P. & Saraiva, A. M. (2006) "Meta-data architecture for a species distribution modeling system". In: 5th International

Conference on Ecological Informatics - ISEI5, Santa Barbara, USA. Delegate Manual: Elsevier / International Society for Ecological Informatics (ISEI), Oxford, UK.

Fook, K., Monteiro, A. M. V. & Câmara, G. (2006) "Web Service for Cooperation in Biodiversity Modeling". VIII Brazilian Symposium on GeoInformatics, GeoInfo2006. Campos do Jordão, SP, Brazil. (http://www.geoinfo.info/geoinfo2006/papers/p40.pdf)

Fook, K.D., Amaral, S., Monteiro, A.M.V., Câmara, G. & Casanova, M.A. (2008) "Sharing executable models through an Open Architecture based on Geospatial Web Services: a Case Study in Biodiversity Modelling". X Simpósio Brasileiro de Geoinformatica. Rio de Janeiro, RJ. Oral presentation. (http://www.geoinfo.info/geoinfo2008/papers/p29.pdf)

Fook, K.D., Monteiro, A.M.V., Câmara, G., Casanova, M.A. & Amaral, S. (2009) "GeoWeb Services for Sharing Modelling Results in Biodiversity Networks". Transactions in GIS, v.13, n.4, p. 379-399(21).

Giannini, T.C.; Santos, I.A. & Saraiva, A.M. (2009) "Ecological niche modeling and geographical distribution of pollinator and plants: a case study of *Peponapis fervens* (Smith, 1879) (Eucerini: Apidae) and Cucurbita species (Cucurbitaceae)". Ecological Informatics. doi:10.1016/j.ecoinf.2009.09.003

Giannini, T.C.; Takahasi, A.; Medeiros, M.C.M.P.; Saraiva, A.M. & Santos, I.A. (2009) "Análise de Componentes Principais (PCA) das variáveis climáticas da área de ocorrência de Krameria Loefl. (Krameriaceae)". Anais do IX Congresso de Ecologia do Brasil, 13 a 17 de Setembro de 2009, São Lourenço – MG, p. 1-5.

Giannini, T.C.; Santos, I.A. & Saraiva, A.M. (2008) "Geographical distribution modeling of plants and pollinator: a case study". In: 6th International Conference on Ecological Informatics – ISEI6. 2-5 Dezembro. Cancun, México, p.19.

Giannini, T.C.; Santos, I.A. & Saraiva, A.M. (2008) "Variáveis que afetam a modelagem da distribuição geográfica: um estudo em *Centris trigonoides* (Centridini, Apidae)". In: Anais do Encontro sobre Abelhas de Ribeirão Preto. Ribeirão Preto: FUNPEC, p. 555. 23-26 Julho. Ribeirão Preto. Brasil. USP-Ribeirão Preto. With poster.

Giannini, T.C.; Saraiva, A.M. & Santos, I.A. (2008) "Modelagem da distribuição geográfica por meio do openmodeller: estudo de caso de plantas e polinizadores". In: Anais do Encontro sobre Abelhas de Ribeirão Preto. Ribeirão Preto: FUNPEC, p. 632. 23-26 Julho. Ribeirão Preto. Brasil. USP-Ribeirão Preto. With poster.

Giovanelli, J.G.R. (2009) "Modelagem de nicho ecológico de anuros da Mata Atlântica". Master Thesis in Zoology (UNESP) – Universidade Estadual Paulista Júlio de Mesquita Filho.

Gomes, P., Ferreira, M. C., Lingnau, C., Bolfe, E. & Siqueira, M. F. (2008) "Segmentação e classificação de dossel florestal em imagens Quickbird". Ambiência (UNICENTRO), v. 4, p. 35-46.

Gonzales, N.M., Domingues, R.P. & Corrêa, P.L.P. “Programando em C++ com Qt Toolkit: Um guia prático de programação em C++ com Qt Toolkit voltado para a aplicação open source: openModeller”. Editora EPUSP. 1st Edition. ISBN: 978-85-86686-54-2.

Iwashita, F. (2007) “Sensibilidade de modelos de distribuição de espécies à qualidade do posicionamento de dados de coleta”. Master Thesis in Remote Sensing (INPE) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

Kawashima, R.S., Siqueira, M.F. & Mantovani, E. (2007) “Dados do monitoramento da cobertura vegetal por NDVI na modelagem da distribuição geográfica potencial do lobo-guará (Chrysocyon bracyurus)”. XIII Simpósio Brasileiro de Sensoriamento Remoto. Florianópolis, SC. v.13. p.3983 – 3990.

Koch, I., Peterson, A.T. & Shepherd, G. (2005) “Distribuição geográfica potencial de espécies de Rauvolfia (apocynaceae) e projeções para cenários climáticos do passado”. 56º. Congresso Nacional de Botânica, Curitiba, PR, Outubro 2005.

Kuniyoshi, M.A. & Correa, P. L. P. (2007) “Aplicação de Testes Unitários no openModeller”, Anais do 15º Simpósio Internacional de Iniciação Científica da USP, São Carlos. Abstract. Poster presentation.

Lorena, A. C., Siqueira, M. F., Giovanni, R., Carvalho, A. C. P. L. F. & Prati, R. C. (2008) “Potential Distribution Modelling Using Machine Learning”. In: The Twenty First International Conference on Industrial, Engineering & Other Applications of Applied 16 Intelligent Systems (IEA/AIE), Wroclaw, Poland. Lecture Notes in Artificial Intelligence, v. 5027, Springer-Verlag, v. 5027. p. 255-264.

Meireles, L.D., Shepherd, G.J., Koch, I. & Siqueira, M.F. (2005) “Modelagem da distribuição geográfica de Araucaria angustifolia com projeções para cenários climáticos do passado”. 56º. Congresso Nacional de Botânica, Curitiba, PR, Outubro 2005.

Muñoz, M.E.S., Giovanni, R., Siqueira, M.F., Sutton, T., Brewer, P., Scachetti, R.S., Canhos, D.A.L. & Canhos, V.P. (2009) “openModeller: a generic approach to species’ potential distribution modelling”. GeoInformatica. DOI: 10.1007/s10707-009-0090-7.

Murakami, E., Santana, F.S., Stange, R.L. & Saraiva, A.M. (2009) “The integration of open standards for Enterprise Service Bus in a solution for agribusiness and environmental applications development”. In: Joint International Agricultural Conference / 7th Biennal Conference of the European Federation of IT in Agriculture, 2009, Wageningen, Netherlands.

Pereira, R. S. & Siqueira, M. F. (2007) “Algoritmo Genético para Produção de Conjunto de Regras (GARP)”. Megadiversidade, v.3, p. 46-55. (http://www.conservacao.org/publicacoes/files_mega3/6algoritmogenetico.pdf)

Rodrigues, E.S.C., Rodrigues, F.A. & Rocha, R.L.A. (2008) “Algoritmo paralelo de entropia máxima aplicado à modelagem de nicho ecológico”.

7th International Information and Telecommunication Technologies Symposium, Foz do Iguaçu, PR, Brazil.

Rodrigues, E.S.C., Rodrigues, F.A. & Rocha, R.L.A. (2009) “Dispositivo adaptativo na análise de modelos de distribuição de espécies”. In: III Workshop de Tecnologia Adaptativa.

Rodrigues, F.A., Avilla, A.O., Rodrigues, E.S.C, Corrêa, P.L.P., Saraiva, A.M. & Rocha, R. L. A. (2009) “Species distribution modeling with neural networks”, e-Biosphere 2009, London, UK. Abstract & Poster.

Rodrigues, F.A., Rodrigues, E.S.C., Sato, L.M., Midorikawa, E.T., Corrêa, P.L.P. & Saraiva, A.M. (2008) “Parallelization of the jackknife algorithm applied to a biodiversity modeling system”. 7th International Information and Telecommunication Technologies Symposium, Foz do Iguaçu, PR, Brazil. p. 58-65.

Santana, F.S. (2009) “Uma Infraestrutura Orientada a Serviços para a Modelagem de Nicho Ecológico”. Doctoral Thesis in Computing Engineering and Digital Systems (EPUSP) – Escola Politécnica da Universidade de São Paulo, SP, Brasil. April, 2009. 141p.

Santana, F.S., Barberato, C., Stange, R.L., Neto, J.J. & Saraiva, A.M. (2009) “Application of Adaptive Decision Tables to Enterprise Service Bus Service Selection”. In: III Workshop de Tecnologia Adaptativa, São Paulo, SP.

Santana, F.S., Fonseca, R.R., Saraiva, A.M. Correa, P.L.P., Bravo, C. & De Giovanni, R. (2006) “openModeller - an open framework for ecological niche modeling: analysis and future improvements”. World Congress on Computers in Agriculture and the Environment proceedings. July, 2006. Orlando.

Santana, F.S., Giannini, T.C., Gomi, E.S., Santos, I.A. & Saraiva, A.M. (2008) “The implementation of an OWL-based ontology for relating *Peponapis* and *Cucurbita* genera”. In: 6th International Conference on Ecological Informatics ISEI2008, Cancún, Mexico.

Santana, F.S., Giannini, T.C., Santos, I.A. & Saraiva, A.M. (2008) “A comparative study applying GARP and their parallel versions for ecological niche modelling”. In: 6th International Conference on Ecological Informatics ISEI2008, Cancún, Mexico.

Santana, F.S., Gushiken, I.Y., Stange, R.L., Murakami, E. & Saraiva, A.M. (2009) “Evolution of a SOA-based architecture for agro-environmental purposes integrating GIS services to an ESB environment”. In: Joint International Agricultural Conference / 7th Biennal Conference of the European Federation of IT in Agriculture, 2009, Wageningen, Netherlands.

Santana, F.S., Murakami, E., Saraiva, A.M. & Correa, P.L.P. (2007) “A comparative study between precision agriculture and biodiversity modeling systems”. 6th Biennal Conference of the European Federation of IT in Agriculture and the World Congress on Computers in Agriculture, EFITA/WCCA 2007, Glasgow, UK.

Santana, F. S., Murakami, E., Saraiva, A. M., Bravo, C. & Correa, P. L. P. (2007) “Uma arquitetura de referência para sistemas de informação para modelagem de nicho ecológico”, Anais do 6º Congresso Brasileiro de Agroinformática – SBIAgro 2007, Campinas: Embrapa Informática Agropecuária. Editors: S.Tiernes, L.H.A. Rodrigues. p. 101-105. Oral presentation.

Santana, F.S., Pariente, C.A.B., Saraiva, A.M. & Corrêa, P.L.P. (2006) “P-GARP (Parallel Genetic Algorithm for Rule-set Production) for clusters applications”. In: 5th International Conference on Ecological Informatics - ISEI5, Santa Barbara, USA. Delegate Manual: Elsevier / International Society for Ecological Informatics (ISEI), Oxford, UK.

Santana, F. S., Pinaya, J.L.D., Saraiva, A. M., Correa, P. L. P., Becerra, J.L.R. & Bravo, C. (2007) “Aplicação de SOA para identificação de serviços em sistemas de modelagem de nicho ecológico e GIS”, I2TS'2007 Proceedings of the 6th International Information and Telecommunication Technologies Symposium, Brasília: IEEE R9. Editors: Fundação Bardall de Educação e Cultura, Boukerche, A., Loureiro, A.A.F., Melo, A.C.M.A. and Gondim, P.R.L.

Santana, F.S. & Saraiva, A.M. (2009) “SOC & SOA in Biodiversity: Discussion and Case Studies”. In: World Conference on Computers in Agriculture and Natural Resources, WCCA 2009, Reno, Nevada, USA.

Santana, F. S., Siqueira, M. F., Saraiva, A. M. & Correa, P. L. P. (2008). “A reference business process for ecological niche modelling”. Ecological Informatics Journal, v. 3 p. 75-86.

Santana, F.S., Siqueira, M.F., Saraiva, A.M. & Correa, P.L.P. (2006) “A meta-model for species spatial distribution modeling process based on ecological niche concepts”. 5th International Conference on Ecological Informatics. December, 2006. Poster presentation.

Santana, F.S.; Stange, R. L.; Giannini; T.C.; Santos, I.A. & Saraiva, A.M. (2009) “Novas abordagens de algoritmos genéticos para modelagem de nicho ecológico.” Anais do IX Congresso de Ecologia do Brasil, 13 a 17 de Setembro de 2009, São Lourenço – MG, p. 1-5.

Santana, F.S., Stange, R.L., Saraiva, A.M. & Corrêa, P.L.P. (2008) “Implementation of a management process in a SOA-based ecological niche modelling software package”. In: 6th International Conference on Ecological Informatics, ISEI, December 2008, Cancún, Mexico.

Santos I.A.; Giannini, T.C.; Naxara, S.R.C. & Saraiva, A.M. (2007) “Using openModeller to analyze the geographical distribution of the Centridini bees (Apidae, Hymenoptera)”. In: Ecological Society of America and The Society for Ecological Restoration International, Joint Meeting, San Jose, EUA.

Santos, A.L., Wanderley, M.G.L., Bestetti, C.B. & Amaral, S. (2007). “Diversidade da tribo Cynodonteae (Poaceae: Chloridoideae) no Estado de São Paulo”. Paper presented at Congresso Nacional de Botânica, São Paulo, SP.

Saraiva, A.M., Corrêa, P.L.P., Sato, L.M., Rodrigues, F.A, Santana, F.S, Rodrigues, E.S.C., Stange, R.L., Murakami, E., Giovanni, R., Canhos, D.A.L. & Canhos, V. P. (2009) “A service-based framework for species distribution modeling”, e-Biosphere 2009, London, UK. Abstract & Poster.

Saraiva, A.M.; Giannini, T.C. & Santos, I.A. (2008) ”Modelagem da distribuição geográfica de polinizadores por meio do openmodeller”. In: Anais do Encontro sobre Abelhas de Ribeirão Preto. Ribeirão Preto: FUNPEC, p. 557. 23-26 Julho. Ribeirão Preto. Brasil. USP-Ribeirão Preto. With poster.

Siqueira, M.F. & Durigan, G. (2007). “Modelagem da distribuição geográfica de espécies lenhosas de cerrado no Estado de São Paulo”. Revista Brasileira de Botânica. v.30. p239-249.

Siqueira, M.F, Durigan, G., De Marco Jr, P. & Peterson, A.T. (2009) “Something from nothing: Using landscape similarity and ecological niche modeling to find rare plant species”. Journal for Nature Conservation. V17(1):25-32.

Souza, M.G. (2007) “Distribuição geográfica conhecida e potencial de *Hennecartia omphalandra* Poisson e *Macropeplus ligustrinus* (Tul.) Perkins (Monimiaceae)”. Master Thesis in Botanics. Instituto de Pesquisa Jardim Botânico do Rio de Janeiro.

Stange, R.L., Buani, B., Santana, F.S., Corrêa, P.L.P., Hirakawa, A.R. & Saraiva, A.M. (2009) “Integration of ecological niche modelling systems and IABIN-PTN using Web Services”. In: World Conference on Computers in Agriculture and Natural Resources, WCCA 2009, Reno, Nevada, USA.

Stange, R.L., Giannini, T.C., Santana, F.S., Neto, J.J. & Saraiva, A.M. (2009) “Estudo comparativo entre algoritmos genéticos adaptativos e não-adaptativos para a modelagem ambiental de *Peponapis* e *Cucurbita*”. In: III Workshop de Tecnologia Adaptativa, São Paulo, SP.

Stange, R.L., Santana, F.S. & Saraiva, A.M. (2008) “Applying J2EE patterns to develop a SOA-based architecture for ecological niche modelling”. In: IEEE R9 International Information and Telecommunication Technologies Symposium, Foz do Iguaçu, PR, Brazil.

Sutton, T., Giovanni, R. & Siqueira, M.F. (2007) “Introducing openModeller - A fundamental niche modelling framework”. OSGeo Journal Volume 1. ISSN 1994-1897. (http://www.osgeo.org/files/journal/final_pdfs/OSGeo_vol1_openModeller.pdf)

Final Comments

Species potential distribution models are important tools that can deal with a wide range of global issues, which include predicting the impact of climate changes, preventing the spread of invasive species and disease vectors, improving agricultural productivity among many others. This project gave a substantial contribution to the development of a set of tools that we believe can help address any of such issues over the next years. It also served as a

unique opportunity for many people to learn and produce high quality research, from undergraduate students to postdoctoral scientists and researchers.

The publicity gained from regular software releases, publications and interactions with other individuals and institutions has resulted in a number of potential areas for future collaboration with the wider scientific community. These include:

- An informal offer from Dr. Neil Caithness at the University of Oxford (UK) to host openModeller services at the OxGrid Campus Grid Computing Centre and the National Grid Service for the UK.
- Informal discussions with various people from the American Museum of Natural History and NatureServe on how we can help them to integrate openModeller into their current niche modeling processes.
- Informal discussions with Brian Hamlin (UC Berkeley, USA) towards including openModeller in future large scale modeling experiments they are planning.

openModeller was also selected to be used by the GBIF web portal and by the second generation of the LifeMapper project being developed by the University of Kansas. Another initiative using openModeller is the GEOSS demonstration project developed by GBIF and the Italian National Research Council. The result of this demonstration project is expected to be used by another project called Ecological Model Web[38] being developed by the Ecological Forecasting Program at NASA.

In addition, we have been able to engage with users of our software from various countries around the world through our users' mailing list, IRC presence and meetings. As an example, this has enabled the introduction of a Taiwanese translation of openModeller Desktop which was contributed by one of our users. Another example was the development of another web interface for openModeller by a Phd student at the Institute of Zoology from the Chinese Academy of Sciences[39].

There was always a major concern related with the continuity of all developments produced by this project. Being an open source initiative based on accessible tools and open standards, if funded, development can easily continue, even by other institutions. There are many opportunities that can be further explored with the results of this project, such as performing large sets of modeling experiments for many species with a specific conservation purpose, or building species distribution model repositories that could be directly used in decision making.

We would like to thank Fapesp for all opportunities that this project has provided, and we hope the results of this work represent a significant contribution to the fields of biodiversity and computing research.

[38] http://www.ieeexplore.ieee.org/xpl/freeabs_all.jsp?isnumber=4422708&arnumber=4423343&index=634

[39] http://sourceforge.net/mailarchive/message.php?msg_name=AFD7D882A5F540FC986655EA37F84A81%40qiaohjhp