Sharing executable models through an Open Architecture based on Geospatial Web Services: a Case Study in Biodiversity Modelling

Karla Donato Fook[1,2], Silvana Amaral[1], Antônio Miguel Vieira Monteiro[1], Gilberto Câmara[1], Marco A. Casanova[3]

[1]Image Processing Division – National Institute of Space Research (INPE)
São José dos Campos – SP – Brazil
[2]Computer Science Department
[2]Centro Federal de Educação Tecnológica do Maranhão (CEFET–MA)
São Luís – MA – Brazil
[3]Informatics Department – PUC-Rio
Rio de Janeiro – RJ – Brazil

{karla, silvana, miguel, gilberto}@dpi.inpe.br,

casanova@inf.puc-rio.br

Abstract. *Biodiversity researchers develop predictive models for species occurrence and distribution which are useful for biodiversity conservation policies. Species distribution modelling tools need to locate and access large amount of data in different sources and produces results from different algorithms. In this scenario, collaboration is an essential feature to improve this research area. Scientists need to share models, data and results to get new discoveries. This paper presents advances in Web Biodiversity Collaborative Modelling Services (WBCMS) development. These services support sharing of modelling results and information about its generation. WBCMS also enable researcher to make new experiments based in previous one. Scientists can use WBCMS to compare experiments, to make new inferences and to improve their studies. A case study explains the model instance usage.*

1. Introduction

Biodiversity research uses tools that allow performing inferences about diversity and abundance of species in different areas. Species distribution models combine *in situ* species data with geographical data. Their results support biodiversity protection policies, are useful to forecast of the impacts of climate change, and help detect problems related to invasive species. Since such data sets may be archived by different institutions, the scientist needs to locate the data sets and make them interoperate. These points create challenges that lead to data representation, management, storage, and access problems. In addition, the scientist would like to share his experiments results with the community and compare it with similar work done elsewhere.

This scenario points to the need for a computational infrastructure that supports collaborative biodiversity studies, allowing sharing data, models and results [Ramamurthy 2006]. Each of these three aspects needs a different strategy. Sharing data needs information about the location of repositories and archival formats. Sharing models needs understanding about the applicability of each algorithm to the species being modelled; it also requires a good documentation explicit and implicit assumptions behind the model. Sharing results needs communication of the species distribution maps as well as producing reports and adding comments. In this context, metadata are useful in order to disambiguate the data and enable reuse. One kind of metadata is provenance, which records data about scientific experiments [Simmhan, Plale and Gannon 2005]. Provenance metadata allows researchers to capture relevant information about scientific experiments, and to assess the experiment quality and timeliness of results [Greenwood, Goble, Stevens et al. 2003; Marins, Casanova, Furtado et al. 2007].

This paper reports advances on development of the Web Biodiversity Collaborative Modelling Services (WBCMS). They are geospatial web services that support cooperation on a species distribution modelling network, including sharing modelling results and its provenance, and enabling researchers to perform new experiments based in previous ones. Prototypes were implemented. An early prototype stored algorithms information in the database and does not produce the model instance. For more details, see [Fook, Monteiro and Câmara 2007]. A new prototype was developed. This prototype is more robust than the early prototype. The main differences between WBCMS prototypes are that the current prototype composes the model instance, and also enable researcher to reuse model instance data to produce new experiments. The WBCMS architecture is part of the OpenModeller[1] Project, a framework for collaborative building of biodiversity models [Muñoz 2004; Giovanni 2005; OpenModeller 2005].

This work is organized as follows. Section 2 presents related work. Section 3 describes WBCMS in detail. In Section 4, we show the current prototype by model instance use example. Finally, section 5 presents final remarks and further work.

[1] http://openmodeller.cria.org.br/

2. Related Work

Trends have enabled a new generation of data services in the scientific community. Web services stand out to support distributed applications in geospatial domain, where geographical application are divided in a series of tasks, organized in a workflow. Bernard et al. (2003) have developed a "road blockage" service, which solve more complex tasks by static chaining several simple services. WS-GIS approach is an SOA-based SDI which aims to integrate, locate, and catalog distributed spatial data sources [Leite-Jr, Baptista, Silva et al. 2007].

The *GeoCatalog* is a tool that implements a software architecture for automated geographic metadata annotation generation [Leme, Brauner, Casanova et al. 2007]. Díaz et al. (2007) designed a gvSIG[2] extension to collect automatically metadata. This application aids users to publish imagery or cartographic data in a Spatial Data Infrastructure. The Earth System Science Workbench (ESSW) is a metadata management and data storage system for earth science researchers. Their infrastructure captures and keeps provenance information for proving credibility of investigator-generated data [Frew and Bose 2001].

In biodiversity field, Best et al. (2007) use geospatial web services to automate the scientific workflow process in marine mammal observations from OBIS-SEAMAP[3]. Web Service Multimodal Tools for Biodiversity Research, Assessment and Monitoring Project (WeBIOS) provides scientists with a system that supports exploratory multimodal queries over heterogeneous biodiversity data sources [WeBios 2005]. BioWired project proposes a P2P grid architecture that supports biodiversity data access by large distributed database [Alvarez, Smukler and Vaisman 2005]. BiodiversityWorld project intends to make available heterogeneous data sources and biodiversity analytic tools in a Grid [Jones, White, Pittas et al. 2003; Pahwa, White, Jones et al. 2006].

The approaches above aim to integrate and share geographical data and tools. However, they do not aim to share modelling results. Our approach aims to support sharing descriptive information about spatial data, and relevant information objects. Our goals are to publish modelling experiments and their provenance, to make it available

[2] www.gvsig.gva.es

into catalogues, and to enable researchers to perform new models based in catalogued model instances.

3. WBCMS description

This section presents the Web Biodiversity Collaborative Modelling Services (WBCMS), a set of geospatial Web services that enables sharing of modelling experiments, and reusing of these data in new experiments.

This approach aims to capture the explicit and implicit information inserted in a biodiversity experiment, in our case, a species distribution modelling. A key idea behind WBCMS is a model instance. It includes data and metadata related to models, results and algorithms and describes an experiment as a whole. The idea is that the researcher examines model instances and be able to understand how a result was produced. He can then compare experiment results to reproduce them, and to use them for his own models. He can get answers for queries such as *"What species are being modelled?"*, *"Where does the data come from?"*, *"Which environmental variables are used?"*, and *"If I have a question, how can I look for similar results?"*. So, consider a distributed environment in which researchers perform species distribution modelling locally, and wish to share their experiments (Figure 1).

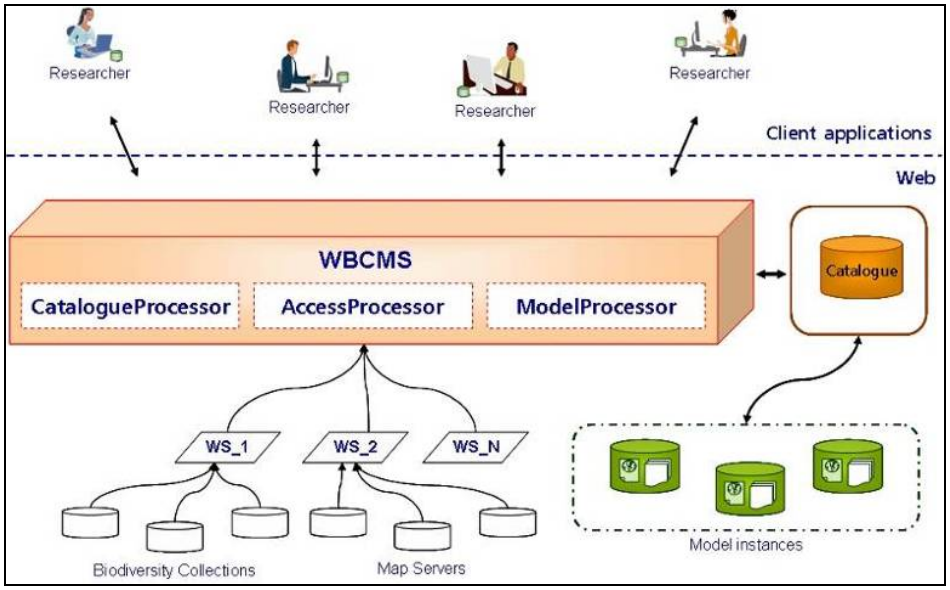

Figure 1. WBCMS Architecture

[3] http://seamap.env.duke.edu

Briefly, researchers can use the WBCMS to (a) share their modelling experiment, (b) access and evaluate experiments, and (c) perform new models based in catalogued models. Therefore, WBCMS builds a catalogue of model instances and holds processors to handles with each activity: *Catalogue Processor*, *Query Processor*, and *Model Processor*. These processors include a set of web and geoweb services. The model instance catalogues can be in different institutions and holds information related to different kind of model, such as environmental and urban models. Therefore, one challenge in this approach is to specify the model instance, since it must provide researchers with the necessary information for a better understanding of an experiment. We present our idea of a model instance in next subsection.

3.1 Model instance outline

This subsection describes the model instance in WBCMS architecture. It aims to describe a modelling experiment as whole. The model instance idea includes several types of models such as Land Use and Coverage Change, and Natural Hazards models. In our case, we are working with species distribution modelling where the modeled object is a species. The model instance includes data and metadata about the model, its generation process, and experiment results (see Figure 2).

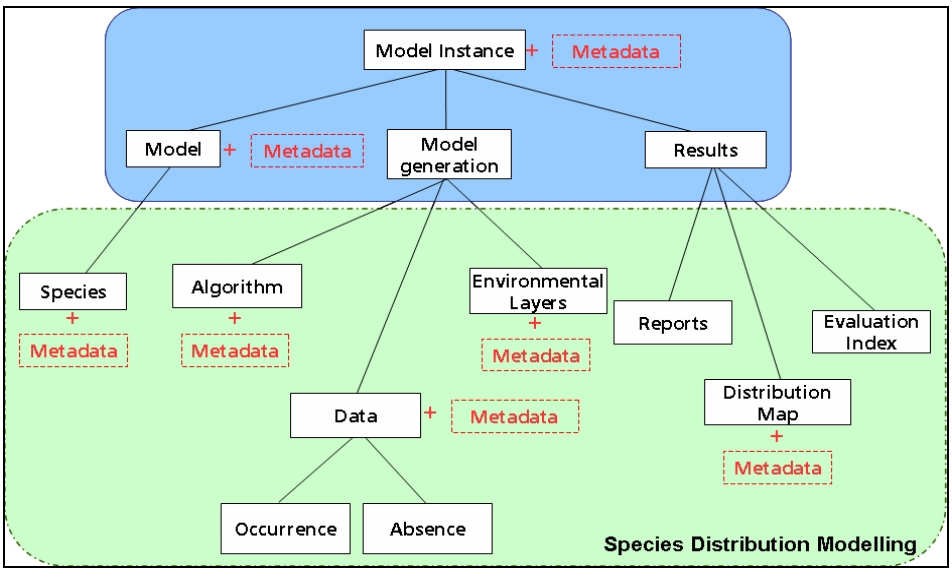

Figure 2. Model Instance Diagram

Figure 2 shows the model instances diagram and highlights that each element of model instance contains their own metadata. The model instance includes information as shown in Table 1.

Table 1. Model instance elements

Element	Desccription	Information
Model instance	Global information related to modelling experiment, and researcher's notes that help scientists in experiment analysis.	Model instance name, title, description, author, affiliation, creation date, running time, and modelling motivation (or question) comments, confidence degree and its justification
Model	Information about the used model, and information related to modelled object, in this case, modelled species.	Model author, description and version. Species: taxonomic data (classification), and their metadata (author, status, online resource, and reference date).
Model generation	Input data and used algorithm, as well as metadata such as execution time and messages.	Species occurrence and absence points (latitude and longitude), and environmental layers are input data examples. It also includes algorithm parameters and metadata like description, version, author, and contact.
Results	Set of modelling result files.	Reports, georeferenced maps, and model evaluation indexes.

Besides metadata about experiment results, a model instance includes other information such as species taxonomic data (see Table 1). Species-occurrence presents different reliability degree to biodiversity researchers, because these records have different sources and methods. Therefore, make it available is not enough to assure their use by the community. The minimum requirements for a species occurrence record are its geographical positioning, and its taxonomic identification together with metadata such when, and details about where the specimen was collected [Guralnick, Hill and Lane 2007].

We used the ISO19115 standard [ISO 2003] to describe the model instance. It includes the experiment provenance, and provides evaluation features for accessing the experiment. The model instance has a set of metadata to describe itself globally, and to describe model instance elements. Therefore, there are metadata copies to different components, for instance use the reference date metadata to points to different dates:

experiment performing, experiment cataloguing and species data recovering. WBCMS extracts part of metadata from result files and recovers another part from web. On the other hand, the researcher needs to inform extra metadata related to experiment in client application, as description and lineage. Next subsection describes WBCMS processors in detail.

3.2 WBCMS Processors

This subsection presents the WBCMS processors. The WBCMS *Catalogue Processor* publishes a model instance. The researcher uses a catalogue application to send basic experiment elements to WBCMS. The *Catalogue Processor* receives modelling result data, accesses remote data, and composes model instance. Then, the WBCMS inserts a model instance into the repository. Figure 3 details the WBCMS *Catalogue Processor*.

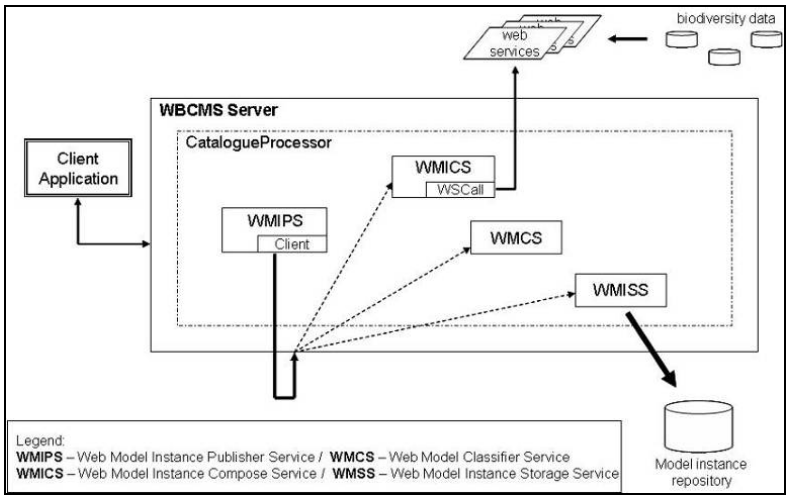

Figure 3. WBCMS Catalogue Processor

The *Catalogue Processor* includes following services: WMIPS – Web Model Instance Publisher Service, WMICS – Web Model Instance Compose Service, WMCS – Web Model Classifier Service, and WMISS – Web Model Instance Storage Service. The WMIPS is an orchestration service that controls the other catalogue processor services. WMCS uses model metadata to perform a model instance classification. WMICS recovers biodiversity data and metadata from web to complement the model instance. Finally, WMISS inserts a model instance into a repository.

A researcher uses the WBCMS *Access Processor* to retrieve model instances. This processor uses the OGC WFS – Web Feature Service [OGC 2005] and two

services: WMIQS – Web Model Instance Query Service, and WMIRS – Web Model Instance Retrieval Service (see Figure 4).

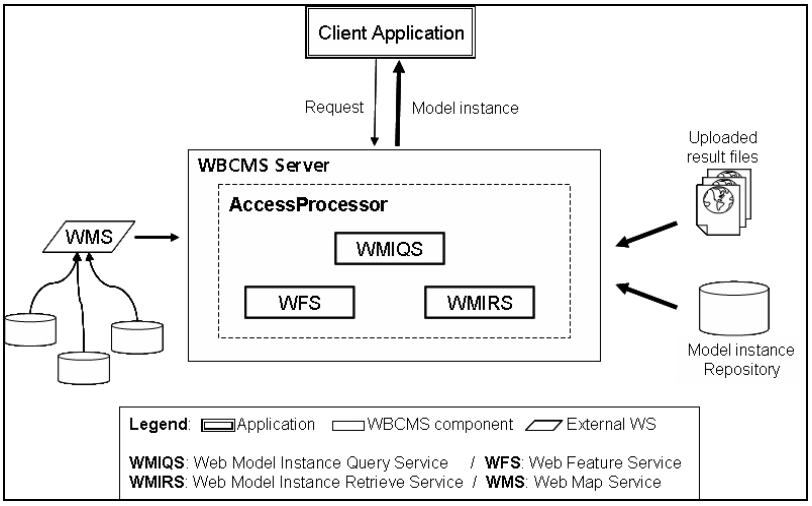

Figure 4. WBCMS *Access Processor*

The WBCMS *Access Processor* receives requests from a client application and uses WMIQS to handle queries, and WMIRS and WMS [OGC 2006] to recover the model instance, and make it available.

The researchers can reuse catalogued data to execute remotely new models using the WBCMS *Model Processor*. This processor includes the WMRS – Web Model Run Service, and uses the OMWS – OpenModeller Web Service. The WMRS is responsible to: (a) prepare input data and allows user to change algorithm parameters, (b) call OMWS to perform the new model, and (c) increment the model instance run count at each model instance reuse. The last activity allows a statistic evaluation of the instance model reuse. We use the UML communication diagram to show the WBCMS *Model Processor* usage (Figure 5).

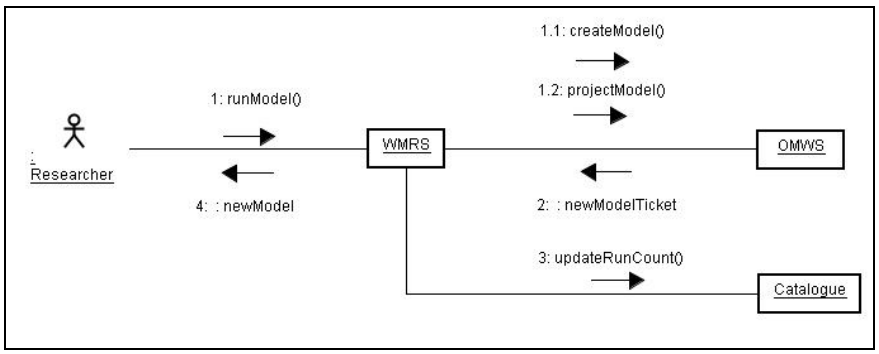

Figure 5. WBCMS *Model Processor*

The WMRS receives the researcher request to perform a new model, and send the necessary request and input data to OWMS produce it. The OMWS receives occurrence data and algorithm parameters from client, performs the model, and returns the produced species distribution model [Giovanni 2005; Sutton, Giovanni and Siqueira 2007]. We developed a prototype as proof of concept of our approach. Figure below shows WBCMS class diagram.

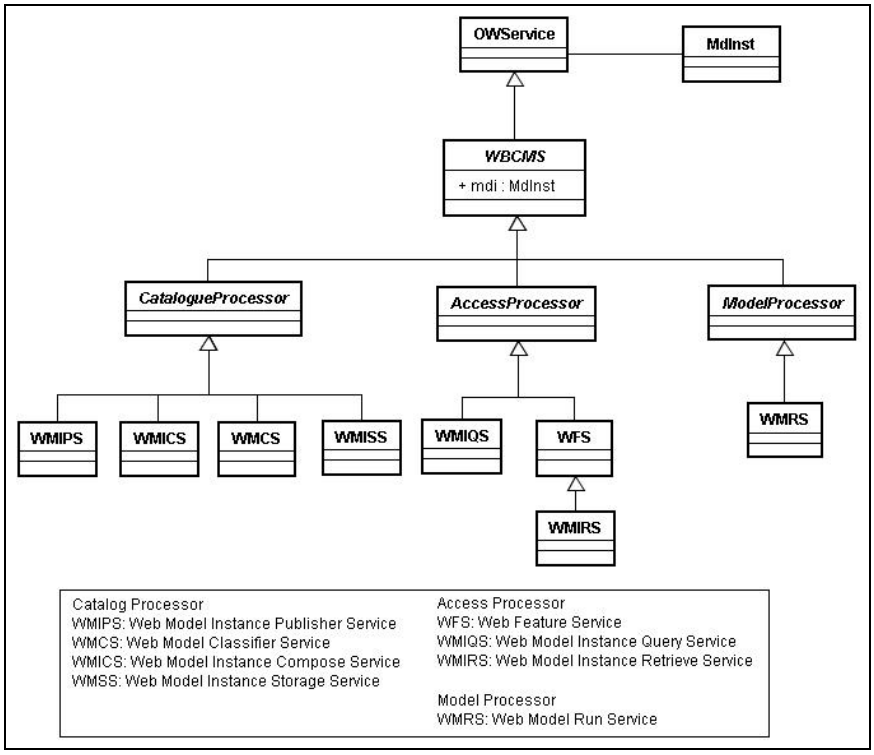

Figure 6. WBCMS Class Diagram

Figure 6 shows web and geoweb services of proposed architecture. There is an association relation between WBCMS and MdInst (*Model Instance*) classes. Next section presents an example of the WBCMS prototype functionalities.

4. WBCMS Prototype: A model instance usage example

This section presents an example that shows how the WBCMS makes a model instance available and how a researcher can produce new species distribution models. The example considers the *Coccocypselum erythrocephalum* Cham. & Schltdl. Species. The genus *Coccocypselum* belongs to Rubiaceae family, one of the most important families in the tropics.

In this example, we show the model instance **md_CErythr**. Initially, the researcher uses the OpenModeller Desktop [Giovanni 2005; Sutton, Giovanni and Siqueira 2007] to produces the species distribution model. This model consists of several result files, such as distribution map, reports and configuration files. The researcher uses the *Model Instance Catalogue* application to capture provenance information from result files, to inform personal comments about the experiment, and to publish the model instance into the catalogue.

The researcher can access **md_CErythr** model instance using the *Model Instance Access* application. This application enables the scientist to visualize each model instance element, and to perform new models based in previous ones. Figure 7 illustrates the modelling results visualization.

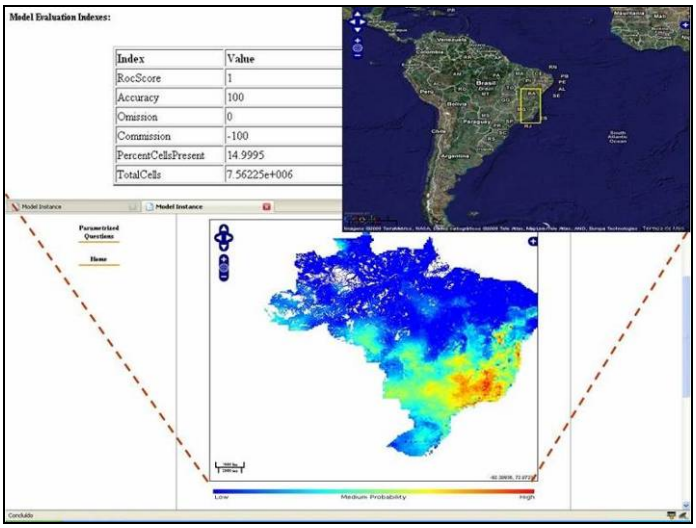

Figure 7. Model instance access application – Modelling result

Besides **md_CErythr**'s species distribution map, the form displays model evaluation indexes, and map with bounding box showing the species area (Figure 7). The evaluation indexes and author comments about the experiment help the user to capture relevant aspects of the species distribution model. The *Model Instance Access* application also makes available metadata about algorithms and model instance authors. The researcher can use WBCMS to perform new models reusing catalogued model instance data. Figure 8 displays **md_CErythr**'s algorithm metadata and parameters. The researcher can change algorithm parameters and layers to produce different models using OMWS (OpenModeller Web Service). So, several models can be produced (Figure 9).

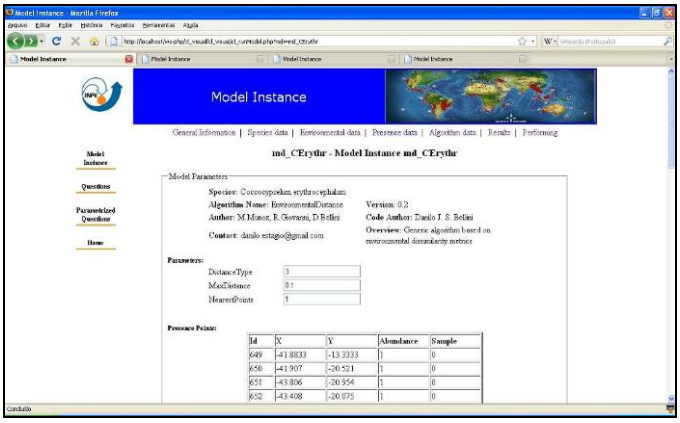

Figure 8. md_CErythr model instance reuse form

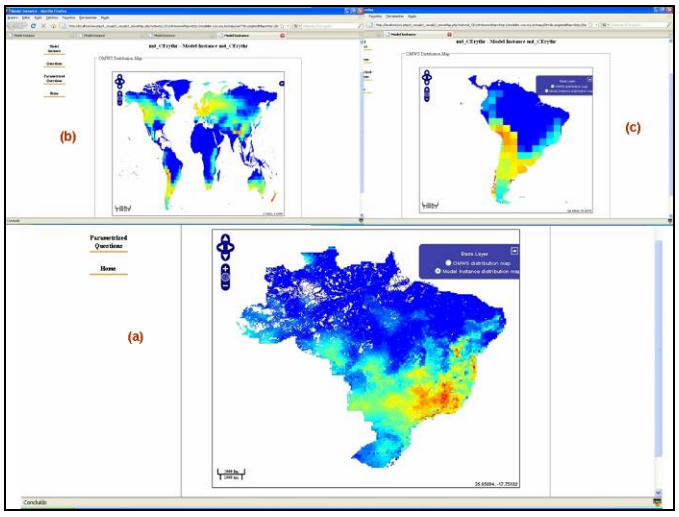

Figure 9. Species distribution maps based on md_CErythr model instance

Figure 9 shows the **md_CErythr** species distribution map (a), and distributions maps based in this model instance (b, and c labels). Algorithm parameters and layers were changed to produce these models. A detailed discussion of distribution maps analysis is beyond the scope of this paper. The objective is enables scientist to compare different distribution maps and to make new inferences about his studies.

5. Final Remarks

We presented in this paper advances in the development of Web Biodiversity Collaborative Modelling Services (WBCMS), a set of geospatial web services that aim at making it available modelling experiment results in a species distribution network, and enable researchers to perform new models based in previous ones.

We introduced the model instance idea that aims at describing an experiment as whole. Then, a set of ISO metadata elements were selected to describe a model instance.

We used compliant OGC web services to show model instance elements. However, existent specifications are not enough to work with the sharing of model description and results. In addition, we developed web services to handle with model instance complexity. We also included in the paper a model instance example illustrating the WBCMS use.

Our experiments, have demonstrated the validity of the proposals and ideas presented in this paper. We consider this line of work promising, even though more tests with a larger volume of modelling experiments are required. Finally, we remark that we will to improve WBCMS to handle more complex query predicates, and to provide model instance reuse statistics.

Acknowledge

Special thanks go to Dr. Cristina Bestetti Costa for their relevant comments and species occurrence data. We also thanked OpenModeller Project (FAPESP process: 04/11012-0), and FAPEMA[4] (In Portuguese: Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão) for partially supporting this research.

References

Alvarez, D., Smukler, A. and Vaisman, A. A. (2005). "Peer-To-Peer Databases for e-Science: a Biodiversity Case Study." Proceedings 20th Brazilian Symposium on Databases and 19th Brazilian Symposium on Software Engineering.

Bernard, L., Einspanier, U., Lutz, M., et al. (2003). Interoperability in GI Service Chains-The Way Forward. 6th AGILE Conference on Geographic Information Science, Muenster.

Best, B. D., Halpin, P. N., Fujioka, E., et al. (2007). "Geospatial web services within a scientific workflow: Predicting marine mammal habitats in a dynamic environment." Ecological Informatics **2**: 210-223.

Díaz, L., Martín, C., Gould, M., et al. (2007). Semi-automatic Metadata Extraction from Imagery and Cartographic data. Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS 2007), Barcelona (Spain).

Fook, K. D., Monteiro, A. M. V. and Câmara, G. (2007). Web Service for Cooperation in Biodiversity Modeling. Advances in Geoinformatics. C. Davis and A. M. V. Monteiro, Springer**:** 203-216.

Frew, J. and Bose, R. (2001). Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products. 13th International Conference on Scientific and Statistical Database Management (SSDBM), Virginia, USA.

Giovanni, R. D. (2005). The OpenModeller project. BiodiversityWorld GRID workshop, Edinburgh, e-Science Institute.

[4] http://www.fapema.br/

Greenwood, M., Goble, C., Stevens, R., et al. (2003). Provenance of e-Science Experiments - experience from Bioinformatics. 2nd UK e-Science All Hands Meeting, Nottingham, UK.

Guralnick, R. P., Hill, A. W. and Lane, M. (2007). "Towards a collaborative, global infrastructure for biodiversity assessment." Ecology Letters **10**: 663-672.

ISO (2003). ISO 19115 Geographic Information - Metadata. Geneva, International Organization for Standardization (ISO).

Jones, A. C., White, R. J., Pittas, N., et al. (2003). BiodiversityWorld: An architecture for an extensible virtual laboratory for analysing biodiversity patterns. UK e-Science All Hands Meeting, Nottingham, UK, Cox, S.J.

Leite-Jr, F. L., Baptista, C. S., Silva, P. A., et al. (2007). WS-GIS: Towards a SOA-Based SDI Federation. Advances in Geoinformatics. C. Davis and A. M. V. Monteiro, Springer**:** 247-264.

Leme, L. A. P., Brauner, D. F., Casanova, M. A., et al. (2007). A Software Architecture for Automated Geographic Metadata Annotation Generation. 2007 e-Science Workshop.

Marins, A., Casanova, M. A., Furtado, A., et al. (2007). Modeling Provenance for Semantic Desktop Applications. SEMISH - XXXIV Seminário Integrado de Software e Hardware, Rio de Janeiro, RJ, SBC.

Muñoz, M. (2004). openModeller: A framework for biological/environmental modelling. Inter-American Workshop on Environmental Data Access, Campinas, SP. Brazil.

OGC. (2005). "OpenGIS Web Feature Service (WFS) Implementation Specification." Panagiotis A. Vretanos. from http://portal.opengeospatial.org/modules/admin/license_agreement.php?suppressHeaders=0&access_license_id=3&target=http://portal.opengeospatial.org/files/index.php?artifact_id=8339

OGC (2006). OpenGIS Web Map Server Implementation Specification, OGC - Open Geospatial Consortium Inc.

OpenModeller. (2005). "openModeller: Static Spatial Distribution Modelling Tool." Retrieved agosto/2005, from http://openmodeller.cria.org.br/.

Pahwa, J. S., White, R. J., Jones, A. C., et al. (2006). Accessing Biodiversity Resources in Computational Environments from Workflow Applications. The Workshop on Workflows in Support of Large-Scale Science.

Ramamurthy, M. K. (2006). "A new generation of cyberinfrastructure and data services for earth system science education and research." Advances in Geosciences **8**: 69-78.

Simmhan, Y. L., Plale, B. and Gannon, D. (2005). "A survey of data provenance in e-Science." SIGMOD Record **34**(3): 31-36.

Sutton, T., Giovanni, R. d. and Siqueira, M. F. d. (2007). "Introducing openModeller - A fundamental niche modelling framework." OSGeo Journal **1**.

WeBios (2005). WeBios: Web Service Multimodal Tools for Strategic Biodiversity Research, Assessment and Monitoring Project, http://www.lis.ic.unicamp.br/projects/webios.