Journal of Computational Interdisciplinary Sciences (2011) 2(1): 37-46

Printed version ISSN 1983-8409 / Online version ISSN 2177-8833
http://epacis.org

TaxonomyBrowser: a biodiversity data management system

Denison L.M. Tavares[1], Samantha C. Cañete[1], Rafael Henkin[1], Pedro C. Estrela[2], Thales R.O. Freitas[3], Renata Galante[1] and Carla M. Dal Sasso Freitas[1]

Manuscript received on August 11, 2010 / accepted on March 10, 2011

ABSTRACT

This paper describes the main components of a biodiversity information system aimed at integrating visualization and data analysis tools with data management functions provided by relational databases. The system TaxonomyBrowser has the goal of aiding biologists in the management of data about specimens they collect in field work as well as serving to catalog specimens found in museums collections. The integration with map visualization and data analysis tools, through the Google Maps API and the R System, respectively, widens the application of the system.

Keywords: biodiversity management systems, information visualization.

1 INTRODUCTION

Biodiversity information systems are a broad category of applications including data management systems that support biologists in storing information about specimens they collect during field work [1, 2]. Considerable part of these systems aims at maintaining data about specimens belonging to scientific collections in museums and other research institutions. Some biodiversity information systems are available through the web (for example, see [3]), but the great majority of scientific databases supporting the work of biologists are based on conventional data base systems or even spreadsheets restricted to the group of local users.

Recently, however, due to the *Taxonomic Databases Working Group* (TDWG) initiatives and the growing use of geo-referenced information based mainly on Google Maps/Earth applications, there has been a surge of many systems intended to share biodiversity data through the web. Darwin Core [4] is a standard for describing data about the occurrence of species and specimens in scientific collections, used by DiGir (*Distributed Generic Information Retrieval*), which is a protocol for communication and publication of data stored in different data bases. As an example of a web system using both the Darwin Core standard and Google Maps, we can cite MaNIS (*Mammal Networked Information System*) [5], which allows access to data about mammals stored in several scientific collections in North-American museums. In Brazil, the most important contribution related to biodiversity data communication is SinBiota [6], a data management system of collection records and observational data restricted to the state of São Paulo. SinBiota has a graphical interface for displaying specimens' distribution and temporal information on a map of that state.

Without any doubt, tracing strategies for biodiversity conservation requires data about the specimens that inhabit the target region. Depending on goals of their studies, biologists might need from a few morphological measures and anatomic characteristics observed from the collected specimens, and information about

Correspondence to: Carla M. Dal Sasso Freitas – E-mail: carla@inf.ufrgs.br
[1]Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.
[2]Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, RJ, Brazil.
[3]Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.

the habitat, to images taken from skulls and complex DNA sequences descriptions obtained in laboratory. Therefore, during different phases of their work, biologists who are users of scientific collections catalogue basic information about collected specimens, retrieve data textually or in maps, process data depending on the objectives of their studies, and produce data either in lab procedures or using computer applications, which should be stored in the existing data bases. All these activities are accomplished with the aid of different computational tools, without integration, being up to the users to provide the data flow between them. In this context, one identifies the need of integrating, in a single computational platform, tools for: (a) managing data, (b) supporting queries of different levels of complexity, including data analysis, and (c) providing data communication tools that allow information to be displayed in different levels of details.

This work describes TaxonomyBrowser, a biodiversity information system developed to comply with the above-mentioned requirements. The system was developed following a participatory design methodology with the involvement of real users from the beginning of the project up to now. Here we extend our previous work [7] by providing more details about the system and presenting new features that were added to the analysis interface. The rest of the paper is organized as follows. We discuss the solutions provided by similar systems in Section 2, and the proposed data model is described in Section 3. Section 4 presents the user interface concentrating in the information visualization approach on which we based our solution both for communicating query results and data analysis. Finally, Section 5 concludes our paper by analyzing the present results and drawing future work.

2 RELATED WORK

There are a large number of projects that aims at managing and publishing biodiversity data on the Web. The Biota project [8] was one of the first projects with such goals, providing support for storing and communicating occurrence records – those that contain observations and measurements taken by biologists during fieldwork. The projects closest to TaxonomyBrowser are Sinbiota, WeBios (*Web Service Multimodal Tools for Biodiversity Research, Assessment and Monitoring*) and BioCORE (*Biodiversity and Computing Research*) [9]. SinBiota [8] is a data management system of collection records and observational data for the state of São Paulo, Brazil. SinBiota has a graphical interface for displaying specimens distribution and temporal information on a map of São Paulo, Brazil. WeBios [9] is a biodiversity information system developed as a joint initiative of biodiversity and computer science researchers at UNICAMP, Brazil. Its goal is to provide bio-scientists with a system that supports exploratory queries over heterogeneous and distributed biodiversity data sources on the Web. It has a service-oriented architecture and employs semantic Web technologies. The BioCore project [10] is a web biodiversity system also developed in a joint effort between computer scientists and biologists. The queries are centered on two kinds of biodiversity data: (i) occurrence records, containing observations about specimens collected or observed during field trips; and (ii) catalog records, containing information from museums' collections.

Several systems available on the Web provide users ways to find the specific set of information that they are interested in without the need of posing complex queries considering different specimens or species. Species2000 [3] provides a collections catalogue, which allows querying specimens by using the taxonomic tree or their scientific names. Projects like GBGIF (*Global Biodiversity Information Facility*) [11], ITIS (*Integrated Taxonomic Information System*) [12] and TDWG (*Taxonomic Database Working Group*) [4] are directing their efforts to establish standards and infrastructure for the integration and interoperability of data among biological collections, making them available on the Web. *Tree of Life* [13] is another considerable set of biodiversity applications that deals with the management of taxonomic data and geographic distribution of species.

Museum collections are also available on the web. MaNIS (*Mammal Networked Information System*) [5] is a web system using both the Darwin Core standard and the Google Maps API to allow accessing data about mammals from several scientific collections in North-American museums. The query results are displayed in the Google Maps interface and exported to both relational and XML format. Arctos [14] is an ongoing effort to integrate access to specimen data, collection-management tools, and external resources on the Web. The users can pose queries considering the taxonomic classification and/or the geographic distribution of species or taxonomic entities. The query results are also kept in the database for future accesses.

Although TaxonomyBrowser shares with other systems the same general goal of aiding biologists in storing and managing their data about specimens, its main contribution is the integration of several visualization and analysis tools in a single environment accessible through the web. Data management as well as data visualization and analysis are available through the same web interface, and the levels of access to the stored information depend on the user roles. An important difference between TaxonomyBrowser and the other systems is the possibility of obtaining several visual representations for the results, depending on what the user needs to observe.

3 TAXONOMY BROWSER BIODIVERSITY DATA MODEL

This section describes the data model designed for the TaxonomyBrowser application. This data model intends to support the representation of biodiversity data as well as querying the dataset, based on the attributes of the taxonomic entities as well as on their collection sites. The data model was conceived during many design sessions with the collaboration of biologists and tests on previous prototyped systems.

3.1 TaxonomyBrowser Architecture

Figure 1 shows an overview of the architecture of the TaxonomyBrowser. The *Collection Manager Interface* (A) allows managing all data that describes the collected (or observed) specimens by providing the four basic operations (insertion, updating, removal and visualization). A dual visual interface (B) integrates the *Query* and the *Map Interfaces*. They are based on the Google Maps/Earth API, and allow performing geo-referenced queries on the dataset. The *R-based Analysis Interface* provides analytical functionalities based on scripts written in the R language, playing the role of advanced queries needed by the biologists in their data analysis procedures. These advanced queries are performed (transparently) by an external application, the System R, which executes the scripts, the results being returned to the user as an image. The database representing the scientific collection (D) stores all the collection data, and the mapping layer translates the data to each interface. The system also provides different views to users, depending on their access rights (or roles), from lay users to the collection manager. Lay (anonymous) users have access to the collection data but are not able to execute scripts, while a user registered as a group member can perform advanced queries, and the collection manager has complete control of the data stored in the database.

3.2 Data Collection Overview

The database was modeled to store data about specimens collected or observed by biologists in order to support the queries they need for data analysis purposes as well as browsing and specific searches performed by anonymous users, either for entertaining and educational purposes. Data are represented as taxonomy nodes, which represent specific levels in the taxonomy tree, the taxonomic classification system for living species in Earth. Taxonomy nodes are organized in a hierarchical way, resulting in a taxonomy tree as illustrated in Figure 2.

A specific path in this tree generates the scientific classification of a specimen. For example, Figure 2 shows the taxonomic classification of a small rodent that inhabits many regions in the southern part of South America: 1 – Animalia (Kingdom), 2 – Chordata (Phylo), 3 – Mammalia (Class), 4 – Rodentia (Order), 5 – Hystricomorpha (Suborder), 6 – Ctenomyidae (Family), 7 – *Ctenomys* (genus), *Ctenomys minutus* (species).

3.3 Database Model

Figure 3 presents a simplified E-R diagram of our data model. The entities in the model represent taxonomic nodes, specimens, their characteristics, the measures associated to characteristics, and bibliography nodes.

Each taxonomic node has a reference to a parent node to build the hierarchy. Among other attributes, a taxonomic node has the name of the taxonomic level it represents (Kingdom, Phylo, Class, Order, Suborder, Family, Genus, Species) and its scientific name. The model includes the elements of the Darwin Core standard, and other characteristics of interest of our expert users. The Darwin Core standard was adopted aiming at the future interoperability of our system and other similar projects mentioned in Section 2.

The entity *Specimen* represents a specific animal that was observed, sampled or entirely collected. It contains the identification of this "object" in the collection, the name of the collector and date of collection as well as the geographical location of the collection site. Specimens have attributes (*Characters* = characteristics) that can be defined at different taxonomic levels. This introduces the concept of inheritance of characters along the taxonomic hierarchy. Characters are categorized in specific groups: morphometric, taxonomic, phylogenetic and others. *Characters* that are physical measures are associated to a special entity that represents a *Measure* in the International Measurement System, facilitating conversions among different ways of recording data. Finally, a *Taxonomy Node* can also have a list of bibliographic references (entity *Bibliography*) which allows the storage of literature about the respective taxonomic level. Besides this, *Bibliography* entities can be associated to specimens, representing the publications that used or cited those specimens.

4 TAXONOMYBROWSER VISUAL INTERFACE

As introduced in Section 3 (see Fig. 1), TaxonomyBrowser is accessible through a web interface, which provides the user

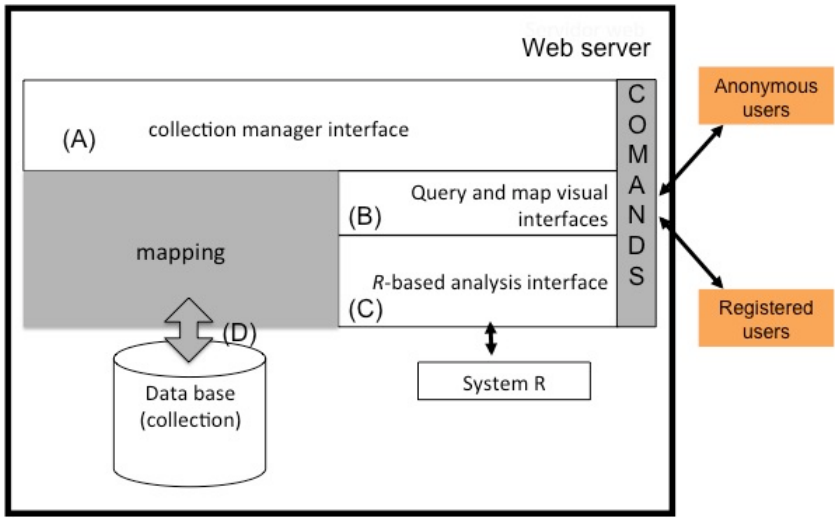

Figure 1 – TaxonomyBrowser architecture.

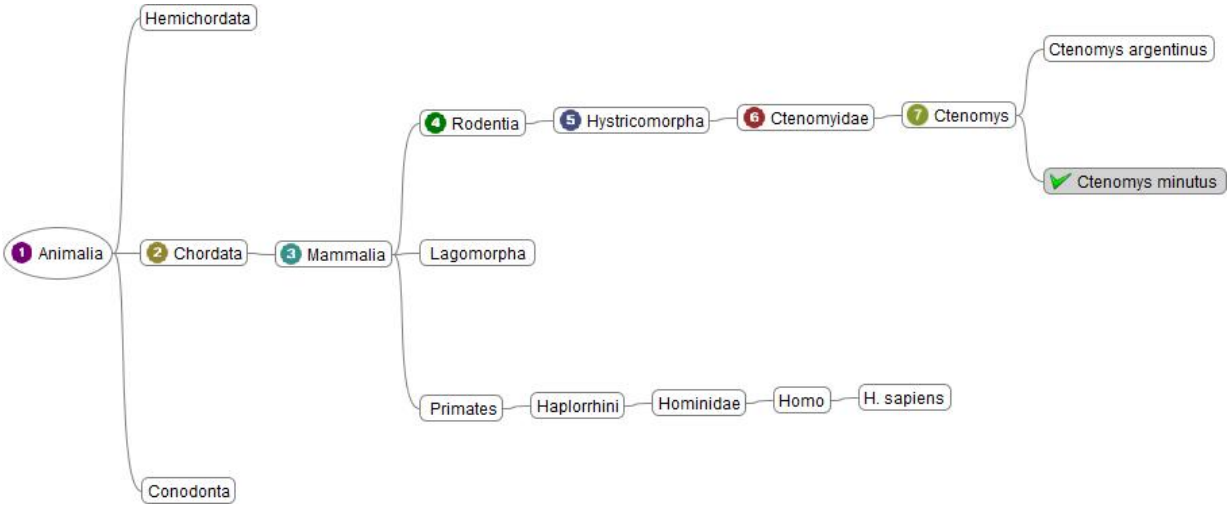

Figure 2 – Example of a taxonomy tree, part of the huge taxonomic tree that represents the classification of living species.

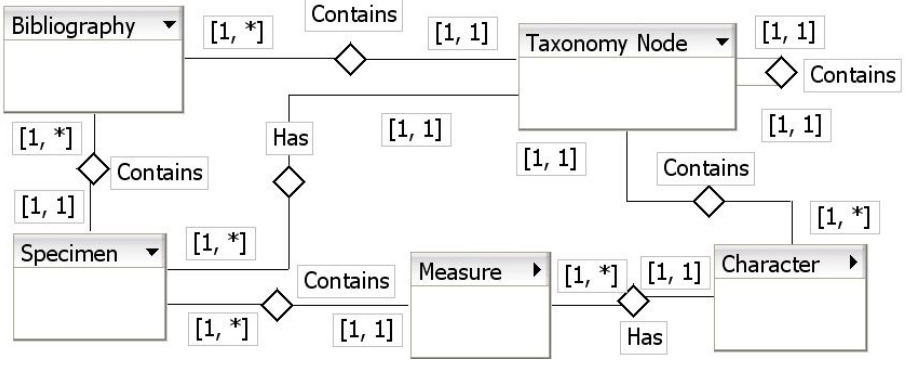

Figure 3 – Simplified Entity-Relationship diagram of the database model.

with several functions, depending on their individual access permissions, or roles. Four user categories are planned: collection manager, group member, collaborating researcher and guest (anonymous) user. As to access levels, the collection manager has full control of the collections records being allowed to include, modify and delete specimens' records, taxonomy nodes and characters descriptions. The group member is allowed to include, modify and delete specimens' records as well as scripts written in the R language for data analysis. The collaborating researcher has access to all the information in the collection including the execution of available scripts. Finally, the guest user is only allowed to query basic information about specimens which occurrence or observation is recorded in the collection.

The TaxonomyBrowser interface is divided in four main parts: the administrative interface, the query (viewing) interface, the map interface and the analysis interface. The tabs at the top of the window provide access to functions related to each element of the data model. The query, map and analysis interface are all accessed through the *analysis* tab.

The administrative interface allows the collection manager to define nodes in the taxonomic tree, their descriptive characteristics and, if necessary, measurement units associated to physical characteristics, and bibliographic data. The overall layout of the administrative interface can be observed in Figures 4 and 5. Figure 4 shows the options available in the *taxonomy* tab and Figure 5 is a snapshot of the characteristics' editing form, which is available through the *characters* tab.

In Figure 4, the current level of the taxonomic tree is shown on the left pane. The user can add a new taxonomic node, using the + sign on the right side of the top node. The user can also view, edit and remove taxonomic nodes using this interface. Figure 5 is an example of how TaxonomyBrowser allows editing information of all entities of our model. The *specimens* tab allows inserting collected specimens or samples as well as browsing through the records already stored in the database.

Since one of the main contributions of the TaxonomyBrowser is to integrate the query, the map and the analysis interfaces, they are presented in the next sections in detail. Although they are semantically divided in three interfaces, they are all accessed through the *analysis* tab.

4.1 Query Interface

The query interface allows the user to specify queries by building search expressions using specimens' characteristics (*characters*) and operators like $<$, $>$, $\leq$, $\geq$, $\equiv$, $\simeq$, *between* and *like* related to values. The operator *like* in an expression *rank name like %con* is used to retrieve all specimens with rank name ending with "con". Figure 6 shows an example where the query is built to retrieve all specimens of a certain species. The specification of a query can be saved for later use either for redisplaying the same results or modifying it. The left pane in the interface shown in Figure 6 contains the list of search specifications already stored in the database: "Lami grafico" is the name of the current search. Figure 7 shows a map with icons indicating the collection locations of specimens selected using a query built to retrieve specimens from the same species ("lami") but with character *comprimento da cauda* less than 12 (see the topmost part of the image). One can choose the color of the icons to allow displaying the results from more than one query in a single map. The interface can also represent the region covered by the results by connecting the locations of all specimens that comply with the query, forming the smallest convex polygon representing the search (not shown in this image).

4.2 Map Interface

The map interface is built using the Google Maps API. It allows the user to interact with the data shown as icons on a map. Using the basic functions from Google Maps, one can obtain detailed information about a selected specimen. The user can use special tools to draw lines and polygons in order to select a group of specimens for further processing. For example, the polygons depicted in the map shown in Figure 8 allows the user to submit that group of specimens as input to a data analysis process implemented as a script written in R. Notice the small operation icons at the bottom of the map providing such geometrical form of selection.

4.3 Analysis Interface

The analysis interface was firstly designed to provide access to scripts written in R and to the System R itself. Users with certain privileges can upload scripts to be used in data analysis procedures. The scripts should refer directly to characters names as defined in the database. Therefore, such interface is mainly for expert users, registered as group member or collaborating researcher. Data to be fed to scripts can be selected from the database in two ways. The most direct way is to specify a query (see Section 4.1) and submit the results to a script chosen from the list of the scripts available in the database. Figure 9 shows a

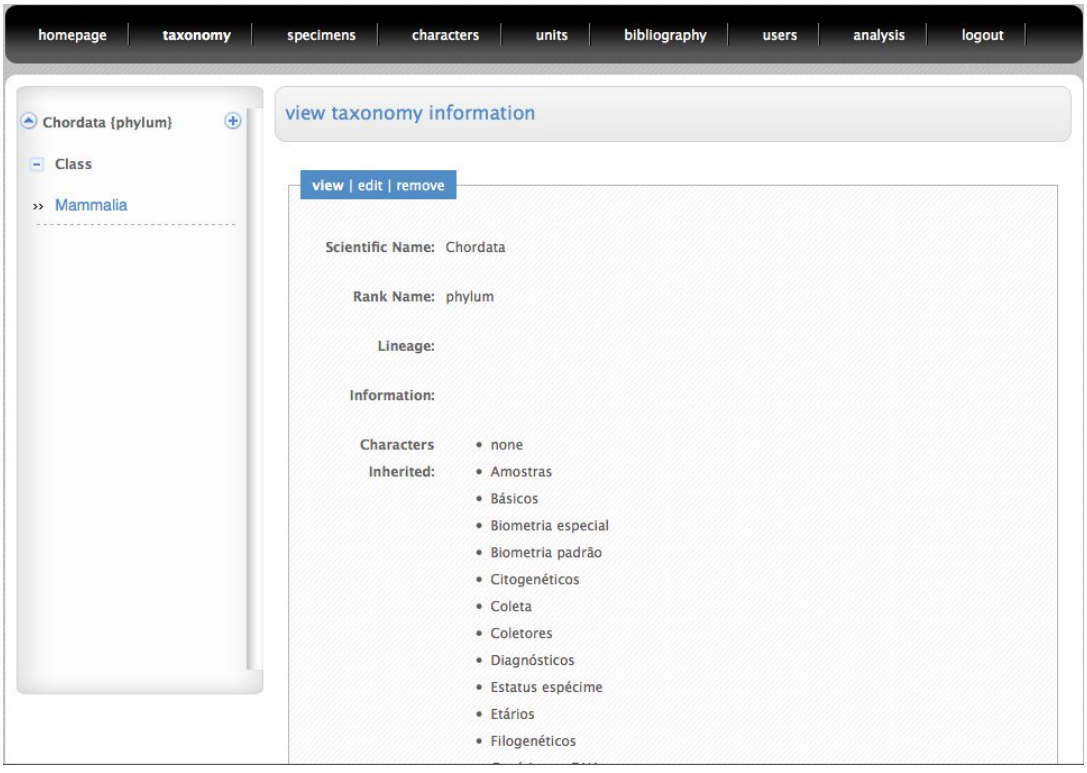

Figure 4 – Operations for managing the taxonomy tree.

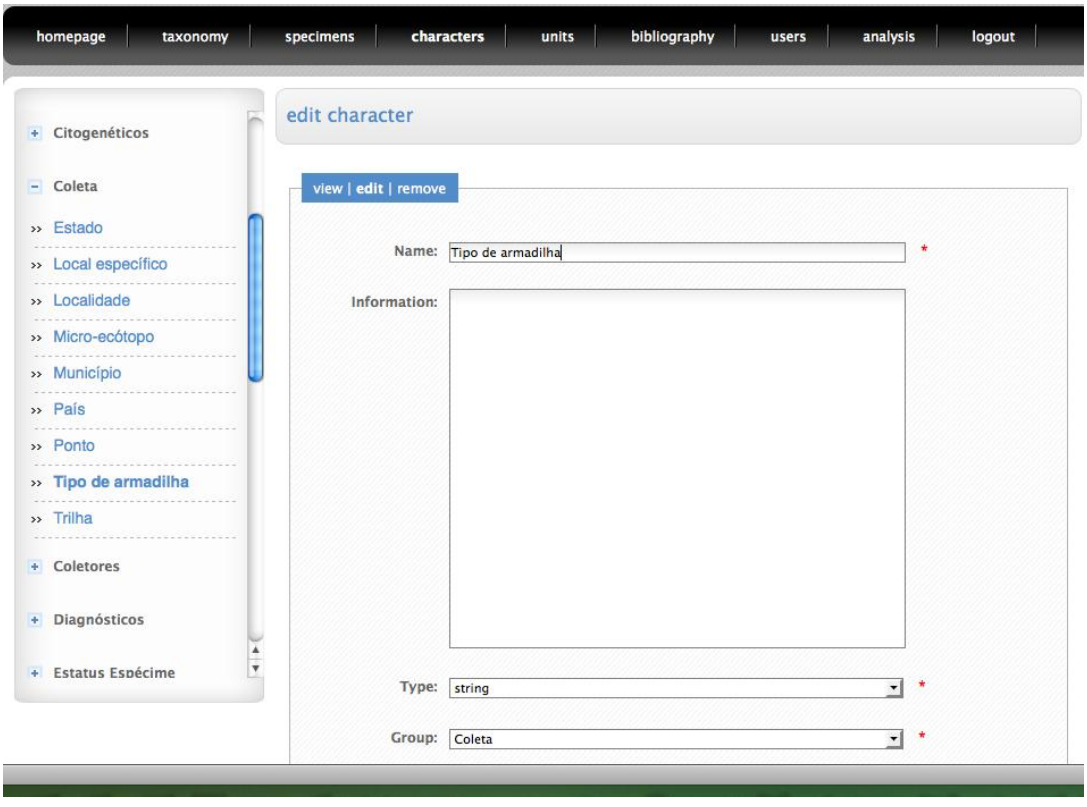

Figure 5 – Editing a characteristic using the *character* tab.

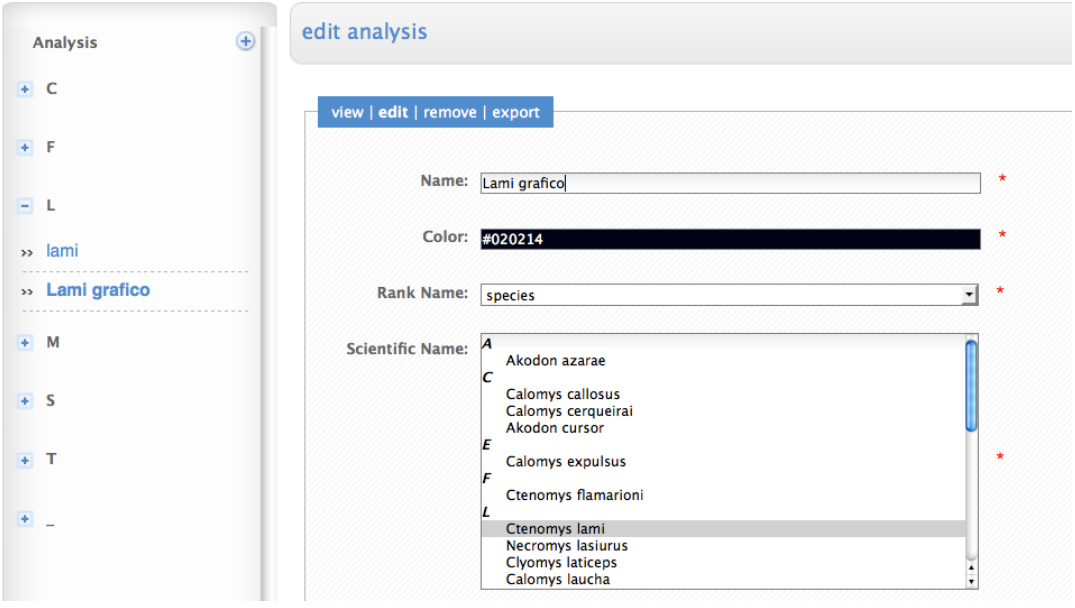

Figure 6 – Specification of a query that retrieves all the specimens of a given species.

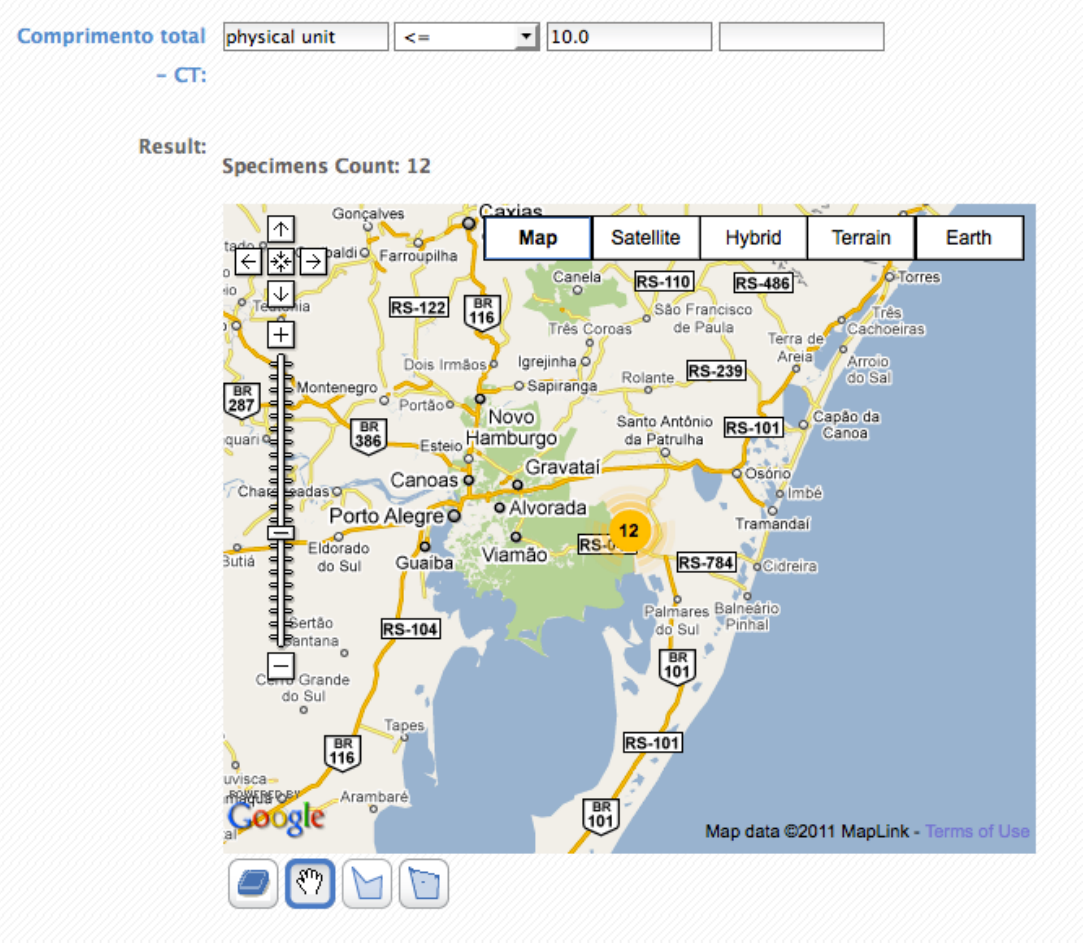

Figure 7 – Results of a query shown in a map of the collection region.

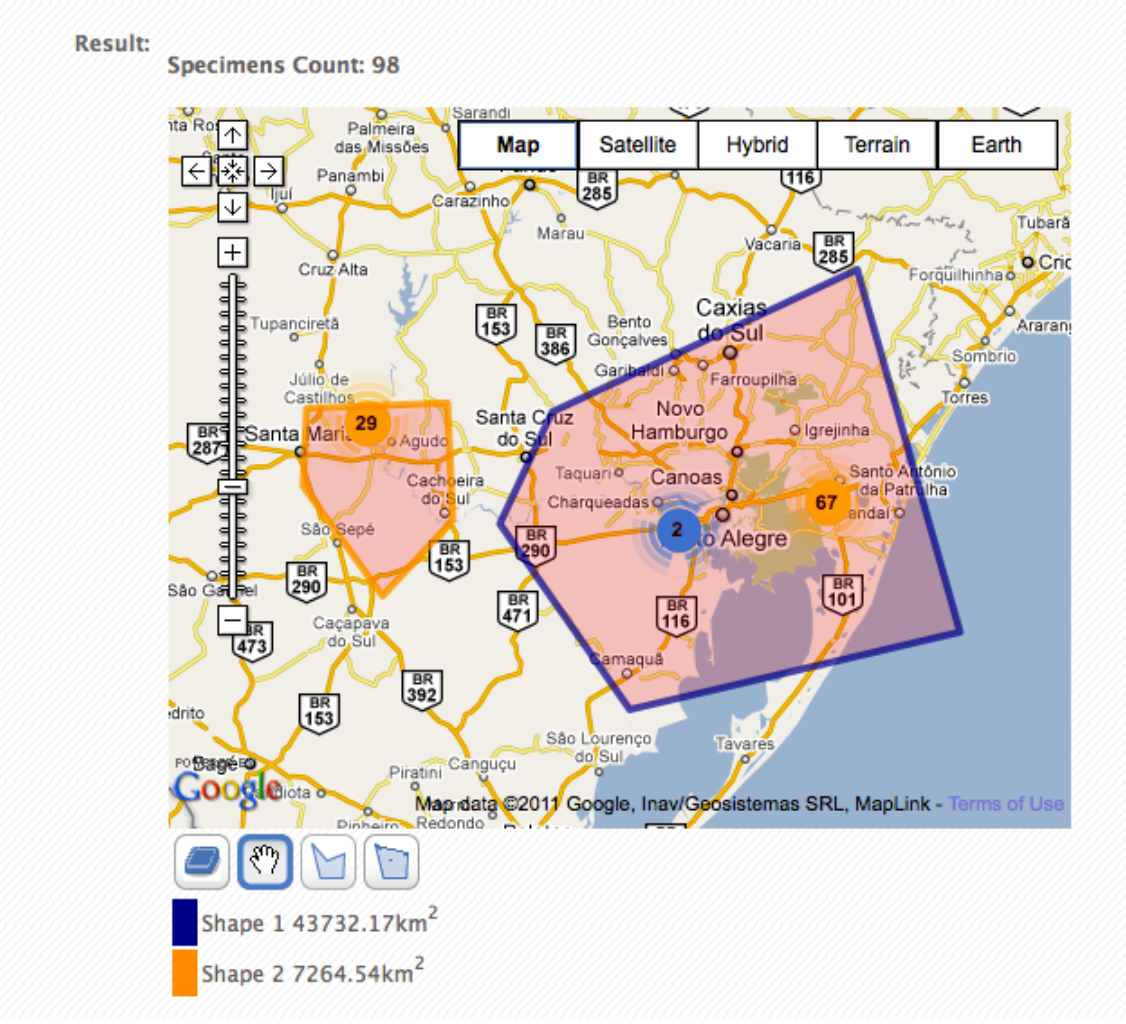

Figure 8 – Map interface with geometric tools for specifying entities to be submitted to analysis.

query result being used as input to a script selected from the list. To be used in a script the result of a query should be exported as a .csv file, which is sent to system R for processing.

Besides the analysis of query results with scripts, one can use them with visualization techniques implemented with the Google Charts API. Actually, the biologists would first visualize their (raw) data with some technique, and then decide which analysis script to run, for checking correlation between characteristics, for example.

Figures 10 and 11 are examples of such visualizations. From the interface that allows exporting data to be used as input by a script, instead of selecting a script, the user can choose a visualization technique (Fig. 10). Figure 11 shows the resulting images where the values of some (chosen) characteristics of the selected specimens are displayed as parallel coordinates and bar chart.

5 CONCLUSIONS

We showed the main components of TaxonomyBrowser, a web application aimed at supporting both biologists' research projects and communication of information concerning to animal species occurrence to the general public. TaxonomyBrowser was developed on top of a relational database, using the Google Maps/Earth API for the basic visualization interface. It also provides a series of visualization techniques based on the Google Charts API. Our application is also integrated with System R for the execution of scripts essential to the data analysis procedures the biologists perform. This allows a smooth transition between navigation, selection and retrieval of data from the database to the data analysis phase.

The integration of all these features is the main contribution of TaxonomyBrowser. The same application and database can be

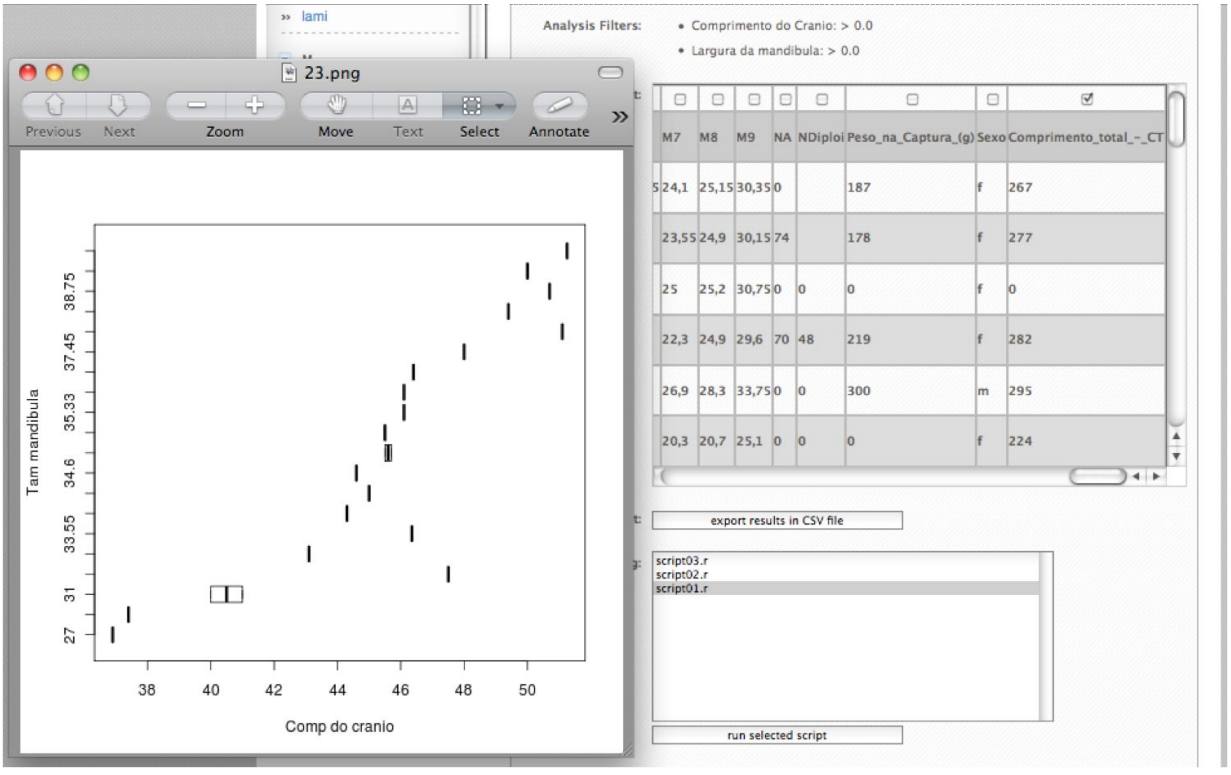

Figure 9 – Selection of a script to be executed using the data in the results table as input. Results of the execution of the R script are shown as the superimposed window.

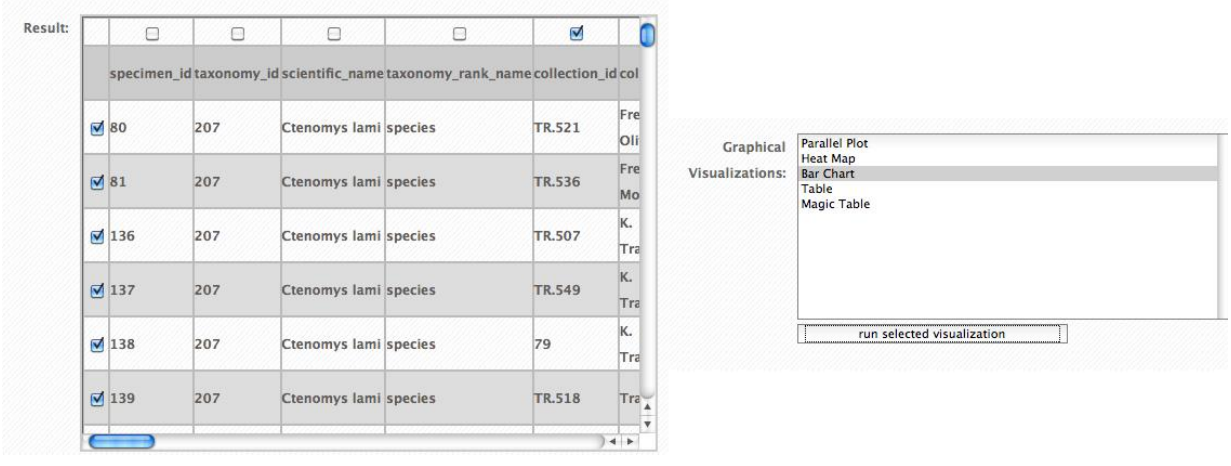

Figure 10 – Selecting specimens and characteristics from a query results table and a visualization technique (from the list) to display the selected specimens. The table shown here is the same as in Figure 9, and the list of visualization techniques is actually below the scripts list also shown in Figure 9.

used by biologists, as their scientific data management system, as well as by lay users, as an educational or information portal. Although there are many systems that address these needs separately, as described in Section 2, as far as we know there is no system described in the literature, which integrates web-based scientific data management with analysis and visualization tools.

Future works include new geometric tools for selecting specimens in separate regions on the map, for analyzing them with the same script, for example, and integrating the results in a

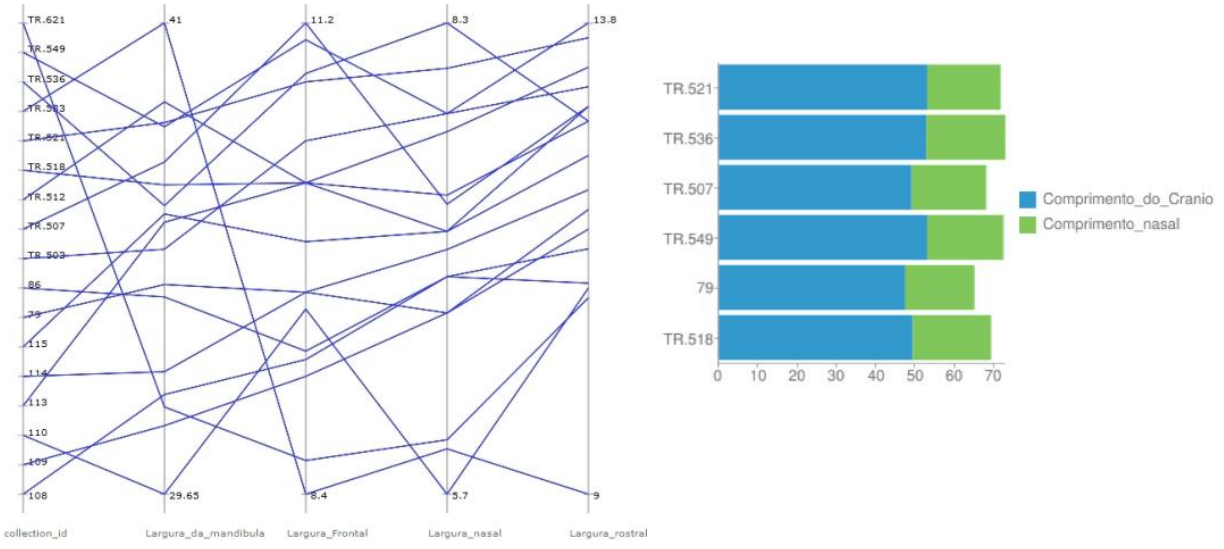

Figure 11 – Displaying characteristics of selected specimens as parallel coordinates and bar charts. For the parallel coordinates chart, several specimens and three characters were chosen in the table of Figure 10, while for the bar chart only two characters and few specimens were selected.

comparative image. We also want to provide more powerful tools to build scripts using specimens' data to improve the analytical power of our system.

ACKNOWLEDGMENTS

This work was funded by CNPq CT-INFO 2007 grant. We also thank Jeronimo Silva and Fabiano Fernandes for their collaboration in a previous project that served as proof-of-concept for the development of TaxonomyBrowser.

REFERENCES

[1] FLEMONS P, GURALNICK RP, KRIEGER J, RANIPETA A & NEUFELD D. 2007. A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA), Ecological Informatics, 2(1): 49–60.

[2] SOBERÓN J & PETERSON T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions: Biological Sciences, 359(1444): 689–698.

[3] SPECIES2000. http://www.sp2000.org (2010).

[4] TDWG. Taxonomic Databases Working Group. 2009. http://www.tdwg.org/standards/450/.

[5] MANIS. http://manisnet.org/ (2010).

[6] COLWELL R. 1996. Biota: The Biodiversity Database Manager.

[7] CANETE SC, TAVARES DLM, CORDEIRO-ESTRELA P, FREITAS TRO, HENKIN R, GALANTE R & FREITAS CMDS. 2010. Integrando visualização e análise de dados em sistema de gerenciamento de dados de biodiversidade. In: IV e-Science Workshop – Proceedings of the 2010 Congress of the Brazilian Computer Science, SBC, Belo Horizonte, Brazil.

[8] BIOTA. http://www.biota.org.br/ (2010).

[9] http://www.lis.ic.unicamp.br/projects. 2010.

[10] MALAVERRI JG, VILAR BSCM & MEDEIROS CB. 2009. A Tool based on Web Services to Query Biodiversity Information. In: WEBIST 2009 – Proceedings of the Fifth International Conference on Web Information Systems and Technologies, INSTICC Press, Lisbon, Portugal, pp. 305–310.

[11] Global Biodiversity Information Facility. http://www.gbif.gov/ (2010).

[12] Integrated Taxonomic Information System. http://www.itis.gov/ (2010).

[13] MADDISON D & SCHULZ K. 2007. The tree of life web project. http://tolweb.org/tree/.

[14] ARCTOS. http://arctos.database.museum/home.cfm (2010).