Using Visualization Techniques for Supporting Interaction with the TaxonomyBrowser Database

Marina Rey
Instituto de Informática - UFRGS
Porto Alegre, Brazil
mfrey@inf.ufrgs.br

Carla Maria Dal Sasso Freitas
Instituto de Informática - UFRGS
Porto Alegre, Brazil
carla@inf.ufrgs.br

***Abstract*—TaxonomyBrowser is a biodiversity information system developed to manage data collected by biologists during field work as well as data resulting from further analyses. It was developed as a web application and stores data on a database organized according to the taxonomic tree of leaving species, i.e., specimens collected or observed during field work are recorded in the database classified as pertaining to some level in the taxonomic tree. In its first version, TaxonomyBrowser had a common, form-based user interface for managing the database. This work presents the new visualization-based interface of the system, describing the features that are available for entering data from collected specimens as well as accessing data through different visualization techniques. The interface was assessed by means of a survey with remote users in order to evaluate the provided features and validate the concept of a visualization-based interface for such kind of system. Participants considered the application to be intuitive, and most of them provided positive feedback.**

***Index Terms*—data visualization, human-computer interaction, graphical user interfaces, biodiversity**

I. Introduction

The process of collecting biological data is a manual and systematic work, which demands that the researcher collects basic data from the specimen on the field, catalogues and stores them for further analysis. Although usually such data are recorded in separate files, there are adequate systems for that, the so-called biodiversity information systems. These data records can be later modified when new laboratory studies are performed on the collected samples. So, it is important to have an unified biodiversity information system capable of providing easy access to previously stored specimens' data either for retrieving or updating the related information.

As observed in previous studies [1], [2], most solutions for biodiversity information systems have unnecessarily complex user interfaces. Such systems usually do not provide any overview of the stored data or any simple method of browsing their entries, only showing a list of all specimens recorded on the database. There are applications that focus on the visualization of taxonomic and phylogenetic trees especially built for biological data analysis, yet they normally do not comprise own databases for storing the information, requiring that users import their data on every use.

In a previous research project, a team composed of computer scientists and biologists developed TaxonomyBrowser, a biodiversity data management system with the purpose of maintaining data about mammals collected by biologists all over the state [2]–[6]. Like other systems, that system does not provide an overview of all data stored in the database nor an easy way of accessing a specific specimen. Moreover, all specimens and specimens' information are displayed as lists.

The present work describes the new interface for the TaxonomyBrowser, which uses information visualization techniques as the primary means for user interaction. We based our work on the display of the taxonomic hierarchy to provide an overview of the specimens recorded in the database and allowing the interaction with the data in a clear and intuitive way. In doing so, we aim at reducing the number (and complexity) of interactions required from the user to perform certain actions on the database, and ultimately improving user tasks.

In a previous paper [7], we described the main aspects of the visualization-based interface for the TaxonomyBrowser and reported the results from a remote survey with users with different levels of knowledge about the application domain. The results showed the tool was well understood and praised by most of the participants.

Herein, we extend that work by providing:

- detailed description of new features that were added to the interface,
- expanded presentation of features that allow adding and modifying specimens' records,
- improved comparison with related work, including other applications not presented before, and
- thorough discussion of results obtained from the remote survey.

The remainder of this paper is organized as follows. First, we present the application domain concepts that are important for understanding our work (Sec. II), and briefly describe applications and on-line portals related to TaxonomyBrowser at a certain extent (Sec. III). Then, in Section IV, we give an overview of the system, summarizing its data model, architecture and main features. Section V gives details about the visualization-based interface, including the new features, while section VI presents the results from user tests we conducted for assessing the new version of the TaxonomyBrowser interface. Finally, Section VII concludes the paper with a discussion about open issues and future work.

II. Background

Some concepts are needed for making easier the understanding of the terminology we use in this work. Such concepts include the notion of taxonomy and phylogenetic trees, and the definition and usage of a biodiversity information system.

A. Taxonomy Tree

Etymologically, the word taxonomy is derived from Greek *taxis*, meaning 'arrangement or division', and *nomos*, meaning "law". According to Enghoff [8], taxonomy can thus be understood as "laws of arrangement and division". Such taxonomies are composed of taxonomic units known as *taxa* (singular: *taxon*), frequently arranged in a hierarchical structure and related to one another by "parent-child" relationships, constituting a taxonomic tree.

The taxa for living beings are distributed in a Linnaean classification, where groupings receive a rank, such as Kingdom, Phylum, Class, Order, Family, Genus, and Species (in decreasing order of inclusiveness) [9]. The taxon's attributes are inherited by its children nodes, which in turn can have other attributes that are inherited by their children, and so on. Taxonomy also consists of the interpretation of names and the way we believe that the taxa are phylogenetically related to each other, being able to evolve as taxa are discovered or altered.

B. Phylogenetic Tree

Formally a phylogenetic tree is a construction that attempts to form the ancestors and descendants relationships for a set of entities [10]. They represent a clear notion of evolution from ancestors to current-day entities. An important characteristic of phylogenetic trees is that the descendants (leaf nodes of the tree) represent present-day entities, while common ancestors represent parents that existed in the past. For this reason, internal nodes are rarely changed, while leaf nodes vary more constantly.

C. Biodiversity Information Systems

Biodiversity Informatics includes the application of information technologies to the management, exploration, analysis and interpretation of primary data regarding life, particularly at the species level of organization [11].

Biodiversity information systems are built around a database that stores taxonomic information from a particular area or group of living organisms, mainly storing individual specimen's and species' information. The collection of these materials is performed during field works, when information about the captured specimens is usually written down in a conventional (paper) notebook. Samples from collected specimens are often stored physically, having a description of their location, for example, which box in which room it is kept, and attributes saved in the database. These samples can be tissue, blood, bones, DNA, organs and even the entire fluid-preserved or taxidermied animal body.

III. Related Work

There is a considerable number of applications and portals that aim at providing access to biodiversity information all over the world. Most of them provide information about specimens or species in table format, some show specimens or species plotted in maps depending on the existence of georeferenced information. Regarding the use of visualization techniques, which is our goal in this work, some applications use phylogenetic or taxonomic trees to represent information about species. In this section, we restrain ourselves to briefly summarize related work that provide some kind of visualization, at least a map, for the display of data about species or collected specimens.

The currently largest data portal regarding to biodiversity is the Global Biodiversity Information Facility (GBIF) [12], a intergovernmental organization that manages an interoperable network of biodiversity databases and tools in order to make the world's biodiversity data universally available on the web [13], [14]. This portal provides access to hundreds of millions of species occurrence records, stored in thousands of datasets made available by different institutions, which are "publishers". All data is available using an online interface, which allows users to view specimens in table format, access images related to each of them, filter entries and plot them on a map. GBIF networks has nodes that aggregates publishers per participant country, and in Brazil, SiBBR - the Brazilian Biodiversity Information Facility is the GBIF node. Thus, SiBBR is a national-wide system for biodiversity data with the purpose of expanding the knowledge on Brazilian biodiversity by providing tools to support scientific research [15]. The system currently integrates more than 300 datasets from 93 publishers, which allow access to more than 10 million records. These records can be accessed on-line using filters on the species of interest, being able to show them plotted on a map or listed as table entries [16]. Some of the publishers within SiBBR are also direct publishers integrated to GBIF.

Another large infrastructure for providing access to biodiversity data is SinBiota ("Sistema de Informação Ambiental do Biota"), which was created within the Biota project (FAPESP - Fundação do Amparo Pesquisa do Estado de São Paulo). The Biota project consolidates data obtained by researchers in their research projects, and make biodiversity information from the region of São Paulo, Brazil, readily available [1]. SinBiota is able to store and handle a substantial amount of data. The system's main feature is to provide a map populated with the location of collected specimens, showing all provided information when selecting each marker displayed. To visualize the collected specimens the user must first either select a certain area of interest or filter the specimens by a certain parameter.

Applications providing visualization techniques that we find more related to our approach are iTOL [17], Treevolution [18], PhyloJIVE [19], Dendroscope [20] and Krona [21]. The Interactive Tree of Life (iTOL) is an on-line tool for the display, annotation and management of phylogenetic trees [17]. Users can manage multiple trees and share their workspace with other researchers, all available through an on-line interface developed using Javascript and HTML5. The system has three types of tree visualizations available: standard indented tree visualization, circular, and unrooted (radial) layout. Another important feature is the possibility of creating a pruned tree by selecting each node manually from the original tree. Besides the standard taxonomic information, specific measures can be displayed for each species entered in the database. The values are displayed in a bar linked to each leaf node, thus creating a bar chart that uses the tree visualization as one of its axis.

Treevolution is a tool developed for the representation and exploration of phylogenetic trees, creating highly interactive visualizations to improve the exploration and analysis by users [18]. This tool provides a higher degree of interaction than other similar tools (e.g., iTOL), being able to expand and prune branches, trace the history, zoom and allowing search the tree by text input. Information can be represented in a bar chart, a linear dendrogram and a radial dendrogram, where families are clustered using color. Treevolution was developed using Java and, thus, must be downloaded for use.

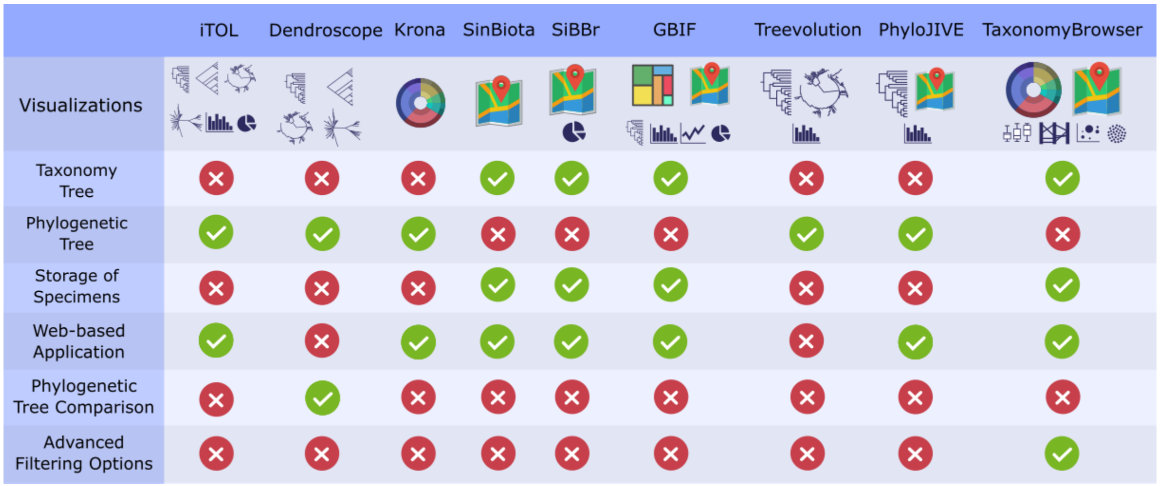
Phylogeny Javascript Information Visualiser and Explorer (PhyloJIVE) integrates biodiversity data with the Tree of Life using an open source web application where users are able to interact with the phylogenetic tree and associated data [19]. Users can also view the nomenclature, photos and other information related to each node of the tree, as well as the geographical location of up to 32 taxa plotted on a map visualization, each one color coded. The information visualized comes from data uploaded by the user or from multiple publicly available online sources.

Dendroscope is designed as an all-round tree visualization tool that can handle trees with hundred thousands of taxa [20]. The tree can be displayed in seven different ways, including: circular cladogram, radial phylogram, rectangular phylogram and slanted cladogram. The system can handle the display and correlation of multiple phylogenetic trees, including features like zooming in certain parts of the tree, reshape, re-root, reorder, extract a sub-tree or network, and even attach images to be displayed next to corresponding nodes.

Krona is a visualization tool that allows intuitive exploration of relative abundances and confidences within the complex hierarchies of metagenomic classifications [21]. It uses a radial, space filling visualization, which subdivides classes into sectors and places them depending on their biological lineage. The sectors are labeled with the scientific name of each taxon and, even though most would not fit in the space given for each partition, the system has an algorithm to increase textual information by orienting leaf node labels along the radial configuration, and internal ones along the tangent of the partition.

We compared these systems, and show a summary of their features in Fig. 1. We have found that iTOL, Krona and Dendroscope are focused on presenting an overview of the data, lacking more complex functions for analysis and management of information. SinBiota, SiBBr and GBIF, on the other hand, target mainly the storage of data, missing a tool for an overall observation of the collected specimens. They only provide a map of occurrences and data in table format. Krona, iTOL, SinBiota, SiBBr, GBIF and PhyloJIVE have very clean and modern user interfaces, allowing users to access them online. Dendroscope and Treevolution are quite the opposite, with a common interface for Windows programs and requiring to be downloaded and installed.

In our work we have focused on providing most features shown in Fig. 1 by means of a web-application that aggregates tools for managing the information, with the additional characteristic that we integrate them with visualization techniques that allow for an overview of the data as well as details when needed, and provide advanced filtering options not present in the surveyed works.



	iTOL	Dendroscope	Krona	SinBiota	SiBBr	GBIF	Treevolution	PhyloJIVE	TaxonomyBrowser
Visualizations									
Taxonomy Tree	✗	✗	✗	✓	✓	✓	✗	✗	✓
Phylogenetic Tree	✓	✓	✓	✗	✗	✗	✓	✓	✗
Storage of Specimens	✗	✗	✗	✓	✓	✓	✗	✗	✓
Web-based Application	✓	✗	✓	✓	✓	✓	✗	✓	✓
Phylogenetic Tree Comparison	✗	✓	✗	✗	✗	✗	✗	✗	✗
Advanced Filtering Options	✗	✗	✗	✗	✗	✗	✗	✗	✓

Figure 1. Comparison between iTOL, Dendroscope, Krona, SinBiota, SiBBr, GBIF, Treevolution and PhyloJIVE and TaxonomyBrowser (the solution presented in this paper) systems regarding their features.

IV. Overview of TaxonomyBrowser

As mentioned before, TaxonomyBrowser was developed in a previous research project in order to store biological data based on a taxonomy tree [2]–[6]. It comprises a database and a web-based interface. In this section we briefly describe the data model and give an overview of the (new) visualization-based interface as well as some implementation details.

A. Data Model

The database structure is based on the taxonomic tree, which ranks each node based on the Linnaean classification: Kingdom, Phylum, Class, Order, Family, Genus, and Species, in decreasing order of inclusiveness [9]. Each collected specimen is recorded in the database as children of its species node, which is children of a higher rank and so on, forming the taxonomic hierarchy. Depending on its taxonomic classification, different characteristics can be recorded for each specimen. The values for the characteristics can be either numerical measurements or textual attributes. If a taxon has a certain characteristic, all its descendants (other taxon or specimens) will inevitably inherit it.

The design of the database was not within the scope of this work since it was already developed in the previous version of TaxonomyBrowser. However, some modifications had to be made in order to add new features, such as the user authentication needed to access specimens' data.

B. Visualization-based Interface

In a seminal paper, Shneiderman [22] describes seven tasks for information visualization (overview, zoom, filter, details-on-demand, relate, history and extract) and seven data types (1-, 2-, 3-dimensional data, temporal and multidimensional data, and tree and network data). He also defines the "Visual Information Seeking Mantra" as the basic principle of visual design: "overview first, zoom and filter, then details on demand".

These principles were followed in the design of the new interface (Fig. 2), which is based on a Sunburst [23] visualization that allows for representing and managing the taxonomic tree. As each species is selected in the Sunburst visualization, the specimens classified as belonging to that species are exhibited as small circles in the center of the layout, and four types of visualizations can be created with the specimens' measures. Following the Visual Information Seeking Mantra, the Sunburst visualization represents the overview of the data, while the specimens' visualization serve as zooming and filtering, and finally, the parallel coordinates, scatterplot, box plot and geospatial visualizations present details on demand.

C. Application Architecture and Dataflow

TaxonomyBrowser is a web-based application for providing wide availability and readiness of use. The main features of the application were implemented using JavaScript, while the connection to the MySQL database was developed using PHP. Widely known libraries such as jQuery and Bootstrap were used, mainly for developing the graphical user interface. The main tool used for implementing all the interactive visualizations is Data-Driven Documents (D3) [24], a JavaScript library created for manipulating documents that held the data to be displayed. Other libraries were used for specific parts of the interface, such as the Intro.js library [25], only needed for the tours and hints, and the Toolbar.js plugin [26], used only for the tooltips provided within the Sunburst visualization.

In order to display information from recorded specimens, data is obtained from the database, and then encoded in JSON, a lightweight data-interchange format that is syntactically identical to JavaScript objects. Due to this, data in JSON can be imported to a JavaScript module with standard JS functions, which leads to a better performance compared to other approaches.

When the page loads, the data acquired from the database is displayed as the (main) Sunburst visualization, where the user is able to freely interact with all taxa currently recorded in the database. This provides the overview of the whole dataset. From the Sunburst visualization it is possible to select species for further inspection. All specimens from the selected species are added to a new list of objects and displayed in the Specimen's View, in the central area of the interface. Moreover, the specimens from the selected species can be filtered by their characteristics and visibility. Visibility means that a specimen can be available publicly, or can be private to a certain user or visible only to members of a certain research group. The specimens contained in this new list will be used for the visualizations shown in the right area of the interface. Each of these features will be further explained in the next section.

V. The TaxonomyBrowser Interface

From the complete overview of the database to the details of each specimen, the visualizations presented are a fundamental part of this work. The interactive features provided within the visualizations allow a user to fully manage all database records as well as to obtain views for analysis purposes.

A. Visualizing the taxonomic tree with Sunburst

Due to the organization of the provided data, a hierarchical visualization capable of displaying a full overview of the database was required. Even though the standard tree visualization would be the safest choice since it is widely used for viewing such datasets, this approach would not enable a complete overview of all data in a constrained area. Radial, Space Filling (RSF) techniques for hierarchy visualization have several advantages over traditional node–link diagrams, including the ability to use the display space efficiently while effectively representing the hierarchy structure [27] and allowing to analyze and alter in detail a variety of regions simultaneously without loosing the overview of the dataset. The RSF Sunburst uses a radial configuration where the inner circle represents the root of the hierarchy and deeper levels are layered around this central node. We chose Sunburst also because an analysis reported by Stasko et al. [28] suggested that participants strongly preferred this tool, citing better

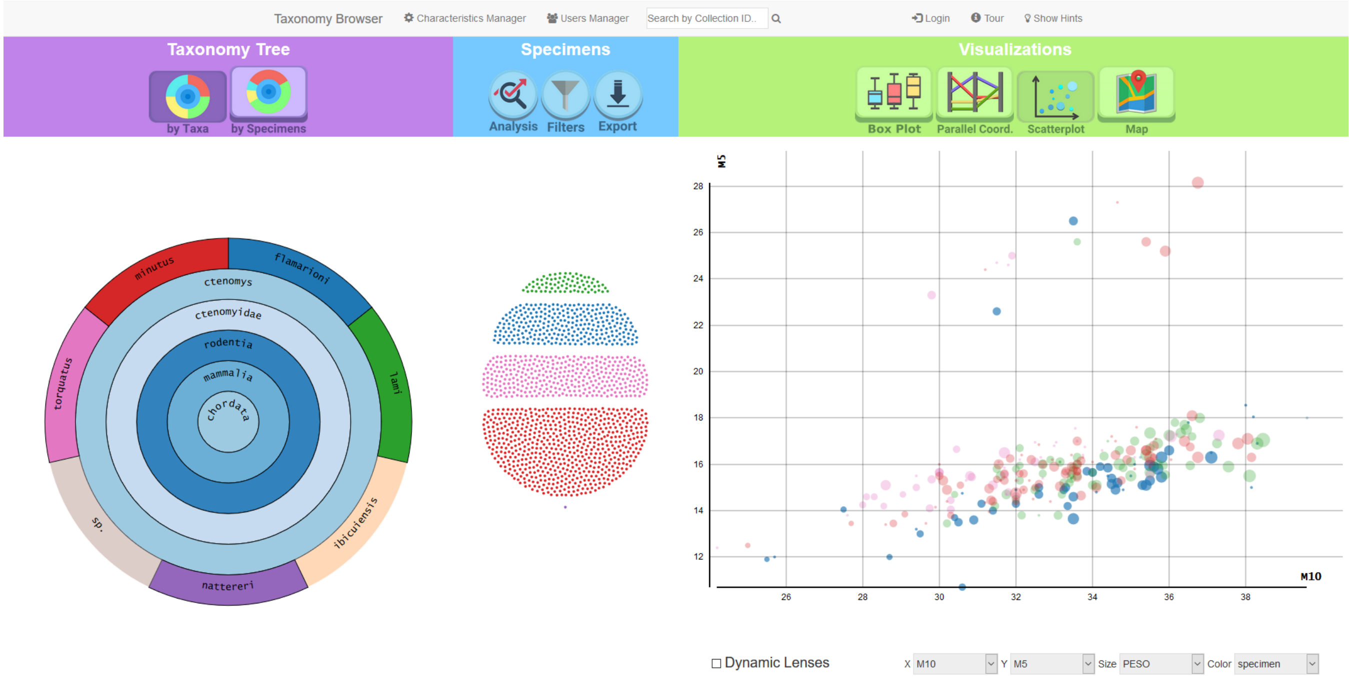


Figure 2. The interface developed for the TaxonomyBrowser. Left: Sunburst visualization showing all species represented in the database starting from the highest level (*Chordata*, in this case), with species *minutus*, *torquatus*, *flamarioni*, *lami* and *nattereri* selected. Center: all specimens selected in the Sunburst visualization, with or without further filtering. Right: example of a visualization available for displaying data about selected specimens. Top: options for the taxonomic tree visualization, specimens manipulation, and data attributes visualizations (box plots, parallel coordinates, scatterplot, and map). Above those buttons there is a header with options for managing the characteristics and users, a search bar for finding specimens by their collection ID, a log-in button and the tour of all the features provided by the tool.

ability to convey structure and hierarchy. They also were more successful and presented a faster performance in the tasks they were asked to complete.

The Sunburst visualization implemented in TaxonomyBrowser allows the following interactions:

- *Hover*: Hovering on each node triggers a toolbar-styled tooltip, with the node's name at the top and icons representing options that allow the user to show information, add children nodes and edit the hovered node (Fig. 3A).
- *Change Partition Size*: By clicking the buttons on top of the Sunburst visualization the user can define if the partitions' sizes are defined by the number of species of each taxon or by the number of specimens recorded (in the database) for each taxon (Fig. 3A and 3B).
- *Zoom*: When right-clicking a node, it becomes the main node displayed in the Sunburst view, showing only its children and therefore presenting them in more detail (Fig. 3C).
- *Select*: In order to select specimens for further inspection or visualization, the user must use the left mouse button on the desired node. If a node of higher rank is selected, all of its children will also be selected. The result will be shown to the right of the Sunburst as small circles, each one representing one specimen, as can be seen in Fig. 2 (center) and Fig. 4.

The current database still does not contain many taxa, with only 7 species of the same taxonomic group. Fig. 5 demonstrates how the visualization would look like with many taxa stored. It also shows how color is distributed in the inner nodes of the hierarchy: if a taxon has more than one child, a new color is assigned to it in order to simplify inspection.

B. *Visualizing specimens as particles in a force-based layout*

This view was designed for showing all specimens selected in the Sunburst visualization. When a filter is applied, only the specimens that fit all filters' criteria are displayed. The visualization is based on a force layout, circles of the same color attracting each other and forming groups (as shown in Fig. 4). Each circle represents a single specimen, with its color representing its species (the same color code used in the Sunburst partition for representing that taxon).

The user can click on each of them in order to display, edit or delete all the specimen's information. Hovering over the view also shows the number of specimens from that species currently selected and the species' name. Depending on the number of specimens selected, the size of each circle and their proximity with each other is altered. This visualization is important to give the user a notion of the number of selected elements, particularly for checking how many remain selected after applying a filter.

C. *Visualizing specimens' attributes using Parallel Coordinates*

Parallel Coordinates is a visualization technique introduced by Inselberg and Dimsdale [29] for plotting multidimen-

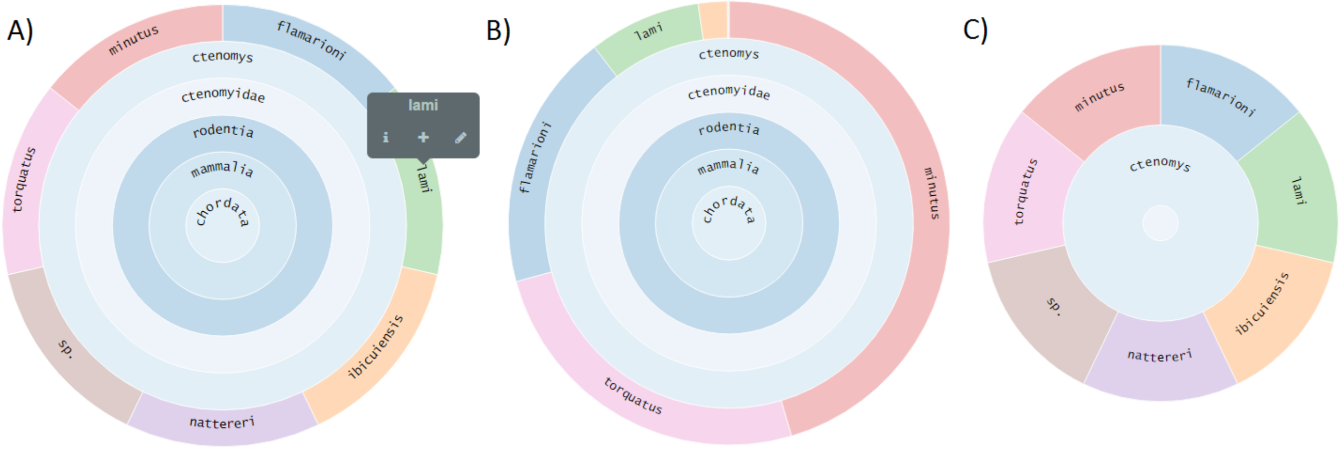


Figure 3. Examples of Sunburst visualization with the current database information: A) initial Sunburst layout, partitioned by the number of taxa and hovered on the *lami* partition, showing its tooltip; B) Sunburst layout when partitioned by specimens; C) Sunburst partitioned by taxa when zoomed on the *Ctenomys* taxon.

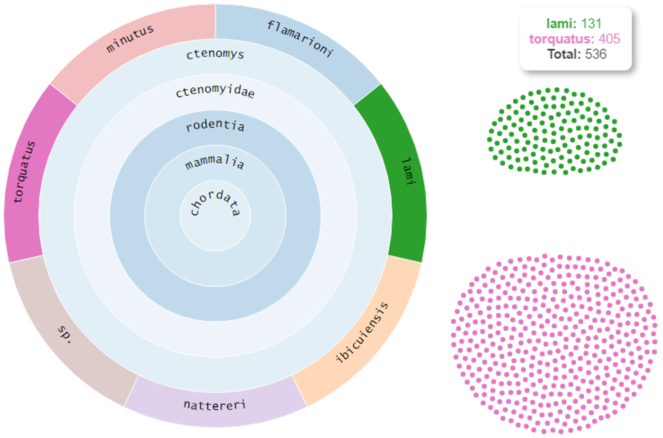


Figure 4. Sunburst visualization with the *torquatus* and *lami* species selected. Circles representing the specimens associated to these species are displayed on the right with a force-based layout. The tooltip displays the number of selected specimens when hovering over this area of the interface.

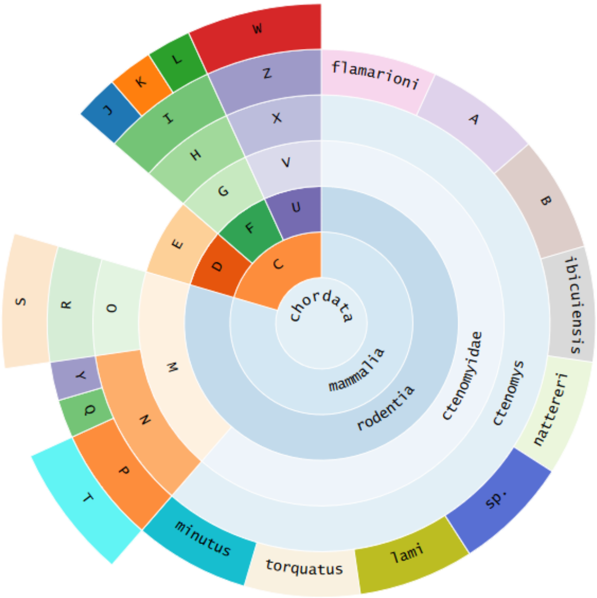


Figure 5. Sunburst visualization with taxa inserted to demonstrate how the visualization will look like when the database is properly populated. Nodes labeled 'C', 'N', 'minutus', 'lami' and 'sp' have been selected.

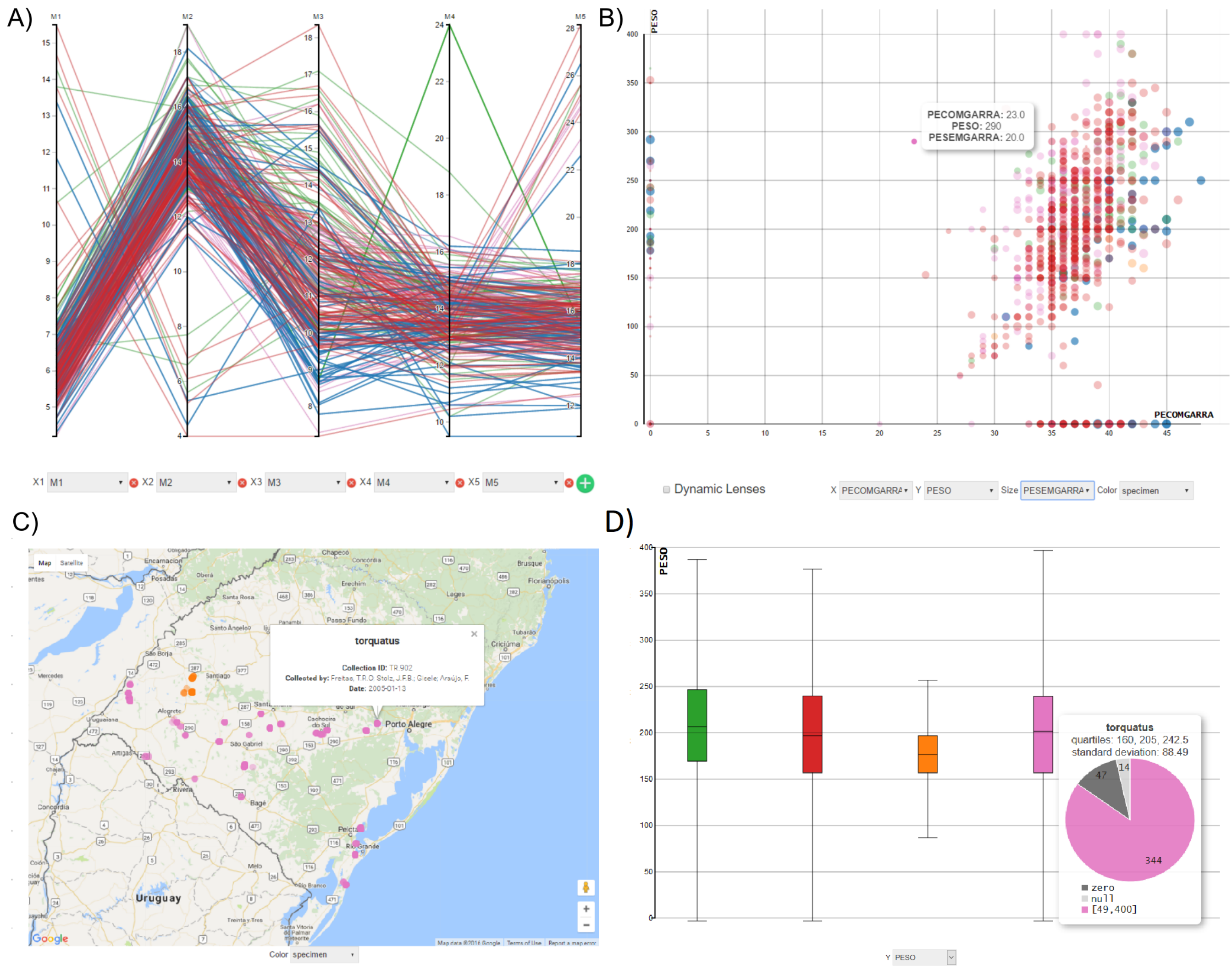
sional data. In this approach, each dimension, representing an attribute, is drawn as a vertical (or horizontal) line, and each multidimensional point is visualized as a polyline that crosses each axis at the appropriate position (depending on the corresponding attribute's value) to reflect the nD position [30].

Accordingly, in this work, each vertical axis represents a single taxon characteristic. The user can select the characteristic for each axis from the checkboxes at the bottom of the chart area. Axes can be added dynamically by the user and viewed simultaneously, as can be seen in Fig. 6A. Every line corresponds to one of the selected specimens that possess all the selected characteristics.

As with other implemented visualizations, the color represents the species of the specimen as defined on the Sunburst view and, when clicked, more information and possible actions are shown in a pop-up window.

D. Visualizing specimens' attributes using Scatterplot

According to Friendly and Denis [31], "of all the graphic forms used today, the scatterplot is arguably the most versatile, polymorphic, and generally useful invention in the history of statistical graphics". The most used scatterplot is a plot of two variables, usually indicated as x and y, measured independently to produce bi-variate pairs (xi, yi), and displayed as individual points on a coordinate grid, typically defined by Cartesian axes, where there is no necessary functional relation between x and y.



PECOMGARRA: 23.0
PESO: 290
PESEMGARRA: 20.0

torquatus
Collection ID: TR.902
Collected by: Freitas, T.R.O; Stolz, J.F.B.; Gisele; Araújo, F.
Date: 2005-01-13

torquatus
quartiles: 160, 205, 242.5
standard deviation: 88.49

Figure 6. A) Parallel coordinates visualization with vertical axes representing characteristics 'M1', 'M2', 'M3', 'M4' and 'M5'. Each polyline crossing the axes represents a specimen. B) Scatterplot visualization allows selecting characteristics for each axis and circular marker attributes. In the example, the X axis encodes the 'PECOMGARRA' characteristic, the Y axis contains the 'PESO' characteristic and the size of each circle represents the 'PESEMGARRA' characteristic. Hovering on a circle shows a tooltip with the values of the three selected characteristics. C) Map visualization using Google Maps API. Circles represent specimens and, when hovering over a circle, a tooltip with the main information regarding that specimen is displayed. D) Box plot visualization with an analysis of the characteristic 'PESO' for each species selected. When hovering a box plot additional information is shown, such as the value of each quartile, the standard deviation of the measures and a pie chart which shows an overview of how the specimens are populated.

This visualization was implemented to be used as a ready-to-use solution for quickly analyzing characteristics of selected specimens. Users can choose three characteristics to be displayed, one for each axis and another for defining the size of each marker. Each circle represents a specimen and can be hovered for displaying the exact values of the selected characteristics, as can be seen in Fig. 6B. Users can also click on a circle to show a pop-up window with all the specimen's information. From this pop-up it is also possible to edit or completely delete the specimen. The colors of each point, as it occurs in all implemented visualizations, represent the color of the specimen's species, the same used in the Sunburst visualization.

E. Visualizing the geospatial distribution of specimens

A geospatial visualization is very important for biologists in order to show the exact location of each captured specimen. For the implementation of this view we chose the Google Maps API, a widely used and simple solution for displaying maps on the web. If a specimen is selected and has longitude and latitude valid values, a point (with the same color as its species in the Sunburst view) is displayed at the corresponding location. When hovered, these points show the specimen's basic information: its collection ID, who collected it and the date of collection. When clicking on the marker, the same pop-up used in all visualizations for displaying specimen's information is shown, also allowing the user to change information and delete the specimen record from the database. This visualization is shown in Fig. 6C.

F. Comparing samples with Box Plot visualization

Box plot is a widely used graphical tool for displaying analytic information about a certain dataset. Values normally included are: the maximum and minimum values, upper and lower quartiles and median. This visualization was added to the TaxonomyBrowser interface recently, after the evaluation described in Section VI. Our first implementation displayed this information as text as a result of the "Analysis" button, but since this button will be used to integrate new analysis methods, we decided to provide such data as box plots.

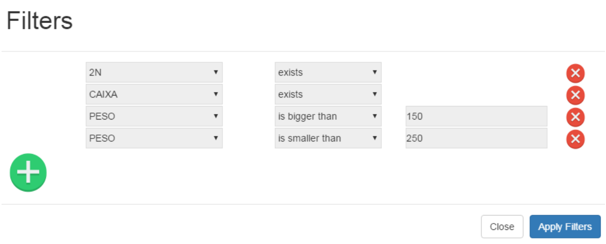
Users can choose one characteristic to be plotted. Then, for each species currently selected, a box plot will be displayed, as shown in Fig. 6(D). This box plot is generated with the data from all currently selected specimens of each species, and not the total specimens stored in the database. When the number of specimens selected is not sufficient to generate a comprehensive analysis, the box plot for that species is not created.

Our box plot visualization provides a feature that allows the user to assess missing data regarding the chosen characteristic in the set of currently selected specimens of a particular species. When hovering a box plot, a tooltip is shown in the form of a pie chart that represents the percentages of the selected specimens of the corresponding species that have the value equal to zero, null or in the range of significant values for the chosen characteristic. This tooltip also shows the precise value for each quartile and the standard deviation of the selected characteristic. This feature associated to our box plot visualization is a first attempt to deal with the problem of data quality and missing data.

The increasing number of biodiversity datasets shared by researchers and organizations and integrated into open-access platforms, like GBIF and SiBBR, raises concerns about data quality (DQ) because it impacts the results obtained from analyses and, ultimately, decisions based on such analyses [32]. So, any support (like showing missing data) for improving the management and assessment of data quality is important to assure that data analyses will provide results with an accuracy and correctness we can estimate.

G. Filtering

Users have highly varied needs for filtering features. By allowing users to control the contents of the display, they can quickly focus on their interests by eliminating unwanted items [22]. In this work, filtering is applied on specimens selected in the Sunburst visualization. When clicking on the filtering button, located above the selected specimens' visualization, a pop-up window with all filtering options is displayed (Fig. 7). A filter can be defined for a specific characteristic of the specimens or their visibility (showing only specimens that are public, private or belonging to a certain research group). There is no limit in the number of filters that can be used simultaneously.



Filters

2N	exists	
CAIXA	exists	
PESO	is bigger than	150
PESO	is smaller than	250

Figure 7. Filtering pop-up with four filters specified. The first and second lines indicate that all entries should have some value recorded for characteristics '2N' and 'CAIXA' (i.e., filter out those specimens with null or zero values for these attributes). The third and fourth lines limit the value of the attribute 'PESO' between 150 and 250.

H. Search Bar

Even though many different filters can be applied using the filtering tool, after using the interface, many users asked for a quicker way to search specific specimens only by their collection ID. This need was fulfilled by adding a small search field on the top bar of the application interface. When entered an existing collection ID, it displays a pop-up window with all the retrieved specimen's information.

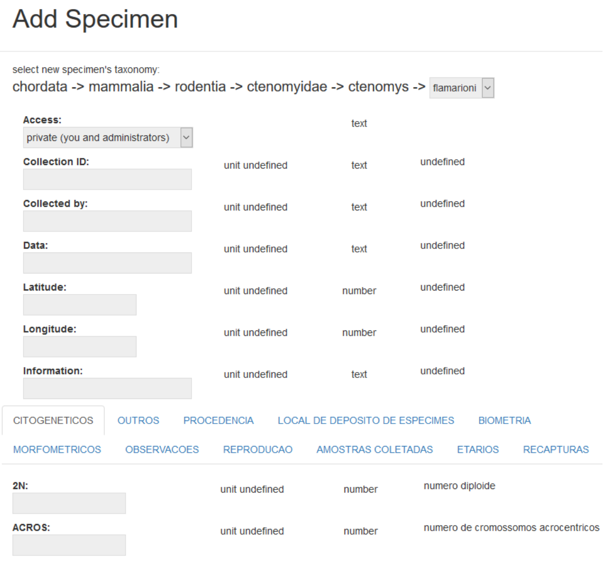
I. Updating the database

As important as retrieving data from the database is the recording of data about collected specimens. TaxonomyBrowser allows inserting data about a new specimen through the Sunburst view. Referring to Fig. 3(A), one can observe the tooltip over a specific species. By clicking on the "+" sign, the user is presented with a form (Fig. 8), where data about a collected specimen classified as belonging to that species can be entered. The form shows all the characteristics defined for the species, and the user enters the values of those attributes that were taken for the specimen being recorded. Users can also define the level of access of each entry, classifying them between private (data can only be seen by the user and the system administrators), public (anyone can access, even without a login on the website) or belonging to one of the previously defined research groups (only users that belong to the research group and administrators can access the data).

VI. Evaluation

In order to assess the visualization-based interface, a remote survey was conducted. This evaluation aimed at measuring the users' understanding of the system and their ability to perform tasks unassisted. The evaluation was performed with the version prior to the implementation of the box plot visualization and search tool, but the tasks were devised to assess the approach of using visualization techniques and not specific tools. Although results from this remote survey had already been reported in the previous paper [7], in this section we provide further insights from observations made by subjects during the evaluation.

The assessment involved 40 participants, 75% male and 25% female, with age between 19 and 58 years old. They were recruited by contacting students mainly from Biology and



Add Specimen

select new specimen's taxonomy:
chordata -> mammalia -> rodentia -> ctenomyidae -> ctenomys -> flamarioni

Access: private (you and administrators)		text	
Collection ID:	unit undefined	text	undefined
Collected by:	unit undefined	text	undefined
Data:	unit undefined	text	undefined
Latitude:	unit undefined	number	undefined
Longitude:	unit undefined	number	undefined
Information:	unit undefined	text	undefined

2N:	unit undefined	number	numero diploide
ACROS:	unit undefined	number	numero de cromossomos acrocentricos

Figure 8. Form for entering data describing a new specimen. Attributes such as the collection ID and geographical location are common for all species. Other attributes are inherited from the species description. Values for these attributes uniquely describe the specimen being added.

Computer Science courses at UFRGS (Universidade Federal do Rio Grande do Sul). 52.5% of these participants are from the field of Computer Science, 32.5% are from Biology, and 15% are from other fields, such as Engineering, Health and Social Sciences. The participants' education levels were very varied: 47.5% were undergraduate students, 22.5% were already graduated, 20% had a MSc degree, and 10% a PhD degree. All participants had experience with web-based systems, while 40% had some experience with biological information systems. 85% of the participants believed they knew what a taxonomy tree is.

The procedure for the evaluation included a participation agreement form and a small introduction to the concepts of taxonomic trees that would be necessary for the usage of the tool. Then, the participants were asked to navigate the website and use the tour provided to familiarize themselves with the application before answering the questions. The entire survey, including the tour through the website, lasted 40 minutes on average.

The first part of the survey had practical tasks to be accomplished by the subjects using the system in the form of questions to be answered. These tasks were as simple as finding which species had the largest number of specimens, selecting and deselecting species for visualization, and the name of the specimen with the higher or lower value in some measure. It was also asked how the users reached their results. Then, several sentences about the user's satisfaction with the interface features and visualizations were presented to be rated.

The final part of the survey was envisioned to measure the usability of the system by means of the System Usability Scale (SUS) [33]. This scale consists of 10 questions whose purpose is to provide a subjective assessment of the overall usability of the interface.

A. Results

Participants provided correct answers to most of the tasks/questions. As questions become more complex, the success rate tends to drop, as expected. No correlation was found between users' profile and incorrect answers. The main results from the practical section of the questionnaire are summarized in Fig. 9 and Table I in the Appendix. All users were able to find the name of the species with the largest number of specimens (T1), where 47.5% used the Sunburst's partitioning by specimens and 22.5% compared the number of circles from the Specimens' View after selecting all species. When describing how they managed to find how many specimens belong to a species (T2), 27.5% of users used the information button from the Sunburst's tooltip to check the number, 30% checked the value by hovering over the Specimens' View, and 20% used the Analysis button to obtain some basic statistics.

Then, a filtering was asked to be performed by the users, and soon after that, the total of the remaining specimens selected was asked (T4). 40% of the users hovered over the Specimens' View to check the number of specimens after the filter, and 20% used the Analysis button. A small number of participants did not understand how to answer the question correctly. Only 0.75% of users had an erroneous idea of some of the application's features. One participant thought that zooming on a taxon provided the same effect as selecting, while another participant thought the answer was the number of different colors shown in the Specimens' View. One participant did not add any filters before answering the question.

The tasks labeled T6 and T7 asked for the participants the collection ID of the lightest and heaviest specimens, respectively. 80% checked on the Scatterplot which circles were on the extremities of the graph and clicked on them. One of them used the dynamic axis to make sure he wasn't selecting wrongly.

In the final task (T9), users had to count how many specimens had the measure 'PESO' between 150 and 250 grams. 90% of the participants used the filters to answer the question. Considering all participants, 40% filtered and then opened the Analysis pop-up to check the sum of all specimens selected, and only 10% hovered over the Specimens' View to check the value after filtering. Only 0.75% tried to count manually the number of circles on the Specimens' View, all failing to reach the correct answer. The remainder of participants did not provide a complete answer, only specifying that they used two filters and citing their parameters. Tasks T3, T5, T8 and T10 asked the users to explain how they reached the answers in the respective preceding tasks.

The results from questions about the user satisfaction with the interface design decisions (Section 3 of the questionnaire) are summarized in Fig. 10 and Table II in the Appendix. In general, users liked and understood the proposed visualizations and the layout of the application. They particularly enjoyed the color scheme, which had 97.5% of approval by the participants. The only question with slightly inconsistent user

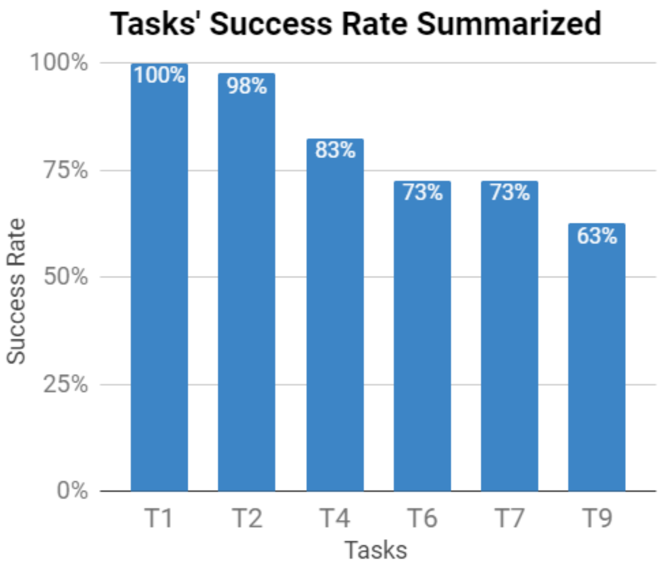


Figure 9. Summary of success rate in the practical tasks performed by remote users. Answers were asked for tasks listed in the questionnaire shown in Table I in the Appendix.

responses was the system's response time, with 72.5% of participants satisfied, 15% neutral and 12.5% non-satisfied with the performance. This can be partially related to the variety of hardware possibly used for testing the tool, since it was a remote survey.

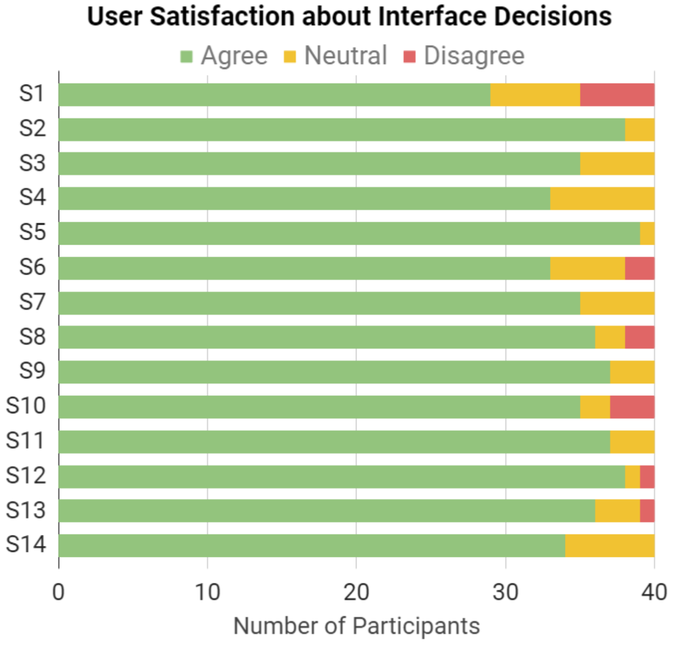


Figure 10. Summary of results from questions related to the user satisfaction about interface design decisions. The complete sentences and percentages can be observed in Table II in the Appendix.

Regarding the System Usability Scale, the average SUS score was 78.3. This value is above the average of 68 and considered an acceptable score, between good and excellent by the adjective ratings established by SUS. The results for each SUS sentence can be observed in Fig. 11. Feedback was mostly positive, specially regarding how well integrated the system's features were and its overall consistency. Also, 87.5% thought that they did not need to learn many things before they could operate with the system. The only truly controversial sentence was SUS.1: "I think that I would like to use this system frequently", with only 55% of agreement. This result can be explained since a significant number of participants were not biologists, and therefore would have no practical application for using TaxonomyBrowser. When analyzing only the answers by participants with a Biology background, 75% agreed they would like to use the system frequently, 12.5% were neutral and only 12.5% disagreed.

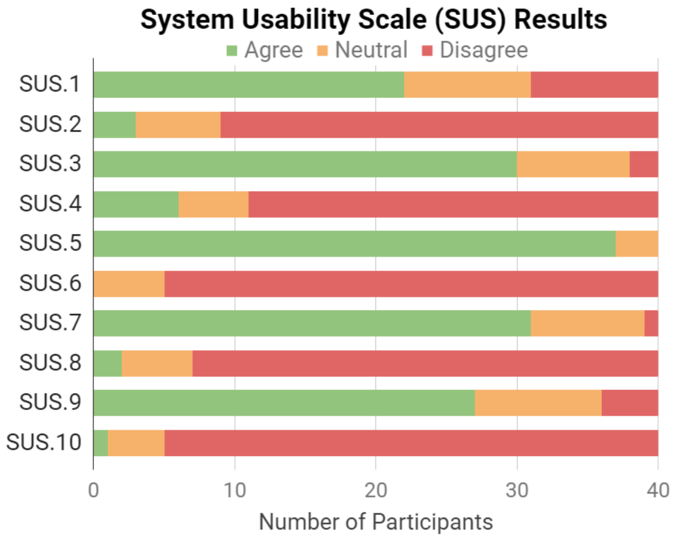


Figure 11. Summary of the answers obtained in the SUS questionnaire. In sentences 1, 3, 5, 7, and 9 it's considered a positive feedback when users agree to the statement, while sentences 2, 4, 6, 8 and 10 it's considered a positive feedback when users disagree. The complete sentences and percentages can be observed in Table III in the Appendix.

From these results we can conclude that the tool was generally visually interesting and understandable to most users, regardless of their field of study or age. Moreover, participants with practical applications for the tool were more interested in using the system frequently. The tasks results can be considered relevant, especially since it was the users' first experience with the application.

B. Discussion

In order to identify which methods users employed to perform the practical tasks, we analyzed the answers provided for questions T3, T5, T8 and T10. We assessed the explanation given to each answer and verified that a considerable amount of wrong answers were probably due to the participant not reading the question thoroughly. For example, in T6 the user is asked to select all species before answering the question. 70% of users that answered the question incorrectly skipped this step and only took into consideration the *Ctenomys lami* species, that was selected for the previous question. We made this assumption since the answers correctly described the activities required for the question, but the collection ID provided coincides with the lightest and heaviest specimen of the *lami* species.

Moreover, we could also confirm that most of the users followed the steps we thought they would follow to perform the practical tasks.

The analysis of additional comments left by 18 participants allowed us to understand how they feel about the tool, and what modifications they perceived as important. One participant remarked that the feedback from selection on the Sunburst was not clear. To help in this matter a stroke was added to the contour of each partition selected on Sunburst, as can be seen in Fig. 12.

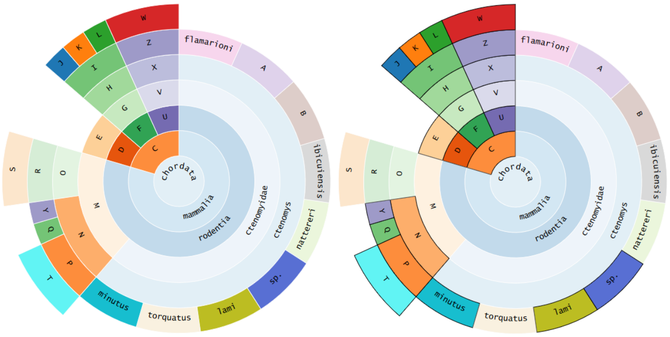


Figure 12. Previous style of visual feedback for selection (left) compared to the new proposed style on the Sunburst visualization (right).

Many users complained about performance issues when using their personal computers, particularly when selecting many specimens. This issue has been noticed and investigated, and we concluded that the number of DOM elements in the website, created for each circle displayed on the screen, was greatly slowing performance. This issue was addressed by changing the Specimens' View to use canvas instead of SVG when there is a considerable amount of specimens selected. The canvas approach, even though has a lower quality and cannot handle event interactions on each circle, is considered as a single DOM element, substantially improving the overall performance. When there is a large amount of specimens selected, interactions such as clicking for checking specific information, and dragging were considered not substantially relevant to the user; the responsiveness obtained by this alteration was deemed worthwhile.

One participant suggested adding the number of parent nodes in the central circle of the Sunburst after zooming, to give an idea of how many of them are not being displayed. To solve this issue, a breadcrumb approach was used above the Sunburst visualization, adding all parent nodes after any zoom is applied, as in Fig. 13. When clicking on each one of these labels, the user can go back to the visualization of the selected taxon, reversing the zoom more efficiently than clicking on the center of the Sunburst for each level.

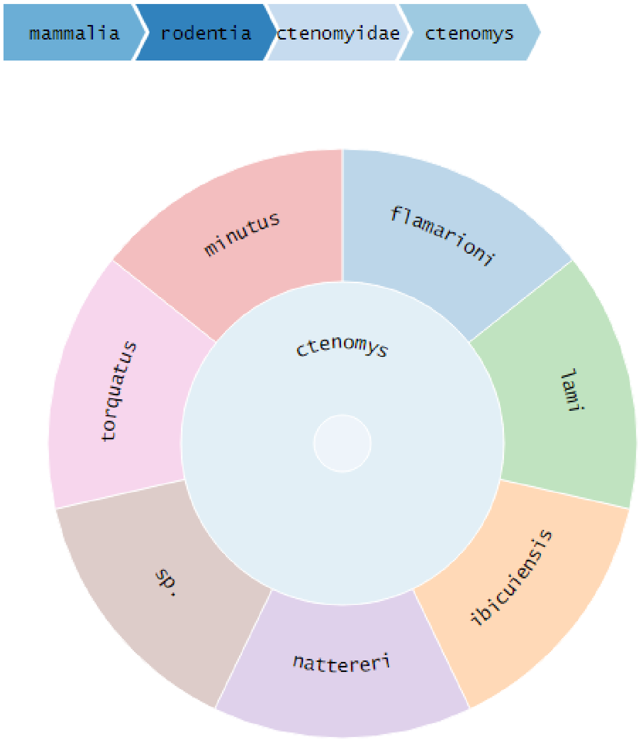


Figure 13. Breadcrumbs and the Sunburst visualization, showing all parent nodes of the zoomed-in taxa.

VII. Conclusions

This work presented a new interface for the Taxonomy-Browser based on Shneiderman's Visual Information Seeking Mantra: "overview first, zoom and filter, then details on demand". A Sunburst visualization was implemented for representing the taxonomic tree as an overview of the dataset, and for managing the database information, such as each taxon and recorded specimens. Once the user selects and/or filters specimens, the resulting data can be viewed on different visualizations, besides being displayed as small circles, for a better understanding of the size and characteristics of the set of filtered specimens. An overview of the whole system is available as a video posted at *https://youtu.be/eYmcUOPDr50*.

The interface was assessed by means of a remote survey based on specific tasks and questionnaires. The evaluation yielded promising results, specially considering that most participants were unfamiliar with the tool and had no external assistance. The answers and suggestions provided by the participants have already allowed improvements in the application, as the box plot visualization. Considering the wide range of hardware used by the participants, it was possible to receive feedback on performance issues and tweak the system accordingly.

These preliminary results demonstrate that even users without any experience with biological databases could use and obtain satisfactory results from the tool. This also indicates that the interface is overall intuitive to potential users.

In order to improve this work, more visualization techniques can be easily integrated to the application. An interesting and important future work would be a visualization for representing the quality of data, not only the missing data. When analyzing all the systems and platforms available for biodiversity data, one realizes that there is no consistent solution for describing, assessing and managing the quality of biodiversity data [32]. For sure, this is an important issue for the outcomes of research using data from these systems.

A more pragmatic future work, is to build a mobile version of the interface, in order to facilitate the recording of new specimens during field work, thus avoiding mistakes when entering data into the database from annotations made on the field.

Finally, further evaluation addressing specific features, performed with the presence of a supervisor in the room and/or including video recording, is also important as future work to allow capturing any remaining usability issues.

Acknowledgment

We thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPERGS (Fundação do Amparo à Pesquisa do Estado do Rio Grande do Sul), federal and state funding agencies, respectively, for financial support. We are grateful to the members of the team that built the first version of TaxonomyBrowser, which helped us to shape the new interface, and also to the subjects that assessed our application and provided insightful comments. Finally, we thank our lab colleagues and the anonymous reviewers for providing many suggestions that improved this paper.

References

[1] V. P. Canhos, “Sistema de Informação Distribuído para Coleções Biológicas : A Integração do Species Analyst e SinBiota,” Relatório Técnico Final do projeto Species Link, pp. 2–51, October 2005. [Online]. Available: http://splink.cria.org.br/docs/outubro2005.pdf

[2] D. Tavares, S. Canete, P. Estrela, R. Henkin, T. Freitas, R. Galante, and C. Freitas, “TaxonomyBrowser: a biodiversity data management system,” *Journal of Computational and Interdisciplinary Science*, vol. 2, pp. 37–46, 2011. [Online]. Available: http://www.inf.ufrgs.br/ dlmtavares/JCIS11-art.30.pdf

[3] J. Silva, “Projeto e Desenvolvimento de Sistema Web para Armazenamento de Coleções de Espécimes,” Porto Alegre, INF/UFRGS (BSc Dissertation), 2007.

[4] R. Henkin, “Interface de consultas analíticas para bases de dados de biodiversidade,” INF/UFRGS (BSc Dissertation), 2010.

[5] S. C. Cañete, “Interface de gerenciamento e consultas visuais em banco de dados de biodiversidade,” PPGC da UFRGS (MSc Dissertation), 2011. [Online]. Available: http://www.lume.ufrgs.br/handle/10183/29014

[6] R. Henkin, “A study on visual analysis of georeferenced haplotype networks,” PPGC da UFRGS (MSc Dissertation), 2013.

[7] M. Rey and C. Freitas, “A visualizaton-based approach for the taxonomybrowser interface,” in *Proceedings of the IHC’2017*. ACM, 2017.

[8] H. Enghoff, “What is taxonomy? - An overview with myriapodological examples,” *Soil organisms*, vol. 81, no. 3, pp. 441–451, 2009.

[9] B. Lee, C. S. Parr, D. Campbell, and B. B. Bederson, “How users interact with biodiversity information using taxontree,” in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI ’04. ACM, 2004, pp. 320–327. [Online]. Available: http://doi.acm.org/10.1145/989863.989918

[10] S. Tahir and M. T. Afzal, “A novel phylogenetic tree data visualization application for researchers,” *Proceedings of 2014 Science and Information Conference, SAI 2014*, no. August, pp. 93–99, 2014.

[11] J. Soberón and T. Peterson, “Biodiversity informatics: managing and applying primary biodiversity data,” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 359, no. 1444, pp. 689–698, 2004.

[12] GBIF, “Global biodiversity information facility,” https://www.gbif.org/, (Accessed on 29/01/2018).

[13] C. Yesson, P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A. Bisby, and A. Culham, “How global is the global biodiversity information facility?” *PLoS ONE*, vol. 2, no. 11, 2007.

[14] “GBIF Strategic and Operational Plans 20072011: From Prototype towards Full Operation,” https://www.gbif.org/document/80522/gbif-strategic-and-operational-plans-2007-2011-from-prototype-towards-full-operation, (Accessed on 29/01/2018).

[15] D. Dias, C. Baringo Fonseca, L. Correa, N. Soto, A. Portela, K. Juarez, R. J. Tumolo Neto, M. Ferro, J. Gonçalves, and J. Junior, “Repatriation Data: More than two million species occurrence records added to the Brazilian Biodiversity Information Facility Repository (SiBBr),” *Biodiversity Data Journal*, vol. 5, p. e12012, 2017. [Online]. Available: http://bdj.pensoft.net/articles.php?id=12012

[16] “Sibbr. data explorer ocurrences and species,” http://gbif.sibbr.gov.br/explorador/en/search, (Accessed on 24/01/2018).

[17] I. Letunic and P. Bork, “Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation,” *Bioinformatics*, vol. 23, no. 1, pp. 127–128, 2007.

[18] R. Santamaría and R. Therón, “Treevolution: Visual analysis of phylogenetic trees,” *Bioinformatics*, vol. 25, no. 15, pp. 1970–1971, 2009.

[19] L. Kreft, A. Botzki, F. Coppens, K. Vandepoele, and M. Van Bel, “PhyD3: A phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization,” *Bioinformatics*, vol. 33, no. 18, pp. 2946–2947, 2017.

[20] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp, “Dendroscope: An interactive viewer for large phylogenetic trees.” *BMC bioinformatics*, vol. 8, no. August 2016, p. 460, 2007.

[21] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, “Interactive metagenomic visualization in a Web browser,” *BMC Bioinformatics*, vol. 12, no. 1, p. 385, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/385

[22] B. Shneiderman, “The eyes have it: a task by data type taxonomy for information visualizations,” *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343, 1996.

[23] J. Stasko and E. Zhang, “Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations,” in *Proceedings of the IEEE Symposium on Information Vizualization 2000*, ser. INFOVIS ’00. IEEE Computer Society, 2000, pp. 57–65. [Online]. Available: http://dl.acm.org/citation.cfm?id=857190.857683

[24] M. Bostock, V. Ogievetsky, and J. Heer, “D3 data-driven documents,” *IEEE transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[25] Intro.js, “Step-by-step guide and feature introduction for your website,” http://introjs.com/, 2016, (Accessed on 18/11/2016).

[26] Toolbar.js, “A jquery plugin that creates tooltip style toolbars,” http://paulkinzett.github.io/toolbar/, 2016, (Accessed on 18/11/2016).

[27] J. Yang, M. O. Ward, E. A. Rundensteiner, and A. Patro, “InterRing: A Visual Interface for Navigating and Manipulating Hierarchies,” *Information Visualization*, vol. 2, no. 1, pp. 16–30, 2003. [Online]. Available: http://ivi.sagepub.com/cgi/content/abstract/2/1/16

[28] J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald, “An evaluation of space-filling information visualizations for depicting hierarchical structures,” *International Journal of Human-Computer Studies*, vol. 53, no. 5, pp. 663–694, 2000.

[29] A. Inselberg and B. Dimsdale, “Parallel coordinates: A tool for visualizing multi-dimensional geometry,” in *Proceedings of the 1st Conference on Visualization ’90*, ser. VIS ’90. IEEE Computer Society Press, 1990, pp. 361–378. [Online]. Available: http://dl.acm.org/citation.cfm?id=949531.949588

[30] K. T. McDonnell and K. Mueller, “Illustrative parallel coordinates,” *Computer Graphics Forum*, vol. 27, no. 3, pp. 1031–1038, 2008.

[31] M. Friendly and D. Denis, “The early origins and development of the scatterplot,” *Journal of the History of the Behavioral Sciences*, vol. 41, no. 2, pp. 103–130, 2005.

[32] A. K. Veiga, A. M. Saraiva, A. D. Chapman, P. J. Morris, C. Gendreau, D. Schigel, and T. J. Robertson, “A conceptual framework for quality

assessment and management of biodiversity data," *PLoS ONE*, vol. 12, no. 6, pp. 1–20, 2017.

[33] J. Brooke, "SUS - A quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996. [Online]. Available: http://hell.meiert.org/core/pdf/sus.pdf

APPENDIX

The following tables contain the success rate shown by subjects in performing the tasks (Table I), the sentences included in our survey and respective results (Table II), and the results for the SUS questionnaire (Table III).

Table I
SUCCESS RATE IN PERFORMING THE TASKS: RESULTS ARE THE PERCENTAGE OF CORRECT ANSWERS TO THE QUESTIONS POSED TO SUBJECTS THAT PARTICIPATED IN THE ASSESSMENT.

Task	Summarized Question	Success Rate
T1	What is the species with most specimens?	100%
T2	How many specimens this species has?	97.5%
T3	Describe briefly how you reached this answer	—
T4	Select only the species 'Ctenomys lami'. Add a new filter testing if the parameter 'Data' exists. How many specimens match the filter?	82.5%
T5	Describe briefly how you reached this answer	—
T6	Remove the filter. Select all species. Use the Scatterplot graph to view the parameter 'PESO' (weight). What is the collection ID of the lightest specimen (ignoring null and '0' values)?	72.5%
T7	What is the collection ID of the heaviest specimen?	72.5%
T8	Describe briefly how you reached this answer	—
T9	Out of all specimen registered in the database, how many have the parameter 'PESO' between 150 and 250 (including specimen with exactly 150 and 250)?	62.5%
T10	Describe briefly how you reached this answer	—

Table II
RESULTS FROM THE SURVEY MEASURING USER SATISFACTION IN RELATION TO INTERFACE DECISIONS.

ID	Sentence	Agree	Neutral	Disagree
S1	I think the tool has good response time.	72.5%	15%	12.5%
S2	I think the chosen visualizations are adequate.	95%	5%	0%
S3	I think the proposed visualizations allow a good understanding of the data base	87.5%	12.5%	0%
S4	I think the menus for each visualization are adequate	82.5%	17.5%	0%
S5	I think the color scheme is pleasant.	97.5%	2.5%	0%
S6	I liked the taxonomy tree visualization (Sunburst).	82.5%	12.5%	5%
S7	I liked the visualization of the selected specimens (circles on the center of the screen).	87.5%	12.5%	0%
S8	I liked and understood the Scatterplot visualization.	90%	5%	5%
S9	I liked and understood the map visualization.	92.5%	7.5%	0%
S10	I found the filtering method intuitive.	87.5%	5%	7.5%
S11	I found the options of visualization, search and comparison of data adequate.	92.5%	7.5%	0%
S12	I think the layout of the system is appealing.	95%	2.5%	2.5%
S13	I found the tool interesting.	90%	7.5%	2.5%
S14	I found the tool effective in its features.	85%	15%	0%

Table III
RESULTS FROM THE SUS QUESTIONNAIRE.

ID	Sentence	Agree	Neutral	Disagree
SUS.1	I think that I would like to use this system frequently.	55%	22.5%	22.5%
SUS.2	I found the system unnecessarily complex.	7.5%	15%	77.5%
SUS.3	I thought the system was easy to use.	75%	20%	5%
SUS.4	I think that I would need the support of a technical person to be able to use this system.	15%	12.5%	72.5%
SUS.5	I found the various functions in this system were well integrated.	92.5%	7.5%	0%
SUS.6	I thought there was too much inconsistency in this system.	0%	12.5%	87.5%
SUS.7	I would imagine that most people would learn to use this system very quickly.	77.5%	20%	2.5%
SUS.8	I found the system very cumbersome to use.	5%	12.5%	82.5%
SUS.9	I felt very confident using the system.	67.5%	22.5%	10%
SUS.10	I needed to learn a lot of things before I could get going with this system.	2.5%	10%	87.5%