

Integrando visualização e análise de dados em sistema de gerenciamento de dados de biodiversidade

Samantha C. Cañete¹, Denison L. M. Tavares¹, Pedro C. Estrela^{2,3}, Thales R.O. Freitas², Rafael Henkin¹, Renata Galante¹, Carla M.D.S. Freitas¹

¹ Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)

² Departamento de Genética – Universidade Federal do Rio Grande do Sul (UFRGS)

³ FIOCRUZ – Rio de Janeiro – RS – Brazil.

{sccanete, dlmtavares, galante, carla}@inf.ufrgs.br,
thales.freitas@ufrgs.br, pedroestrela@yahoo.com

Abstract. *This paper describes the main components of a biodiversity information system aimed at integrating visualization and data analysis resources and conventional data management functions provided by relational databases. The system we call TaxonomyBrowser has the goal of aiding biologists in the management of data about specimens they collect in fieldwork 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 our system.*

Resumo. *Este artigo descreve os componentes principais de um sistema de informações de biodiversidade cujo objetivo é integrar recursos de visualização e de análise de dados e funções convencionais de gerenciamento encontradas em bancos de dados. O sistema, denominado TaxonomyBrowser, visa auxiliar os biólogos no gerenciamento de dados a respeito de espécimes por eles coletados assim como exemplares já catalogados em coleções de museus. A integração com visualização de mapas e com uma ferramenta de análise de dados (API do Google Maps e o sistema R, respectivamente) ampliam a aplicabilidade do sistema.*

1. Introdução

Sistemas computacionais para auxiliar o gerenciamento de dados que descrevem espécimes coletados por biólogos fazem parte do que se convencionou chamar de sistemas de informação de biodiversidade [Guralnick e Neufeld 2005][Soberón e Peterson 2004]. Uma boa parcela desses sistemas é destinada a manter as informações sobre exemplares depositados em coleções de museus, de instituições ou grupos de pesquisa. Alguns desses sistemas estão disponíveis na web; a grande maioria, entretanto, é baseada em bancos de dados convencionais e é restrita a usuários locais.

Recentemente, entretanto, principalmente por causa da “universalização” do georeferenciamento de informações, devida em grande parte à iniciativa da Google¹, e das iniciativas de padronização de transmissão e compartilhamento de dados de

¹ A API do *Google Maps* provê suporte à exibição de mapas dentro de qualquer página de *sites* na internet.

biodiversidade, como os padrões Darwin Core e DiGir², tem-se verificado um número crescente de iniciativas direcionadas para o compartilhamento de informações da biodiversidade através da web. No que diz respeito ao Brasil, são poucas as iniciativas conhecidas, sendo o SinBiota³ a mais difundida delas.

Sem dúvida, qualquer política ou ação de conservação da biodiversidade requer dados de espécies que habitam a região alvo. Frequentemente, mais do que conhecer para decidir como e porque conservar uma determinada espécie, é a própria evolução de um conjunto de espécies que motiva o estudo sobre um conjunto de dados. O estudo pode abranger desde poucas medidas morfológicas e características anatômicas até sequências de DNA obtidas de exemplares coletados. Além disso, durante o processo de análise de dados, novas informações são geradas, as quais podem ser importante armazenar. Assim, tipicamente, usuários de tais coleções de dados: catalogam informações básicas sobre exemplares coletados, recuperam (em mapas ou textualmente) dados, processam esses dados dependendo do objetivo do estudo e geram (em laboratório ou computacionalmente) novas informações que podem ser agregadas à base de dados. Neste contexto, caracteriza-se a necessidade de integrar gerenciamento de dados, consultas em níveis variados de complexidade, funções de análise de dados e apresentação de resultados em níveis diferenciados de detalhamento em uma única plataforma computacional.

O presente trabalho descreve um sistema com tais características, desenvolvido com base nas necessidades reais de um grupo de pesquisa em evolução de vertebrados. A seção 2 relaciona trabalhos correlatos e a seção 3 descreve o modelo (e a base) de dados. A seção 4 apresenta as interfaces de consulta e análise, através de exemplos. Finalmente, a seção 5 conclui e elenca os próximos desenvolvimentos.

2. Trabalhos relacionados

Este artigo está relacionado a um projeto de desenvolvimento de um sistema web que visa tanto o apoio ao gerenciamento e análise de dados de espécimes coletados por biólogos (ou registrados em coleções de museus) como a disponibilização pública dessas informações.

O conjunto de publicações mais relacionado ao presente trabalho é aquele produzido no âmbito dos projetos Sinbiota (Sistema de Informação Ambiental do Biota), WeBios (*Web Service Multimodal Tools for Biodiversity Research, Assessment and Monitoring*)⁴ e BioCORE (*Biodiversity and Computing Research*)⁵. O objetivo do Sinbiota é integrar as informações geradas pelos pesquisadores vinculados ao Programa Biota/Fapesp e relacioná-las a uma base cartográfica digital⁶. Foi desenvolvido o Atlas/Biota, um portal⁷ onde os resultados são exibidos em um mapa do estado de São Paulo, sendo possível também visualizar em detalhes as ocorrências, que incluem referências bibliográficas e uma breve descrição de como foi feita a coleta do espécime.

² Darwin Core [TDWG 2009] é um padrão de metadados para descrição de informações a respeito de ocorrência de espécies e exemplares de museus, utilizado pelo protocolo DiGir (*Distributed Generic Information Retrieval*) para transmissão e publicação de informações armazenadas em bases distribuídas.

³ SinBiota é um sistema destinado a gerenciar registros de coletas e observações de campo no estado de São Paulo mostrando informações espaciais e temporais em mapas.

⁴ <http://www.lis.ic.unicamp.br/projects/webios>

⁵ <http://www.lis.ic.unicamp.br/projects/biocore>

⁶ <http://www.biota.org.br/>

⁷ <http://sinbiota.cria.org.br/>

O objetivo do projeto WeBios é um sistema que permite a especificação de consultas multimodais sobre fontes heterogêneas de dados de biodiversidade (por exemplo, fotos, dados geográficos, ontologias e metadados específicos de domínio). O sistema BioCORE [Malaverri, Vilar e Medeiros 2009] teve sua origem no projeto WeBios, mas além de permitir gerenciar dados de coletas da natureza, também incorpora informações do Museu de Zoologia da UNICAMP. Para permitir o acesso e a recuperação de informação de forma transparente, o banco de dados foi encapsulado em um serviço web, combinando registros de ocorrências de espécies e registros de catálogos.

Outros sistemas estão disponíveis na forma de portais, provendo acesso a conjuntos específicos de informações, mas sem permitir consultas mais complexas que permitam relacionar dados de diferentes espécimes ou espécies. O projeto Species 2000 (<http://www.sp2000.org>) fornece um catálogo de várias espécies, sendo a navegação realizada através de uma árvore taxonômica ou busca pelos nomes científicos e comuns das espécies. O GBIF (*Global Biodiversity Information Facility*) é um portal agregador de dados, no qual as consultas podem ser feitas sobre um determinado país, nível taxonômico ou nome de espécie. Após efetuar uma busca, é possível aplicar filtros, procurando por outras espécies em um determinado local, por exemplo.

Coleções de museus também estão disponíveis na Internet. O MaNIS⁸ (*Mammal Networked Information System*) importa registros que estejam no formato Darwin Core e facilita o acesso a dados de coleções através de navegadores. Os resultados das consultas podem ser visualizados na interface do Google Maps, exportados em tabelas ou no formato XML. O Arctos⁹ permite a execução de consultas considerando localização geográfica, características do espécime ou elemento taxonômico, além de salvar os resultados de consultas para que sejam acessados posteriormente.

Nosso trabalho diferencia-se pela integração de recursos de visualização e de análise de dados às funções comumente encontradas nos sistemas de biodiversidade. A integração com visualização de mapas, pela API do Google Maps, e com uma ferramenta de análise de dados (o sistema R) ampliam a aplicabilidade do sistema.

3. Modelo Proposto para o Armazenamento de Dados de Biodiversidade

Esta seção descreve o modelo de dados proposto para o sistema *TaxonomyBrowser* que permite o armazenamento e consulta georeferenciada de dados de biodiversidade. Para a especificação do modelo de dados contamos com a participação de biólogos parceiros do projeto e para a validação utilizamos dados reais fornecidos por eles.

3.1 Arquitetura do Sistema *TaxonomyBrowser*

A Figura 1 ilustra a arquitetura geral do *TaxonomyBrowser*, que pode ser separada em quatro componentes distintos. A *Interface de Gerência da Coleção (A)* permite a manutenção de todos os dados de espécimes através das quatro operações básicas (visualização, inserção, alteração e remoção). A *Interface Visual baseada no Google Maps/Earth (B)* permite consultas georeferenciadas aos dados da coleção. A *Interface de Consulta Otimizada (C)* permite a execução de consultas avançadas para análise de dados por parte dos biólogos. As consultas otimizadas são processadas através de aplicações externas (por exemplo, programas no sistema R), que são

⁸ <http://manisnet.org/>

⁹ Projeto do Museu de Zoologia de Vertebrados da Universidade de Berkeley (Califórnia, EUA)
<http://arctos.database.museum/home.cfm>

integradas no sistema *TaxonomyBrowser* de forma transparente ao usuário. A *base de dados da coleção (D)* armazena todos os dados das coleções, sendo que a camada de mapeamento faz a tradução dos dados da coleção para cada uma das interfaces de consulta. O sistema ainda permite o acesso a usuários com diferentes visões. Por exemplo, o biólogo tem acesso pleno ao sistema, podendo atualizar os dados e processar qualquer tipo de consulta. O usuário final tem acesso às consultas georeferenciadas e aos dados da coleção¹⁰, podendo acionar *scripts* de análise no sistema R¹¹.

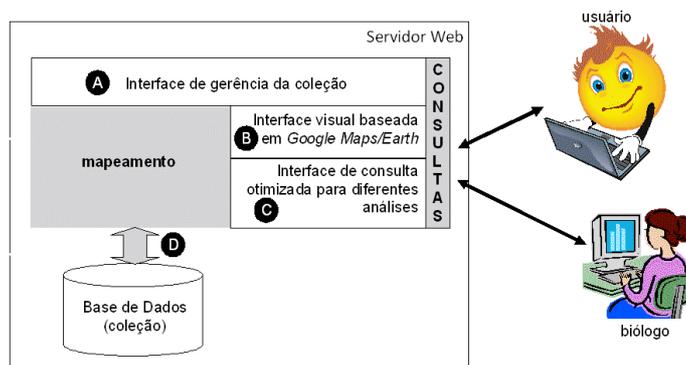


Figura 1. Arquitetura do Sistema TaxonomyBrowser

3.2 Visão Geral dos Dados da Coleção

A base de dados foi modelada para armazenar informações de espécimes coletadas por biólogos, processar consultas para análise de dados bem como disponibilizar os dados da coleção para o público em geral. Os dados são representados por nodos taxonômicos, que representa um nível específico dentro do sistema taxonômico de classificação dos organismos vivos. A organização de vários nodos é feita de forma hierárquica, gerando uma árvore taxonômica como ilustrado na Figura 2.

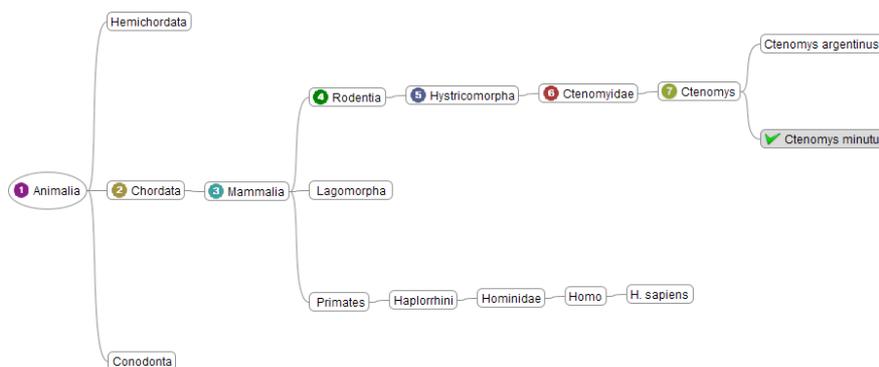


Figura 2. Exemplo de Árvore Taxonômica.

Um caminho específico dentro dessa árvore taxonômica gera uma classificação científica de um espécime. Por exemplo, a classificação científica de um roedor conhecido como *tucu-tuco*, comumente encontrado nas planícies da América do Sul, é dado por: 1- *Animalia* (Reino), 2- *Chordata* (Filo), 3- *Mammalia* (Classe), 4- *Rodentia*

¹⁰ O sistema *TaxonomyBrowser* foi implementado sobre o SGBD *MySQL*, tendo em vista sua portabilidade para diferentes sistemas operacionais, um vasto conjunto de ferramentas que facilitam o desenvolvimento e administração de sistemas e também por sua licença *GNU General Public License*.

¹¹ O sistema R (<http://www.r-project.org/>) está instalado no mesmo servidor e os scripts escritos pelos usuários são também armazenados no banco de dados.

(Ordem), 5- *Hystricomorpha* (Subordem), 6- *Ctenomyidae* (Família), 7- *Ctenomys* (Gênero), *Ctenomys minutus* (Espécie) (Figura 2).

3.3 O Modelo de Banco de Dados Proposto

A Figura 3 ilustra o diagrama E-R (Entidade-Relacionamento) simplificado do modelo proposto. Cabe ressaltar que o projeto do modelo do banco de dados incorpora os elementos que são parte do padrão Darwin Core [TDWG 2009], acrescentando outras características de interesse de nossos usuários alvo. O Darwin Core foi utilizado visando principalmente a interoperabilidade futura do trabalho com projetos semelhantes, descritos na seção 2, e também por ser um padrão que possui uma vasta documentação, sendo facilmente adaptável ao nosso modelo.

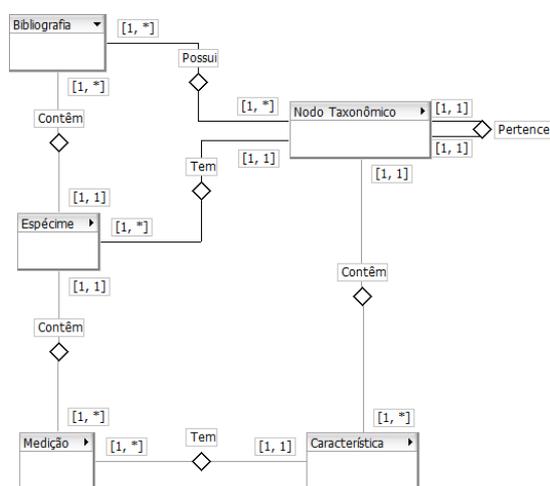


Figura 3. Diagrama E-R do Sistema *TaxonomyBrowser*.

Cada nodo taxonômico armazena uma referência a um nodo “pai”, para construir a hierarquia, o nome do nível taxonômico (Reino, Filo, Classe, Ordem, Subordem, Família, Gênero ou Espécie), o nome científico deste nível taxonômico (por exemplo, um nome científico para o Gênero pode ser *Ctenomys*), além de informações extras que podem ser adicionadas quando houver necessidade. O nodo taxonômico pode ainda conter uma lista de referências bibliográficas (entidade *Bibliografia*) que identifica as obras que dizem respeito ao nível taxonômico que está sendo descrito. A entidade *Bibliografia* armazena o título da obra, o autor, o editor e o ano de publicação. As características devem estar relacionadas a um grupo específico, que pode ser do grupo de características morfométricas, taxonômicas e filogenéticas (entidade *Característica*). Uma característica possui também um tipo específico para armazenagem, que pode ser: unidade física, nome literal, número inteiro, número real, arquivo de imagem, arquivo de som, arquivo de texto e arquivo de vídeo. Uma unidade física está ligada a uma unidade de medida (entidade *Medição*) no sistema internacional de medidas (SI). As unidades físicas são importantes para facilitar conversões de medidas entre pesquisas feitas no sistema. A entidade *Espécime* designa um exemplar coletado ou uma amostra de um ser vivo. A entidade *Espécime* contém informações de identificação do exemplar, coletor, data de coleta, e, ainda, pode conter informações geográficas sobre o local de coleta.

4. Interface visual

O sistema possui uma interface visual que permite aos usuários realizar todas as operações mencionadas na seção 3.3, dependendo de suas permissões de acesso. Foi utilizada a arquitetura *Model-View-Controller* onde a aplicação é separada por camadas: dados (*model*), layout (*view*) e a comunicação entre eles (*controller*). A camada que gerencia a lógica da aplicação foi desenvolvida utilizando classes *Controllers*. Para cada entidade do sistema, uma classe *Controller* foi implementada. Estas classes são responsáveis por responder às ações dos usuários vindas da camada de visualização, alterar a camada de dados dependendo da entrada dos usuários e ainda, validar dados gerados pelos usuários do sistema.

Para um melhor gerenciamento das funcionalidades do sistema, a interface visual foi subdividida em quatro interfaces principais, sendo elas a interface administrativa, a interface de consulta, a interface sobre mapa e a interface de análise. A interface administrativa permite inserir nodos taxonômicos, incluindo a definição de suas características (nome de apresentação, tipo, unidade de medida) e a inclusão, alteração ou remoção de espécimes vinculados a um nodo taxonômico folha. Tendo em vista o foco do presente trabalho, serão detalhadas apenas as interfaces de consulta e de análise.

4.1. Interface de consulta

A interface de consulta possibilita aos usuários pesquisar por espécimes registrados no banco de dados, usando características relevantes, e apresentando-os num mapa, quando possuírem coordenadas geográficas de localização/coleta.

Um editor de buscas permite determinar os critérios de busca de informações usando as características dos espécimes em expressões com uma operação (que pode ser $<$, $>$, $<=$, $>=$, $==$, $!=$, *between* e *like*) e dois valores, como podemos ver na Figura 4(a). A operação *between* é a única que necessita de dois valores; as outras utilizam apenas o primeiro valor. Por exemplo, “comprimento da cauda” *between* 10 e 15 retornará todos os espécimes que tem o campo “comprimento de cauda” com valores entre 10 e 15. A operação *like* é utilizada como, por exemplo, *like* “%con”, retornando todos os espécimes cujo nome termina em “con”. Pode-se escolher a cor do ícone que representará os espécimes resultantes da busca sobre o mapa, permitindo diferenciar grupos de espécimes. A especificação de uma busca é armazenada para utilização posterior, sendo possível sua modificação ou a re-exibição do resultado.

4.2. Interface sobre mapa

O tratamento de mapas é realizado com a API do Google Maps, a qual consiste basicamente de um conjunto de classes JavaScript, que fornecem a interface necessária para construir aplicações sobre mapas, realizar consultas por endereços em coordenadas geográficas, realizar funções de *zoom*, acrescentar pontos de referência e descrições no mapa, dentre outras possibilidades.

A interface permite interagir com os dados apresentados na forma de ícones. Marcações na forma de linhas e polígonos permitem dividir ou agrupar os espécimes resultantes das buscas selecionando-os para uma operação posterior. Também é possível gerar a envoltória convexa de cada consulta, demarcando a área onde se encontram todos os espécimes resultantes da busca.

A Figura 4(b) apresenta o resultado da busca especificada em 4(a) com todos os exemplares de *calomys sp* com peso entre 9 e 22 g cadastrados até o momento. Estes espécimes estão representados no mapa com ícones verdes e a região onde se encontram marcada em amarelo. Para realizar a análise de um subgrupo desses espécimes, podemos separá-los desenhando uma linha, e considerar apenas os espécimes que estão abaixo ou acima da linha, por exemplo, para submissão a um procedimento de análise.

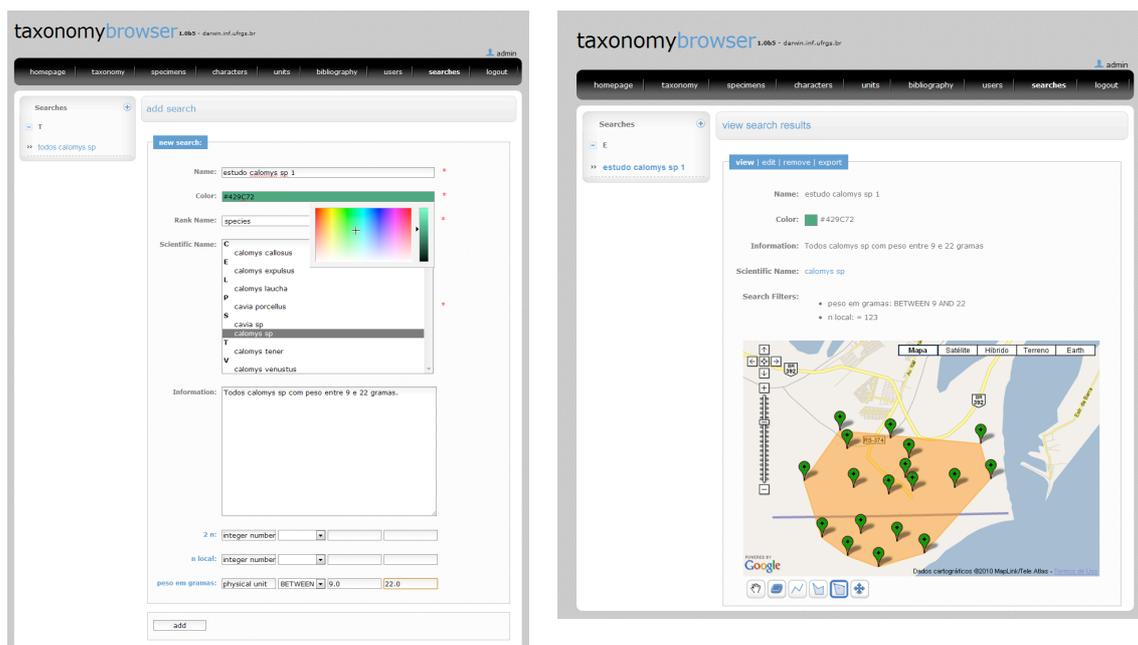


Figura 4. Interface visual: (a) editor de buscas; (b) resultado de busca

4.3. Interface de análise

A interface de análise permite aos usuários submeter os dados de espécimes resultantes de uma busca (seções 4.1 e 4.2) a um *script* na linguagem do sistema R. O exemplo da Figura 5 mostra o resultado de uma busca, a escolha de um *script* (“analise.r”) dentre duas possibilidades, a escolha da função desejada e o resultado da execução do *script* exibido numa janela separada.

O sistema oferece uma opção para o usuário salvar os dados de uma consulta para continuar a análise em um momento posterior, além de permitir que sejam enviados novos *scripts* para o sistema. Estes *scripts* enviados pelos usuários devem seguir uma formatação adequada para que sejam processados e oferecidos para um usuário que não tenha experiência com a linguagem R.

6. Conclusões e Trabalhos Futuros

Este artigo descreveu os principais componentes do sistema *TaxonomyBrowser* destinado ao gerenciamento de dados de espécimes coletados por biólogos. A utilização da API Google Maps permite a visualização de mapas e a integração com o sistema R permite analisar os dados sem a mudança de ambiente. A generalidade do nosso modelo também permite que seja utilizado como repositório de descritores de coleções de museus.

Os trabalhos futuros concentram-se no aprimoramento da interface sobre o mapa, com novas possibilidades de interação e vinculação com *scripts*, para análises diferentes em regiões selecionadas, por exemplo, e integração de resultados.

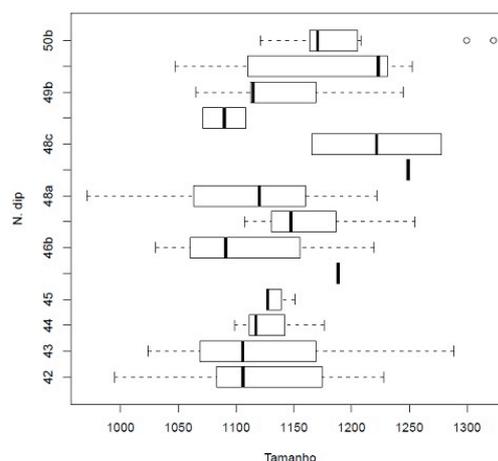
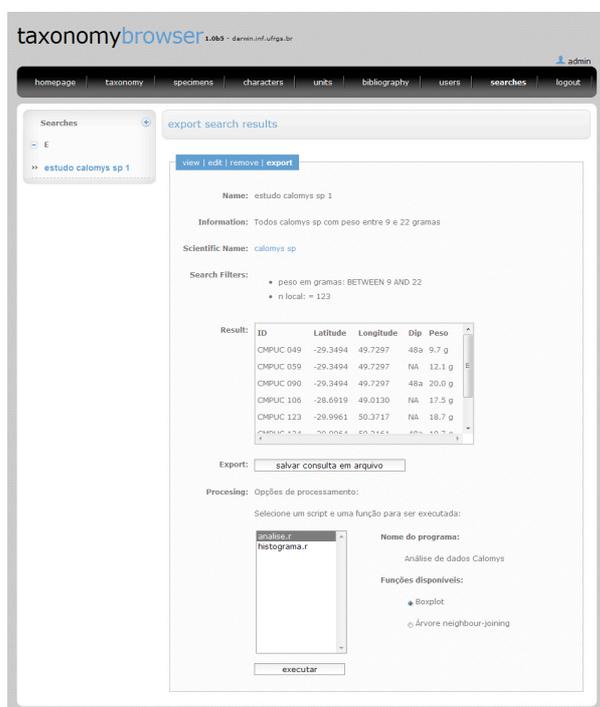


Figura 5. Interface de análise (a) e (b) resultado do processamento de *script* R.

Agradecimentos

Trabalho desenvolvido com apoio do CNPq.

Referências

- Guralnick, R. e Neufeld, D. (2005) “Challenges Building Online GIS Services to Support Global Biodiversity Mapping and Analysis: Lessons from the Mountain and Plains Database and Informatics project.” *Biodiversity Informatics*, 2:56–69.
- Jones, M., Berkley, C., Bojilova, J. e Schildhauer, M. (2001) “Managing Scientific Metadata”. *IEEE Internet Computing*, 5(5):59–68.
- Malaverri, J. G., Vilar, B. S. C. M., e Medeiros, C. B. (2009) “A Tool based on Web Services to Query Biodiversity Information”. In *WEBIST*, p. 305-310.
- Soberón, J. e Peterson, T. (2004) “Biodiversity informatics: managing and applying primary biodiversity data”. *Philosophical Transactions: Biological Sciences*, 359(1444):689–698.
- TDWG – Taxonomic Databases Working Group (2009) Darwin Core. <http://www.tdwg.org/standards/450/>