SiBBr: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira

Luiz M. R. Gadelha Jr.[1], Pedro Guimarães[1], Ana Maria Moura[1], Debora P. Drucker[2], Eduardo Dalcin[3], Guilherme Gall[1], Jurandir Tavares Jr.[1], Daniele Palazzi[1], Maira Poltosi[1], Fabio Porto[1], Francisco Moura[1], Wagner Vieira Leo[1]

[1]Laboratório Nacional de Computação Científica (LNCC)
Av. Getúlio Vargas, 333 – 25.651-075 – Petrópolis – RJ
{lgadelha,dpalazzi,fmoura,gmgall,jurandir,maira}@lncc.br
{pedrodpg,wagner}@lncc.br, anamaria.moura@gmail.com

[2]Embrapa Monitoramento por Satélite
Av. Soldado Passarinho, 303 – 13.070-115 – Campinas – SP
debora.drucker@embrapa.br

[3]Instituto de Pesquisas Jardim Botânico do Rio de Janeiro (JBRJ)
Rua Pacheco Leão, 915 – 22.460-030 – Rio de Janeiro – RJ
edalcin@jbrj.gov.br

Abstract. *In this article we describe the Brazilian Biodiversity Information System, which aims to provide an infrastructure for gathering, integrating, and analyzing data produced by various institutions in this area. Both its architecture and the process for harvesting and indexing data on species occurrences and checklists, one of the already implemented components, are described. An implementation of a scalable scientific workflow for species distribution modeling, one of the most used applications for analyzing biodiversity, is presented. Finally, we describe current work on integrating socioeconomic data and on managing ecological data.*

Resumo. *Neste artigo descrevemos o Sistema de Informação sobre a Biodiversidade Brasileira, que tem como objetivo fornecer uma infraestrutura para coleta, integração e análise de dados produzidos e disponibilizados por diversas instituições da área. A arquitetura de software do sistema é descrita, bem como o fluxo de coleta e indexação de dados de ocorrências e listas de espécies, um dos componentes já implementados. É apresentada a implementação de um workflow científico escalável para a modelagem da distribuição de espécies, uma das aplicações mais utilizadas na análise da biodiversidade. Finalmente, são descritos trabalhos em andamento como a integração de dados socioeconômicos e o gerenciamento de dados ecológicos.*

1. Introdução

Diversos estudos [Chapin et al. 2000] [Cardinale et al 2012] mostram que existe um forte relacionamento entre as atividades humanas, mudanças globais, biodiversidade e os processos e serviços ecossistêmicos. Chapin et al. [Chapin et al. 2000] observam que variáveis da biodiversidade, como o número de espécies e indivíduos de cada espécie,

bem como quais espécies estão presentes em um local, além das interações que ocorrem entre elas e o ambiente, determinam as características e atributos funcionais das espécies que afetam os processos ecossistêmicos. Mudanças globais, muitas vezes causadas pelo homem, como a propagação de espécies invasoras e mudanças no uso da terra, podem alterar significativamente essas variáveis da biodiversidade e consequentemente a expressão das características e atributos funcionais das espécies. Isso, por sua vez, pode afetar serviços ecossistêmicos essenciais para a humanidade. Esforços para abordar esse problema resultaram na assinatura da Convenção sobre a Diversidade Biológica (CBD) [CBD 2014] em 1992 durante a Cúpula da Terra (Rio-92) no Rio de Janeiro. O Plano Estratégico para Biodiversidade 2011-2020, proposto pelos signatários do CBD, define ações que os países devem realizar para cumprir vinte Metas de Biodiversidade de Aichi até 2020. Para monitorar as mudanças na biodiversidade é essencial coletar, documentar, armazenar e analisar indicadores sobre a distribuição espaço-temporal das espécies, além de obter informações sobre como elas interagem entre si e com o ambiente em que vivem [Michener et al. 2012]. O desenvolvimento e implantação de mecanismos para produzir esses indicadores dependem do acesso a dados confiáveis obtidos em expedições de campo, por sensores automáticos, em coleções biológicas e na literatura acadêmica. As metodologias e técnicas usadas para gerenciar e analisar esses dados definem uma área de pesquisa frequentemente chamada de *Informática na Biodiversidade* [Soberón e Townsend Peterson 2004] [Hobern et al. 2013].

Neste trabalho, apresentamos o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), que está sendo desenvolvido para integrar e disseminar dados coletados e publicados por diversas instituições brasileiras, como universidades, institutos de pesquisa e agências governamentais. O SiBBr desempenha também o papel de nó brasileiro do Global Biodiversity Information Facility (GBIF) [GBIF 2013]. Os dados disponibilizados pelo SiBBr servirão de base para estudos de análise e síntese realizados por cientistas. Os resultados destes estudos subsidiarão o governo em suas decisões, como por exemplo, a definição de áreas prioritárias de conservação ou a avaliação do potencial impacto ambiental de grandes obras governamentais.

A seção 2 deste artigo apresenta um panorama global sobre as infraestruturas disponíveis para coleta e integração de dados de biodiversidade; a arquitetura proposta para o desenvolvimento do SiBBr é apresentada na seção 3; a seção 4 descreve o processo de coleta e indexação de dados no SiBBr; a seção 5 destaca a importância das ferramentas de análise e síntese de dados e como podem ser aplicadas no contexto de workflows científicos; sessão 6 conclui o artigo apresentando trabalhos futuros.

2. Trabalhos Relacionados

O *Global Biodiversity Information Facility* (GBIF) [Edwards et al. 2000] [GBIF 2014] é uma infraestrutura global para coleta e integração de dados de Biodiversidade. As instituições detentoras destes dados podem expô-los utilizando o padrão de publicação de dados Darwin Core [Wieczorek et al. 2012] e o padrão de metadados *Ecological Metadata Language* (EML) [Fegraus et al. 2005]. O GBIF recupera e indexa os dados providos nestes formatos pelas diversas instituições que participam da sua rede e os disponibiliza através do seu portal de dados [GBIF 2014]. Como nó brasileiro do GBIF, o SiBBr utiliza estes mesmos padrões para colher e agregar dados sobre coleções biológicas, ocorrências e listas de espécies de instituições do Brasil ou do exterior, que

publiquem dados sobre a biodiversidade brasileira. O *Data Observation Network for Earth* (DataONE) [Michener et al. 2012] [DataONE 2014] é uma rede de observação de dados para a Terra, a qual provê uma infraestrutura distribuída para gerenciamento de dados ecológicos. Alguns provedores de dados participantes do DataONE, também utilizam o EML como padrão de metadados como o *Knowledge Network for Biocomplexity* (KNB) [KNB 2014] e o *Long Term Ecological Research Network* (LTER) [LTER 2014]. No entanto, em função da heterogeneidade dos conjuntos de dados ecológicos, estes são publicados no formato original, seja em planilha ou em arquivos texto separado por vírgulas, o que torna mais complexos os trabalhos de síntese envolvendo dados de diversas pesquisas ecológicas. Para facilitar trabalhos de análise, o SiBBr pretende implementar técnicas de *integração tardia* [Halperin et al. 2013], onde os dados de planilhas, típicas da rotina de pesquisa dos ecólogos, são extraídos e armazenados em bancos de dados relacionais para fins específicos, muitas vezes sem a preocupação de utilizar um esquema de dados bem-definido, de modo permitir uma integração de dados colaborativa. Outros sistemas relacionados ao SiBBr incluem o *Atlas of Living Australia* [ALA 2014], o SiB Colombia [SiB Colômbia 2014], o speciesLink [speciesLink 2014] e a EUBrazilOpenBio Hybrid Data Infrastructure [Amaral et al. 2014]. O SiBBr se diferencia destas outras iniciativas pelo escopo mais amplo: além de integrar dados socioeconômicos, disponibilizará ferramentas de computação de alto desempenho para modelagem de distribuição de espécies.

3. Arquitetura de Software do SiBBr

A descrição arquitetural vigente do SiBBr aqui apresentada, lista tanto componentes já implementados como os que serão desenvolvidos. O SiBBr é organizado de forma modular, onde boa parte da funcionalidade está disponível através de serviços web. Dessa forma, novas funcionalidades podem ser agregadas gradualmente e os componentes existentes podem ser modificados e adaptados de forma flexível. A visão geral da arquitetura do SiBBr é apresentada na Figura 1. Os dois principais componentes para gerenciamento de dados são providos pelo *Portal de Espécies e Ocorrências* e pelo *Portal de Dados Ecológicos*. O primeiro [Portal de Espécies e Ocorrências 2014] foi desenvolvido com o apoio do SiB Colombia, reutilizando código-fonte do portal de dados do GBIF. Ele é responsável pela coleta, indexação e disseminação de conjuntos de dados, registros de ocorrência e listas de espécies disponibilizadas por diversos publicadores de dados. Esses conjuntos de dados são publicados utilizando o padrão Darwin Core [Wieczorek et al. 2012] com metadados no padrão EML [Fegraus et al. 2005]. A funcionalidade de publicação de dados é provida pela ferramenta *Integrated Publishing Toolkit* (IPT) [IPT 2013], que permite que dados disponíveis localmente nas instituições participantes, como por exemplo, em planilhas ou bancos de dados relacionais, sejam mapeados para o padrão Darwin Core e coletados pelo Portal de Ocorrência e Listas de Espécies. Atualmente estão publicados mais de 533 mil registros.

O Portal de Dados Ecológicos é responsável pelo recebimento, armazenamento e disseminação de conjuntos de dados de, por exemplo, Pesquisas Ecológicas de Longa Duração (PELD) [Michener et al. 2011]. Diferentes abordagens em ecologia, aliadas às tradições de pesquisa distintas, tanto em suas subdisciplinas como em áreas afins, levam à produção de dados altamente heterogêneos (Drucker 2011). Tais dados podem ser,

entre outros, contagens de indivíduos, medidas de variáveis ambientais ou representações de processos ecológicos. As terminologias utilizadas também variam de acordo com a linha de pesquisa, bem como a forma de estruturar os dados digitalmente (Jones et al. 2006). Também foi adotado o padrão de metadados EML para a descrição dos conjuntos de dados ecológicos. Os conjuntos de dados em si, em função da heterogeneidade, são publicados no formato original, através de planilhas ou arquivos textuais com valores separados por vírgula. Como o SiBBr adotará o mesmo conjunto de padrões e ferramentas do DataONE, eventualmente poderá integrar esta rede.

O SiBBr iniciou sua integração à Infraestrtura Nacional de Dados Espaciais (INDE) [INDE 2014], que agrega dados espaciais sobre o Brasil publicados por diversas instituições, para fornecer dados georeferenciados sobre biodiversidade. Serão utilizados a ferramenta Geonetwork [Geonetwork 2014], o BioMetadata [BioMetadata 2014] e uma extensão proposta ao Padrão de Metadados para Informações Geográficas (ISO 19115:2003). Um componente de visualização de dados está sendo desenvolvido para permitir a exibição em mapas e gráficos de múltiplas camadas de dados providas pela INDE e pelo SiBBr.

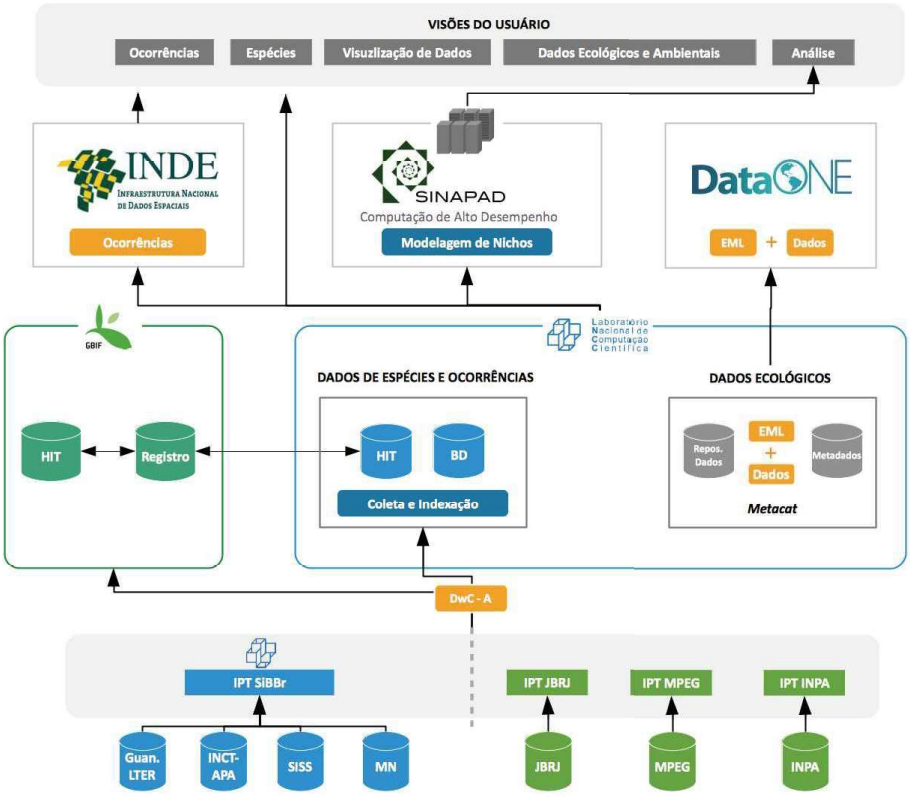

Figure 1. Arquitetura de software do SiBBr

Rotinas frequentes de análise e síntese serão apoiadas por sistemas de gerenciamento de workflows científicos [Deelman et al. 2008]. Técnicas de computação paralela e distribuída serão utilizadas na execução dessas análises, a exemplo da modelagem de distribuição de espécies [Townsend Peterson 2011], através de recursos

computacionais de alto desempenho do Sistema Nacional de Processamento de Alto Desempenho (SINAPAD) [SINAPAD 2014].

4. Coleta e Integração de Dados de Ocorrências e Listas de Espécies

As instituições que desejam publicar dados sobre ocorrências e listas de espécies no SiBBr recebem um registro globalmente único que as identifica tanto no escopo do SiBBr quanto do GBIF. Cada instituição pode instalar e cadastrar instâncias do IPT para publicar os dados disponíveis na sua instituição. Dados disponíveis em diversos formatos são mapeados para o padrão Darwin Core e os metadados correspondentes são criados no padrão EML. A partir daí ambos arquivos são empacotados pelo IPT e transformados em um arquivo único no formato Darwin Core Archive (DwC-A). Atualmente, o processo de coleta e indexação é feito através do *Harvesting and Indexing Toolkit* (HIT) [HIT 2013]. Esta é uma ferramenta web produzida pelo GBIF que acessa o endereço dos IPTs cadastrados na ferramenta, realiza o *download* dos DwC-A dos conjuntos de dados e faz a extração dos dados indexando-os à base de dados do Portal de Ocorrência e Listas de Espécies.

Em colaboração com outros nós do GBIF, esse processo será aprimorado para contemplar o seguinte processo: busca dos dados nas entidades publicadoras (*coleta*); e extração dos dados compactados no formato DwC-A e armazenamento nas bases de dados (*indexação*). Cada vez que um conjunto de dados é criado ou atualizado, o IPT atualiza sua fila de notificações, que funciona no modelo publicador-assinante. O processo de coleta possui um componente que monitora as notificações publicadas pelos IPTs. É acionado então um componente coletor, que baixa os arquivos DwC-A disponíveis nesses IPTs, validando sua estrutura. O coletor então encaminha os registros do conjunto de dados para o componente de indexação, através de uma fila de processamento distribuído. As mensagens da fila são encaminhadas para o componente coordenador de validações, que aciona uma série de validações disponíveis pré-configuradas pelo administrador do processo. Estas validações estão relacionadas à semântica: verificam dados temporais, coordenadas geográficas e correspondente descrição textual e a taxonomia. Os resultados das validações são adicionados aos dados, mantendo os valores originais. O resultado desse fluxo de validação retorna ao publicador com o objetivo de melhorar a qualidade dos dados. Ao fim das validações, são geradas notificações que liberam o registro para indexação no banco de dados e publicação pelo Portal de Espécies e Ocorrências.

5. Ferramentas de Apoio à Análise de Dados de Biodiversidade

Ferramentas de análise e síntese de dados de biodiversidade, a exemplo da modelagem de distribuição de espécies (MDE) [Townsend Peterson et al. 2011], são amplamente utilizadas. Essas análises normalmente empregam diversas aplicações distintas, executadas de forma fracamente acoplada, sendo um caso típico para a utilização de sistemas de gerenciamento de workflows científicos [Deelman et al. 2009]. Por exemplo, no caso da MDE, dados climatológicos globais são recuperados de provedores de dados ambientais, enquanto dados de ocorrências de espécies são obtidos de portais como o GBIF. É comum que esses dados tenham que ser adaptados com ferramentas de gerenciamento de informações geográficas ou filtrados com ferramentas de controle de qualidade. Após esses passos de pré-processamento, algoritmos para MDE, a exemplo do Maxent [Phillips et al. 2004], são aplicados para prever a distribuição potencial de

espécies utilizando os dados ambientais e os de ocorrência de espécies adquiridos e manipulados nos passos de pré-processamento. Finalmente, uma etapa de pós-processamento é realizada, onde ferramentas estatísticas e de visualização de dados são utilizadas para analisar o resultado da modelagem. Tal processo é computacionalmente demandante, o que torna importante o uso de ferramentas que sejam escaláveis.

O SiBBr implementou um protótipo de workflow científico para MDE [SiBBr Github 2014], que permite a execução de diversos algoritmos de MDE disponíveis na biblioteca openModeller [Muñoz 2011]. A implementação foi realizada no Swift [Wilde et al. 2011], um sistema de gerenciamento de workflows científicos com ênfase em paralelismo e distribuição. A estratégia de paralelização implementada utiliza uma *thread* de execução por espécie modelada no workflow científico. Outras oportunidades de paralelismo que podem ser exploradas incluem a execução de uma *thread* por algoritmo do MDE, por cenário climatológico previsto pelo Painel Intergovernamental sobre Mudanças Climáticas (IPCC) [IPCC 2014], ou ainda por conjunto de parâmetros. O workflow foi executado em uma máquina de memória compartilhada com 72GB de memória e 24 núcleos de processamento. O primeiro gráfico (Figura 2.a) mostra o tempo de execução do workflow científico quando o número de espécies varia de 1 a 24 e o número de *threads* de execução é 24. O segundo gráfico (Figura .b) mostra o tempo de execução quando o número de espécies é 24 e o número de *threads* varia de 1 a 24. Para calcular o tempo médio de execução, foram realizadas três execuções para cada configuração utilizada. Uma vantagem de utilização do Swift na implementação do workflow científico para MDE é o seu suporte nativo ao registro de dados de proveniência [Gadelha et al. 2012] das modelagens executadas, o que facilita tanto a reprodutibilidade do experimento computacional como a sua análise e validação.

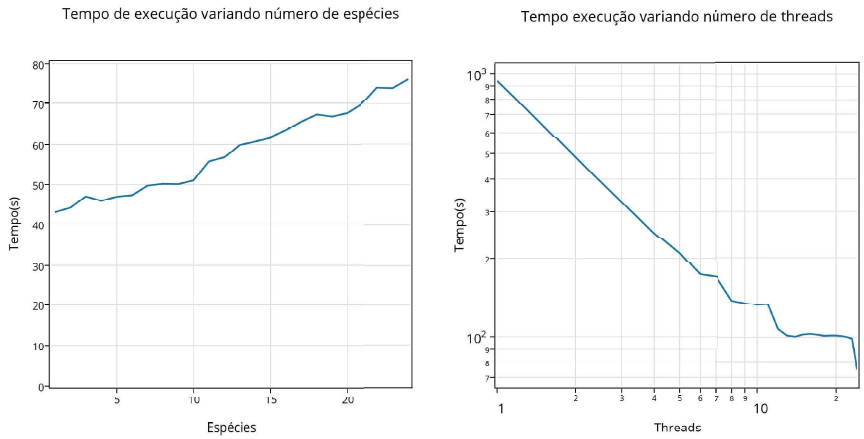

Figura 2. Tempo de execução do workflow científico para MDE.

6. Conclusão

O desafio de monitorar as mudanças na biodiversidade e avaliar o impacto das mesmas na sustentabilidade e desenvolvimento humanos requer desde a utilização de técnicas avançadas para integração de dados sobre biodiversidade e ecologia até o emprego de modelos preditivos computacionais. Nesse trabalho mostramos que o SiBBr já permite a agregação de dados de espécies e ocorrências disponibilizadas por diversas instituições acadêmicas e de pesquisa bem como de órgãos governamentais. Esses dados serão utilizados como insumo para estudos de análise e síntese realizados por cientistas e especialistas, que poderão, a curto e médio prazo apoiar as decisões governamentais

relativas à conservação da biodiversidade e seu uso sustentável e socialmente justo. Um primeiro protótipo de workflow científico para modelagem de distribuição de espécies permite uma execução escalável e com registro de informações de proveniência.

Como trabalhos futuros podemos citar: (i) expansão da capacidade de análise através de consultas aos dados por meio de uma linguagem declarativa do tipo SQL, que permita extrair, agregar e ordenar dados segundo filtros específicos, possibilitando também criar visões diferenciadas sobre conjuntos de dados e reutilizá-las posteriormente em novas consultas; (ii) Oferecer os workflows científicos já desenvolvidos para MDE através de uma interface web e integrados ao ferramental de portais científicos do SINAPAD [Gomes et al. 2014], possibilitando a execução simultânea dos modelos em diferentes centros de computação de alto desempenho.

Referências Bibliográficas

Amaral, R., et al. (2014). Supporting biodiversity studies with the EUBrazilOpenBio Hybrid Data Infrastructure. Concurrency and Computation: Practice and Experience.

ALA. http://www.ala.org.au. Acessado em abril de 2014.

BioMetada. http://service.ncddc.noaa.gov/rdn/www/metadata-standards/documents/BIO - Metadata.pdf. Acessado em abril de 2014.

Cardinale, B. J et al. (2012). Biodiversity loss and its impact on humanity. Nature, 486(7401), 59–67.

CBD, http://www.cbd.int. Acessada em abril de 2014.

Chapin III, F. Stuart et al. Consequences of Changing Biodiversity. Nature 405, no. 6783 (2000): 234–242.

DataONE, http://www.dataone.org. Acessado em abril de 2014.

Deelman, E., D. Gannon, M. Shields, and I. Taylor. Workflows and e-Science: An Overview of Workflow System Features and Capabilities. Future Generation Computer Systems 25, no. 5 (2009): 528–540.

Drucker, D. P. (2011, Julho). Avanços na Integração e Gerenciamento de Dados Ecológicos. Natureza & Conservação, 9(1), pp. 115-120.

Edwards, J. L. (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science, 289(5488), 2312–2314.

Fegraus, E. et al. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. Bull. of the Ecological Society of America 86 (2005): 158–168.

Gadelha, L. et al. (2012). MTCProv: a practical provenance query framework for many-task scientific computing. Distributed and Parallel Databases, 30(5-6), 351–370.

GBIF, http://www.gbif.org. Acessado em abril de 2014.

GeoNetwork, http://geonetwork-opensource.org/. Acessado em abril de 2014.

Gomes, A. T. A., Bastos, B. F., Medeiros, V., & Moreira, V. M. (2014). Experiences of the Brazilian national high-performance computing network on the rapid prototyping of science gateways. Concurrency and Computation: Practice and Experience.

Halperin, D. et al. (2013). Real-time collaborative analysis with (almost) pure SQL: a case study in biogeochemical oceanography. In Proceedings of the 25th International Conference on Scientific and Statistical Database Management – SSDBM. ACM.

Hobern, D. et al. Global Biodiversity Information Outlook - Delivering Biodiversity Knowledge in the Information Age. GBIF Secretariat, 2013.

HIT, https://code.google.com/p/gbif-indexingtoolkit/. Acessado em abril de 2014.

INDE, http://www.inde.gov.br. Acessado em abril de 2014.

IPCC, http://www.ipcc.ch. Acessado em abril de 2014.

IPT, https://code.google.com/p/gbif-providertoolkit/. Acessado em abril de 2014.

Jones MB et al., 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics, 37:519-44.

KNB, https://knb.ecoinformatics.org/. Acessado em abril de 2014.

LTER, http://www.ilternet.edu/. Acessado em abril de 2014.

Michener, W., J. Porter, M. Servilla, & K. Vanderbilt. Long Term Ecological Research and Information Management. Ecological Informatics 6, no. 1 (2011): 13–24.

Michener, W. K. et al. Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics 11, 5–15 (2012).

Muñoz, M. et al. openModeller: a generic approach to species' potential distribution modelling. GeoInformatica 15, 111–135 (2011).

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In Twenty-first International Conference on Machine Learning - ICML '04 (p. 83). ACM Press.

Portal de Espécies e Ocorrências, SiBBr (versão de homologação). http://gbif.sibbr.gov.br/portal. Acessado em abril de 2014.

SiBBr Github. http://gihub.com/sibbr. Acessado em abril de 2014.

SiB Colombia. http://www.sib-colombia.net. Acessado em abril de 2014.

SINAPAD, http://www.lncc.br/sinapad. Acessado em abril de 2014.

speciesLink. http://splink.cria.org.br. Acessado em abril de 2014.

Soberón, Jorge, e Townsend Peterson. Biodiversity Informatics: Managing and Applying Primary Biodiversity Data. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359, no. 1444 (2004): 689–698.

Townsend Peterson, A. et al. (2011). Ecological Niches and Geographic Distributions (p. 328). Princeton University Press.

Wieczorek, John et al.. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7, no. 1 (2012): e29715.

Wilde, M. et al. (2011). Swift: A language for distributed parallel scripting. Parallel Computing, 37(9), 633–652.