Georreferenciamento de dados biológicos legados do INPA

Elizziane M. B. Farias, José L. Campos dos Santos

Instituto Nacional de Pesquisas da Amazônia – INPA
Programa de Grande Escala da Biosfera-Atmosfera da Amazônia – LBA
Caixa Postal 447 – 69.080-971 – Manaus – AM – Brasil

{elizziane.farias,lcampos}@inpa.gov.br

Abstract. *Biological collections are the primary sources of knowledge regarding biodiversity. In Brazil, such information have been historically gathered and maintained by research institutions in a traditional way. To register geographic information associated with biological data is essential for the curatorial work for later sharing. For legacy data, they need to align solutions with the new technological time of geospatial world. This work, under development, presents an architectural proposal for a system to improve the process of georeferencing legacy data of INPA's biological collections through an infrastructure of a gazetteer in a Web environment.*

Resumo. *Coleções biológicas são fontes primárias de conhecimento sobre a biodiversidade. No Brasil, tais informações vêm sendo historicamente coletadas e mantidas pelas instituições de pesquisa de maneira tradicional. Registrar informações geográficas associadas aos dados é essencial para o trabalho de curadoria para posterior compartilhamento. Para os dados legados, precisam-se de soluções de alinhamento com o novo momento tecnológico do mundo geoespacial. Este trabalho, em desenvolvimento, apresenta uma proposta de arquitetura de um sistema para auxiliar o processo de georreferenciamento de dados legados de coleções biológicas do INPA através de uma infraestrutura de um gazetteer em ambiente Web.*

1. Introdução

A Amazônia ocupa uma posição de destaque em relação à biodiversidade mundial (PPBIO, 2015). Informações primárias são encontradas nas coleções biológicas, que são bancos de materiais (espécimes ou exemplares) vivos ou preservados, associados a dados biológicos e espaciais, nos quais são ferramentas imprescindíveis para o trabalho de especialistas em biodiversidade e apoio indispensável para outras áreas de conhecimento. Os dados biológicos legados nas instituições da Amazônia ainda representam aproximadamente 98% (levantamento da última década) do total mantido em suas bases digitais de dados (Campos dos Santos, 2003). Um dos principais problemas é que grande parte desses dados coletados encontram-se ainda inadequados para processamento em ambientes de Sistemas de Informações Geográficas (SIG) e consequentemente para análises espaciais. Isso ocorre devido à falta de informações consistentes de posicionamento geográfico, que na maioria das vezes descritas em formato textual, por exemplo: *"Lago de Balbina em Presidente Figueiredo, Estado do Amazonas, Brasil*", *"Lago Mamori", "Vila do Araça, Careiro, Amazonas, Brasil", "15 KM da AM010 – Próximo a Reserva Adolpho Ducke, Manaus, Amazonas, Brasil".*

Georreferenciamento de dados legados envolve o processo de derivação da informação de descrição textual de localidades em informações geoespaciais através de coordenadas geográficas juntamente com suas estimativas de incerteza. Pode-se prever a utilização de técnicas de recuperação de informação (RI) enfatizando a recuperação e indexação de dados geográficos e espaciais, denominada Recuperação de Informação Geográfica (RIG) e utilização de recursos de referência espacial indireta como a adoção de *gazetteers* (dicionários de nomes geográficos associados a coordenadas geográficas). Diante do problema, o presente trabalho apresenta uma proposta de desenvolvimento de uma arquitetura baseada nas técnicas de RIG para georreferenciamento e conversão de dados biológicos legados em dados georreferenciados associado à infraestrutura de um *gazetteer* para disseminação.

2. Trabalhos Relacionados

Gazetteers podem fornecer informações adicionais sobre a história do local, dados populacionais, pronúncias de lugares e etc. (Kessler et. al, 2009). Atualmente existem vários sistemas disponíveis que fornecem informações de localizações espaciais. Dentre os principais sistemas existentes, destaca-se: *Gettty Thesauri*[1], *Fuzzy Gazetteer*[2], *GeoNames*[3], *Global Gazetteer*[4], *GEONet*[5] entre outros. O *Gettty Thesauri of Geographic Names* é um *thesaurus* (lista de sinônimos) que contêm um extenso vocabulário de dados de nomes geográficos implementado através *gazetteer*, reunindo informações de diversos lugares do mundo associadas a nomes de lugares geográficos com suas localizações espaciais. *Fuzzy Gazetteer* é um dicionário de nomes geográficos que permite pesquisas para nomes com variações de escrita de vários lugares do mundo. Sua base de dados possui mais de 7 milhões de nomes de lugares. *GeoNames* é uma base de dados geográficos que disponibiliza vários serviços web, possuindo mais de 8 milhões de nomes geográficos correspondentes a mais 6,5 milhões de elementos geográficos. Os dados armazenados incluem latitude, longitude, altitude, população, subdivisão administrativa e código postal. *Global Gazetteer* conta com uma base de dados geográficos de aproximadamente 3 milhões de elementos geográficos. Dentre suas funcionalidades, disponibiliza informações sobre altitude, previsão do tempo, topografia do local consultado. *GEONet Names Server* é uma base de dados de nomes de elementos geográficos com uma cobertura geográfica de vários países do mundo com exceções de lugares dos Estados Unidos e Antártica. Sua base de dados possui cerca de 5,5 milhões de nomes geográficos.

3. Arquitetura proposta

Após levantamento de requisitos que envolve o processo de georreferenciamento de dados legados, enumerou-se uma lista de requisitos funcionais básicos do *gazetteer*, que são: **(1)** O processo de aquisição e processamento de informações será implementado com funcionalidades que permitam consultas, georreferenciamento de localizações descritivas, e visualizações de informações geográficas; **(2)** O sistema possuirá uma base de dados de nomes geográficos exclusiva e unificada, referente a locais geográficos na região amazônica; **(3)** O sistema disponibilizará interface para buscas, permitindo ao usuário pesquisador/curador realizar consultas de locais geográficos sem coordenadas de suas bases de dados legadas, consultando na base de dados de nomes geográficos e retornando coordenadas geográficas relativa as localidades consultadas; **(4)** O sistema

[1] http://www.getty.edu/research/tools/vocabulary/tgn
[2] http://dma.jrc.it/services/fuzzyg
[3] http://www.geonames.org/
[4] http://www.fallingrain.com/world/
[5] http://geonames.nga.mil/namesgaz/

gazetteer disponibilizará um mapa integrado ao banco de dados com recurso interativo, no qual permite consultas e visualização de dados georreferenciados; **(5)** A arquitetura do *gazetteer* será implementada em ambiente cliente/servidor apoiada por tecnologias de software livre, como o SGBD PostgreSQL e extensão PostGIS, Biblioteca JavaScript OpenLayers e linguagem PHP; e **(6)** Quanto ao processamento de informações, a entrada de dados passa por uma estruturação antes da sua busca na base de dados do sistema.

A Figura 1 apresenta a visão geral da arquitetura proposta, incluindo o fluxo de informações do *gazetteer* durante o georreferenciamento, além de contar com a estruturação e execução de funções básicas. A principal funcionalidade do sistema, diferenciada entre demais ferramentas existentes, é a capacidade de realizar processos de georreferenciamento de dados biológicos por meio de tratamento da localização em formato descritivo textual associada a uma base de dados de nomes geográficos de localidades. O módulo indica atributos de localidades não georreferenciadas de uma base de dados, permitindo consultas a bases de dados e realizando comparações precisas ou parciais entre as informações de cada descrição textual de localidade encontrada, retornando coordenadas geográficas, como resultado e saída de dados. Este processo é realizado através do módulo de *parsing* integrado ao *gazetteer*.

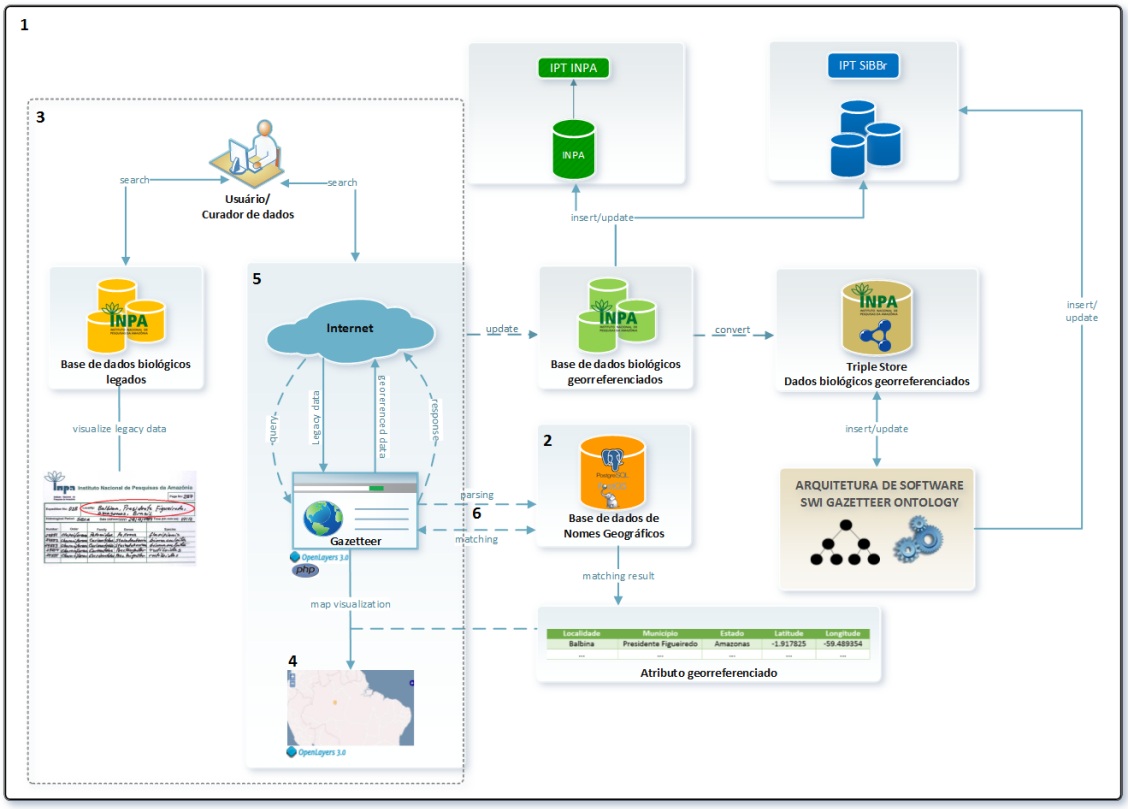

Figura 1. Visão geral da arquitetura do *gazetteer*.

Quando requisitado o georreferenciamento de dados pelo usuário, o processamento de *parsing* é ativado, no qual o módulo analisa a entrada de dados a ser georreferenciado, analisa a descrição textual nos seus componentes gramaticais, semânticos e léxicos e realiza comparação na base de dados de nomes geográficos do *gazetteer*. A imposição da entrada do dado legado em um formato pré-estabelecido diminuirá a incidência de erros na fase do *parsing*. Solicitando uma consulta, o módulo de *matching* executa uma sequência de tentativas para retornar coordenadas geográficas, partindo da possibilidade de georreferenciamento mais preciso através dos registros de localidades geográficas armazenadas no *gazetteer*.

Sem a intervenção do usuário, a base de dados georreferenciada resultante do processo será convertida a uma base de triplas, que poderá atualizar e fornecer bases de dados georreferenciadas para integração automática junto a sistemas do tipo *SWI Gazetteer* (Cardoso et al., 2014). O principal objetivo da arquitetura do sistema de georreferenciamento é obter posições espaciais de ocorrências de espécimes a partir de descrições textuais de localizações geográficas de dados biológicos legados e disponibilizar suas fontes de dados recém-georreferenciadas em formato do padrão LOD (*Linked Open Data*) para o repositório semântico do *SWI Gazetteer*. Esse repositório possui um foco particular sobre relacionamentos semânticos e correção de dados para validação de dados recém-georreferenciados, além de permitir a interação com o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr) através do seu *Integrated Publishing Toolkit* (IPT), o mesmo acontecendo nas bases do sistema *Specify*[6] e IPT do INPA que garantirá a curadoria dos dados georreferenciados pelo sistema proposto.

3. Considerações Finais

Apesar da disponibilidade de ferramentas com a finalidade de sugerir coordenadas geográficas a dados legados de coleções, constata-se grandes necessidades de proposições que auxiliem este tipo de georreferenciamento. Portanto, a arquitetura proposta de um sistema para georreferenciamento através de infraestrutura de *gazetteer* poderá reduzir a redundância de estudos e custos, por revelar a existência prévia de informações relevantes sobre os ecossistemas amazônicos e sua evolução biogeográfica. Pois, dados biológicos georreferenciados auxiliam a identificação de padrões e respondem questões em larga escala, tanto temporal quanto geográfica, gerando produtos úteis para cientistas, educadores e tomadores de decisão.

Referências

Cardoso, S. D, Amanqui, F. K., Serique, K. J, Campos dos Santos, J. L, Moreira, D. A. (2014) "SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data". In: Future Generation Computer Systems. In press.

Campos dos Santos, J. L. (2003) "A biodiversity information system in an open data/metadatabase architecture". 254 p. Tese (Doutorado em Computer Science) – University Twente, International Institute for Geo-Information Science and Earth Observation - ITC Printing Department,The Netherlands, ISBN: 90-6164-214-0.

Gadelha Jr., L. M. R., Guimaraes, P., Moura, A. M. C., Drucker, D., Dalcin, E., Gall, G., Tavares Jr, J., Palazzi, D. C., Poltosi, M., Porto, F., Moura, F., Leo, W. V. (2014) "SiBBr: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira." In: VIII Brazilian e-Science Workshop (BRESCI 2014), Brasília.

PPBIO (2015) Programa de Pesquisa em Biodiversidade. Disponível em: http://ppbio.inpa.gov.br. Acessado em: 24 de março 2015.

Kessler, C., Janowicz, K., Bishr, M. (2009) "An agenda for the next generation gazetteer: Geographic information contribution and retrieval". In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information System, pp. 91–100.

[6] *Specify* é um sistema de banco de dados que gerencia informações de espécies e espécime para informatizar coleções biológicas e para publicação de seus dados na Web - http://specifyx.specifysoftware.org/