SWI: Um gazetteer interativo para dados sobre
biodiversidade com suporte a web semântica

Silvio Domingos Cardoso

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Silvio Domingos Cardoso

SWI: Um gazetteer interativo para dados sobre biodiversidade com suporte a web semântica

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Dilvan de Abreu Moreira

USP – São Carlos
Julho de 2015

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

C268s Cardoso, Silvio Domingos
SWI: Um gazetteer interativo para dados sobre
biodiversidade com suporte a web semântica / Silvio
Domingos Cardoso; orientador Dilvan de Abreu Moreira.
- São Carlos - SP, 2015.
121 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática Computacional)
- Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2015.

1. Web Semântica. 2. Informação Geográfica
Voluntária. 3. Gazetteer. I. Moreira, Dilvan de
Abreu, orient. II. Título.

Silvio Domingos Cardoso

**SWI: An interactive gazetteer for biodiversity data with
semantic web support**

Master dissertation submitted to the Instituto de
Ciências Matemáticas e de Computação – ICMC-
USP, in partial fulfillment of the requirements for the
degree of the Master Program in Computer Science
and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Prof. Dr. Dilvan de Abreu Moreira

USP – São Carlos
July 2015

Este trabalho é dedicado a toda a minha família em especial aos meus pais Sebastião e Irene.

AGRADECIMENTOS

Gostaria de agradecer a todos que me apoiaram durante o desenvolvimento desse trabalho de mestrado. Em especial ao meu orientador Prof. Dr. Dilvan de Abreu Moreira, pelo apoio, paciência e conselhos para desenvolver o trabalho. Também agradeço a todos os funcionários e docentes do ICMC que diretamente ou indiretamente contribuíram para minha formação.

Agradeço muito a minha família pelo apoio nesta etapa da minha vida e por sempre me apoiar nas horas de mais necessidade. Também agradeço aos colegas do ICMC, pelos bons momentos de convívio, estudo e descontração em especial ao Laboratório Intermídia.

E por fim, agradeço a CAPES pelo apoio financeiro, ao Instituto de Ciências Matemáticas e de Computação e à Universidade de São Paulo pela estrutura disponibilizada para realização deste projeto.

*“Conhecimento não é aquilo que você sabe,
mas o que você faz com aquilo que você sabe.”
(Aldous Huxley)*

RESUMO

CARDOSO, S. D.. **SWI: Um gazetteer interativo para dados sobre biodiversidade com suporte a web semântica**. 2015. 121 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

O Brasil é considerado o país da megadiversidade por abrigar diversas espécies de flora e fauna. Dessa forma preservar essa diversidade é extremamente importante, pois a vida no planeta depende dos muitos ecossistemas que compõem essa biodiversidade. Atualmente, vários estudos sobre formas de recuperar e acessar informações sobre biodiversidade vem sendo discutidos na comunidade científica. Muitas instituições importantes têm disponibilizado gratuitamente seus registros de coletas disponíveis abertamente em repositórios online. No entanto, os dados disponibilizados nesses repositórios contêm informações geográficas imprecisas ou ausentes. Isso acarreta vários problemas como, por exemplo, a inviabilidade da realização de planos sistemáticos para preservar áreas para conservação de espécies ameaçadas. O problema principal para a realização desse planos é determinar com precisão a distribuição dessas espécies. Nesse contexto, o problema de pesquisa identificado é a necessidade de melhorar as informações geográficas contidas em dados sobre biodiversidade disponíveis em repositórios online. Para atacar esse problema, o *SWI Gazetteer* foi desenvolvido. Ele usa tecnologias da Web Semântica e técnicas de Recuperação de Informação Geográfica para associar coordenadas geográficas a nomes de lugares. Quando procuram por lugares, usuários podem realizar buscas semânticas que conseguem melhores resultados (em relação à precisão e cobertura de dados) que buscas tradicionais por palavras chaves. O *Gazetteer* também permite a difusão de suas informações usando formatos dos padrões *Linked Open Data*. Os resultados dos experimentos mostram que o *SWI Gazetteer* é capaz de aumentar, em até 102%, o número de registros com coordenadas geográficas em amostras representativas de repositórios de dados sobre biodiversidade bem conhecidos (como o GBIF e SpecieLink).

Palavras-chave: Web Semântica, Informação Geográfica Voluntária, Gazetteer.

ABSTRACT

CARDOSO, S. D.. **SWI: Um gazetteer interativo para dados sobre biodiversidade com suporte a web semântica**. 2015. 121 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Brazil is considered a mega-diversity country for harboring various species of flora and fauna. Therefore preserve this diversity is extremely important, because the life on the planet depends on the many ecosystems that comprise this biodiversity. Currently, several studies on how to recover and access biodiversity information are being discussed within the academic community. Various important institutions have made their biological collection records openly available in online repositories. However, the data available in these repositories contain inaccurate or missing geographic information. This leads to various problems, such as the impossibility of carrying out systematic plans to preserve areas for endangered species. The main problem in realizing these plans is to accurately determine the geographic distributions for these species. In this context, the identified research problem is the need to improve geographic information contained in biodiversity data available in the online repositories. To tackle this problem, the Semantic Web Interactive Gazetteer (SWI Gazetteer) was developed. It uses Semantic Web technologies and Geographic Information Retrieval techniques to associate geographic coordinates to place names. When searching for places, users can perform semantic searches that achieve better results (in terms of accuracy and data coverage) than traditional keyword search. The gazetteer also allows the dissemination of its information using standard Linked Open Data formats. Experiment results shown that the SWI Gazetteer is able to increase, in up to 102%, the amount of records with geographical coordinates in representative data samples from well know biodiversity sites (such as GBIF and SpeciesLink).

Key-words: Semantic Web, Volunteered geographic information, Gazetteer.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplos de tipos de ambiguidades. Fonte: (GOUVÊA, 2009)	42
Figura 2 – Ambiguidade presente nas entidades geográficas do TGN. Fonte: Gouvêa (2009)	43
Figura 3 – Expressões de Contexto em Português. Fonte: (GOUVÊA, 2009)	44
Figura 4 – Exemplo de tipos de evidências para desambiguação de topônimos. Fonte: (GOUVÊA, 2009)	44
Figura 5 – Arquitetura da Web Semântica. Fonte: (BERNERS-LEE; HENDLER; LASSILA, 2001)	48
Figura 6 – Exemplo de triplas em um documento RDF. Fonte: (SERIQUE, 2012)	52
Figura 7 – Mecanismos de um <i>Reasoner</i> . Fonte:(AMANQUI, 2014)	54
Figura 8 – Relacionamentos <i>Simple Feature</i> , GeoSPARQL. Fonte: (OGC, 2012). . . .	57
Figura 9 – Exemplos de representação de geometrias em WKT. Fonte: (KOUBARAKIS <i>et al.</i> , 2012)	60
Figura 10 – Quantidade de triplas na <i>Web of Data</i> em 2011. Fonte: (MOURA; JR., 2013)	66
Figura 11 – Número de contribuintes e precisão posicional. Fonte: (HAKLAY <i>et al.</i> , 2010)	69
Figura 12 – Número de contribuidores por erro na precisão posicional. Fonte:(HAKLAY <i>et al.</i> , 2010)	69
Figura 13 – Arquitetura do <i>SWI Gazetteer</i> . Fonte: Cardoso <i>et al.</i> (2015)	72
Figura 14 – Tela principal do <i>SWI Gazetteer</i>	75
Figura 15 – Tela mostrando a funcionalidade de busca e aprimoramento de coordenadas.	76
Figura 16 – Demonstração do método de sumarização de coordenadas geográficas, onde os valores que ocorrem com mais frequência são identificados e atribuídos para as demais coordenadas. Fonte: (CARDOSO <i>et al.</i> , 2014)	79
Figura 17 – União da ontologia LinkedGeoData com a GeoSPARQL.	80
Figura 18 – Mapeamento das informações geográficas para a especificação <i>asWKT</i> (Representação simplificada). Fonte: (CARDOSO <i>et al.</i> , 2014)	80
Figura 19 – Representação da ligação entre um indivíduo, que representa uma localidade, com outro que representa uma geometria, usando o <i>object property hasGeometry</i> da ontologia GeoSPARQL.	81
Figura 20 – Representação dos grafos do <i>SWI Gazetteer</i> evidenciando seus indivíduos, conexões entre geometrias, clusters criados e municípios da DBpedia. A Área 1 contém os indivíduos a Área 2 os clusters.	82

Figura 21 – Gráfico da qualidade dos dados do SpeciesLink.	90
Figura 22 – Gráfico da qualidade dos dados do GBIF.	91
Figura 23 – Representação das coordenadas do SpeciesLink e GBIF coletadas em Fevereiro de 2014	92
Figura 24 – Coordenadas referentes à Reserva Florestal Adolpho Ducke contidas nos dados do SpeciesLink (a) e GBIF (b). Pontos vermelhos representam coordenadas geográficas erradas para a reserva. O ponto verde representa a coordenada geográfica correta para a reserva.	92
Figura 25 – Qualidade das informações geográficas na amostra de dados referente ao SpeciesLink (novembro de 2014). Fonte: (CARDOSO <i>et al.</i> , 2015)	93
Figura 26 – Coordenadas imprecisas na amostra coletada em novembro de 2014.	94
Figura 27 – Representação dos centroides criados pelos algoritmos de agrupamento.	95
Figura 28 – Número de registros com coordenadas geográficas antes e depois da utilização do <i>SWI Gazetteer</i>	96
Figura 29 – Distribuição de registros com coordenadas para coletas de acordo com os anos. Fonte: (CARDOSO <i>et al.</i> , 2015)	97
Figura 30 – Resultados de precisão e revocação para os três <i>endpoints</i> . Fonte: (CARDOSO <i>et al.</i> , 2015)	99
Figura 31 – Resultados de precisão, revocação e F-measure para as consultas realizadas pelo algoritmo de busca semântica geoespacial, criado no módulo de reformulação de consultas.	102
Figura 32 – Média dos Resultados de precisão, revocação e F-measure para o algoritmo de busca semântica geoespacial criado no módulo de reformulação de consultas.	102

LISTA DE ALGORITMOS

Algoritmo 1 – GeoParser semântico para realizar as consultas inseridas pelo usuário utilizando palavras chave	83
--	----

LISTA DE CÓDIGOS-FONTE

Código-fonte 1 – Representação de um WKT em GeoSPARQL	58
Código-fonte 2 – Consulta para encontrar cachoeiras que estão a uma distância de 1000 metros de alguma represa.	81
Código-fonte 3 – Consulta geo-espacial semântica resultante para a informação fornecida pelo usuário “reservas próximas a 100 km da cidade de Manaus”.	85
Código-fonte 4 – Consulta por fazendas dentro de áreas protegidas.	100
Código-fonte 5 – Consulta Não Topologica utilizando a função convexHull.	119
Código-fonte 6 – Consulta Não Topologica utilizando a função buffer.	119
Código-fonte 7 – Consulta para verificar junções espaciais.	119
Código-fonte 8 – Consulta para verificar junções espaciais.	120
Código-fonte 9 – Consulta para verificar seleções espaciais.	120
Código-fonte 10 – Consulta para verificar seleções espaciais.	120

LISTA DE TABELAS

Tabela 1 – Comparativo entre os trabalhos relacionados.	65
Tabela 2 – Valores identificados pelo algoritmo de GeoParser	86
Tabela 3 – Resultados da execução das consultas geoespaciais nas <i>triple store</i> Strabom, uSeekM e Parliament.	104

LISTA DE ABREVIATURAS E SIGLAS

ACR		Ambiguidade da Classe do Referente
ARC		Ambiguidade da Referência
ART		Ambiguidade do Referente
CSV		Comma-Separated Values
GML		Geography Markup Language
GPS		Global Positioning System
HTML	...	Hypertext Transfer Protocol
INPA		Instituto Nacional de Pesquisas da Amazônia
ITN		Rede Integrada de Transporte
KB		Knowledge Base
LOD		Linked Open Data
OGC		Open Geospatial Consortium
OMS		Open Street Maps
OWL		Web Ontology Language
RDF		Resource Description Framework
RDFS		Resource Description Framework Schema
REM		Reconhecimento de Entidades Mencionadas
RIG		Recuperação de Informação Geográfica
RT		Resolução de Topônimos
SPARQL	.	Simple Protocol and RDF Query Language
SWI		Semantic Web Interative
TGN		Getty Thesaurus of Geographic Names
URI		Uniform Resources Identifiers
VGI		Volunteered Geographic Information
W3C		World Wide Web Consortium
WKT		Well-Known Text
WSD		Word Sense Disambiguation
XML		Extensible Markup Language

SUMÁRIO

1	INTRODUÇÃO	27
1.1	Justificativa	28
1.2	Objetivo	30
1.3	Organização	30
2	RECUPERAÇÃO DE INFORMAÇÃO GEOGRÁFICA	33
2.1	Utilização de VGI no contexto da biodiversidade	33
2.2	Recuperação de Informação Geográfica	36
2.3	<i>Gazetteers</i>	37
2.4	<i>Volunteered Geographic Information</i>	38
2.5	GeoTAGs	39
2.6	GeoParsing	40
2.7	GeoCoding	41
2.8	Resolução de Topônimos	41
2.8.1	<i>Estratégias para Resolução de Topônimos</i>	43
2.9	Considerações Finais	45
3	WEB SEMÂNTICA	47
3.1	Introdução	47
3.2	Ontologias	49
3.2.1	<i>Geo-Ontologias</i>	49
3.3	<i>Linked Open Data</i>	50
3.4	RDF e RDF-Schema	52
3.5	Web Ontology Language - OWL	53
3.6	<i>Reasoners</i>	54
3.7	Busca Semântica Geoespacial	55
3.8	SPARQL e GeoSPARQL	56
3.8.1	<i>Well-Known Text</i>	57
3.9	Considerações Finais	59
4	TRABALHOS RELACIONADOS	61
4.1	Estado da Arte	64
4.2	Precisão de dados utilizando VIG	67
4.3	Considerações Finais	70

5	ARQUITETURA DO SWI GAZETTEER	71
5.1	Camada de Apresentação	74
5.1.1	<i>Protótipo do SWI</i>	74
5.2	Camada de Negócios	77
5.2.1	<i>Módulo de Clustering</i>	77
5.2.2	<i>Módulo de Mapeamento</i>	78
5.2.3	<i>Módulo de Reformulação de Consultas</i>	83
5.2.4	<i>Algoritmo GeoParser</i>	83
5.3	Camada de Dados	85
5.4	Considerações Finais	87
6	EXPERIMENTOS	89
6.1	Dados utilizados	89
6.1.1	<i>Amostra de dados coletada em fevereiro de 2014</i>	90
6.1.2	<i>Amostra coletada em Novembro de 2014</i>	91
6.2	Verificação dos Dados Agrupados	93
6.3	Quantidade de registros aprimorados pelo <i>SWI Gazetteer</i>	95
6.4	Comparação com Outras Fontes na LOD	98
6.5	Avaliação do Algoritmo de <i>GeoParser</i>	100
6.6	<i>GeoSPARQL Triple Stores</i>	103
6.6.1	<i>Comparação das buscas Geoespaciais entre triplesotres</i>	104
6.7	Considerações Finais	106
7	CONCLUSÃO	107
7.1	Produções Científicas	108
7.2	Dificuldades e Limitações	108
7.3	Trabalhos Futuros	109
	REFERÊNCIAS	111
APÊNDICE A	CONSULTAS UTILIZADAS PARA TESTAR AS <i>TRIPLE STORE</i>	119

INTRODUÇÃO

Com o crescimento e popularização da Web, ela tem se tornando uma das formas mais difundidas para se obter e compartilhar dados científicos. Um grande volume de informações em praticamente todas as áreas do conhecimento vem sendo gerado constantemente. Para que esses dados sejam úteis, é necessário dispor de ferramentas para que eles sejam analisados e organizados, permitindo assim que usuários possam acessar e buscar informações relevantes para suas tarefas.

Na área de biodiversidade isso não é diferente. Atualmente, vários estudos sobre formas de uso, recuperação e acesso a informações sobre biodiversidade vêm sendo discutidos amplamente dentro das agendas acadêmicas de pesquisa. Esse grande interesse pela área de biodiversidade se dá pelo fato da mesma ser uma importante fonte econômica, abrangendo atividades agrícolas, pesqueiras, florestais e ainda por permitir várias possibilidades de pesquisas (ALHO, 2008).

Devido a esse fato, várias iniciativas de investigação para possíveis soluções de como lidar com essas informações biológicas têm surgido por meio de várias instituições de pesquisa ao redor do mundo. A essa aplicação de técnicas de informática a informações sobre biodiversidade para melhorar seu gerenciamento, apresentação, descoberta, exploração e análise se dá o nome de *Biodiversity Informatics* (BISBY, 2000). Além de apresentarem uma grande variedade de informações, dados sobre biodiversidade também têm um alto nível de complexidade, apresentando parâmetros espaço-temporais, falta de uma estrutura bem definida para representação, vocabulário expresso por uma linguagem particular aos biólogos, ausência de informações, entre outros (AMANQUI *et al.*, 2013).

Observando esse quadro e consultando informações na literatura é possível ver que os principais desafios enfrentados pela área de biodiversidade, utilizando-se da computação, são: (i) lidar com grandes volumes de informações, (ii) realizar a interoperabilidade de informações de diferentes fontes e formatos, (iii) manipular dados e imagens, (iv) manipular informações

geográficas (GIL; KOZIEVITCH; TORRES, 2010).

Com o intuito de amenizar os problemas gerados pelo grande volume de informações e a tarefa laboriosa de recuperar e acessar essas informações manualmente, repositórios de dados que integram informações primárias sobre biodiversidade de museus, herbários e coleções microbiológicas vêm sendo criados e disponibilizados gratuitamente na Internet. Entretanto uma falha importante desses repositórios é a falta de informações geográficas espaciais sobre a ocorrência de espécies em determinadas regiões. Isso acarreta vários problemas como, por exemplo, a inviabilidade de realização de planos sistemáticos para a conservação de espécies. Esses planos visam, entre outras coisas, propor os melhores locais para a conservação e manejo da biodiversidade dentre aqueles disponíveis, satisfazendo requisitos como adequação, abrangência, eficiência e representatividade (LEMES *et al.*, 2011).

A principal lacuna em realizar esse planejamento é o fato de que, para uma infinidade de espécies, as distribuições geográficas são pouco conhecidas e possuem inúmeras falhas, como, por exemplo, a imprecisão do seu georreferenciamento, problema esse conhecido como déficit Wallaceano (LEMES *et al.*, 2011).

Nesse contexto, o problema de pesquisa identificado é que as informações geográficas, contidas em dados sobre biodiversidade de repositórios *online*, estão muitas vezes faltando ou são imprecisas. Considerando esse problema, a hipótese levantada é que o desenvolvimento de um dicionário geográfico colaborativo (*Gazetteer*), utilizando tecnologias da Web Semântica, pode auxiliar na Recuperação de Informações Geográficas de dados sobre biodiversidade.

Este *Gazetteer* tornaria possível, com sua utilização, a busca por informação referente a locais de coleta de espécimes mais precisa. Isso será alcançado por meio da recuperação de coordenadas geográficas ausentes e da contribuição de usuários com novos locais e coordenadas.

1.1 Justificativa

Atualmente, diversas buscas na Web referem-se a entidades geográficas, como, por exemplo, cidades, ruas e países, demandando assim ferramentas que trabalhem para associar essas entidades a coordenadas geográficas, os *Gazetteers*. Tal necessidade também existe na área de Biodiversidade, onde é necessário verificar a ocorrência de populações de espécies ameaçadas de extinção em determinados locais para protegê-las, por meio da realização de um planejamento sistemático preciso.

Para recuperar as informações georreferenciadas sobre coleções de espécimes, é possível acessar repositórios sobre biodiversidade disponíveis na web e obter dados de coletas de espécimes. Dentre os repositórios de dados sobre biodiversidade disponíveis *online*, podemos listar dois importantes o *SpeciesLink* (2014), que possui um grande número de coleções sobre as espécies brasileiras, e o *GBIF* (2014), que contém coleções de vários países.

O [SpeciesLink \(2014\)](#) é um sistema distribuído que integra, em tempo real, informações sobre dados primários de 310 coleções estrangeiras e nacionais. Diariamente, o *SpeciesLink* trás informações estatísticas sobre seus dados, e, em abril de 2015, o mesmo registrava 7,507,883 registros *online*. Contudo, apenas 3,187,385 registros eram georreferenciados, ou seja, somente 42% de todos os dados possuíam informações sobre referências geográficas. Já o portal de dados do [GBIF \(2014\)](#) tem cerca de 416 milhões de registros sendo que 363 milhões possuem coordenadas. Seriam só 13% dos dados sem informações geográficas, mas isso apenas reflete o fato de que o GBIF contém mais registros recentes que o *SpeciesLink*. Coletas feitas antes do advento da tecnologia *Global Positioning System* GPS (portanto quase nunca georreferenciadas) são importantíssimas para se entender o passado e a evolução dos ecossistemas, deixando evidente a lacuna desse tipo de informação e a necessidade de se recuperar os registros ausentes.

A importância de se realizar o planejamento sistemático com uma taxa de precisão significativa é de que, ao demarcar uma reserva ou parque florestal para preservação de uma espécie ameaçada de extinção, é importante verificar se essa espécie realmente ocorre naquele local, pois caso ele seja demarcado como zona protegida, e a espécie não ocorra lá, esse erro de julgamento poderá levar à extinção de populações dessa espécie em regiões fora da área demarcada para preservação, onde de fato a espécie ocorre ([LEMES *et al.*, 2011](#)).

Outro cenário, que é ainda pior do que o mencionado anteriormente, é o da “extinção por ignorância”. Nesse caso, o número de locais evidenciados para a implementação do planejamento sistemático é substancialmente reduzido, pois os locais de ocorrência da espécie são desconhecidos, levando a uma menor taxa de eficiência do planejamento e possível extinção das populações desconhecidas ([LEMES *et al.*, 2011](#)).

Utilizando um *Gazetteer* colaborativo, é possível auxiliar os biólogos na realização desse planejamento sistemático com uma taxa de precisão significativa, pois usuários podem colaborar para aperfeiçoar o conteúdo das informações geográficas e realizar consultas em grandes bases de dados. Além disso, é possível associar coordenadas geográficas a informações de coletas realizadas antes do advento da tecnologia GPS, sendo assim de grande utilidade para se entender o passado e a evolução dos ecossistemas.

Além disso, a aplicação da Web Semântica no desenvolvimento desse *Gazetteer* possibilita auxiliar na resolução de diversos problemas presentes na área de Recuperação de Informações Geográficas (RIG), como a desambiguação de localidades e a realização de consultas complexas que consistam em relações, além da representação espacial ou administrativa. Esses são problemas que requerem *reasoning* sobre os locais, como, por exemplo, *to find all lakes in wildlife reserves near Seattle* ([KESSLER; JANOWICZ; BISHR, 2009](#)).

1.2 Objetivo

Este trabalho tem como objetivo construir um *Gazetteer* colaborativo baseado em tecnologias da Web Semântica, denominado *SWI Gazetteer* (*Semantic Web Interactive Gazetteer*), para auxiliar a Recuperação de Informações Geográficas sobre biodiversidade. Para isso, serão usados os dados do acervo biológico do SpeciesLink e do GBIF, sobre regiões de coletas de espécies realizadas na Amazônia, ou seja, reservas, rios, lagos, igarapés, entre outros.

O objetivo do *SWI Gazetteer* é promover acesso *online* a informações geográficas (locais) relevantes a catálogos de dados biológicos (coleções). Com a utilização do *SWI Gazetteer*, é possível aos usuários colaborarem com informações para o enriquecimento de sua base de dados, inserindo novas informações e aprimorando as coordenadas geográficas. Com o intuito de obter melhores resultados em buscas por informações geográficas, o *SWI Gazetteer* utiliza a semântica envolvida nas informações sobre localidades. Por exemplo, ao buscar pela consulta: “locais dentro do Parque Nacional do Jaú” é possível recuperar informações que não contém o nome do “Parque Nacional do Jaú”, recuperando assim rios, igarapés, lagos, entre outras localidades listadas pelos biólogos dentro do parque.

Para atingir o objetivo proposto os seguintes objetivos específicos são necessários:

1. Análise dos dados geográficos (locais referentes a reservas, rios, comunidades indígenas, entre outros) de bancos de dados sobre coleções biológicas (foram usados os repositórios *SpeciesLink* e GBIF, como exemplos representativos);
2. Desenvolvimento de uma arquitetura que possibilita o uso de tecnologias da Web Semântica para a construção do *Gazetteer*;
3. Disponibilização de um protótipo do *Gazetteer* que possibilita aos usuários buscar, inserir e recuperar informações geográficas;
4. Disponibilização de um *endpoint GeoSPARQL* para utilização de terceiros.

Tendo esses objetivos traçados, demonstramos que a utilização de tecnologias da Web Semântica como ontologias, linguagens RDF, OWL e GeoSPARQL, descritas no Capítulo 3, podem auxiliar a RIG e aos biólogos a recuperar informações geográficas sobre locais de coleta de espécimes. Como objetivo secundário, esse projeto visa fortalecer a colaboração entre a Universidade de São Paulo-Instituto de Ciências Matemáticas e de Computação (ICMC-USP) e o Instituto Nacional de Pesquisas da Amazônia (INPA).

1.3 Organização

Este trabalho está estruturado da seguinte forma:

Capítulo 2: Apresenta os conceitos referentes a recuperação de informação geográfica como, por exemplo, *Gazetteers*, *Geoparses* e VGI (*Volunteered Geographic Information*).

Capítulo 3: Apresenta a fundamentação teórica sobre Web Semântica, evidenciando os principais conceitos que serão utilizados no desenvolvimento do trabalho, como ontologias, buscas semânticas, *Linked Open Data*, entre outros.

Capítulo 4: Este capítulo apresenta os trabalhos relacionados, evidenciando o uso de ontologias, busca semântica e as principais investigações quanto ao estado da arte para o desenvolvimento de *Gazetteers*.

Capítulo 5: Apresenta a arquitetura do SWI *Gazetteer*: implementação do protótipo, mapeamento de registros geográficos para ontologias, algoritmo de busca semântica.

Capítulo 6: Este capítulo mostra os resultados obtidos com a utilização do SWI *Gazetteer*: análise dos dados utilizados, busca semântica, recuperação de coordenadas geográficas para os registros do SpeciesLink e GBIF.

Capítulo 7: Este capítulo conclui a dissertação de mestrado e exhibe as produções científicas obtidas com o desenvolvimento do trabalho, as limitações encontradas e os trabalhos futuros para aperfeiçoamento do SWI *Gazetteer*.

RECUPERAÇÃO DE INFORMAÇÃO GEOGRÁFICA

Para melhor contextualizar o problema alvo deste trabalho, é preciso compreender como Informação Geográfica Voluntária (*Volunteered Geographic Information*, VGI) pode ser utilizada no contexto da biodiversidade e como é realizado o processo de identificação de entidades geográficas. Assim, este capítulo aborda os principais conceitos referentes a utilização de VGI no contexto da biodiversidade na Seção 2.1. Na Seção 2.2 os conceitos sobre Recuperação de Informações Geográficas (RIG) são descritos. Na Seção (2.3 e 2.4) é demonstrado o conceito de *Gazetteer* e Informação Geográfica Voluntária. A Seção (2.5 , 2.6 e 2.7) descreve os processos para extrair e demarcar entidades geográficas. E por fim, na Seção 2.8 é contextualizada a Resolução de Topônimos, buscando compreender as características e estratégias associadas ao georreferenciamento de textos.

2.1 Utilização de VGI no contexto da biodiversidade

Informação Geográfica Voluntária atualmente é uma área em constante crescimento, devido à sua versatilidade. Embora o termo, descrito por Goodchild e Hill (2008), seja relativamente novo e fortemente adotado para práticas de geração de conteúdo por usuários na web, há uma longa tradição de pessoas contribuindo com informação geográfica de forma voluntária como, por exemplo, o relato da ocorrência e números de espécimes populacionais ao longo das décadas na área de biodiversidade (KLINKENBERG, 2013).

Nesse contexto, a VGI assume um importante papel na área de biodiversidade, pois permite o monitoramento de espécies que estão em extinção, podendo fornecer um apoio substancial para a pesquisa em biodiversidade (KLINKENBERG, 2013).

As questões emergentes em VGI atualmente se acercam em: i) quão confiável são suas informações, ii) como pesquisadores podem utilizá-las, iii) como VGI pode ajudar a monitorar

espécies envolvendo aspectos de distribuição, iv) como a tecnologia da informação pode ajudar na prática de VGI, v) como manter e aprimorar as bases de conhecimento geradas e vi) quão válido são as informações geográficas voluntárias para pesquisas em biodiversidade (KLINKENBERG, 2013).

As respostas para essas questões possuem várias implicações e dependem da precisão das informações fornecidas e quão útil elas serão no futuro. A precisão das informações é de extrema importância. A capacidade de mapeamento tem se alterado rapidamente com o advento das ferramentas *online* e smartphones com GPS. A obtenção de informações localmente precisas e a documentação das mesmas permitem aos pesquisadores reverem registros de dados originais, mantidos em museus e herbários, que são comumente utilizados como fontes de dados para pesquisas (KLINKENBERG, 2013).

Museus e herbários possuem centenas de registros, contendo coleções de plantas e animais. A maioria desses registros representa os esforços voluntários de centenas de pessoas ao longo dos anos. Embora muitas coleções tenham começado há mais de 100 anos, elas fornecem dados valiosos para os pesquisadores compreenderem a linha de vida das espécies (KLINKENBERG, 2013).

Nesse contexto, existem diversas razões pelas quais as coletas voluntárias são historicamente úteis para os pesquisadores, dentre elas Klinkenberg (2013) cita:

- As coleções de espécies possuem informações geográficas que permitem ao pesquisador localizar o registro em um mapa interativo;
- As informações sobre as coletas podem ser inseridas em banco de dados;
- A precisão espacial, mesmo sendo limitada, ou seja, com representações de coleta referindo-se apenas a cidades mais próximas, em geral está dentro dos limites de incerteza e precisão aceitáveis pelo pesquisador;
- A precisão temporal é alta, ou seja, a data da coleta é gravada e pode ser utilizada para pesquisas sobre comportamento e linha de vida das espécies;
- A precisão semântica acerca do nome de uma espécie é alta, uma vez que qualquer alteração no nome científico é documentada na própria coleção.

Atualmente existem várias iniciativas de VGI importantes que fornecem informações críticas para serem utilizadas por pesquisadores. Elas possuem alta “credibilidade”, pois abordam a necessidade de precisão e confiabilidade (KLINKENBERG, 2013).

Dentre essas iniciativas Klinkenberg (2013) relata os seguintes projetos:

- **eBird**: Iniciado em 2002, o projeto acumula um dos maiores recursos de dados sobre biodiversidade, em 2006 os participantes relataram mais de 4,3 milhões de observações

voluntárias de aves em toda a América do Norte. Os coordenadores do eBird tem como objetivo construir um banco de dados que permita a busca por informações passadas e verificação de tendências e comportamentos das espécies através de uma escala geográfica ampla, como, por exemplo, a linha de vida de uma determinada espécie nos últimos 100 anos.

- **Audubon Christmas Bird Count:** Iniciado em 1900, este projeto é baseado na identificação visual de observações de espécies, com dados coletados por várias equipes de voluntários. A precisão geográfica dos dados é relativamente baixa, pois um círculo de contagem de 15 km de diâmetro é utilizado. Assim como a precisão temporal, onde todas as aves avistadas em um prazo de um dia são catalogadas. No entanto, a precisão da observação é relativamente alta, pois a confirmação de observado/não observado é relatada. Ao longo dos anos, as observações veem sendo comparadas e usadas por muitos pesquisadores em trabalhos científicos.
- **Breeding Bird Survey:** É um projeto internacional que começou em 1966 para acompanhar o estado e as tendências das populações de aves norte-americanas. O projeto baseia-se principalmente na identificação do som emitido pelas espécies e os resultados são normalmente coletados por um ou dois indivíduos que percorrem mais de 4100 rotas de pesquisa com 24,5 quilômetros de extensão cada, onde as espécies dos EUA e Canadá são observadas. Embora os resultados sejam de precisão espacial média, as informações sobre as coletas são altamente confiáveis, uma vez que os pesquisadores possuem mecanismos para identificar o canto dos pássaros. Os resultados desse projeto veem sendo comparados ao longo dos anos e utilizados em vários trabalhos de pesquisa.
- **E-Flora BC: *The Atlas of the Plants of British Columbia*:** E-Flora BC é um projeto que descreve um atlas regional biogeográfico e ecológico sobre as espécies de plantas, líquens e fungos em *British Columbia*, Canadá. As páginas do atlas incluem mapeamento, ilustrações, descrições de espécies e informações ecológicas. O E-Flora BC faz parte de uma iniciativa de VGI, pois conta com informações coletadas por botânicos que, em grande parte, são voluntários e contribuem com o envio de fotos dos espécimes, para serem armazenadas em bancos de dados.

Atualmente, mesmo com a perda de espécies em ritmo acelerado, diversas espécies ainda não estão documentadas e novas espécies continuam a ser descobertas. Esse fato ressalta a necessidade crítica da coleta de dados para estudos em biodiversidade. Além disso, existe a necessidade de se compreender mais sobre a abundância das espécies e as tendências de cada população. Desse modo, é necessário que um número significativo de voluntários se tornem aliados dos biólogos na investigação das mudanças ocorridas na biodiversidade, verificando se os principais componentes de precisão, confiabilidade e durabilidade são cumpridos para tornar a prática de VGI viável (KLINKENBERG, 2013).

Para auxiliar a prática de VGI e fornecer mecanismos para suprir as necessidades de precisão, confiabilidade e durabilidade das informações, a área de Recuperação de Informação Geográfica é a base usada para promover suporte para a representação das informações geográficas, temporais e espaciais.

2.2 Recuperação de Informação Geográfica

A Recuperação de Informações na Web é uma tarefa laboriosa, mas por meio de mecanismos de busca automáticos é possível obter informações rapidamente. Tais mecanismos consultam os sites e analisam seu conteúdo, classificando de maneira geral as páginas por meio de métodos como o *PageRank* (PAGE *et al.*, 1998) ou focados em conteúdos, como por exemplo, o CiteSeer (GILES; BOLLACKER; LAWRENCE, 1998).

No entanto, os sistemas de busca tradicionais não apresentam suporte para informações contextuais, como, por exemplo, a verificação de informações semânticas analisando as ligações entre seus *hiperlinks* ou a localidade referenciada pelos textos. Isso impede uma análise com precisão de informações dentro ou próximas a determinadas regiões geográficas (BUYUKOKKTEN *et al.*, 1999) tornando difícil a recuperação de informações relevantes para o usuário (JONES; ALANI; TUDHOPE, 2001).

Uma grande quantidade da informação, presente na Web, tem conteúdos descrevendo espaços geográficos e, como consequência, muitos usuários gostariam de especificar nomes de lugares como parte de suas consultas (JONES; ALANI; TUDHOPE, 2001). Atualmente, um conjunto significativo das consultas submetidas aos mecanismos de busca na Web se referem a nomes de lugares e acidentes geográficos (ex: “praia”, “serra”); tais pesquisas representam um subconjunto significativo das consultas submetidas aos mecanismos de busca (GOUVÊA, 2009). *Web sites* que contém, por exemplo, informações sobre restaurantes, teatros e cinemas são mais interessantes para usuários vizinhos dessas localidades (BUYUKOKKTEN *et al.*, 1999).

No entanto, a forma semi-estruturada da Web dificulta o acesso a esse tipo de informação. Jones e Purves (2008) evidenciam as principais dificuldades no uso da Web como fonte de informações geográficas:

- Descrições por meio de linguagem natural para representar contextos geográficos;
- Nomes de lugares podem ser homônimos e confundidos com nomes de organizações, pessoas, construções e ruas;
- Dificuldade em interpretar relações espaciais como, por exemplo, “próximo”, “ao oeste”, dentre outras;
- Construção de *ranking* específico para definição da relevância geográfica.

Com o intuito de solucionar esses problemas e utilizar as vantagens relacionadas à descrição semântica e geográfica das informações na Web, diversas técnicas têm sido desenvolvidas, tanto em ambiente acadêmico como comercial, com o objetivo de acessar recursos que possuem contextos geográficos (JONES *et al.*, 2002; GEY *et al.*, 2006).

Seguindo esse contexto Larson (1996) define que a RIG está relacionada com a Recuperação de Informação Geográfica em dois tipos: i) buscas determinísticas, onde é possível encontrar todos os documentos relacionados à determinada coordenada geográfica e ii) consultas probabilísticas, onde é possível, realizar buscas por pontos de referencia como, por exemplo, encontrar todos os rios próximos a Manaus.

Diferente dos modelos tradicionais de recuperação de informação, como o modelo Vetorial (SALTON, 1989), que associa documentos por meio de índices de termos utilizando o conteúdo presente no texto para recuperar e organizar documentos de acordo com a posição e frequência das palavras, na RIG os termos extraídos devem ser relacionados a descrições espaciais, ou seja, a entidades geometricamente definidas e localizadas no espaço (GOUVÊA, 2009).

O componente principal em um Sistema de Recuperação de Informações Geográficas é o *Gazetteer*. Ele busca estruturar relacionamentos semânticos presentes nas localidades, sendo organizado geralmente em uma hierárquica com o tipo, a localização geográfica e outras descrições relacionadas (HILL, 2000).

A partir da utilização de *Gazetteers* é possível, por exemplo, reconhecer palavras e frases que se relacionam a alguma localidade, podendo o mesmo auxiliar na desambiguação desses termos, resolvendo as localidades ambíguas e descrevendo a localidade a ser referenciada de forma precisa.

2.3 Gazetteers

Um *Gazetteer* é um diretório geográfico que associa nomes de lugares a coordenadas geográficas. Eles são geralmente implementados como diretórios que contem triplas de nomes de lugares (*place names*) (N), entidades (*feature types*) como, por exemplo, serra, praia, (T) e representações geométricas (*geographic footprints*) (F) com coordenadas geográficas (KESSLER; JANOWICZ; BISHR, 2009).

Eles oferecem funções para mapear nomes de lugares para representações geométricas (N - F) e nomes de lugares para tipos de lugares (N - T). *Gazetteers* são fundamentais para aplicações de serviços de mapeamento baseados na web, motores de busca espaciais e *geo-parsing*, pois, quando trabalham em conjunto com esses serviços, eles provêem recursos para acesso e manipulação de dados geográficos, permitindo que consultas sejam realizadas por RIGs (KESSLER; JANOWICZ; BISHR, 2009).

Gazetteers são comumente construídos a partir da extração de informações de textos. No entanto, atualmente uma nova forma de contribuir com informações geográficas vem sendo amplamente difundida, prática essa que se dá o nome de *Volunteered Geographic Information*.

2.4 *Volunteered Geographic Information*

A variedade e quantidade da informação geográfica disponível na *Web* para uso comum tem crescido rapidamente. Desde a introdução do Google Earth, em 2004, e do Google Maps, em 2006, tem aumentado rapidamente o interesse em se criar ferramentas que permitam pessoas localizar pontos de interesse em aplicativos que forneçam serviços geográficos, tais como, encontrar endereços e selecionar rotas para veículos em mapas, especificamente em áreas urbanas (JR.; VELLOZO; PINHEIRO, 2013).

Com o advento da Web 2.0, muitos exemplos de situações em que usuários podem contribuir com atualizações e novos registros, através de *Web Sites* e aplicativos como Geonames, Wikimapia, Open Street Map e o Flickr, vêm aparecendo. Esse fenômeno é comumente chamado de Informação Geográfica Voluntária (JR.; VELLOZO; PINHEIRO, 2013).

Dispositivos habilitados com GPS tornam possível o georreferenciamento e permitem aos usuários conectar facilmente coordenadas de latitude e longitude a fotos, vídeos, tweets, entre outros. Com o auxílio de tais dispositivos, usuários podem criar recursos geoespaciais mais complexos utilizando ferramentas como o Open Street Map, Wikimapia, ou o Google Maps.

Atualmente, o Open Street Map é a principal fonte de VGI. Ele permite que usuários contribuam com conteúdo geográfico expressando coordenadas geográficas e criando anotações para marcação de recursos, possibilitando que cidadãos, cientistas ou pessoas especializadas contribuam com informações para o enriquecimento dos *Gazetteers* (BEARD, 2012).

A VGI tem mostrado um grande potencial para diversos contextos, abrangendo áreas que necessitam tanto de informações geograficamente mais detalhadas, como as oferecidas por organizações nacionais tais como o IBGE, como informações mais comuns, como locais e anotações sobre paisagens, locais de lazer (por exemplo, locais para caminhadas, ciclismo, canoagem, pesca, etc), flora e fauna locais, eventos locais, entre outros conteúdos (BEARD, 2012).

Além disso, a utilização de VGI permite que coleções de informações geográficas, que estão em constante crescimento, sejam manipuladas por vários usuários, como, por exemplo, na área de biologia, onde os dados de coletas biológicas podem ser atualizados por todos os envolvidos e não apenas por uma pessoa ou um especialista, reduzindo assim a sobrecarga de trabalho e a quantidade de dados nulos ou ausentes (BEARD, 2012).

A prática de VGI permite que o conteúdo inserido e recuperado em um *Gazetteer* seja feito por meio de marcações ou GeoTAGs, realizadas por usuários.

2.5 GeoTAGs

Humanos adicionam diversos itens, como fotos ou livros, em suas coleções individuais por meio de esquemas de ordenação individual com o objetivo de agrupar itens semelhantes e manter itens distintos à parte, para facilitar a recuperação de informação. Um exemplo dessa forma de classificação pode ser representada pela forma que organizamos nossos livros em uma estante, separando-os de acordo com vários critérios, como tema, idade, espessura, ou até mesmo cor. Tais preferências por organizar itens em formato individual também ocorrem em coleções virtuais (KEßLER *et al.*, 2009).

O uso de *tags* - palavras ou combinações de palavras por pessoas permite a associação de itens virtuais. Essa abordagem também é aceita para classificar itens, como por exemplo, em uma prateleira virtual. No entanto, tais marcações, *tags*, podem possuir diversos significados de pessoa para pessoa, necessitando assim de uma definição (KEßLER *et al.*, 2009).

A definição formal de uma *tag*, portanto, deve incluir tanto o usuário quanto a informação do objeto a ser etiquetado. Gruber (2005) sugere para modelar a marcação (*tagging*) como um processo onde $Tagging = \{L, U, I, S\}$, o qual estabelece uma relação imediata ente o rótulo (L), vindo de um usuário (U), o vocabulário associado à informação do Item (I) e uma fonte (S) que permite o compartilhamento entre aplicações.

O processo de marcação estabiliza a relação entre o usuário, a informação do item e o rótulo. Caso a informação do item represente um ou mais objetos geográficos, o rótulo associado pode se referir tanto à dimensão do objeto representado quanto à sua semântica, incluindo um nome próprio do indivíduo, categoria ou a sua extensão espaço-temporal (nomeação, por exemplo, a região que contém) (KEßLER *et al.*, 2009).

Uma GeoTAG estende a noção de *tag* por adicionar uma localização explícita no espaço e tempo para a informação do item. No caso de fotografias digitais, uma marcação de tempo com a data de criação é comumente adicionada pela câmera automaticamente. As coordenadas geográficas são fornecidas por módulos GPS embutidos ou adicionadas manualmente por um usuário (KEßLER *et al.*, 2009).

Com base nessa definição de marcação, proposta por Gruber (2005), o rótulo T e as coordenadas C para a relação, omitindo a fonte S, são adicionados. Dessa forma, o processo de marcação geográfica pode ser representado como: $Geotagging = \{L, U, C, I, T\}$ (KEßLER *et al.*, 2009).

Além da utilização de GeoTAGs para inserir e recuperar nomes de lugares, é possível extrair informações geográficas de registros que já possuem essa informação no texto. Para realizar a extração desses nomes é preciso realizar o GeoParsing de tais entidades geográficas.

2.6 GeoParsing

Para compreender o processo de extração do contexto geográfico de textos é necessário analisar seu conteúdo e como as relações espaciais são determinadas para especificar lugares ou espaços geográficos na Terra. Comumente as relações espaciais são divididas em três categorias Egenhofer (2002):

- Topológicas: Relações que descrevem os conceitos de vizinhança, incidência e sobreposição sem variar com escalas e rotações (ex. *dentro de*).
- Métricas: Relações que descrevem os termos das direções como, por exemplo, *ao norte* e *ao sul*, e que descrevem distâncias como, por exemplo, *perto de* e *próximo de*.
- De Ordem: Relações que expressam a ordem, total ou parcial, dos objetos espaciais, como, por exemplo, *em frente a* e *acima de*.

Para denotar relacionamentos espaciais entre objetos no espaço é necessário utilizar *spatial reasoning*. Esse tipo de raciocínio permite fazer previsões e diagnósticos usando um subconjunto conhecido de relações espaciais. Podendo o mesmo ser classificado como qualitativo ou quantitativo, de acordo com o tipo de informação utilizada no processo de *reasoning* (GOUVÊA, 2009).

O método de inferência espacial quantitativo realiza a diferenciação de diversas relações espaciais, por exemplo, relações topológicas e métricas, e é tipicamente formalizado usando um sistema de coordenadas geográficas. Esse tipo de processamento é claramente distinto da forma como as pessoas interpretam a representação de lugares, por formatos textuais. (BORGES *et al.*, 2007)

As pessoas se comunicam e pensam a respeito de representações geográficas por meio de termos vagos, que são imprecisos ou probabilísticos, como, por exemplo, “centro da cidade”, “perto de”, “ao lado de”. Raramente referências como “o restaurante está a 35,93 metros a oeste” são utilizadas para representar a direção de algum lugar (GOUVÊA, 2009).

Mesmo que vagos e imprecisos, esses termos necessitam ter suas representações de entidades (*features*) (T) associadas aos *Gazetteers*. Exemplos de *features* são “Reserva Adolpho Ducke”, “Ilha da Marchantaria”, entre outros, os quais são representados pelos nomes de lugares (*place names*) (N) “Adolpho Ducke” e “Marchantaria” e pelas *features* de tipo “Reserva” e “Ilha” respectivamente (GOUVÊA, 2009).

Após a verificação das entidades geográficas contidas nos textos é necessário reconhecer o correto contexto geográfico representado por elas. Para realizar tal tarefa é necessário fazer o GeoCoding das entidades geográficas.

2.7 GeoCoding

Segundo [Davis, Fonseca e Borges \(2003\)](#) a fase de geocodificação, ou seja, *geocoding*, é a localização de pontos na superfície da Terra a partir de informações alfanuméricas, envolvendo as seguintes etapas:

- *Parsing*: Esta fase consiste na análise dos caracteres que contém informações de lugares e sua transformação em uma tupla de dados estruturada e padronizada para ser utilizada na fase de *matching*.
- *Matching*: Tenta encontrar, na base de dados, informações de referência geográfica mais precisas que podem ser positivamente associadas com um dado lugar.
- *Locating*: Após determinar, na fase anterior, quais são as informações de referência geográfica mais precisas, coordenadas geográficas (*footprint*) são atribuídas ao lugar.

Um *footprint* é uma representação geométrica de alguma entidade geográfica, sendo expresso em coordenadas do tipo, latitude e longitude. A localização referida por um *footprint* é geralmente definida segundo [Larson e Frontiera \(2004\)](#) por:

- Pontos: Mantém um senso geral da localização sem extensões ou formas.
- Polígonos: Ocorre a identificação da localização, extensão e forma com grau variável de precisão.

Para se realizar a conversão dos nomes de localidades para a sua respectiva posição espacial, torna-se necessário a utilização de *Gazetteers*. Associados aos *Gazetteers*, há a necessidade de adoção de estratégias com objetivo de desambiguar localidades, pois mesmo nomes iguais podem representar espaços geográficos distintos, ou podem conter uma referência para mais que uma localidade. Para isso é necessário realizar a Resolução de Topônimos.

2.8 Resolução de Topônimos

Devido à multiplicidade de representações para espaços geográficos existentes em formas textuais, que podem conter homônimos, sinônimos, entre outros fatores, surge a necessidade de lidar com a ambiguidade ocasionada no texto para se identificar corretamente um topônimo.

Segundo [Garbin e Mani \(2005\)](#), um topônimo é o nome de uma entidade geográfica na superfície da Terra que pode ser representada por alguma especificação geométrica como, por exemplo, uma linha, um ponto ou polígono.

No aspecto geográfico, por exemplo, um topônimo pode representar uma entidade geopolítica e mudar de nome ou extensão ao longo do tempo. Já no aspecto textual, lugares

		Localização	Ambiguidade
Geo/Geo	ARC	Pelotas, RS Rio de Janeiro, RJ	Princesa do Sul Rio
	ART	Bom Jesus, RS Belém, PA	Bom Jesus, RN Belém, PB
Geo/Não Geo	ACR	Pelotas, RS Serra, ES	Rua Pelotas Serra (Governador de São Paulo)

Figura 1 – Exemplos de tipos de ambiguidades. Fonte: (GOUVÊA, 2009)

distintos na Terra podem ter o mesmo nome. Para lidar com esse problema, Leidner (2004) define a área de Resolução de Topônimos (RT), a qual tem o objetivo de possibilitar o mapeamento correto das localidades referenciadas pelos textos, a partir da resolução dos vários tipos de ambiguidades relacionadas a elas.

Para entender como aplicar a RT em textos, é necessário compreender como pessoas referenciam lugares e quais tipos de ambiguidades podem acontecer. Normalmente, nomes de lugares são descritos em uma linguagem natural denotados de expressões vagas com relações imprecisas, como por exemplo, “Norte do Rio Grande do Norte”, “Centro de Manaus” ou “próximo a Rio Grande” (GOUVÊA, 2009).

Os trabalhos presentes na área de RT em geral buscam desambiguar textos considerando seu aspecto linguístico e a extensão espacial, não compreendendo o nível de refinamento necessário para o georreferenciamento. Portanto, do ponto de vista linguístico, a principal dificuldade para desambiguação é a superação das seguintes ambiguidades representadas na Figura 1 (GOUVÊA, 2009):

- Ambiguidades Geo/Geo:
 - Ambiguidade da Referência (ARC - *Reference Ambiguity*) – quando uma determinada localidade pode ser referenciada por vários nomes diferentes.
 - Ambiguidade do Referente (ART - *Referent Ambiguity*) – quando o nome pode ser usado para referenciar outras localidades.
- Ambiguidades Geo/Não Geo:
 - Ambiguidade da Classe do Referente (ACR - *Referent Class Ambiguity*) – quando o nome pode ser usado para referenciar outros tipos de entidades.

Esses tipos de ambiguidade ocorrem com frequência nos nomes de lugares para representar entidades geográficas. Nesse contexto, Smith e Mann (2003) quantificaram o tipo e o grau de ambiguidade dos topônimos examinando a ontologia geográfica TGN (*Getty Thesaurus of*

Continente	% Lugares com Múltiplos Nomes (Sinônimos)	% Nomes com Múltiplos Lugares (Homônimos)
América do Norte e Central	11.5	57.1
Oceania	6.9	29.2
América do Sul	11.6	25.0
Ásia	32.7	20.3
África	27.0	18.2
Europa	18.2	16.6

Figura 2 – Ambiguidade presente nas entidades geográficas do TGN. Fonte: [Gouvêa \(2009\)](#)

Geographic Names), conforme ilustrado na Figura 2, ficando assim evidente que é necessário tratar a ambiguidade dos nomes de lugares ([GOUVÊA, 2009](#)).

A utilização de dicionários toponímicos (*Gazetteers*), para a resolução deste tipo de ambiguidade, demanda portanto a correta representação dos Indicadores de Localidade e sua associação com as localidades identificadas por eles. Para isso é necessário aplicar estratégias para resolução de topônimos.

2.8.1 Estratégias para Resolução de Topônimos

A Resolução de Topônimos compreende dois processos, identificação e desambiguação. Inicialmente as referências geográficas devem ser identificadas de acordo com o seu tipo, por exemplo, cidade ou país, de forma a resolver o problema de ambiguidades *Geo/Não-Geo*. Posteriormente é necessário desambiguar as referências *Geo-Geo*, para que os topônimos sejam descritos com uma extensão espacial única, para assim serem associados a sua localização geográfica em um *Gazetteer* ([GOUVÊA, 2009](#)).

Para cada uma dessas etapas, são necessárias técnicas específicas. A etapa de identificação está relacionada ao Reconhecimento de Entidades Mencionadas (REM) ([LEVELING; HARTRUMPF, 2007](#)), que tem como propósito representar cada palavra ou grupo de palavras em uma determinada categoria pré-definida. Já a segunda etapa, desambiguação, necessita de técnicas de Desambiguação Lexical de Sentido (*Word Sense Disambiguation* ou WSD) que busca considerar tanto ambiguidades relacionadas à homonímia quanto à polissemia, sendo possível assim resolver localidades de acordo com um contexto particular, por exemplo, “Banco de areia” ou “Banco do Brasil” ([GOUVÊA, 2009](#)).

De forma geral, as estratégias de identificação são agrupadas dentro das seguintes categorias:

Baseada em Regras (*Rule-Based*) – Para reconhecer e desambiguar as entidades geográficas utilizam-se regras e heurísticas definidas manualmente, tratando os dois tipos de evidências exemplificadas na Figura 3 [McDonald \(1996\)](#):

Internas (Pelo Próprio Nome incluindo ou não separadores)	(1ª Palavra de Alguma Frase) NOME CIDADE Ex: Pelotas a cidade do doce. PAL. MAIÚSCULO + {/, -} + Sigla de Algum Estado Ex: Pelotas/RS ou Pelotas-RS
Externas (Por Expressões de Contexto relacionadas ao Topônimo)	{em, de, na cidade de, no município de} + PAL. MAIÚSCULO Ex: na cidade de Pelotas {região metropolitana de} + PAL. MAIÚSCULO Ex: região metropolitana de Porto Alegre

Figura 3 – Expressões de Contexto em Português. Fonte: (GOUVÊA, 2009)

Tipo de Expressão	Expressão
Identificadores de Contexto	cidade, município, distrito, rua, avenida, rio, ilha, montanha, vale, país, continente, zona, região, condado, freguesia, deserto, província, povoado, aldeia, monte, vila, república, península
Localização	fora de, nos arredores de, dentro de, entre, em, acima, ao longo, atrás, acima, ao lado, à esquerda, à direita
Distância Relativa	adjacente, longe de, perto de, próximo de
Orientação	leste, norte, sul, oeste, oriente, ocidente, sudeste, sudoeste, nordeste, noroeste
Outras Expressões	“cidades como”, “e outras cidades”, “cidades incluindo”, “cidades especialmente”, “uma das cidades”, “cidades tais como”

Figura 4 – Exemplo de tipos de evidências para desambiguação de topônimos. Fonte: (GOUVÊA, 2009)

- Internas: baseia-se no próprio nome da localidade, por exemplo, procurar no conteúdo do texto todos os nomes de cidades, ou padrões relacionados a letras maiúsculas, siglas, entre outros.
- Externas: busca identificar expressões relacionadas ao contexto no qual a localidade aparece, por exemplo, na cidade de, no município de, etc.). Uma visão geral dos tipos de expressões de contexto utilizadas para desambiguar localidades externas, utilizadas por Gouvêa (2009), pode ser vista na Figura 4.

O problema de se utilizar a abordagem baseada em regras é a necessidade de defini-las manualmente e o fato de que regras fixas são úteis apenas para domínios específicos (GOUVÊA, 2009). Para uma melhor eficácia, outra estratégia que capture automaticamente palavras e expressões para desambiguação de localidades, como a abordagem denominada *Data-Driven*, pode ser utilizada. Essa técnica utiliza aprendizado de máquina e aplica análises estatísticas,

em um corpus de treino com o objetivo de identificar regras e classificadores úteis para a desambiguação de topônimos. Contudo, os custos associados à construção de um corpus de treino com qualidade e ampla cobertura podem ser muito significativos, necessitando abordagens mais aprimoradas onde outros métodos podem ser aplicados, como por exemplo, *bootstrapping*, onde existe uma pequena quantidade de dados classificados com diversos dados não-classificados (GOUVÊA, 2009).

Como o foco do *SWI Gazetteer*, proposto neste trabalho, representa um domínio específico, neste caso os dados geográficos referente a coleções biológicas, será utilizado, por simplicidade, a primeira estratégia de resolução de topônimos, que é a abordagem baseada em regras. Devido ao fato do domínio ser mais restrito, a extração das regras para reconhecer e desambiguar as entidades geográficas pode ser realizada em um tempo razoável. Casos mais genéricos, como, por exemplo, textos sobre notícias, podem exigir uma quantidade maior de esforço.

2.9 Considerações Finais

Este capítulo apresentou a utilização da prática de informação geográfica voluntária na área de biodiversidade e os conceitos bases da área de Recuperação de Informação Geográfica, para o desenvolvimento do *SWI Gazetteer*. Inicialmente foi feita uma breve introdução sobre RIG na Web, evidenciando a forma como as pessoas representam e se referem a conteúdos geográficos. Logo após, o conceito de *Gazetteer* foi explicado e, em seguida, a nova prática para contribuir com atualizações e novos registros para *Gazetteers*, também denominada de *Volunteered Geographic Information*, foi descrita.

As demais seções compreenderam questões sobre a inserção de conteúdos geográficos por usuários (utilizando GeoTAGs), a extração de lugares realizando o GeoParsing de textos e a codificação de suas coordenadas no processo de GeoCoding.

Por fim, para dar suporte a todos esses processos, foram discutidas a necessidade e as formas de como aplicar a Resolução de Topônimos em textos para desenvolver o *Gazetter*.

Estes conceitos serão utilizados no trabalho proposto para desenvolver o *SWI Gazetteer*. Sendo necessário incorporar a prática de VIG para permitir a usuários e biólogos inserirem informações por meio de GeoTAGs e desambiguar os nomes de locais contidos nos dados das coleções biológicas disponíveis no SpeciesLink e no GBIF.

WEB SEMÂNTICA

Visando melhorar a forma de representação da Web atual, também denominada de Web sintática, surge a proposta da Web Semântica, projetada pelo *World Wide Web Consortium* (W3C). A web semântica é definida pelo W3C como a Web da nova geração onde as informações podem ser usadas tanto por humanos quanto por máquinas. Nesse último caso, a informação não é somente utilizada pela máquina para fins de apresentação de conteúdo, mas também para automação, integração, inferências e reutilização em aplicativos (BOLEY; WAGNER, 2001).

Para os seres humanos, compreender a semântica de alguma palavra ou sinal não é uma tarefa extraordinária. O cérebro humano associa os conceitos acumulados ao longo dos anos, para compreender as informações repassadas. No entanto, para as máquinas, semântica não deve ser relacionada à compreensão humana, mas sim, à inferência e dedução de informações que estão em um formato estruturado (AMANQUI, 2014).

Devido à necessidade de se abordar a semântica das localidades para o processo de desambiguação e processamento das consultas, que geralmente são feitas por meio de termos vagos, tecnologias da Web Semântica serão usadas no *Gazetteer*. Essas tecnologias ajudarão a representar entidades geográficas ao se realizar um consultas ao *Gazetteer* e fazer a desambiguação e classificação de locais.

3.1 Introdução

Para atingir os objetivos da web semântica, onde os dados devem ser “compreendidos” tanto por máquinas e humanos, Berners-Lee, Hendler e Lassila (2001) propõem um modelo em camadas para a arquitetura da Web Semântica também conhecida como “bolo de noiva”, a arquitetura lógica dessa arquitetura é apresentada na Figura 5.

A ideia central da arquitetura proposta por Berners-Lee, Hendler e Lassila (2001) é de que cada camada seja melhorada e adicionada gradativamente para trazer novas funcionalidades,

Figura 5 – Arquitetura da Web Semântica. Fonte: (BERNERS-LEE; HENDLER; LASSILA, 2001)

juntamente com aumento de expressividade para representar os dados e possibilidades de se realizar inferências e autenticações. Essa arquitetura promove o reuso utilizando algo já validado, desse modo minimiza a necessidade de se criar novas camadas a partir do início. Essa abordagem se dá pelo fato da estrutura da Web permitir pequenas mudanças, que são fáceis de se implementar, ao contrário de modificações radicais, que são muito difíceis (SERIQUE, 2012).

O modelo em camadas da arquitetura da Web Semântica pode ser descrito da seguinte forma:

- Base: compreendida pelo padrão Unicode, que permite aos computadores representar e manipular caracteres de várias línguas, e URI (*Uniform Resources Identifiers*), que permite identificar unicamente recursos disponíveis na Web. Ambos padrões tem como objetivo facilitar o intercâmbio de dados na Web.
- Camada Sintática: é formada pela linguagem XML (*Extensible Markup Language*), que permite realizar marcações para descrever informações por meio de tags que podem ser facilmente analisadas por máquinas.
- Camada de Dados: utiliza a camada sintática como suporte e provê um modelo lógico para descrição de dados, onde é possível descrever informações sobre um recurso específico.
- Camada Ontologia: estende a camada anterior e provê um maior nível de expressividade para a semântica das informações.
- Camada Lógica: permite definir regras lógicas para inferir novos conhecimentos.

- Camada de Prova e Confiança: provê mecanismos para avaliar o nível de confiabilidade das informações.
- Camada de Assinatura Digital: permite incorporar mecanismos de segurança para garantir a confiabilidade da informação.

Para este trabalho, serão utilizadas as tecnologias da Web Semântica referentes a: ontologias (Seção 3.2), as melhores práticas para representar os dados na Web Semântica, chamadas de *Linked Open Data* (Seção 3.3), linguagens de construção: OWL, XML, RDF, RDFS (Seções 3.4, 3.5 e 3.6), Busca Semântica GeoEspacial (Seção 3.7) e GeoSPARQL (Seção 3.8).

3.2 Ontologias

O termo “ontologia” em sua origem significa o estudo da existência (GUIZZARDI, 2010). De acordo com Smith *et al.* (2007), a ontologia é um ramo da filosofia que procura compreender os tipos, propriedades, estruturas dos objetos e suas relações em todas as áreas da realidade.

Embora o termo ontologia esteja bem difundido atualmente e sua importância seja reconhecida, não existe um consenso sobre sua exata definição. Segundo Gruber (2005) uma ontologia é uma especificação formal e explícita de conceitos que compartilham o mesmo domínio de informação. Já Noy e McGuinness (2001) relatam que ontologias surgiram para compartilhar, organizar e especificar conhecimentos de um determinado domínio.

Uma ontologia, de forma geral, pode ser descrita como uma especificação por meio de componentes básicos: classes, relações, axiomas e instâncias, expressas por meio de uma linguagem de construção (SERIQUE, 2012).

O principal papel das ontologias na Web Semântica é explicitar o vocabulário utilizado e servir como padrão para compartilhamento de informação entre agentes, softwares e aplicações (SERIQUE, 2012).

Com o intuito de fornecer mais expressividade ao contexto geográfico, será abordado a seguir o conceito de Geo-Ontologias para representar um certo domínio.

3.2.1 Geo-Ontologias

Uma ontologia geográfica, ou geo-ontologia, descreve entidades geográficas de duas formas distintas: como geo-campo e geo-objeto. Quando uma entidade relacionada está focada em dados espaciais como um conjunto de distribuição contínua ela é denominada como geocampo, ao contrario de um geo-objeto, que trabalha com dados espaciais discretos e identificáveis espalhados pelo mundo (GIMENEZ; TANAKA; BAIAO, 2013).

Ao contrário das ontologias gerais, geo-ontologias devem representar relacionamentos espaciais, fatores espaciais, fatos físicos, coleções de dados, modelos de computação geoespacial

(GIMENEZ; TANAKA; BAIÃO, 2013). Unificando tais características referentes à geoinformações com conhecimento de especialistas no domínio. Wang, Li e A (2008) definem a seguinte fórmula para descrever geo-ontologias:

$$Geo - ontologia = \{C, R, A, X, I\}$$

Nesta fórmula um conjunto de conceitos referentes a um objeto geográfico é representado por C, a relação e as definições desse conjunto são representadas por R, os atributos do objeto geográfico são representados por A, os axiomas e suas regras de restrição sobre os conceitos são representados por X e as definições de um objeto material são dadas por I.

Na abordagem utilizada por Giunchiglia *et al.* (2012), uma geo-ontologia pode ser considerada como um *facet*, ou seja, uma hierarquia homogênea de termos descrevendo um aspecto do domínio, onde cada termo na hierarquia denota um diferente conceito, como especificado pela formula abaixo:

$$Espaço = \{C, E, R, A\}$$

Nessa abordagem, Giunchiglia *et al.* (2012) procuram descrever o espaço em termos de objetos do mundo real, onde cada conceito descreve uma classe, uma entidade, uma relação ou atributo. Conforme apresentado pela fórmula descrita por Giunchiglia *et al.* (2012), C define as classes (entidades do mundo real, como Lagos, Reservas, entre outros), E define o conjunto de entidades (as instâncias das classes em C), R define um conjunto de relações (relacionamento classe entidade, por exemplo, *part-of*, *instance-of*, entre outros) e A define o conjunto de atributos (descrições quantitativas e qualitativas das entidades).

Neste trabalho será utilizada a definição de espaço proposta por Giunchiglia *et al.* (2012) para manipular as localidades que irão compor o *Gazetteer*. Desse modo, um tipo de lugar, por exemplo, uma reserva, será representado pela classe C (Reserva), que irá possuir um conjunto de entidades E (Egler, Campina, Aldopho Ducke) contendo diversas relações R (*part-of*, *near-of*, etc) com um conjunto de atributos para descrever suas características, representado por A (coordenadas geográficas como, por exemplo, pontos ou polígonos).

O objetivo de representar o *Gazetteer* utilizando essa abordagem de representação de Geo-ontologias, proposta por Giunchiglia *et al.* (2012), é a disponibilização dos dados do *Gazetteer* em um padrão de uso comum que vem sendo largamente utilizado na *Web of Data*. Esse padrão é conhecido como *Linked Open Data*.

3.3 *Linked Open Data*

Tradicionalmente, os dados publicados na Internet são disponibilizados em documentos em diversos formatos como CSV, XML ou hipertextos como HTML. Tais formatos não pre-

servam a semântica dos dados para que os mesmos sejam compreensíveis por computadores e humanos (HEATH; BIZER, 2011).

A estrutura convencional de hipertextos, adotada pela web, faz com que o relacionamento entre documentos e dados contidos nos mesmos seja distinto e implícito. Tal fato ocorre pois as páginas HTML atualmente utilizadas não são suficientemente expressivas para permitir a disponibilização individual de entidades, descrevendo apenas um documento em particular a ser conectado por *links* inseridos por usuários (HEATH; BIZER, 2011). No entanto, atualmente a web está evoluindo para um espaço global de informação onde documentos e dados estão interconectados. O padrão utilizado para o relacionamento desses dados é conhecido como *Linked Open Data* (HEATH; BIZER, 2011).

O termo *Linked Open Data* (LOD) se refere às melhores práticas para publicação e conexão estruturada de dados na Web. Tais práticas vêm sendo adotadas por diversos trabalhos nos últimos anos, como por exemplo, Auer, Lehmann e Hellmann (2009), Parundekar, Knoblock e Ambite (2010), Amanqui *et al.* (2013) levando a criação de um espaço global de dados com bilhões de registros – a *Web of Data*.

A adoção de LOD tem levado a extensão da Web como um espaço global de dados onde diversos domínios estão interconectados, como, por exemplo, pessoas, companhias, livros, publicações científicas, dados científicos, entre outros (HEATH; BIZER, 2011).

Para conectar todas essas informações, Berners-Lee *et al.* (2006) definem um conjunto de regras para publicar os dados na Web de forma que todos os dados publicados sejam parte de um único espaço de dados:

1. Usar URIs para nomear coisas;
2. Usar HTTP URIs para que pessoas possam procurar coisas;
3. Usar padrões RDF, SPARQL para estruturar os dados e consultá-los;
4. Incluir links para outras URIs, sendo assim possível descobrir e relacionar os dados entre si.

Esse conjunto de regras, também conhecido como “princípios de *Linked Open Data*”, provê regras básicas para publicar e conectar dados usando a infraestrutura e arquitetura já existentes na Web atual (HEATH; BIZER, 2011). Neste trabalho serão seguidos os padrões sugeridos por Berners-Lee *et al.* (2006) para publicar e acessar dados na Web; esses padrões servirão para definir a forma de estruturação dos dados do *Gazetteer*. Utilizando esses padrões será possível contribuir para o enriquecimento da *Web of Data*, além de seguir a proposta de colocar os dados referentes a coleções de espécies da Amazônia em um formato utilizado pela Web Semântica, como especificado em Santos *et al.* (2011) .

Figura 6 – Exemplo de triplas em um documento RDF. Fonte: (SERIQUE, 2012)

Para publicar os dados no formato de LOD é necessário utilizar tecnologias da Web Semântica como RDF/RDFS, SPARQL.

3.4 RDF e RDF-Schema

Para representar dados e documentos contidos na web de forma compreensível tanto para computadores quanto para humanos, o W3C, órgão padronizador da web, definiu as linguagens RDF e RDFS.

O *Resource Description Framework* ou RDF é baseado em XML e serve como base para o processamento de metadados permitindo que computadores representem e compartilhem dados semânticos na Web (SILVA; LIMA, 2004).

O RDF possui diversas aplicações na Web, sendo utilizado na busca de recursos para tornar mecanismos de busca mais eficientes, em bibliotecas virtuais na descrição de conteúdos, no comércio eletrônico, em web sites particulares, entre outros. O RDF em si é basicamente uma linguagem simples que permite relacionamentos entre informações (SILVA; LIMA, 2004).

Os documentos RDF são compostos por triplas que são declarações do tipo objeto-atributo-valor, onde cada tripla expressa os termos de um recurso Web (sujeito), suas propriedades (predicado) e o valor da propriedade (objeto) (SILVA; LIMA, 2004).

A Figura 6 mostra um exemplo de código RDF, utilizando XML, e seu respectivo diagrama. Nesse exemplo, o sujeito definido pela URI `http://dme.um.pt/jcardoso/` possui diversos predicados, como, por exemplo, a propriedade título, que possui um objeto com o valor Jorge Cardoso Web Page.

O RDF define um modelo para descrição de relações entre objetos em termos de propriedades e valores, porém não define mecanismos para descrever tais propriedades e relações entre estas propriedades com outros recursos. Para preencher essa lacuna, o W3C especificou a RDF Schema (RDFS), essa linguagem é uma extensão do RDF responsável por prover mecanismos para declaração de relações entre objetos (SILVA; LIMA, 2004).

O RDFS descreve regras para o uso das propriedades do RDF, definindo um vocabulário de domínio onde é possível representar hierarquias entre classes e relacionamentos. Uma im-

portante característica do RDFS é que suas propriedades são definidas separadamente de suas classes. Assim, uma propriedade pode ser declarada e usada com uma, ou múltiplas classes a qualquer momento (SILVA; LIMA, 2004).

Neste trabalho, o formato de dados RDFS será utilizado para descrever os dados do *Gazetteer*. No entanto, somente a representação em RDF não fornece meios para descrever classes, relacionamentos, igualdades ou desigualdades, restrições de cardinalidade e características das propriedades.

3.5 Web Ontology Language - OWL

A *Web Ontology Language* (OWL) é uma linguagem criada pelo W3C para descrever ontologias. Com a OWL, é possível criar representações de dados e documentos na web semântica representando explicitamente os vocabulários de conceitos, taxonomias e relacionamentos de um domínio de conhecimento. A OWL é uma extensão do vocabulário RDF/RDFS sendo assim expressiva para descrever classes, relacionamentos, restrições de cardinalidade, igualdade ou desigualdade de classes e características de propriedades (BECHHOFFER *et al.*, 2004).

Os dados e documentos, representados como instâncias de classes em OWL, podem ser submetidos a mecanismos de inferência, onde é possível realizar classificações de forma automática dos conceitos da ontologia gerada, verificar inconsistências em hierarquias e heranças incorretas e checar as restrições sobre os valores das propriedades e sua cardinalidade (BECHHOFFER *et al.*, 2004).

Atualmente, a OWL encontra-se em sua segunda versão (OWL 2) que é subdividida em três sub-linguagens, OWL EL, OWL QL e OWL RL, cada uma com um poder de expressividade diferente, porém todas permitem a criação de ontologias (BECHHOFFER *et al.*, 2004).

OWL EL é baseado na família EL++ de lógica descritiva, sendo que sua utilização é particularmente útil em aplicações que contém um grande número de propriedades e classes para definir uma ontologia. Além disso, a OWL EL utiliza um padrão comum para ontologias com conceitos e planejamento, ou seja, a combinação de conjunções e qualidades existenciais (BECHHOFFER *et al.*, 2004).

Já a OWL QL, estruturada a partir da família DL-Lite de lógica de descrição (*Description Logic*), foi criada para permitir o raciocínio (*reasoning*) eficiente de grandes quantidades de dados estruturados de acordo com esquemas relativamente simples. Ela fornece vários recursos para capturar modelos conceituais, tais como diagramas de classe UML, diagramas de Entidade de Relacionamento, e esquemas de banco de dados (BECHHOFFER *et al.*, 2004).

Por fim, a OWL RL foi criada para dar suporte a aplicações que exigem raciocínio escalável em troca de alguma restrição de poder expressivo. Através de um subconjunto sintático, é possível implementar o raciocínio (*reasoning*) usando tecnologias baseadas em regras que

Figura 7 – Mecanismos de um *Reasoner*. Fonte: (AMANQUI, 2014)

geralmente são mais escaláveis e fáceis de implementar (BECHHOFFER *et al.*, 2004).

As linguagens da Web Semântica possuem uma semântica formal que possibilita a realização de inferências usando *reasoners* automatizados. Os *Reasoners* são uma das mais importantes ferramentas para realizar inferência em consultas de dados.

3.6 *Reasoners*

Um *reasoner* é um programa que realiza inferências lógicas a partir de um conjunto explicitamente afirmado de fatos ou axiomas. Ele normalmente fornece suporte automatizado para tarefas de raciocínio como, por exemplo, classificação, depuração e consulta (DENTLER *et al.*, 2011).

A finalidade dos *reasoners* na Web Semântica é realizar inferência sobre os dados, utilizando geralmente tecnologias baseadas em regras, com o objetivo de obter novas informações. Para realizar o processo de inferência, os *reasoners* utilizam um motor de inferência associado com um conjunto de regras descritas em linguagens como OWL, RDF, RDFS (AMANQUI, 2014).

Geralmente *reasoners* contêm dois mecanismos internos para processar bases de conhecimento (KB, *knowledge base*), são eles Tbox e Abox (DENTLER *et al.*, 2011). A Figura 7 apresenta ambos mecanismos associados a uma KB:

A seguir ambos mecanismos são descritos conforme Amanqui (2014):

- Tbox: Também conhecido como Parte Terminológica, contém um conjunto de declarações e axiomas para descrever a estrutura de um domínio. Além disso, o Tbox contém frases que descrevem as relações entre conceitos, de forma hierárquica como, por exemplo, a representação a seguir:

$$\text{Homem} \equiv \text{Pessoa} \cap \text{Macho}$$

Nesse exemplo, o conceito (classe) Homem é declarado como a interseção dos conceitos Pessoa e Macho, ou seja, o conceito Homem é formado por todos os indivíduos que compõe as classes Pessoa e Macho simultaneamente.

- Abox: Também conhecido como Parte Declarativa, contém um conjunto de sentenças assertivas sobre os indivíduos que indicam as relações entre indivíduos e conceitos, ou seja, a qual hierarquia os indivíduos pertencem. Um exemplo dessa relação é mostrado a seguir, onde é afirmado que Joaquim é um indivíduo da classe Homem:

$$\text{Homem} \equiv \text{Joaquim}$$

Vale destacar que uma das funções dos *Reasoners* é apoiar o processo de consultas da Web Semântica, considerando que o processo de busca necessita de algum tipo de inferência. No entanto, para realizar buscas, é preciso utilizar alguma linguagem de busca para representar qual tipo de informação é desejada.

3.7 Busca Semântica Geoespacial

Segundo [Egenhofer \(2002\)](#), a busca semântica geoespacial é a forma de processar e solicitar informações envolvendo diferentes tipos de dados geoespaciais.

Na web semântica, o processo de recuperação de informações geoespaciais é definido como Busca Semântica Geoespacial. Esse tipo de busca permite a recuperação de informações que contenham algum significado geográfico, ou seja, que representem alguma entidade ou fenômeno associado a uma localização na Terra ([EGENHOFER, 2002](#)).

Para realizar as consultas, expressas por algum usuário, o mecanismo de busca semântica utiliza-se de ontologias como base para realizar suas inferências por meio de *reasoners*, permitindo aos usuários recuperar os dados desejados utilizando a semântica presente nas expressões ou termos da pesquisa ([EGENHOFER, 2002](#)).

Ao realizar uma busca Semântica Geoespacial, o mecanismo de geo-parsing utiliza uma Geo-ontologia para desambiguar localidades com a finalidade de encontrar resultados mais relevantes. Dessa forma, locais ou documentos relacionados a eles são apresentados como resultado, considerando o sentido do termo utilizado dentro do contexto semântico do documento ([EGENHOFER, 2002](#)).

A utilização de Busca Semântica Geoespacial possibilita encontrar lugares ou informações que necessitam da realização de consultas complexas, ou seja, que requerem *reasoning* sobre os locais, como, por exemplo, “Lagos próximos a Reserva Adolpho Ducke” ([KESSLER; JANOWICZ; BISHR, 2009](#)).

3.8 SPARQL e GeoSPARQL

Para se realizar buscas Semânticas na Web, geralmente é utilizada a linguagem SPARQL (*Simple Protocol and RDF Query Language*). As consultas são realizadas em um repositório de dados RDF (geralmente uma *triplestore*) por meio de um *endpoint* SPARQL. Um endpoint SPARQL é uma interface que permite a humanos e aplicações acessarem os dados armazenados em uma *triplestore*. Esse endpoint pode ser usado por outros sistemas para requisitar dados ou por aplicações (*stand-alone* ou Web) onde humanos podem criar e executar consultas.

Conceitualmente, as consultas realizadas em SPARQL fazem a correspondência de padrões em grafos. Os padrões são como declarações RDF, mas que podem conter nomes de variáveis no lugar de alguns nós (recursos) ou links (propriedades) (AMANQUI, 2014). Todas as triplas RDF que se encaixarem nesse padrão são retornadas como resultado da consulta.

Devido a necessidade de se abordar o relacionamento semântico das entidades geográficas trabalhos como (BATTLE; KOLAS, 2012; BERETA; SMEROS; KOUBARAKIS, 2013) têm feito o uso de GeoSPARQL: uma extensão da linguagem SPARQL para consulta em dados RDF que abordam o domínio geoespacial. Inferência geoespacial é fundamental para inúmeros domínios como, por exemplo, planejamento de transportes, hidrologia, biologia, entre outros. E, nesses domínios, muitas aplicações utilizam-se de bancos de dados relacionais com extensões espaciais.

No entanto, realizar esse tipo de tarefa utilizando banco de dados relacionais gera vários problemas como, por exemplo, consultas com muitas junções entre as entidades, consultas com propriedades variáveis, entre outros. Com a popularização da web semântica e a utilização de soluções em RDF para realizar junções e inferências geoespaciais entre bancos de dados, foi possível solucionar tais problemas. Porém, como vários grupos criaram diversas soluções com fins variados para esse tipo de tarefa, notou-se a necessidade de estabelecimento de um padrão (BATTLE; KOLAS, 2012).

Devido a esse fato, em Setembro de 2012 o Open Geospatial Consortium (OGC) lançou a especificação GeoSPARQL que é uma linguagem de busca geográfica para dados RDF que define uma pequena ontologia para representar geometrias e uma série de predicados e funções de consulta SPARQL (OGC, 2012).

Segundo OGC (2012) GeoSPARQL é especificada em três componentes principais:

1. Definição de um vocabulário para representar relacionamentos topológicos, geometrias e características;
2. Um conjunto de funções específicas de domínio espaciais para uso em consultas SPARQL.
3. Um conjunto de regras de transformação de consulta.

Figura 8 – Relacionamentos *Simple Feature*, GeoSPARQL. Fonte: (OGC, 2012).

Os relacionamentos topológicos disponíveis no GeoSPARQL estão divididos em três grupos: *Egenhofer*, *RCC8* e *Simple Feature*. Um exemplo das funções espaciais implementadas pelo grupo *Simple Feature* (como definido pela (OGC, 2012)) é apresentado na Figura 8. Por exemplo, para saber se dois objetos, A e B, se sobrepõem, podemos usar a função `geo:overlaps`.

A utilização de GeoSPARQL é importante pois segue os padrões propostos pela OGC. A OGC é um conjunto de 472 companhias, agências governamentais e universidades que tem como objetivo desenvolver e publicar padrões geoespaciais, sendo os mesmos largamente utilizados em sistemas de recuperação de dados geoespaciais tais como Battle e Kolas (2012) e Bereta, Smeros e Koubarakis (2013).

Para representar as coordenadas geográficas e então permitir que consultas sejam realizadas pela linguagem GeoSPARQL, a OGC define a representação de geometrias por meio da especificação *Well-Known Text*.

3.8.1 Well-Known Text

Well-Known Text (WKT) é um padrão largamente utilizado pela OGC para representar geometrias. O WKT pode ser usado para representar sistemas de referência de coordenadas geográficas, geometrias e as transformações entre os sistemas de referência de coordenadas geográficas.

A representação WKT é descrita na “*OpenGIS Simple Feature Access - Part 1: Common Architecture*” como uma especificação da OGC que segue os mesmos padrões definidos na ISO 19125-1. Esse padrão consiste na especificação de como representar e manipular uma *Simple Feature*, ou seja, uma propriedade com todos os atributos espaciais descritos por partes, por uma linha reta ou uma interpolação planar entre conjuntos de pontos (Koubarakis et al., 2012).

Geometrias em WKT são escritas em 1 ou 2 dimensões, podendo as mesmas existirem

nos planos *R2*, *R3* ou *R4*. Geometrias pertencentes ao *R2* consistem de pontos com coordenadas *x* e *y*, por exemplo, *POINT(1 2)* na sintaxe definida pelo WKT. Geometrias existentes no *R3* possuem pontos com as coordenadas *x,y* e *z* ou *x, y* e *m*, onde *m* é um valor de medida. Por exemplo, o ponto *POINT(-3 -60 32)* pode ser usado para representar a temperatura da cidade de Manaus em 32° graus Celsius. Nesse exemplo, -3 é a latitude da cidade de Manaus, -60 é a longitude e 32 é a temperatura. As geometrias que existem no *R4* tem pontos com coordenadas *x, y, z* e *m* com semântica similar (KOUBARAKIS *et al.*, 2012).

Segundo Koubarakis *et al.* (2012), as geometrias representadas usando WKT têm as seguintes propriedades :

1. Todas as Geometrias são topologicamente fechadas, isso significa que todos os pontos que integram o limite da Geometria são considerados pertencentes à geometria, mesmo que eles não possam ser representados explicitamente pela geometria;
2. Todas as coordenadas com uma geometria estão em um mesmo sistema de referência de coordenadas geográficas;
3. Para objetos Geométricos existentes no *R3* e *R4*, as operações espaciais apenas trabalham em seu “mapa geométrico”, isso é, são projetados somente no *R2*. Portanto, os valores *z* e *m* não refletem nos cálculos quando são executadas funções espaciais para verificar interseções, junções, ou na geração de novos valores geométricos por meio das funções *buffer*, *convexHull*, entre outras. Assim, a especificação WKT trabalha somente com representações geométricas bidimensionais, mas permite associar valores a essas geometrias.

A representação do padrão WKT na linguagem de busca geográfica GeoSPARQL segue os padrões apresentados anteriormente e acrescenta algumas particularidades como, por exemplo, a necessidade de se criar um tipo de dados juntamente com a sua URI correspondente. Um exemplo de como representar um WKT em GeoSPARQL é exibido no Código-fonte 1. Nesse exemplo, o ponto WKT assume o tipo *wktLiteral* para que possa ser reconhecido pelo GeoSPARQL e uma URI é inserida para determinar o formato da coordenada geográfica juntamente com sua representação.

É importante notar que, conforme se altere essa URI, por exemplo, utilizando a representação especificada pela URI: <http://www.opengis.net/def/crs/EPSG/0/4326>, a representação de latitude e longitude dos pontos do polígono é alterada. Desse modo, o exemplo descrito no código-fonte 1 teria as seguintes coordenadas: “*Polygon((34.1 -83.6, 34.1 -83.2, 34.5 -83.2, 34.5 -83.6, 34.1 -83.6))*”.

Código-fonte 1: Representação de um WKT em GeoSPARQL

```
1 <geo:asWKT rdf:datatype= "http://www.opengis.net/ont/geosparql#  
  wktLiteral">  
2   <![CDATA[ <http://www.opengis.net/def/crs/OGC/1.3/CRS84>  
3   Polygon((-83.6 34.1, -83.2 34.1, -83.2 34.5, -83.6 34.5, -83.6  
  34.1))]>  
4 </geo:asWKT>
```

Alguns exemplos da representação de um WKT são mostrados na Figura 9. Sua sintaxe detalhada de representação pode ser encontrada em (OGC, 2011). Além disso, é possível representar as coordenadas geográficas utilizando a representação GML (*Geography Markup Language*).

Neste trabalho, a representação WKT será utilizada para representar as informações geográficas dos lugares que compõe o *Gazetteer*. Essa escolha se deve ao fato desse padrão ser largamente utilizado pela OGC, promovendo assim fácil interoperabilidade entre diversas ferramentas (GARBUS; KYZIRAKOS; KOUBARAKIS, 2013).

3.9 Considerações Finais

Este capítulo apresentou os conceitos necessários para compreender o uso de web semântica para o desenvolvimento do *SWI Gazetteer*. Inicialmente foram discutidos os conceitos de ontologias e geo-ontologias que possibilitam a *Gazetteers* realizarem a desambiguação de locais. Além disso, o padrão *Linked Open Data*, que padroniza a disponibilização de arquivos em formatos RDF, por meio de endpoints SPARQL, foi discutido no contexto do trabalho proposto.

Devido ao contexto geográfico deste trabalho, a linguagem de busca GeoSPARQL foi escolhida para implementar consultas espaciais que possuem significado semântico como, por exemplo, "Rios próximos a Manaus". Além disso, foi descrito o formato de dados geométricos WKT, para a representação de pontos, linhas, polígonos do *Gazetteer*.

Devido ao recente uso da web semântica para desenvolver *Gazetteers* e a adoção da prática de VIG para se obter coordenadas geográficas, o próximo capítulo tem como objetivo listar os principais trabalhos que abordam os conceitos de web semântica e as práticas de VIG. Além disso, são descritos os principais desafios para a próxima geração de *Gazetteers* que utilizam a web semântica.

Geometry type	WKT representation	Geometry
Point	<code>Point(5 5)</code>
LineString	<code>LineString(5 5,28 7,44 14,47 35,40 40,20 30)</code>
Polygon	<code>Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5))</code>
Polygon	<code>Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5), (28 29,14.5 11,26.5 12,37.5 20,28 29))</code>
MultiPoint	<code>MultiPoint((5 5),(28 7),(44 14), (47 35),(40 40),(20 30))</code>
Geometry Collection	<code>GeometryCollection(Point(5 35), LineString(3 10,5 25,15 35,20 37,30 40), Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5), (28 29,14.5 11,26.5 12,37.5 20,28 29)))</code>

Figura 9 – Exemplos de representação de geometrias em WKT. Fonte: (KOUBARAKIS *et al.*, 2012)

TRABALHOS RELACIONADOS

Com intuito de se verificar na literatura recente os trabalhos na área de RIG que envolvem a construção e disponibilização de *Gazetteers* semânticos e utilização de *Volunteered Geographic Information*, foi realizado um mapeamento sistemático.

Para realizar esse mapeamento, os termos “*Semantic Gazetteer*” e “*Volunteered Geographic Information*” foram utilizados para pesquisar os trabalhos, como resultado 143 artigos foram encontrados. Após uma análise entre a data de publicação e título do artigo, 88 trabalhos publicados entre 2009 e 2014 foram selecionados para análise dos resumos. Ao se verificar o resumo de cada um desses trabalhos, observando se o contexto de *Gazetteer* Semântico e VIG eram abordados, esse número foi reduzido para 38 trabalhos.

Dentre esses 38 trabalhos, foi realizada uma análise do conteúdo abordado por cada um, evidenciando a utilização de ontologias, contexto da biodiversidade e práticas de VIG, sendo que sete trabalhos foram escolhidos por terem similaridade com o projeto do *SWI Gazetteer*. A seguir, esses trabalhos são descritos e por fim uma comparação com o trabalho proposto é realizada.

- [Jr., Vellozo e Pinheiro \(2013\)](#) propõem um *Framework* para a criação de aplicações que lidam com o contexto de *Volunteered Geographic Information*. Esse *Framework* lista como implementar aplicações para VGIs usando interfaces Web e plataformas móveis. A arquitetura do *Framework* proposto tem como objetivo encapsular uma estrutura básica em blocos que pode ser facilmente estendida e reutilizada em novos projetos com a intenção de estimular novos aplicativos VGI dirigidos aos eventos do cotidiano, com objetivo de motivar pessoas a contribuírem.
- O *Gazetteer*, proposto por [Moura e Jr. \(2012\)](#), realiza a expansão do conteúdo de um *Gazetteer* referente a nomes hidrográficos dos dados da Agência Nacional de Águas (ANA), utilizando a mesma técnica proposta por [Machado et al. \(2011\)](#) para realizar a

extração dos nomes e, como resultado, gera um corpus contendo 5384 nomes de rios e 670 bacias.

- O OntoGazetteer, proposto por [Machado et al. \(2011\)](#), utiliza conceitos de ontologia para mapear dados de diversas fontes de dados como, por exemplo, portais de notícias (UAI-Minas) e dados geográficos oficiais (IBGE). O OntoGazetteer foi implementado em Java, usando PostGIS como sistema de gerenciamento de banco de dados espaciais e utiliza expressões regulares para extrair as localidades dos portais de notícias. O *Gazetteer* proposto por [Machado et al. \(2011\)](#) apenas trata a desambiguação de lugares, tendo uma taxa de acerto de 80%.
- O FODGS, proposto por [Peng et al. \(2010\)](#), é um *Gazetteer* que utiliza folksonomias e ontologias para descrever nomes de lugares por meio de tags. Nesse trabalho, cada nome de lugar possui um único *footprint* e uma tag associada a ele. [Peng et al. \(2010\)](#) utiliza os dados referentes às localidades chinesas e implementa seu *Gazetteer* utilizando um Java Web Server para permitir ao usuário realizar buscas em um sistema baseado em XML. Esse sistema codifica as consultas em formato SPARQL para permitir o acesso aos dados, que são armazenados de forma híbrida no *triple store*-Jena TDB e em um SGDB.
- O trabalho proposto por [Gil, Kozevitch e Torres \(2010\)](#) demonstra um Serviço de Anotação Geográfica para Sistemas de Biodiversidade. Nesse trabalho são coletadas informações sobre a localização de armadilhas posicionadas por biólogos para capturar borboletas, e então é realizada uma correlação entre os dados de coletas e os locais em que os espécimes foram encontrados. Nesse trabalho é utilizado o SGBD PostgreSQL para armazenar as informações dos espécimes. Para acessar os dados, os usuários utilizam um Web Service onde é possível inserir as consultas em uma interface ou utilizar mapas de navegação que são implementados usando a API OpenLayers.
- O DIGMAP, proposto por [Manguinhas, Martins e Borbinha \(2009\)](#), é um *Gazetteer* que permite a consulta de nomes de locais históricos. Para realizar as buscas, o usuário utiliza um Web Service que permite consultar locais e exportá-los para XML. O DIGMAP foi implementado em Java e utiliza o SGBD Apache Derby para armazenar suas informações. Para realizar a extração dos nomes de lugares, [Manguinhas, Martins e Borbinha \(2009\)](#) utilizam repositórios como o Geonames e a Wikipédia. Sua desambiguação é feita por meio de ontologias.
- O KIDGS, proposto por [Liu et al. \(2009\)](#), é um *Gazetteer* que permite que usuários consultem nomes de lugares, sendo eles vagos ou não, como, por exemplo, “north of Beijing”, “airport in Beijing”, “Beijing”. Usuários podem montar *queries* utilizando arquivos XML no formato sujeito, predicado, objeto e enviá-las para um Web-Service que processa essa *query* em uma implementação interna do KIDGS. Os dados do KIDGS são armazenados no SGBD PostGIS em 3 tabelas. Uma tabela é encarregada de armazenar

os nomes de lugares e as outras duas são encarregadas de armazenar os metadados para descrever as informações de cada propriedade. Além disso, essas tabelas também têm como objetivo conectar o PostGIS com a ontologia em OWL. No entanto, os autores do KIDGS não especificam qual é a ontologia utilizada.

Ao comparar o trabalho proposto com esses sete trabalhos, as seguintes diferenças foram evidenciadas:

- O trabalho proposto [Jr., Vellozo e Pinheiro \(2013\)](#) relata apenas uma proposta para implementação de um *Framework* para desenvolver futuros *Gazetteers* baseados em VGI. No entanto, seu *Framework* não demonstra nenhum suporte para utilização de web semântica, seja na desambiguação de lugares, realização de buscas, inserção de lugares ou utilização dos princípios de *Linked Open Data*. Isso torna o uso do *Framework* inviável para implementar os futuros desafios da área de RIG listados na seção 4.1.
- Já o trabalho proposto por [Moura e Jr. \(2012\)](#) e o OntoGazetteer proposto por [Machado et al. \(2011\)](#) somente utilizam conceitos de ontologia exemplificados por [Machado et al. \(2011\)](#) para realizar a expansão dos nomes de um *Gazetteer* e a extração de localidades por meio de notícias presentes em web sites. Ambos não fornecem nenhuma base para aplicação de VGI, buscas semânticas ou *Linked Open Data*. Além disso, [Moura e Jr. \(2012\)](#) e [Machado et al. \(2011\)](#) não deixam explícita a ontologia utilizada para desambiguar os nomes de ambos os *Gazetteers*, somente é demonstrado seu modelo.
- No FODGS, proposto por [Peng et al. \(2010\)](#), os autores definem que um lugar possui apenas um único *footprint*. Esse tipo de representação limita a representação espacial de um objeto, visto que diversos nomes podem ter o mesmo *footprint*. Além disso, os lugares do *Gazetteer* não contêm o que eles representam. Por exemplo, Cidade de São Carlos é relatado no *Gazetteer* somente como São Carlos, assim não é possível verificar se esse lugar é uma cidade, um estado, um bairro ou uma rua. Os autores definem esse tipo de representação pois as localidades chinesas não utilizam sufixos de representação, ou seja, Cidade, Parque, Lago, entre outros. Outro fator limitante desse *Gazetteer* é que suas relações espaciais, ou seja, qual ponto pertence a um determinado local, são computadas *offline*, assim sempre que seus dados são atualizados é necessário refazer todo o processo de relações espaciais. Essa abordagem é inviável para um *Gazetteer* colaborativo, pois ele teria que ser re-estruturado a cada nova mudança. Por fim, os autores ainda utilizam bancos de dados relacionais para armazenar informações estatísticas tais como censo, impossibilitando que informações sobre a densidade demográfica de um local fiquem disponíveis para pesquisas que envolvam *Linked Open Data*.
- O trabalho proposto por [Gil, Kozevitch e Torres \(2010\)](#) aborda o contexto biológico, no entanto coordenadas geográficas são manipuladas em um ambiente controlado onde as

informações são precisas. Isso, entretanto, não ocorre com os dados do SWI *Gazetteer*, onde as informações são fornecidas por usuários e sujeitas a erros, necessitando assim de mecanismos para melhorar a acurácia das mesmas. Além disso, o trabalho proposto por [Gil, Kozevitch e Torres \(2010\)](#) não fornece nenhum suporte para utilização de web semântica, seja na desambiguação de lugares, realização de buscas, inserção de lugares ou utilização dos princípios de Linked Open Data.

- O DIGMAP proposto por [Manguinhas, Martins e Borbinha \(2009\)](#), assim como os demais trabalhos, não possui suporte para utilização de VGI. A semântica empregada no DIGMAP vem da utilização da ontologia GeoNetPT para a desambiguação das localidades, desse modo, o mesmo não possui suporte para buscas semânticas e *Linked Open Data*.
- O KIDGS proposto por [Liu et al. \(2009\)](#) utiliza ontologias para desambiguar as localidades e para o processo de busca semântica, no entanto, suas buscas não trabalham com coordenadas geográficas. Além disso, o KIDGS não fornece suporte para utilização de VGI e Linked Open Data, pois somente seus dados estão em RDF e não existe nenhuma serviço SPARQL para requisitar esses dados.

A Tabela 1 demonstra, de forma simplificada, as diferenças relatadas entre o trabalho proposto e os analisados, evidenciando quais trabalhos utilizam ontologias, VGI, busca semântica e Linked Open Data. A seção 4.1 descreve o estado da arte e os principais desafios na utilização desses conceitos em *Gazetteers*.

4.1 Estado da Arte

Dados geoespaciais vêm sendo comumente organizados e disponibilizados no formato de *Linked Open Data*, permitindo assim o uso de informações georreferenciáveis externas para o enriquecimento de aplicações que envolvem dados geográficos. No entanto, ainda são relatados na literatura problemas quanto à utilização dessas fontes de dados, como por exemplo, a confiabilidade dos dados gerados por voluntários e a da desambiguação de nomes para construção de *Gazetteers*. Nesse contexto, é possível listar os desafios a serem abordados pela próxima geração de *Gazetteers*:

Coleta e Integração: A manutenção e atualização dos dados contidos em *Gazetteers* são tarefas muito custosas, principalmente se alguma intervenção manual é necessária. Enquanto grande parte dos *Gazetteers* atuais são mantidos por pequenos grupos bem definidos, a próxima geração de *Gazetteers* deve também incorporar a utilização de contribuições voluntárias ([GOODCHILD; HILL, 2008](#)). Atualmente, isso vem sendo possível por meio de *Gazetteers* que incorporam a utilização de VGI, como por exemplo, Wikimapia, Geonames e Open Street Maps. No entanto, é preciso integrar seus dados e informações voluntárias à *Web of Data*. A VGI, associada com as práticas de *Linked Open Data*, possibilita essa integração. Porém, como

Trabalho	Utiliza Ontologias	Utiliza VGI	Buscas Semânticas	Linked Open Data
Framework proposto por Jr., Vellozo e Pinheiro (2013)	Não	Apenas descreve um modelo de implementação para VGI	Não	Não
<i>Gazetteer</i> proposto por Moura e Jr. (2012)	Somente para desambiguação	Não	Não	Não
OntoGazetteer Machado et al. (2011)	Somente para desambiguação	Não	Não	Não
FODGS Peng et al. (2010)	Sim, mas não realiza desambiguação	Não	Em parte, não permite utilizar coordenadas geográficas	Em parte, dados como censo ainda estão em bases relacionais
Trabalho proposto por Gil, Kozievitch e Torres (2010)	Não	Não	Não	Não
DIGMAP Manguinhas, Martins e Borbinha (2009)	Somente para desambiguação	Não	Não	Não
KIDGS Liu et al. (2009)	Sim	Não	Em parte, não permite utilizar coordenadas geográficas	Em parte, pois não segue todos os princípios, apenas seus dados estão em RDF
<i>SWI Gazetteer</i>	Sim	Sim	Sim	Sim

Tabela 1 – Comparativo entre os trabalhos relacionados.

Domínio	Número de datasets	Triplas	%	Outlinks	%
Mídia	25	1.841.852.061	5,82 %	50.440.705	10,01 %
Geográfico	31	6.145.532.484	19,43 %	35.812.328	7,11 %
Governamental	49	13.315.009.400	42,09 %	19.343.519	3,84 %
Publicações	87	2.950.720.693	9,33 %	139.925.218	27,76 %
Múltiplos domínios	41	4.184.635.715	13,23 %	63.183.065	12,54 %
Ciências da vida	41	3.036.336.004	9,60 %	191.844.090	38,06 %
Conteúdo gerado por usuários	20	134.127.413	0,42 %	3.449.143	0,68 %
TOTAL	295	31.634.213.770		503.998.829	

Figura 10 – Quantidade de triplas na *Web of Data* em 2011. Fonte: (MOURA; JR., 2013)

apresentado na Figura 10, os dados gerados por usuários ainda representam menos de 1% das triplas que compõe a *Web of Data*. A próxima geração de *Gazetteers* deve alavancar a informação geográfica voluntária juntamente com *Linked Open Data* para melhorar a frequência, completude e atualização das informações. No entanto, o desafio consiste em encontrar mecanismos robustos para coleta de VGI que filtrem o ruído inerente a tais informações criado na web social (MOURA; JR., 2013).

Recuperação, Busca e Navegação: A utilização de listas de localidades ou *thesaurus* semi-formais presentes na maioria dos *Gazetteers* atuais, dificulta a descoberta, navegação e a realização de consultas complexas, pois não suporta inferências lógicas como, por exemplo, buscas que necessitam da realização de *subsumption* (verificar se um conceito B é um subconjunto de um conceito A). Além disso, o tipo de hierarquia para representar as localidades e as relações entre elas são dadas de forma estática e pouco intuitiva. Dessa forma, a utilização de ontologias possibilitará solucionar esse tipo de problema, pois tornará possível introduzir relações arbitrárias e descobrir novas relações implícitas, como, por exemplo, semelhanças entre os tipos de localidades, coerência dos dados e a possibilidade de realizar consultas complexas (MOURA; JR., 2013). Dentro desse contexto, Kessler, Janowicz e Bishr (2009) relatam que a construção de ontologias de domínio para a próxima geração de *Gazetteers* será uma tarefa desafiadora, no entanto, as mesmas serão necessárias para compor os padrões de infraestrutura semântica disponibilizados pela OGC.

Resolução de entidades e desambiguação: Em diversos *Gazetteers*, ou em problemas de pesquisa onde nomes são utilizados como referência de lugares, existem problemas de desambiguação. Comumente a RIG trabalha com os tipos de ambiguidade geo/geo, ou seja, um lugar possui o mesmo nome que outro, geo/não-geo, lugares usam nomes próprios ou de outras entidades para serem referenciados (GOUVÊA, 2009). Nesse contexto, a utilização de *Gazetteers* no reconhecimento de possíveis referências a lugares pode auxiliar a resolver a ambiguidade. No entanto, é necessário verificar outros fatores, como a coocorrência de nomes de lugares relacionados ou a presença de termos fortemente associados citados em textos. Fontes, como a DBPedia, podem fornecer elementos para resolver o problema, pois, a partir do seu uso, é

possível ampliar o conteúdo das bases de referência, dando clareza para caracterizar os lugares citados e caracterização semântica de relacionamentos expressos nas triplas RDF (MOURA; JR., 2013).

Dados Temporais: Atualmente os *Gazetteers* representam uma tripla de lugares contendo nome, tipo da entidade e *footprint*. No entanto, pode ser necessário o acréscimo de um atributo temporal ao conjunto de dados espaciais. Um exemplo representativo dessa necessidade é a associação de dados censitários sobre a mudança dos polígonos referentes aos municípios brasileiros, visto que suas fronteiras municipais mudam ao longo do tempo. Em *Gazetteers*, como o GeoNames, por exemplo, cidades possuem um atributo referente à população, mas não existe uma data indicando a validade da informação, logo esse dado se torna incompleto e, dependendo do estudo que está sendo utilizado, pode ser inútil. A utilização de triplas em RDF é capaz de auxiliar na semântica dos relacionamentos, no entanto existem limitações importantes sobre como modelar e implementar séries históricas geoespaciais de dados demográficos pensando em *Linked Open Data* (MOURA; JR., 2013).

Data Fusion e redundância: Um problema da *Web of Data* é a redundância de informações sobre uma mesma entidade do mundo real. Isso ocorre em situações onde diferentes conjuntos de dados criam diferentes URIs para uma mesma entidade. Para tentar solucionar esse problema, a *Web of Data* possui um tipo de relacionamento que caracteriza dois objetos como iguais. No entanto, é difícil estabelecer quando os objetos são similares, pois mesmo que ambos correspondam a uma entidade do mundo real, suas formas de representação no contexto espacial podem ser diferentes como, por exemplo, a variação de escalas (JAIN *et al.*, 2010). Sendo assim, essa é mais uma dificuldade para se conseguir aumentar, de forma segura, a quantidade de conexões entre fontes de dados na *Web of Data* (MOURA; JR., 2013).

Nesse contexto, o *Gazetteer* proposto tenta amenizar e solucionar alguns dos desafios para a próxima geração de *Gazetteers*. Dentre eles, o trabalho tem o foco na coleta e integração de dados, recuperação, busca e navegação, resolução de entidades e desambiguação de entidades, e, por fim, na manipulação de dados temporais. Para tratar da qualidade das informações geográficas presentes no *Gazetteer* é utilizada a Lei de Linus.

4.2 Precisão de dados utilizando VIG

Tendo em vista a coleta e a natureza distribuída dos dados coletados por voluntários, é extremamente importante validar o quão boa é a qualidade das informações que são recolhidas através das atividades geográficas voluntárias. Realizar essa verificação é crucial para determinar a eficácia das atividades de VGI e sua contribuição para diversas aplicações, que vão desde contextos básicos, como aplicações de navegação em mapas, a aplicações mais sofisticadas, tais como escolha e planejamento de locais para construção de indústrias (HAKLAY *et al.*, 2010).

Com objetivo de se verificar a qualidade de coordenadas, Haklay *et al.* (2010) descrevem

em seu trabalho a avaliação de uma rede de estradas do Reino Unido extraídas do Open Street Maps (OSM) comparando os valores das coordenadas geográficas com uma base de teste da agência de mapeamento nacional britânica. Os resultados dessa pesquisa mostram uma sobreposição de cerca de 80% das coordenadas, ou seja, são similares.

Segundo [Haklay et al. \(2010\)](#), esses valores não são surpreendentes, pois, informações fornecidas por muitos participantes em projetos de VGI como o OSM, são similares ao conjunto de dados mantidos por agências governamentais. No entanto, é necessário verificar em qual fase do processo de coleta de dados a qualidade torna-se confiável.

Uma forma de explorar a questão da garantia da qualidade em projetos VGI, como o OSM, é olhar para projetos semelhantes, embora não na área de informação geográfica, como, por exemplo, projetos de código aberto que permitem pessoas colaborarem com informações. Dessa forma, avaliações paralelas podem ser traçadas entre os problemas de qualidade de VGI e a qualidade de software ([HAKLAY et al., 2010](#)).

No contexto da área de qualidade de software, em que a Lei de Linus tem origem, diversos projetos de código livre adotam sua utilização, como, por exemplo, o Apache Web Server. A Lei de Linus é comumente interpretada como “*Given enough eyeballs, all bugs are shallow*”, ou seja, “Dados olhos suficientes, todos os erros são óbvios”. Isso significa que, em projetos de desenvolvimento de código aberto onde vários programadores estão envolvidos no desenvolvimento do código, realizando diferentes situações para testes e aprimoramentos do sistema, o código tende a se tornar cada vez melhor, sem procedimentos e garantias formais de qualidade ([HAKLAY et al., 2010](#)).

Assim é possível traçar um paralelo da Lei de Linus, para verificar a precisão do posicionamento de informações geográficas. A lógica para esta Lei é: Se somente existe um contribuinte em uma área, ele ou ela pode inserir alguns erros, por exemplo, se esquecer de demarcar uma localidade ou inserir uma localização imprecisa. Portanto, mais contribuintes podem notar os dados imprecisos ou erros e reduzir o número de informações inválidas ([HAKLAY et al., 2010](#)).

Para verificar a validade da aplicação da Lei de Linus, no contexto de VGI, [Haklay et al. \(2010\)](#) utilizam áreas geográficas de Londres, referentes à Rede Integrada de Transporte (ITN) e os registros do OSM, com o objetivo de verificar se as coordenadas geográficas de ambos são similares. O resultado da pesquisa é a conclusão de que as áreas referentes às rodovias utilizadas no OSM são bem precisas, chegando a 85% de acurácia ao se utilizar um limiar de no máximo 8 metros de diferença.

Após uma abordagem mais detalhada sobre a viabilidade da Lei de Linus, [Haklay et al. \(2010\)](#) dividem as coordenadas geográficas de seu experimento num *grid* de 1 km^2 e analisam o número de contribuintes e a precisão posicional das coordenadas geográficas. Como resultado, os autores verificaram que ter cinco ou mais contribuintes é capaz de levar a qualidade das informações geográficas acima de 70%, como apresentado na Figura 11, onde mais de 93% dos

Figura 11 – Número de contribuintes e precisão posicional. Fonte: (HAKLAY *et al.*, 2010)

Figura 12 – Número de contribuidores por erro na precisão posicional. Fonte:(HAKLAY *et al.*, 2010)

pontos têm mais de 70% de precisão.

Embora esse valor seja elevado, a qualidade não é mais dependente do número de contribuintes após um determinado número. As coordenadas precisas são editadas por um número de 5 a 20 colaboradores, sendo que, a partir de 13 contribuintes, a precisão das coordenadas se estabiliza, como apresentado na Figura 12.

Conforme relatado por Haklay *et al.* (2010), a Lei de Linus pode ser aplicada ao Open Street Map e em projetos para VGI em geral, mesmo quando o número de contribuintes é relativamente pequeno. No entanto, a relação entre o número de contribuintes e a qualidade dos

dados não é linear. A partir de cinco contribuintes em uma determinada área, uma melhora na qualidade das informações geográficas é notada e quando o número de contribuintes passa de 13 colaboradores essa melhora se estabiliza e há uma pequena oscilação (ruído) da precisão posicional das coordenadas geográficas.

Conclui-se então que é possível considerar a Lei de Linus como um indicador espacial de qualidade dos dados, sem a necessidade do uso de um conjunto de referência, como, por exemplo, bases de dados geográficas do IBGE, para validar a qualidade dos dados fornecidos.

O *SWI Gazetteer* utiliza a Lei de Linus para validar as informações geográficas fornecidas pelos usuários. Além disso, é proposto um método baseado na Lei de Linus que é capaz de aprimorar as coordenadas geográficas imprecisas dos repositórios SpeciesLink e GBIF de forma automática, descrito na seção 5.2.

4.3 Considerações Finais

Este capítulo abordou os principais trabalhos relacionados ao desenvolvimento de *Gazetteers* na área de Recuperação de Informação Geográfica, descrevendo as similaridades e divergências entre o *SWI Gazetteer* e os sete trabalhos relacionados que foram escolhidos. Além disso, foi abordado o atual estado da arte para o desenvolvimento de *Gazetteers* e como suas coordenadas geográficas podem ser qualificadas, sem o uso de um conjunto de dados de referência, por meio da Lei de Linus (que é utilizada na literatura como um indicador espacial de qualidade).

O *SWI Gazetteer* tenta amenizar e solucionar alguns dos desafios para a próxima geração de *Gazetteers*. Dentre eles, o trabalho proposto tem foco na coleta e integração de dados no padrão de *Linked Open Data* por meio da manutenção e atualização dos dados por usuários voluntários. Espera-se que, com a prática de VGI, a qualidade, completude e atualização das informações sobre as localidades das coleções biológicas sejam aprimoradas.

ARQUITETURA DO SWI GAZETTEER

Ao considerar os trabalhos recentes na área de RIG, que envolvem o desenvolvimento de *Gazetteers*, descobriu-se que grande parte deles desenvolveu seus sistemas utilizando camadas, pois elas fornecem fácil escalabilidade, estabilidade e performance para os sistemas. Geralmente, esses trabalhos dividem a implementação de seus *Gazetteers* em três camadas: apresentação, negócios e dados.

Para implementar as funcionalidades presentes nessas camadas, grande parte dos *Gazetteers* atuais utiliza tecnologias que não possuem suporte à aplicação de técnicas da web semântica para abordar os novos desafios da área de RIG e da prática de VGI. A arquitetura do *SWI Gazetteer*, apresentada na Figura 13, foi desenvolvida levando em conta a necessidade de se adotar suporte a web semântica para construção de *Gazetteers* e experiências documentadas na literatura (KESSLER; JANOWICZ; BISHR, 2009; OGC, 2012; BEARD, 2012; BERETA; SMEROS; KOUBARAKIS, 2013). Essa arquitetura possui três camadas: Apresentação (*Presentation Tier*), Negócios (*Business Tier*) e Dados (*Data Tier*). Cada uma delas possui uma determinada função para possibilitar o uso de VIG por usuários e a manipulação de dados em formatos utilizados pela Web Semântica.

A *Camada de Apresentação* contém uma interface Web que possibilita a interação dos usuários por meio de mapas digitais interativos, e agentes de software por meio de *endpoints* GeoSPARQL. Dessa forma, é possível a usuários e agentes de software incluir ou buscar informações no *SWI Gazetteer*.

Para inserir informações, tais como localidade, município, data e coordenadas geográficas, é necessária a comunicação com a *Camada de Negócios*. Essa camada opera de duas formas, por meio da inserção de novos dados e da consulta sobre lugares.

Quando a *Camada de Negócios* recebe uma informação para ser inserida no *SWI Gazetteer*, os módulos de filtro (*Filter*), agrupamento (*Cluster*) e mapeamento (*Mapping*) são utilizados. A seguir, é descrito o funcionamento desses módulos para cada tipo de tarefa executada: inserção

Figura 13 – Arquitetura do SWI Gazetteer. Fonte: Cardoso *et al.* (2015)

ou busca.

1. Inserção de dados na camada de negócios

- a) *Módulo para Filtrar Informações (Filter)*: Esse módulo é responsável por verificar se as informações inseridas pelo usuário são válidas. Após realizar a filtragem dos dados, as informações são repassadas ao módulo de agrupamento de registros.
- b) *Módulo de Agrupamento os Registros (Cluster)*: Nesse módulo, são utilizadas técnicas de GIR, tais como, resolução de topônimos e desambiguação de locais usando fontes externas (como a DBpedia e o Geonames). Ao obter os dados já validados, esse módulo é responsável por realizar o agrupamento dos dados, verificando se existem outras localidades, na base de dados do *SWI Gazetteer*, que possam ser aprimoradas com a informação fornecida. Após todo esse processo, os dados são enviados para o módulo de mapeamento de dados em RDF.
- c) *Módulo de Mapeamento de Dados em RDF (Mapping)*: Esse módulo tem como objetivo transformar os dados enviados pelo módulo anterior em dados no formato RDF e então enviá-los para camada de dados (*Data Tier*). Durante a conversão para RDF, cada tipo de informação disponível para um lugar é mapeado para o tipo apropriado em RDF. Desta forma, coordenadas são mapeadas para literais em WKT, municípios para URIs contidas na DBpedia, etc. Esse mapeamento usa ontologias, tais como LinkedGeoData e GeoSPARQL para mapear os dados.

2. Busca de dados na camada de negócios

- a) *Módulo de Reformulação de Consultas (Query Reformulation)*: Esse módulo tem como função receber as informações referentes às consultas do usuário que foram transmitidas pela camada de apresentação e reformular as mesmas em um formato semântico. Para realizar essa tarefa, inicialmente os processos de *GeoParser* e *GeoCoding* são realizados para criar o planejamento da *query*, isto é, verificar se a mesma se refere a uma consulta com apenas nomes de lugares ou se também possui coordenadas geográficas. Após esse processo de planejamento, esse módulo envia a *query* reformulada para o módulo de mapeamento de dados em RDF.
- b) *Módulo de Mapeamento de Dados em RDF (Mapping)*: Esse módulo recebe as *queries* analisadas e converte as palavras-chave para uma consulta GeoSPARQL. Cada tipo de lugar listado na *query* é mapeado para o tipo apropriado na ontologia LinkedGeoData. Para verificar qual função GeoSPARQL deve ser utilizada na *query*, por exemplo, interseção (*sfIntersects(wkt1, wkt2)*) ou proximidade (*sfWithin(wkt1, wkt2,distance)*), são usadas as propriedades próximo, perto, dentro, entre outras identificadas pelo módulo anterior.

A *Camada de Dados* tem como função armazenar os dados do *SWI Gazetteer* em uma *triple store*. Basicamente, uma *triple store* é um “banco de dados” construído para armazenar e recuperar as triplas no formato RDF. Essa *triple store* fornece *endpoints* SPARQL onde consultas semânticas podem ser feitas. Na arquitetura do *SWI Gazetteer*, a *triple store* também pode responder consultas GeoSPARQL. Atualmente, existem diversas implementações para *triple stores*, incluindo projetos *open source*, como uSeekM, Parliament e Strabon, e projetos comerciais, como Oracle Database 12c.

Para desenvolver a implementação da arquitetura do *SWI Gazetteer*, foi utilizada a linguagem de programação Java e, sempre que possível, a reutilização de componentes e frameworks que se adequavam ao escopo do *SWI Gazetteer*. Nas próximas seções, a implementação da arquitetura do *SWI Gazetteer*¹ é descrita evidenciando-se as ferramentas e técnicas utilizadas em cada módulo.

5.1 Camada de Apresentação

Na Camada de Apresentação, foi desenvolvida uma interface gráfica usando o framework Google Web Toolkit (GWT 2.6). Além dele, a API GWT-OpenLayers foi usada para desenvolver os mapas digitais. Ambas ferramentas permitem o desenvolvimento da camada de apresentação usando-se apenas a linguagem de programação Java, simplificando o desenvolvimento e integração com as outras camadas do protótipo.

5.1.1 Protótipo do SWI

Devido ao fato de contribuições voluntárias atenderem a grupos com interesses específicos em uma determinada área, é importante garantir que os sistemas VGI, os quais esses usuários utilizam, atendam a requisitos específicos. Dentre eles, [Klinkenberg \(2013\)](#) descreve que:

1. Um dos aspectos mais importantes para sistemas que oferecem suporte a VGI deve ser sua forma simplificada de receber novas contribuições, pois sistemas complexos, que requerem diversos passos para a inserção de informações, reduzem o número de contribuições.

2. Os contribuintes devem ser capazes de ver o resultado de suas contribuições, especialmente em mapas.

3. Enquanto mapas, referentes a sistemas GIS interativos, são a melhor maneira de mapear os dados, eles não são necessariamente os mais amigáveis. Os meios pelos quais o público interage com os mapas devem ser tão intuitivos quanto possível, sem a necessidade de tutoriais extensos.

4. Web Sites e interfaces públicas devem disponibilizar suas informações atualizadas, pois, se as informações fornecidas em um site não forem atuais, os usuários e participantes irão

¹ <http://dx.doi.org/10.5281/zenodo.17814>

Figura 14 – Tela principal do *SWI Gazetteer*

se desmotivar com o projeto.

Com intuito de atender a esses requisitos, foi desenvolvido o protótipo do *SWI Gazetteer* que se encontra disponível no endereço: <<http://biomac.icmc.usp.br:8080/swi/>>. Esse protótipo é composto por uma única tela que tem as funcionalidades necessárias para os biólogos colaborarem com o sistema.

A tela inicial do *SWI Gazetteer*, apresentada na Figura 14, contém, na parte superior, um campo de busca que permite aos usuários entrarem com o nome de alguma localidade para ser pesquisada no sistema (Painel 1). Além disso, é possível definir qual a fonte das informações a serem pesquisadas, por meio do campo de escolha de qualidade. A qualidade escolhida permite retornar informações geográficas de acordo com o tipo de usuário que as disponibilizou, como, por exemplo, biólogos do INPA, registros que foram computados ou dados do IBGE.

Logo abaixo da funcionalidade de busca, no Painel 2, o usuário encontra um botão para recuperar localidades que não possuem coordenadas geográficas. Quando o usuário escolhe essa funcionalidade, o sistema recupera as informações e as mesmas são listadas na tabela abaixo desse botão.

Na parte inferior do Painel 2 são exibidos os campos para o usuário contribuir com uma nova localidade ou aprimorar algum registro já existente na base de dados. Para aprimorar alguma localidade, o usuário deve selecionar algum local na tabela de locais e então entrar com as informações no formulário de informações (parte inferior do Painel 2). As informações necessárias são: nome do local, tipo do local, município e as informações geográficas que podem ser inseridas manualmente ou através do mapa.

Ao escolher alguma localidade, o sistema exibe as informações daquele local no Painel 3, onde é possível visualizar: o nome do lugar, município, quantidade de registros que foram agrupados em um mesmo grupo e a URI referente a esse grupo no *triple store* do *SWI Gazetteer*.

Para aprimorar alguma localidade que já possui uma coordenada geográfica, o usuário

Figura 15 – Tela mostrando a funcionalidade de busca e aprimoramento de coordenadas.

deve acessar a funcionalidade de busca apresentada na Figura 15. Nessa funcionalidade, o usuário pode digitar alguma localidade e procurar suas informações, por exemplo, "reservas próximas a 100 km da cidade de Manaus" (Painel 1). Após o *SWI Gazetteer* executar essa consulta, as informações sobre as localidades, que satisfazem a busca, são exibidas na tabela de informações (Painel 2).

Uma vez que os dados estão na tabela de informações, o usuário pode selecionar alguma localidade e as informações referentes ao local selecionado são mostradas no Painel 3. Essas informações são: número de pessoas que concordam com a informação exibida, o nome da localidade, o município da localidade (caso informado), a quantidade de pessoas que já contribuíram para aprimorar aquela informação, a data da informação, o número de registros que foram agrupados e a URI da localidade pesquisada.

Caso o usuário opte por concordar ou discordar da informação existente no *SWI Gazetteer*, ele deve informar sua escolha no Painel 5, onde é exibido um gráfico que mostra o número de contribuições para aquele local e a quantidade de pessoas que concordam com aquela informação. Se o usuário optar por inserir uma nova informação, ele deve informar uma nova coordenada geográfica, clicando em uma das opções do Painel 4, como, por exemplo, pontos, polígonos e linhas. Logo depois é necessário marcar no mapa qual é a nova informação e assim confirmar as informações para o sistema novamente, no Painel 5.

Somente é possível atualizar a informação geográfica e se ter um novo centróide, se a porcentagem de pessoas que concordam dessa informação esteja abaixo de 70%. Caso contrário, a nova coordenada informada pelo usuário somente será armazenada numa lista de candidatos.

A partir de cinco contribuições, a informação posicional das coordenadas geográficas possui precisão acima de 70%, ao se comparar com áreas de *grid* de 1 km^2 , conforme relatado por Haklay *et al.* (2010). Por exemplo, no caso de uma área mapeada pelo IBGE (um parque nacional, floresta, etc), a partir de 5 contribuições existe uma probabilidade de 70% que as

coordenadas estejam dentro da reserva (com erro de 1km^2). Esse é um excelente número de precisão de coordenadas geográficas, conforme relatado pelos especialistas no domínio.

A implementação dessa funcionalidade está utilizando um outro grafo RDF auxiliar, onde as informações candidatas são armazenadas. Para verificar todas as políticas de inserção e aprimoramento de dados no *SWI Gazetteer* e armazenar esses dados em uma *triple store*, os dados precisam ser enviados para a camada de negócios.

5.2 Camada de Negócios

A camada de negócios contém as principais regras de negócio do *SWI Gazetteer*. Implementando as funcionalidades para comunicação entre a camada de apresentação e camada de dados, com o objetivo de aprimorar os registros geográficos. Nesta seção, são descritos os principais módulos dessa camada: agrupamento (*clustering*), mapeamento (*mapping*) e o módulo de reformulação de consultas.

5.2.1 Módulo de Clustering

Este módulo tem como principal objetivo agrupar os lugares que possuem nomes similares. Para realizar esse agrupamento, inicialmente é feita a resolução de topônimos. Uma lista dos tipos de lugares presentes na ontologia LinkedGeoData é carregada e uma expressão regular é criada para cada anotação contida nas classes da ontologia.

Essa expressão regular permite recuperar a informação de um determinado local, conforme descrito por Gouvêa (2009) e Machado *et al.* (2011), da seguinte forma: A expressão $(?i)ilha\b(.+?)$ recupera todas as triplas que contém a palavra ilha no início, onde “ilha” se refere a uma anotação da classe Island. É importante ressaltar que, utilizando-se uma ontologia como base para os tipos de lugares, é possível reaproveitar tipos já relatados em outros idiomas e incluir novos. Além disso, ao refazer o uso da ontologia para resolução de topônimos é possível agregar significado semântico para o local durante a resolução de topônimos.

Após a etapa de resolução de topônimos, o módulo usa algoritmos de agrupamento, como por exemplo, Star Algorithm (ASLAM; PELEKHOV; RUS, 2004) ou K-means (SHAMEEM; FERDOUS, 2009), juntamente com uma lista de *stop words* para remover palavras sem significado linguístico e assim agrupar os locais. Para auxiliar os algoritmos de agrupamento, os lugares são divididos em suas respectivas municipalidades e classes, como Áreas Protegidas, Igarapés, Rios entre outros, que foram identificadas pela resolução de topônimos ao utilizar a ontologia.

Essa separação permite tratar situações onde lugares, com mesmo nome, possuem localidades muito distintas, ou seja, a desambiguação geo/geo (seção 2.8). Além disso, esse módulo também usa dados dos repositórios Geonames e DBpedia para auxiliar na desambiguação

dos nomes, sempre que possível.

Para verificar a similaridade entre as localidades e agrupá-las, foi utilizado o coeficiente de similaridade de Jaccard (YIN; YASUDA, 2006). Ele observa quão parecidas são duas palavras, gerando valores no intervalo de $[0, 1]$, onde o valor 0 representa palavras extremamente diferentes e 1 palavras exatamente idênticas. A utilização desse coeficiente se deve ao fato que nomes, para a mesma localidade, podem ter várias formas de escrita para representar o mesmo local, como relatado por Santos (2003).

Para utilizar o coeficiente de Jaccard durante o agrupamento das informações, um limiar de similaridade entre as localidades foi definido manualmente. A partir de testes iniciais, verificou-se que os melhores limiares a serem utilizados são os valores 0.4, 0.5 e 0.6. A utilização destes *thresholds* permite que o agrupamento dos dados tenha um equilíbrio entre sensibilidade e estabilidade, ou seja, o algoritmo é capaz de agrupar os locais considerando o ruído inerente aos locais, da base de dados, e a capacidade de agrupar locais cujos nomes são sintaticamente similares.

Uma vez que o módulo de agrupamento termina, as coordenadas geográficas agrupadas são aprimoradas usando o método de sumarização proposto por Cardoso *et al.* (2014).

Para satisfazer este método de sumarização, inicialmente é verificado se a localidade possui um polígono nos dados do IBGE, em caso afirmativo, os pontos que estão fora do polígono são descartados e então a mediana dos pontos que estão dentro do polígono é feita. Caso não exista um polígono nos dados do IBGE, para representar a localidade no SWI Gazetteer, apenas são verificados os valores que aparecem com mais frequência no grupo. Após obter o valor que mais se repete, suas informações geográficas são atribuídas ao centróide indicando que aquelas coordenadas assumirão o novo valor para todos os registros (Figura 16). A escolha dessa abordagem se deve ao fato de que, dado um valor que ocorre com mais frequência em um grupo, tende a ser o valor correto para uma localidade. Essa abordagem estende a proposta da Lei de Linus, descrita por Haklay (2008)

Após aprimorar os dados do SWI Gazetteer, são utilizadas ontologias geoespaciais para mapear os dados gerados para uma *triple store*. Esse mapeamento é descrito na próxima seção.

5.2.2 Módulo de Mapeamento

Neste módulo, os dados gerados pela etapa de aprimoramento das coordenadas geográficas são mapeados usando as ontologias LinkedGeoData e GeoSPARQL. O uso dessas ontologias se deve ao fato de ambas já terem sido validadas por especialistas e utilizadas por vários trabalhos, tais como Koubarakis *et al.* (2012), Parundekar, Knoblock e Ambite (2010), Auer, Lehmann e Hellmann (2009). Além disso, a adoção dessas ontologias permite a reutilização de padrões já especificados e validados na área, para reduzir o trabalho de construção do SWI Gazetteer.

Essas duas ontologias existem separadamente e, para conectá-las, foi utilizada a ferra-

Figura 16 – Demonstração do método de sumarização de coordenadas geográficas, onde os valores que ocorrem com mais frequência são identificados e atribuídos para as demais coordenadas. Fonte: (CARDOSO *et al.*, 2014)

menta de edição de ontologias Protégé 5. Para conectar as ontologias, a classe *Feature*, presente em ambas, foi unificada utilizando-se o relacionamento *EquivalentTo* ($\equiv$) que possibilita dizer que as duas classes são equivalentes, conforme mostrado na Figura 17. Assim, ao se unificar as duas ontologias, é possível mapear as informações geográficas, latitude e longitude para o *data property asWKT*, fornecido pela ontologia GeoSPARQL (Figura 18).

Para mapear as informações do *SWI Gazetteer* referentes a localidade, município, data do registro, número de coordenadas (em um grupo) e a quantidade de pessoas que concordam com a informação geográfica, foram adicionados os *data properties* (apresentados na Figura 17) à ontologia LinkedGeoData. Essa ontologia não continha esses *data properties*. Já as coordenadas geográficas, pontos e polígonos foram mapeados para o *data property asWKT*, presente na ontologia GeoSPARQL.

Após representar os dados do *SWI Gazetteer* (Figura 19 área “Dados do SWI Gazetteer”), é necessário criar indivíduos para representarem as localidades e as geometrias que compõem as coordenadas geográficas. Para realizar essa representação, é necessário criar um indivíduo para uma localidade e conectá-lo com um indivíduo que representa uma geometria, por meio do *object property hasGeometry* (Figura 19). Isso permite que a *triple store* possa realizar inferências geográficas sobre os dados, utilizando as funções definidas pelo GeoSPARQL.

Os indivíduos do *SWI Gazetteer*, de uma forma geral, são estruturados conforme representação da Figura 20. Nessa representação, é possível visualizar os dois grafos utilizados para construir o *SWI Gazetteer*: o principal e o auxiliar. O Grafo Auxiliar armazena as coordenadas geográficas candidatas para uma determinada localidade avaliada pelo usuário, conforme descrito

Figura 17 – União da ontologia LinkedGeoData com a GeoSPARQL.

```

1. <!-- Gazetteer:Lake713 -->
2. <owl:NamedIndividual rdf:about="&Gazetteer;Lake713">
3. <rdf:type rdf:resource="dbp:Lake"/>
4. <Gazetteer:locality>Lago Tiaracá</Gazetteer:locality>
5. <Gazetteer:county>Novo Airão</Gazetteer:county>
6. <geosparql:hasGeometry rdf:resource="&Gazetteer;/Geometry/713"/>
7. </owl:NamedIndividual>
8. <!-- Gazetteer:Geometry/713 -->
9. <owl:NamedIndividual rdf:about="&Gazetteer;Geometry/713">
10. <geosparql:asWKT rdf:datatype="geow:wktLiteral">
11. <![CDATA[http://www.opengis.net/def/crs/OGC/1.3/CRS84 Point(0.20000000298023224 -66.0)]]>
12. </geosparql:asWKT>
13. </owl:NamedIndividual>

```

Figura 18 – Mapeamento das informações geográficas para a especificação *asWKT* (Representação simplificada).

Fonte: (CARDOSO *et al.*, 2014)

Figura 19 – Representação da ligação entre um indivíduo, que representa uma localidade, com outro que representa uma geometria, usando o *object property hasGeometry* da ontologia GeoSPARQL.

na seção 5.1. Já o Grafo Principal tem duas áreas: na Área 1 são representados os indivíduos do *SWI Gazetteer*, ou seja, os centroides dos agrupamentos, e na Área 2 os registros referentes ao SpeciesLink ou outro repositório de dados.

A conexão entre os indivíduos da Área 1 com os registros da Área 2 é feita por meio de *blank nodes* (nós anônimos). A utilização desses *blank nodes* deve-se ao fato dos mesmos serem utilizados para representar listas em RDF (MALLEA *et al.*, 2011). Na Figura 20, eles são representados pelos valores “*cluster:ProtectAI*” e “*cluster:LakeII*”.

Uma vez que os dados do *SWI Gazetteer* foram mapeados é possível ao usuário realizar buscas que possuem significado semântico, como a consulta GeoSPARQL exibida no Código-fonte 2. Nessa consulta, é possível localizar todas as cachoeiras que estão a menos de 1000 metros de alguma represa.

Código-fonte 2: Consulta para encontrar cachoeiras que estão a uma distância de 1000 metros de alguma represa.

```

1 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns/#>
2 PREFIX geo:<http://www.opengis.net/ont/geosparql/#>
3 PREFIX geof:<http://www.opengis.net/def/function/geosparql/>
4 PREFIX lgdo:<http://linkedgeodata.org/ontology/>
5 SELECT ?p ?a
6 WHERE {
7     ?p rdf:type lgdo:Dam .
8     ?p geo:hasGeometry ?g2 .
9     ?g2 geo:asWKT ?w2 .
10    ?a rdf:type lgdo:Waterfall .
11    ?a geo:hasGeometry ?g1 .
12    ?g1 geo:asWKT ?w1 .
13    FILTER(geof:distance(?w1,?w2,units:metre)<1000)
14 }
```

Figura 20 – Representação dos grafos do *SWI Gazetteer* evidenciando seus indivíduos, conexões entre geometrias, clusters criados e municípios da DBpedia. A Área 1 contém os indivíduos a Área 2 os clusters.

5.2.3 Módulo de Reformulação de Consultas

Embora seja possível realizar buscas semânticas no *SWI Gazetteer* usando *queries* SPARQL, essa ainda seria uma tarefa laboriosa para os usuários (biólogos). Sendo assim, foi implementado este módulo que reformula consultas com *string* de busca para *queries* SPARQL.

Este módulo tem como função receber as informações referentes às consultas do usuário (que foram transmitidas pela camada de apresentação) e reformular as mesmas em um formato semântico. Para realizar essa tarefa, é realizado o processo de *GeoParser* para criar o planejamento da *query*, conforme descrito no algoritmo *GeoParser* (Seção 5.2.4).

5.2.4 Algoritmo GeoParser

O algoritmo *GeoParser* (Algoritmo 1) foi desenvolvido com base no trabalho de [Zhu et al. \(2012\)](#), onde é relatado que, para realizar buscas sobre localidades, é necessário identificar a <região alvo, região de referência, direção>. Por exemplo, na consulta “reservas próximas à cidade de Manaus” os valores de região alvo, região de referência e direção são mapeados da seguinte maneira: <reservas, cidade de Manaus, próximas>.

Algoritmo 1: *GeoParser* semântico para realizar as consultas inseridas pelo usuário utilizando palavras chave

```

Input: consulta do usuário : query
1  OntClass,OntProperty,result ← Array()
2  Ontology ← carregarOntologia()
3  OntClass ← carregarClasses(query,Ontology)
4  OntProperty ← carregarPropriedades(query,Ontology)
5  if sizeOf(OntClass) > 1 then
6      identified ← unificarClassesComNomesDeLugares(query,OntProperty,OntClass)
7      moreGeneric ← buscaClasseMaisGenerica(OntClass)
8      moreSpecific ← buscaClasseMaisEspecificas(OntClass)
9      result ← buscaPorInterseção(moreGeneric,moreSpecific,OntProperty, identified,query)
10     result ← buscaPorProximidade(moreGeneric,moreSpecific,OntProperty,
        identified,query)
11 else
12     result ← buscaUnicoLocal(OntClass,query)
13 return results

```

Tendo como base esse conceito, o algoritmo desenvolvido no *SWI Gazetteer* opera da seguinte forma:

1. Inicialmente o algoritmo recebe o parâmetro inicial para a consulta, que é uma *string* com o nome da localidade informado pelo usuário. Logo após os atributos básicos: ontologia utilizada no *SWI Gazetteer*, classes e propriedades presentes na consulta do usuário são carregados.

2. Para carregar as classes e propriedades referentes à consulta, linhas 3 e 4, são utilizadas as funções *carregarClasses* e *carregarPropriedades*. Nessas duas funções é realizada a tokenização da consulta do usuário, remoção de *stopwords* e a busca por *tokens* similares às anotações contidas nas classes e propriedades da ontologia. Como exemplo de classes reconhecidas temos os valores listados na tabela 2, na coluna “Classes descobertas para realizar a busca”.
3. Uma vez que as classes e propriedades são carregadas, o algoritmo verifica qual tipo de consulta deve ser realizada, linha 5. Caso exista mais de uma classe identificada na consulta do usuário, o *geoparser* classifica essa consulta como uma consulta complexa, ou seja, necessita de dois tipos de lugares reconhecidos para executar as funções de planejamento da *query* (*buscaPorInterseção* e *buscaPorProximidade*). Caso contrário, a consulta é classificada como simples e será realizada pela função *buscaUnicoLocal*.
4. Antes de realizar o planejamento de *query*, o *geoparser* utiliza a função *unificarClassesComNomesDeLugares* para encontrar quais partes léxicas da consulta necessitam ser verificadas juntamente com os as classes, linha 6. Por exemplo, para a consulta “reservas próximas a 100 km da cidade de Manaus” a classe *ProtectedArea* não necessita nenhuma verificação léxica. Já para a classe *City*, é necessário verificar se os resultados possuem similaridade com a palavra “Manaus”, conforme listado no exemplo da Tabela 2.
5. Uma vez que as partes léxicas e semânticas da consulta são identificadas, é preciso verificar qual é a classe mais genérica da consulta, ou seja, a *região alvo* e qual é a classe mais específica, a *região de referência*, linhas 7 e 8. Isso é feito para que as funções de planejamento de *query* (*buscaPorInterseção* e *buscaPorProximidade*) encontrem pontos semelhantes à *região de referência* para iniciar as buscas.
 - a) Para identificar essas duas classes, as funções *buscaClasseMaisGenerica* e *buscaClasseMaisEspecificas* são utilizadas. Basicamente, essas funções utilizam os conceitos listados por [Zhu et al. \(2012\)](#), onde o *lugar alvo* sempre é o primeiro tipo informado e a *região de referência* é a região mais específica após o *lugar alvo*. É necessária a utilização de uma ontologia para verificar qual classe é a mais específica dentre as informadas.
6. Após todas essas informações serem identificadas, o algoritmo de *geo-parser* faz as consultas GeoSPARQL para as funções de planejamento de *query* (*buscaPorInterseção* e *buscaPorProximidade*).
 - a) Para verificar qual consulta deve ser realizada, as funções de planejamento de *query* utilizam as propriedades identificadas na consulta do usuário. Para o exemplo da consulta “reservas próximas a 100 km da cidade de Manaus”, a propriedade “near” é utilizada. Desse modo, a função *buscaPorInterseção* não irá realizar nenhuma

consulta GeoSPARQL. Assim, cabe à função *buscaPorProximidade* realizar o plano de *query* e encontrar os resultados.

- b) Inicialmente, a função *buscaPorProximidade* procura por alguma informação de distância na *query* do usuário, nesse caso a informação de “100 km” é encontrada e convertida para metros. Essa conversão é necessária devido à utilização da função *geo:buffer*, presente no GeoSPARQL, para consultar por locais “próximos”.
- c) Para realizar o plano de *query*, inicialmente é consultada na *triple store* a *região de referência* que consiste no indivíduo mais similar com a classe *City* e a palavra “Manaus”.
- d) Uma vez que as informações referentes à *região de referência* e *direção* foram identificadas, a função *buscaPorProximidade* constrói a *query* final para encontrar a *região alvo* que irá compor os resultados, como o exemplo mostrado no Algoritmo 3. Assim que os resultados são recuperados da *triple store*, os mesmos são exibidos para o usuário na interface do *SWI Gazetteer*, como descrito na seção 5.1 .

Código-fonte 3: Consulta geo-espacial semântica resultante para a informação fornecida pelo usuário “reservas próximas a 100 km da cidade de Manaus”.

```

1 PREFIX geof:<http://www.opengis.net/def/function/geosparql/>
2 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3 PREFIX geo:<http://www.opengis.net/ont/geosparql#>
4 PREFIX opengis: <http://www.opengis.net/def/uom/OGC/1.0/>
5 PREFIX swi: <http://www.semanticweb.org/ontologies/Gazetteer#>
6
7 SELECT * WHERE {
8     ?s a <http://dbpedia.org/ontology/ProtectedArea> .
9     ?s geo:hasGeometry ?s1 .
10    ?s1 geo:asWKT ?o1 .
11    ?s swi:locality ?locality .
12    FILTER(geof:sfWithin(?o1,geof:buffer("POLYGON();http://www.
    opengis.net/def/crs/EPSG/0/4326"^^<http://strdf.di.uoa.gr/ontology
    #WKT>,100000,opengis:metre))).
13 }
```

5.3 Camada de Dados

Após o mapeamento dos dados do *SWI Gazetteer*, os registros são armazenados em uma *triple store*, que tem como função servir de *endpoint* para realização de consultas SPARQL. O armazenamento dos dados numa *triple store* promove a fácil disponibilização dos dados como LOD. Um dos objetivos deste projeto é expor, compartilhar e conectar dados, informação e

Consulta Executada	Classes descobertas para realizar a busca	Propriedades Descobertas para auxiliar na busca	Parte léxica da consulta
reservas próximas a 100 km da cidade de Manaus	Genérico: dbpedia: ProtectedArea Específico: dbpedia: City	swi: near	— / Manaus
Campo da Catuquira	geonames:Camp	—	Catuquira
Locais dentro do Parque Nacional do Jau	Genérico: geonames#Feature Específico: dbpedia: ProtectedArea	geosparl: sfWithin	— / Jau
Rios do estado do Amazonas	Genérico: dbpedia: River Específico: dbpedia: State	—	— / Amazonas
Places part-of município de Manaus	Genérico: geonames: Feature Específico: dbpedia: City	swi: partyOf	— / Manaus
<p>Prefixos utilizados:</p> <p>PREFIX dbpedia: < http://dbpedia.org/ontology/></p> <p>PREFIX geonames: <http://www.geonames.org/ontology#></p> <p>PREFIX swi:<http://www.semanticweb.org/ontologies/Gazetteer></p> <p>PREFIX geosparql: <http://www.opengis.net/ont/geosparql#></p>			

Tabela 2 – Valores identificados pelo algoritmo de GeoParser

conhecimento na web semântica sobre dados referentes a localizações geográficas relacionadas a biodiversidade. A LOD basicamente define as melhores técnicas para isso (HEATH; BIZER, 2011).

Durante o desenvolvimento do *SWI Gazetteer* somente foi possível encontrar três implementações de *triple store open source* que permitem trabalhar com representações geoespaciais fornecidas pela linguagem de busca GeoSPARQL, Parliament, Strabon e UseekM. Assim, a seção 6.6 descreve o experimento conduzido para avaliar o desempenho dessas três *triplestores* com intuito de escolher a que possui melhor performance para o *SWI Gazetteer*.

5.4 Considerações Finais

Neste capítulo, foi descrita a implementação da arquitetura do *SWI Gazetteer*, evidenciando-se o funcionamento e a implementação de cada um dos módulos presentes na arquitetura.

A arquitetura do *SWI Gazetteer* apresenta um novo processo para desenvolvimento, baseado nos pontos ressaltados por Kessler, Janowicz e Bishr (2009), OGC (2012), Beard (2012), Bereta, Smeros e Koubarakis (2013), para atender os desafios da próxima geração de *Gazetteers*. Para isso, essa arquitetura utiliza práticas de VGI juntamente com tecnologias disponíveis na web semântica.

O próximo capítulo descreve os resultados obtidos com a implementação do *SWI Gazetteer*, ressaltando as principais contribuições e limitações da pesquisa.

EXPERIMENTOS

Neste capítulo, serão apresentados os resultados que foram alcançados com o *SWI Gazetteer*. A seção 6.1 descreve a análise dos dados referente aos repositórios SpeciesLink e GBIF e, as demais seções, os resultados obtidos com a implementação do *SWI Gazetteer*.

6.1 Dados utilizados

Para desenvolver este trabalho, foram utilizados dados referentes às coletas biológicas de espécimes realizadas pelo INPA e disponibilizadas nos repositórios SpeciesLink e GBIF.

O [SpeciesLink \(2014\)](#) é um sistema distribuído que integra, em tempo real, informações primárias sobre biodiversidade de museus, herbários e coleções microbiológicas, sendo possível obtê-las gratuitamente pela Internet. O repositório SpeciesLink possibilita aos curadores das coleções, ou seja, biólogos responsáveis pelas coleções, criarem filtros para buscar e inserir dados, sendo que, nesse último caso, o curador pode escolher não tornar os dados abertos de imediato por considerá-los sensíveis para futuras pesquisas ([SPECIESLINK, 2014](#)).

O Global Biodiversity Information Facility [GBIF \(2014\)](#) é um repositório de dados que fornece acesso livre aos dados científicos sobre biodiversidade mundial, através da *Internet*, utilizando Web Services. Atualmente, o GBIF conta com cerca de mais de 500 milhões de registros referentes à história natural de espécimes e coletas de campo, tornando-o o maior repositório de dados de biodiversidade na *Internet*.

Esses repositórios foram escolhidos por possuírem um grande número de dados sobre biodiversidade e seus arquivos poderem ser acessados facilmente em arquivos “csv” (Comma Separated Values) e “xls” (Excel).

Figura 21 – Gráfico da qualidade dos dados do SpeciesLink.

6.1.1 Amostra de dados coletada em fevereiro de 2014

Inicialmente foi coletada uma amostra de dados dos repositórios SpeciesLink e GBIF, em fevereiro de 2014. Nesse período, foi definido um critério de qualidade para avaliar a precisão dos dados geográficos. Basicamente, esse critério de qualidade é dividido em níveis, como mostrado na Figura 21 e Figura 22. Nos registros coletados, é possível notar que apenas 24,85% de todos os dados do SpeciesLink e 16,80% do GBIF apresentam todas as informações geográficas, ou seja, contém o nome do local e município que um espécime foi recolhido e suas coordenadas geográficas (latitude e longitude), conforme mostrado na Figura 21 e Figura 22.

Outro ponto interessante verificado em relação a esses dados é que entre os registros de coletas marcados com qualidade de informação 2 e 3, verificou-se que 31,56% são referentes a coletas muito antigas, datadas entre 1850 e 1979. Essas coletas não poderiam ser georreferenciadas pela ausência de aparelhos GPS e pela dificuldade de se determinar coordenadas no meio da floresta sem tais dispositivos. Essa afirmação também é comprovada por Santos (2003), onde é mostrado o modelo de coleta utilizado por biólogos. Esses registros são muito importantes, pois mostram a ocorrência de espécies ao longo de mais de um século.

Como o objetivo do SWI é tratar os dados que contêm informações geográficas imprecisas e aprimorá-las usando coletas com registros confiáveis, foram utilizados os dados de qualidade 2 a 4 para desenvolver o *Gazetteer*. Ou seja, 80% dos registros do SpeciesLink e 88% do GBIF. Como os registros remanescentes não apresentam informações de local de coleta, latitude, longitude ou município, eles foram descartados por serem inutilizáveis. Não há como associar a eles informações de localização (além do fato de terem sido coletados no estado do Amazonas). Desta forma, 20% dos registros do SpeciesLink e 12% do GBIF foram retirados dos experimentos iniciais do *SWI Gazetteer*.

Além da verificação da quantidade dos dados, foi feita uma análise para verificar a

Figura 22 – Gráfico da qualidade dos dados do GBIF.

imprecisão dos registros geográficos presentes nessa amostra de dados, como as exibidas na Figura 23a e Figura 23b. Uma vez que toda amostra é apenas para o estado do Amazonas, foi notado que várias coordenadas apontavam para regiões fora dos limites do estado, por exemplo, no mar, na Argentina, e em países vizinhos ao Brasil que fazem limite territorial com a Amazônia, como, por exemplo, Venezuela, Colômbia e Peru.

Esses valores errôneos aparecem ao longo do mapa distribuídos de forma horizontal ou vertical em relação ao Amazonas, como visualizado nas Figura 23a e Figura 23b. Esses erros se devem provavelmente ao fato de usuários digitarem as coordenadas geográficas manualmente, em tabelas eletrônicas, sem o auxílio de dispositivos computadorizados para transmitir tais dados automaticamente dos dispositivos GPS, ou de programas que testem sua validade (por exemplo, se as coordenadas estão dentro do município da coleta).

Além da imprecisão dos dados listados fora dos limites do estado do Amazonas, foram detectadas, depois de uma análise mais minuciosa das localidades, vários lugares com informações de latitude e longitude imprecisas como, por exemplo, a situação mostrada na Figura 24. Nessa figura, é possível visualizar que a Reserva Florestal Adolpho Ducke, representada pelo pino verde na sua localização correta, aparece em alguns registros com coordenadas geográficas muito fora dessa posição, marcadas pelos pinos vermelhos.

Essas informações evidenciam a necessidade de aprimoramento das coordenadas geográficas dos repositórios SpeciesLink e GBIF. Além desses dados, uma outra amostra, referente ao repositório do SpeciesLink, foi obtida em Novembro de 2014 com intuito de verificar se os dados ainda eram imprecisos, como notado nas amostras coletadas em Fevereiro de 2014.

6.1.2 Amostra coletada em Novembro de 2014

Na amostra de dados, referente ao mês de Novembro de 2014, foi verificado que em torno de 31% de todos os dados de coletas realizadas pelo INPA, no estado do Amazonas, continham coordenadas geográficas.

(a) Dados do SpeciesLink. Fonte: (CARDOSO *et al.*, 2014)

(b) Dados do GBIF. Fonte: (CARDOSO *et al.*, 2014)

Figura 23 – Representação das coordenadas do SpeciesLink e GBIF coletadas em Fevereiro de 2014

(a) Coordenadas geográficas referentes à Reserva Adolpho Ducke presentes nos dados do SpeciesLink. Fonte: (CARDOSO *et al.*, 2014)

(b) Coordenadas geográficas referentes à Reserva Adolpho Ducke presentes nos dados do GBIF. Fonte: (CARDOSO *et al.*, 2014)

Figura 24 – Coordenadas referentes à Reserva Florestal Adolpho Ducke contidas nos dados do SpeciesLink (a) e GBIF (b). Pontos vermelhos representam coordenadas geográficas erradas para a reserva. O ponto verde representa a coordenada geográfica correta para a reserva.

No entanto, para comparar a precisão das informações geográficas nas duas amostras de dados, o mesmo critério de qualidade utilizado em fevereiro de 2014 foi mantido, Figura 25. Ao analisar as informações, é possível verificar que somente 20% dos registros do SpeciesLink, fornecidos pelo INPA, possuem informações geográficas, ou seja, contém o nome do lugar, latitude, longitude e município onde os espécimes foram coletados (nível 4 de qualidade). Assim como na amostra anterior, esse número demonstra a falta de informações geográficas nesses dados de biodiversidade.

Nessa amostra de dados, também foram considerados apenas os registros nos níveis 2 a 4 que correspondem a 76% do total de registros do SpeciesLink. Os demais registros, 24%, foram descartados por serem imprecisos. Assim como na amostra anterior, também foi observado que várias localidades contém informações imprecisas de latitude e longitude, como mostrado na

Figura 25 – Qualidade das informações geográficas na amostra de dados referente ao SpeciesLink (novembro de 2014). Fonte: (CARDOSO *et al.*, 2015)

Figura 26a. Nessa figura, é possível observar a localização do Parque Nacional do Jaú, área em amarelo, e as localizações imprecisas representadas pelos pinos em vermelho. Analisando todos os dados da amostra, Figura 26b, é possível observar que várias coordenadas geográficas também se encontram fora do estado do Amazonas, em países vizinhos ao Brasil e no oceano.

Assim como na amostra coletada em fevereiro de 2014, é possível observar que os registros possuem dois problemas principais com as coordenadas geográficas associadas à coleção de espécimes:

1. Alguns registros não contêm coordenadas geográficas.
2. Alguns registros contêm coordenadas erradas ou imprecisas.

Considerando esses problemas, o *SWI Gazetteer* foi desenvolvido com o objetivo de aprimorar as coordenadas geográficas e amenizar esses dois principais problemas, que ocorrem nos registros de coletas biológicas. As próximas seções descrevem os resultados obtidos com o *SWI Gazetteer* ao utilizar as bases de dados descritas.

6.2 Verificação dos Dados Agrupados

Uma vez que os dados do SpeciesLink e GBIF foram agrupados pelo módulo de clustering (seção 5.2), foi necessário verificar a precisão dos locais. Para se realizar essa verificação, foi feita uma análise com uma amostra de 100 grupos selecionados aleatoriamente, para os testes realizados com o Star Algorithm (ASLAM; PELEKHOV; RUS, 2004) e K-means (SHAMEEM; FERDOUS, 2009). Após essa seleção, cada valor foi verificado manualmente. Nessa verificação, foi analisada a porcentagem de locais agrupados corretamente e o número de centróides.

Ao realizar a análise dos grupos, foi possível verificar que a coerência das localidades associadas foi bem significativa. Em média, 84% das localidades estiveram associadas corretamente para o Star Algorithm (ASLAM; PELEKHOV; RUS, 2004). Nenhum dos limiares escolhidos

(a) Exemplos de coordenadas geográficas para o Parque Nacional do Jaú contidas na amostra do SpeciesLink (novembro de 2014). Os pontos em vermelho representam as coordenadas imprecisas para o parque nacional. A área em amarelo representa a localização correta para o parque. Fonte: (CARDOSO *et al.*, 2015)

(b) Distribuição das coordenadas geográficas dos registros do SpeciesLink para o estado do Amazonas (novembro de 2014). Fonte: (CARDOSO *et al.*, 2015)

Figura 26 – Coordenadas imprecisas na amostra coletada em novembro de 2014.

para verificar a similaridade dos locais apresentou grandes variações quanto à precisão dos dados associados, com os valores de precisão de 84,3% para o limiar 0.4, 83,8% para o limiar 0.5 e 84,1% para o limiar 0.6.

Realizando o mesmos testes para o algoritmo K-means (SHAMEEM; FERDOUS, 2009), foi possível verificar que, em média, 90% das localidades estão associadas corretamente e nenhum dos limiares escolhidos apresenta grandes variações quanto a precisão dos dados, com valores 88,9% para o limiar 0.4, 91.73% para o limiar 0.5 e 89.85% para o limiar 0.6. É importante ressaltar que o algoritmo K-means depende do valor de K escolhido para um bom agrupamento dos dados. Para escolher esse valor, o *SWI Gazetteer* está usando a métrica $k \approx \sqrt{N/2}$ (CHIANG; MIRKIN, 2010), onde N é o número de localidades que pertencem a um município juntamente com seu tipo específico, como, por exemplo, áreas protegidas, rios, lagos, entre outros.

Outro ponto avaliado, durante o agrupamento das informações, foi a validação do número de centróides criados, onde uma amostra de 100 centróides foi escolhida aleatoriamente. Ao se verificar esses dados, foi possível visualizar que vários centróides possuem nomes semelhantes, ou seja, podem passar a ideia de mesma localidade, como, por exemplo, os centroides representados pela Área 1 na Figura 27.

No entanto, especialistas do domínio relataram, ao avaliar o *SWI Gazetteer*, que, apesar dos nomes serem parecidos, existe a possibilidade desses lugares serem distintos, como, por exemplo, os locais da Área 2. Isso ocorre porque as regiões na Amazônia são vastas e os coletores podem ter utilizado apenas o nome em comum da região para especificar as coletas. Desse modo,

	Exemplo de centroides criados	Coordenadas
Área 1	reserva florestal adolpho ducque	-2.957475 -59.931843
	reserva florestal adolpho ducque floresta primária de platô área do projeto team parcela da sede subparcela 9 indivíduo nº 616	-2.928889 -59.94639
	reserva floraducque terra firme	-2.916667 -59.983334
	reserva florestal ducque adolpho ducque	-2.883333 -59.966667
	reserva florestal adolpho ducque terra firme	-2.883333 -59.966667
	reserva ducque	-2.883333 -59.966667
Área 2	reserva florestal ducque próximo ao alojamento solo argiloso adolpho ducque	-2.957475 -59.931843
	reserva ducque igarapé do acará floresta de baixio	-3.633333 -60.566666
	reserva florestal ducque na área limpa do viveiro adolpho ducque	-2.883333 -59.966667
	reserva florestal ducque igarapé barro branco solo úmido na floresta perto do igarapé adolpho ducque	-2.883333 -59.966667
	reserva florestal ducque próxima a torre de observação solo argiloso adolpho ducque	-2.883333 -59.966667
	reserva florestal ducque próximo a piscina ig do barro branco adolpho ducque	-2.883333 -59.966667

Figura 27 – Representação dos centroides criados pelos algoritmos de agrupamento.

as próximas implementações do *SWI Gazetteer* irão incluir novos atributos e técnicas de Active Learning (MCCONNELL, 1996) para auxiliar no agrupamento das informações.

6.3 Quantidade de registros aprimorados pelo *SWI Gazetteer*

Após o agrupamento das localidades pelo módulo de cluster (seção 5.2) o *SWI Gazetteer* utiliza o método de sumarização para aprimorar os dados do SpeciesLink e do GBIF. Inicialmente, os registros coletados em fevereiro de 2014 dos repositórios SpeciesLink e GBIF continham respectivamente 24,85% e 16,60% de registros com coordenadas geográficas (longitude e latitude) para o estado do Amazonas. Após aplicar a técnica de sumarização de coordenadas juntamente com o Star Algorithm (ASLAM; PELEKHOV; RUS, 2004), foram obtidos os resultados apresentados na Figura 28a. O número de registros com coordenadas geográficas no SpeciesLink foi aumentado de 60.786 para 91.298 (um aumento de 50,2%) e no GBIF de 25.687 para 47.625 (um aumento de 85,4%). Esses números representam um aumento significativo, de cerca de 61%, no total do número de registros com coordenadas geográficas.

Figura 28 – Número de registros com coordenadas geográficas antes e depois da utilização do *SWI Gazetteer*.

Usando esse mesmo processo, para os registros do SpeciesLink referentes à amostra de novembro de 2014, foi possível aumentar o número de registros com coordenadas geográficas de 30,3% para 57,5% (38.774 registros foram adicionados). Isso representa um aumento de 90% no número de registros com coordenadas geográficas. A amostra inicial continha 43.128 registros com coordenadas geográficas, após utilizar o *SWI Gazetteer*, esse número foi para 81.872, conforme exibido na Figura 28b.

Utilizando a técnica de sumarização (CARDOSO *et al.*, 2014) com o algoritmo K-means (SHAMEEM; FERDOUS, 2009) na amostra de novembro de 2014, foi possível adicionar coordenadas geográficas a 44.321 registros. Isso representa um aumento de 102%. Conforme mencionado anteriormente, a amostra do SpeciesLink de novembro de 2014 continha 43.128 registros com coordenadas geográficas utilizando o método de agrupamento K-means, esse número foi para 87.449. Assim, podemos afirmar que o uso do *SWI Gazetteer* (com a técnica de sumarização Cardoso *et al.* (2014) e agrupamento) pode levar a um aumento significativo da informação geográfica em dados típicos sobre biodiversidade.

Devido ao fato do algoritmo K-means obter melhores resultados na precisão das localidades associadas corretamente e conseguir um maior número de coordenadas aprimoradas, 102%, a versão final do *SWI Gazetteer* utiliza esse algoritmo de agrupamento juntamente com o método de sumarização de coordenadas.

Outra avaliação realizada para validar o *SWI Gazetteer* foi a verificação da distribuição das coordenadas ao longo dos anos, onde os registros do SpeciesLink, referentes a novembro de 2014, foram agrupados de acordo com o ano da coleta. Na figura 29, é apresentada uma comparação do número de registros com coordenadas antes dos testes com o *SWI Gazetteer* (em vermelho) e após (em azul). A Figura 29a mostra a distribuição de coordenadas para coletas antes da disponibilidade de dispositivos GPS e a Figura 29b para depois. Em ambos os casos, é possível visualizar que um grande número de registros tiveram suas coordenadas geográficas

(a) Coletas realizadas antes da disponibilidade de dispositivos GPS

(b) Coletas realizadas após a disponibilidade de dispositivos GPS

Figura 29 – Distribuição de registros com coordenadas para coletas de acordo com os anos. Fonte: (CARDOSO *et al.*, 2015)

recuperadas independentemente da disponibilidade de aparelhos de GPS.

Outra avaliação foi a comparação dos registros originais do SpeciesLink 2014 que já tinham coordenadas com os registros depois de serem processados pelo *SWI Gazetteer* (modificados ou não). Foi selecionada aleatoriamente uma amostra de 100 registros de cada conjunto. Durante a análise das mesmas, foi verificado que 67 dos registros originais do SpeciesLink tinham coordenadas corretas e 33 tinham coordenadas imprecisas, tais como apresentado na seção 6.1. Sendo que da amostra do *SWI Gazetteer*, esse número é de 86 coordenadas corretas e 14 imprecisas.

Isso demonstra que o *SWI Gazetteer* é capaz de adicionar coordenadas mais precisas para as localidades. Além disso, essas 14 coordenadas imprecisas indicavam locais próximos às localidades corretas, o que não acontece para as 33 coordenadas imprecisas da amostra do SpeciesLink 2014, que estão em áreas muito distantes dos locais ou até mesmo marcadas no oceano.

Para analisar estatisticamente esses resultados, foi utilizado o Teste-T (ANDERSON; SWEENEY; WILLIAMS, 2013) para comparar duas amostras pareadas com $\alpha = 0,05$ e as seguintes hipóteses:

H_0 : A quantidade de registros corretos não difere significativamente, antes e depois do *SWI Gazetteer*.

H_1 : A quantidade de registros corretos difere significativamente, antes e depois do *SWI Gazetteer*.

Comparando o valor de p obtido com aplicação do Teste-T ($p = 0,000005$) com o nível de significância adotado de $\alpha = 0,05$, verificamos que $\alpha > p$ e assim rejeitamos a hipótese H_0 e concluímos que a quantidade de registros corretos difere significativamente, antes e depois do uso do *SWI Gazetteer*.

Esses dados evidenciam que, de fato, o *SWI Gazetteer* contribui para amenizar os problemas de acurácia existentes em informações geográficas voluntárias na área de biodiversidade. Além disso, com a construção do *SWI Gazetteer*, especialistas podem extrair informações geográficas de grandes quantidades de dados de forma mais rápida. Essa tarefa era muito mais laboriosa quando os dados estavam armazenados em arquivos “csv”, com várias informações imprecisas e difíceis de serem analisadas.

6.4 Comparação com Outras Fontes na LOD

Com intuito de comparar os dados do *SWI Gazetteer* com outros *endpoints* disponíveis na LOD para verificar a presença das mesmas informações, foram selecionados três repositórios que contém entidades geográficas: Geonames, WikiMapia e Wikipédia. No entanto, verificou-se que dois desses repositórios não possuem implementações de *endpoints* SPARQL para seus dados. Dessa forma, foi necessário encontrar outros repositórios que contivessem cópias dos dados do Geonames e WikiMapia e disponibilizassem os mesmos através de *endpoints* SPARQL.

Inicialmente, foi realizada uma análise de *endpoints* SPARQL/GeoSPARQL no W3C. Devido ao fato do WikiMapia não possuir dados em RDF e não ser possível encontrar nenhum repositório que tenha feito a conversão de seus dados para esse formato, esse *Gazetteer* foi descartado do experimento. Sendo assim, foram selecionados somente *endpoints* que possuíam informações sobre os repositórios Geonames e Wikipédia.

Junto a essa análise, também foi verificado se os *endpoints* disponibilizavam implemen-

Amount of georeferenced data

Figura 30 – Resultados de precisão e revocação para os três *endpoints*. Fonte: (CARDOSO *et al.*, 2015)

tações de GeoSPARQL para realizar consultas utilizando as funções geoespaciais. Por exemplo, verificar se um polígono está dentro de outro através da função *geo:sfwithin*.

A busca por essas fontes de dados resultou em três *endpoints*: Dbpedia¹, Factor² e GeoSPARQL³. Sendo que dentre eles, o Factor e o GeoSPARQL contém informações sobre o Geonames e o Dbpedia sobre a Wikipedia.

Ao se iniciar as consultas semânticas, foi verificado que o *endpoint* GeoSPARQL continha apenas informações referentes aos municípios brasileiros, desprezando locais como reservas naturais, rios, lagos, entre outros. Também foi possível constatar que, dentre os três *endpoints*, somente ele permitia realizar funções geoespaciais. No entanto, como seus dados não abordavam reservas, lagos, rios, como os outros, ele foi descartado do experimento.

Para realizar as *queries*, uma amostra de 60 localidades, listadas na amostra do SpeciesLink e GBIF de fevereiro de 2014, foi selecionada e, com ela, foi criada uma base de dados confiável sobre quais registros eram relevantes para uma determinada localidade nos *triple stores* escolhidos.

Após a construção dessa base de dados, as consultas foram submetidas aos *endpoints*. Os resultados de precisão e revocação de cada um são apresentados na Figura 30. Nesses resultados, é possível verificar que, com a utilização da busca semântica, é possível obter valores de precisão e revocação próximos, o que não acontece em sistemas de busca por palavras chaves (AMANQUI *et al.*, 2013). No SWI Gazetteer foi possível obter 0,566 de precisão e 0,643 de revocação.

No entanto, devido ao fato desses repositórios não conterem todos os locais pesquisados, a revocação dos dados foi de, no máximo, 0,54. Evidenciando assim que a criação de outro *endpoint* para disponibilizar dados sobre localidades relacionadas a coletas biológicas é válida. Outro motivo que torna válida a criação do SWI Gazetteer e a disponibilização de seus dados

¹ <http://dbpedia.org/sparql>

² <http://factforge.net/>

³ <http://www.geosparql.org/>

em um *endpoint* GeoSPARQL, é a evidente falta de mecanismos que contenham informações sobre localidades brasileiras em *endpoints* e que possibilitem o uso de funções geoespaciais. Fato esse que foi evidenciado na busca por *endpoints* que contivessem dados do Geonames, WikiMapia e Wikipédia, realizada neste trabalho. Foi somente possível recuperar informações sobre municípios brasileiros em um único *endpoint* que realizava consultas geoespaciais.

A disponibilização de um *endpoint* GeoSPARQL é importante devido à necessidade de se realizar consultas geoespaciais que contenham significado semântico. Isso fica claro no exemplo da consulta (Código-fonte 4) que mostra uma busca por todas as fazendas que estão dentro de uma reserva florestal. Tal consulta, utilizando-se sistemas de busca por palavras-chave, não é eficiente, conforme relatado por [Amanqui et al. \(2013\)](#). Além disso, a realização manual desse tipo de consulta por especialistas seria uma tarefa muito laboriosa, logo o uso de buscas semânticas é extremamente útil para consultas complexas como essa.

Código-fonte 4: Consulta por fazendas dentro de áreas protegidas.

```

1 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns/#>
2 PREFIX geo:<http://www.opengis.net/ont/geosparql/#>
3 PREFIX geof:<http://www.opengis.net/def/function/geosparql/>
4 PREFIX dbp:<http://dbpedia.org/ontology/>
5 PREFIX lgdo:<http://linkedgeodata.org/ontology/>
6 SELECT ?p ?a ?w1 ?w2
7 WHERE {
8     ?p rdf:type lgdo:Farm .
9     ?p geo:hasGeometry ?g2 .
10    ?g2 geo:asWKT ?w2 .
11    ?a rdf:type dbp:ProtectedArea .
12    ?a geo:hasGeometry ?g1 .
13    ?g1 geo:asWKT ?w1 .
14    FILTER geof:sfWithin(?w2, ?w1)
15 }
```

6.5 Avaliação do Algoritmo de *GeoParser*

Para verificar a eficiência do algoritmo de *GeoParser* (seção 5.2.4) quanto a precisão e revocação dos dados [Manning, Raghavan e Schütze \(2008\)](#), as consultas listadas na Tabela 2 foram avaliadas. Nessa avaliação uma coleção de referência contendo quais dados do *SWI Gazetteer* são relevantes para cada consulta da Tabela 2 foi criada pelos especialistas. Como resultado dessa avaliação, foi possível concluir que o algoritmo de busca semântica geoespacial conseguiu bons resultados de revocação para a maioria dos casos listados na Tabela 2, conforme exibido na Figura 31. Isso demonstra que o algoritmo é capaz de encontrar as informações relevantes para o usuário que estão armazenadas no *SWI Gazetteer*.

Analisando as consultas detalhadamente, foi possível notar que a *query*, “Locais dentro do parque nacional do Jaú” com valor de revocação 0.5, menor dentre as outras consultas, obteve esse resultado devido ao fato dos especialistas marcarem os locais, que devidamente ocorrem dentro do parque, com coordenadas geográficas apontando para as proximidades do polígono descrito pelo IBGE. Desse modo, ao executar o algoritmo de busca semântica e computar as coordenadas que estão dentro do polígono, esses locais ficam fora dos resultados retornados. As próximas versões do *SWI Gazetteer* irão amenizar a ocorrência desse comportamento nas informações geográficas criadas pelos usuários.

Verificando a precisão das informações, o *SWI Gazetteer* conseguiu bons resultados para maioria das consultas testadas, sendo que os valores mais baixos são de 0.71 e 0.75, conforme exibido na Figura 31. Assim, é possível concluir que o algoritmo de busca semântica é capaz de recuperar informações relevantes e também precisas para as consultas do usuário.

Analisando essas duas consultas, que tiveram valores de precisão 0.71 e 0.75 (“Rios do estado do Amazonas” e “Places part-of município de Manaus”), foi possível notar que o algoritmo retornou resultados inconsistentes devido ao fato de algumas informações estarem mapeadas de forma imprecisa. Os casos onde esses erros ocorrem serão aprimorados com a colaboração dos usuários por meio da interface criada na seção 5.1. Nela é possível inserir a classe correta para localidades armazenadas no *SWI Gazetteer*, amenizando assim a recuperação de locais imprecisos.

Além da verificação das consultas separadamente, foi realizada a análise da média dos valores de precisão, revocação e F-measure para as consultas da Tabela 2, exibido na Figura 32. Os resultados obtidos mostram que o algoritmo de busca semântica geoespacial possui bons resultados para todas as medidas. Sendo assim, foi utilizado o valor de F-measure obtido, 0.824, que demonstra que o algoritmo é capaz de recuperar com maior eficácia e exatidão os itens considerados relevantes para a consulta do usuário. Situação essa que não acontece com sistemas de busca por palavras chaves, conforme relatado por Amanqui *et al.* (2013). Vale ressaltar que, o algoritmo de busca semântica proposto por Amanqui *et al.* (2013), somente é capaz de recuperar espécimens, por exemplo, tipos de insetos, aves, mamíferos entre outros. Portanto, o mesmo não foi comparado com o algoritmo de busca semântica geoespacial proposto para recuperar localidades.

A busca por localidades é essencial para que biólogos possam aprimorar os registros geográficos do *SWI Gazetteer*, sendo essa uma das funcionalidades frequentemente utilizadas no sistema. Para reduzir o tempo de resposta das consultas GeoSPARQL, criadas pelo módulo de reformulação de consultas, sub-seção 5.2.3, as três principais *triple stores* open source (que implementam funções geoespaciais), descritas por Garbis, Kyzirakos e Koubarakis (2013), foram testadas para se determinar a mais rápida.

Figura 31 – Resultados de precisão, revocação e F-measure para as consultas realizadas pelo algoritmo de busca semântica geoespacial, criado no módulo de reformulação de consultas.

Figura 32 – Média dos Resultados de precisão, revocação e F-measure para o algoritmo de busca semântica geoespacial criado no módulo de reformulação de consultas.

6.6 GeoSPARQL Triple Stores

Para testar as *triple stores* Parliament⁴, Strabon⁵ e UseekM⁶ foram utilizadas buscas semânticas que continham funções GeoSPARQL não topológicas, junções espaciais e seleções espaciais. Para realizar esse teste, uma base de dados contendo 40 mil pontos sobre localidades referentes ao repositório do Geonames⁷ foi selecionada.

Com intuito de testar a disponibilidade do serviço GeoSPARQL nessas *triple stores*, um *Web Client* Java usando requisições RESTful (acessando os dados das *triple stores* usando requisições HTTP) foi desenvolvido⁸. Esse cliente permite realizar as solicitações às *triple stores* segundo os padrões de LOD (HEATH; BIZER, 2011).

A máquina utilizada para executar os experimentos possui as seguintes configurações: processador: Intel Xeon X5560 2.80GHz com 8 cores e 8 MB de cache, 8 GB memória e disco rígido de 1 TB com 7200 rpm. Nessa máquina foram instaladas as *triple stores* Parliament 2.7.4, Strabon 3.2 e uSeekM 1.2.1, todos utilizando o Tomcat 6.

Para medir o tempo de resposta para cada consulta, foi usado o tempo necessário desde o envio de uma consulta por requisição HTTP até o retorno dos resultados para o *Web Client*. Além disso, um tempo limite de 30 minutos para a execução de cada requisição foi definido.

Para iniciar os testes foram implementadas 6 consultas com funções GeoSPARQL divididas em 3 categorias:

1. Funções não topológicas, ou seja que não descrevem conceitos de vizinhança, incidência e sobreposição. Para essas consultas, foram utilizadas as funções *geof:convexHull* e *geof:buffer*, ambas retornam uma geometria referente a algum objeto contido na *triple store*.
2. Junções espaciais, definidas pelas funções *geof:sfWithin* e *geof:sfIntersects*. Essas funções tem como objetivo verificar se, dado dois objetos A e B, o objeto A está contido ou intercepta o objeto B.
3. Consultas de seleção espacial, ou seja, verifica-se se um polígono ou ponto, informados manualmente, contém ou intercepta algum ponto ou polígono, armazenado na *triple store*.

Cada uma dessas implementações de consultas está descrita no Apêndice A.

⁴ <http://parliament.semwebcentral.org/>

⁵ <http://www.strabon.di.uoa.gr/>

⁶ <https://dev.opensahara.com/projects/useekm>

⁷ <http://geographica.di.uoa.gr/datasets/geonames.tar.gz>

⁸ <http://dx.doi.org/10.5281/zenodo.12981>

Query	Tipo	Strabon	uSeekM	Parliament
1	Funções Não Topológicas	1.51 seg	1.21 seg	15.33 seg
2		6.68 seg	0.294 seg	26.6 seg
3	Junções Espaciais	2 seg	> 30 min	> 30 min
4		1.85 seg	> 30 min	> 30 min
5	Seleção Espacial	0.114 seg	0.364 seg	10.27 seg
6		0.111 seg	0.207 seg	15.94 seg

Tabela 3 – Resultados da execução das consultas geoespaciais nas *triple store* Strabon, uSeekM e Parliament.

6.6.1 Comparação das buscas Geoespaciais entre triplesotres

O primeiro conjunto de consultas testadas foram as não topológicas, *queries* 1 e 2 no Apêndice A. Nessas consultas, a *triple store* uSeekM obteve melhor desempenho em relação às outras duas. Esse resultado se deve ao fato das funções não topológicas serem computacionalmente intensivas e, neste ponto, a implementação do uSeekM trabalha melhor com dados de I/O, como relatado por Garbis, Kyzirakos e Koubarakis (2013).

Nas consultas referentes às junções espaciais, a *triple store* Strabon obteve os melhores resultados. Esse resultado se deu devido ao fato das *triple store* uSeekM e Parliament necessitarem realizar um produto cartesiano de suas coordenadas para assim verificar se um objeto A está contido ou intercepta um objeto B. Além disso, vale ressaltar que ambas, uSeekM e Parliament, não concluíram a consulta no tempo hábil definido, como apresentado na Tabela 3.

Segundo Kolas, Emmons e Dean (2009), o desempenho de uma consulta para uma única tripla no Parliament depende do número de elementos que estão vinculados no padrão sujeito, predicado e objeto. No pior caso, para o processamento de uma consulta, quando o Parliament realiza um produto cartesiano de seus elementos, a operação para procurar um elemento necessita de $O(n^{2/3})$ comparações. Visto que em alguns casos o Parliament utiliza o processo de busca para inserir novos elementos, no pior caso sua inserção também será $O(n^{2/3})$.

Por fim, no conjunto de consultas referentes às seleções espaciais, as *triple store* Strabon e uSeekM obtiveram resultados próximos, sendo que a Strabon foi um pouco melhor. Isso é devido ao fato que ambos utilizam o SGBD PostGIS para realizar o processo das *queries* que utilizam as seleções espaciais, sendo que, após esse passo, a uSeekM continua a realizar seu plano de consulta usando a *triple store* nativa do Sesame e isso acarreta em um pouco mais de sobrecarga.

Além da verificação das consultas, foi realizado um estudo sobre a arquitetura e funcionalidades implementadas pelas três *triple stores*. Nesse estudo, foi verificado se elas implementavam as funcionalidades referentes a RDFS+OWL (OWL 2, OWL 2 RL ou DL) ou algum suporte a regras.

Ao se verificar a *triple store* uSeekM, foi possível constatar algumas limitações, como:

1. A necessidade de todas as geometrias especificarem a utilização do sistema de referência espacial 4326/WGS84.
2. A utilização da função ORDER BY é diferente da especificada na linguagem SPARQL (ações para corrigir isso estão em aberto).
3. Algumas funções geométricas não possuem total suporte:
 - a) A função geobuffer apenas computa as coordenadas que estão assumidas em um mesmo plano cartesiano.
 - b) A função getSRID possui alguns problemas para retornar URIs referentes aos pontos.
4. Não é possível realizar *Query Rewrite Extension* e os desenvolvedores não possuem planos para suportá-la.
5. A indexação de declarações com BNodes não é suportada, ou seja, não é possível criar um nó em um grafo RDF sem que sua URI seja especificada.
6. Não possui suporte a OWL ou regras em seu *backend* padrão.
7. A API Jena não consegue se comunicar diretamente com seu serviço SPARQL, sendo necessário realizar modificações.

Ao se verificar a *triple store* Strabon foi possível constatar que:

1. Ela não possui suporte a OWL ou regras.
2. Não é possível realizar *Query Rewrite Extension*, no entanto os desenvolvedores possuem planos para implementar essa funcionalidade.
3. O modelo da arquitetura, em que o Strabon foi implementado, não permite que sua SAIL RDBMS seja “empilhável”, ou seja, não é possível usar outras SAILS em conjunto com a implementação do Strabon.
4. A inserção de arquivos contendo triplas precisa ser realizada localmente, caso contrário o Strabon barra a inserção de conteúdo.
5. A API Jena pode se comunicar diretamente com seu serviço SPARQL.

Ao se verificar a *triple store* Parliament foi possível constatar que:

1. O Parliament possui a vantagem de utilizar arquivos para implementar o armazenamento de suas triplas, sendo assim, é mais fácil a realização de backups ou deslocamento de seus dados. No entanto, esses arquivos são fáceis de ser corrompidos e isso acarreta inviabilização de seus dados para uso.

2. Possui suporte para regras.
3. A API Jena pode se comunicar diretamente com seu serviço SPARQL.
4. A criação de grafos não pode ser feita remotamente, ou seja, ao inserir triplas é necessário implementar um método para criá-las. Para isso, é necessário usar funções disponíveis em sua API, ou seja, não é possível criar grafos diretamente com a API Jena.

Com estes resultados, é possível verificar que todas as *triple stores* necessitam de ajustes. Sendo que a *triple store* Strabon possui os melhores resultados quanto ao tempo de resposta para as consultas. Portanto, a implementação do *SWI Gazetteer* utiliza essa *triple store* para o armazenar os dados e realizar as consultas semânticas geoespaciais.

6.7 Considerações Finais

Neste capítulo, foram descritos os resultados obtidos com o desenvolvimento deste trabalho. Inicialmente, uma análise dos dados dos repositórios SpeciesLink e GBIF foi realizada, nessa análise foi mostrado como os dados geográficos sobre biodiversidade são imprecisos.

Após a utilização do *SWI Gazetteer* para aprimorar as coordenadas geográficas dos repositórios GBIF e SpeciesLink, foi possível obter um aumento de 102% no número de registros com coordenadas geográficas, por meio do agrupamento de dados utilizando algoritmo K-means e a técnica de sumarização de coordenadas.

Ao se comparar os registros originais do SpeciesLink com os processados pelo *SWI Gazetteer*, foi possível verificar que o *SWI Gazetteer* foi capaz de obter mais registros com coordenadas precisas, 86%, que os registros originais, 67%.

Também foi possível verificar que o algoritmo de geoParser obteve bons resultados de F-measure (0.8), sendo assim útil para os biólogos realizarem consultas complexas no *SWI Gazetteer*.

Por fim, foi realizada uma análise das principais *triple store* open source que trabalham com funções geográficas (por meio de GeoSPARQL). Foram evidenciados a performance e os principais pontos a serem abordados para o aprimoramento desses sistemas.

CONCLUSÃO

Nesta dissertação de mestrado, foi desenvolvido um dicionário geográfico colaborativo (*Gazetteer*) utilizando tecnologias da web semântica para recuperar e aprimorar informações referentes à dados de biodiversidade (locais das coletas biológicas). Por meio desse *Gazetteer*, o *SWI Gazetteer*, foi possível aprimorar as coordenadas geográficas das localidades contidas em amostras significativas de registros do SpeciesLink e GBIF.

As principais contribuições deste trabalho são detalhadas a seguir:

- **Arquitetura do SWI Gazetteer:** A arquitetura do *SWI Gazetteer* apresenta um novo processo, para desenvolver *Gazetteers*, baseado nos pontos ressaltados por [Kessler, Janowicz e Bishr \(2009\)](#), [OGC \(2012\)](#), [Beard \(2012\)](#), [Bereta, Smeros e Koubarakis \(2013\)](#) para atender os desafios da próxima geração de *Gazetteers*. Para isso, essa arquitetura utiliza práticas de VGI juntamente com tecnologias disponíveis na web semântica.
- **Aprimoramento das coordenadas geográficas:** Utilizando técnicas de agrupamento de informações, juntamente com o método de sumarização de coordenadas, é possível obter uma melhora significativa no número de registros com coordenadas geográficas em conjuntos representativos de registros de coletas. Após realizar a análise das localidades associadas pelo *SWI Gazetteer*, constatou-se uma precisão de 86% nas suas coordenadas.
- **Algoritmo para realizar consultas semânticas geoespaciais:** Ele permite a realização de consultas complexas (buscas semânticas), inseridas pelo usuário em linguagem natural, utilizando as ontologias presentes no *SWI Gazetteer*. Uma análise desse algoritmo mostrou bons resultados de revocação e precisão. A possibilidade de criar consultas, que buscam informações com base nas relações entre dados (ontologias), oferece vantagens em relação a consultas baseadas apenas em *strings* ([AMANQUI et al., 2013](#)). Vale ressaltar que esse é um dos desafios para a próxima geração de *gazetteers*, segundo [Kessler, Janowicz e Bishr \(2009\)](#), [Moura e Jr. \(2013\)](#), que já é suportado pelo *SWI Gazetteer*.

- **Mapeamento das informações para LOD:** O método de mapeamento de localidades, dentro da arquitetura proposta, permite a integração de grandes conjuntos de dados geográficos e de contribuições voluntárias à LOD, sendo esse outro desafio para a próxima geração de *Gazetteers* (MOURA; JR., 2013). Além disso, após os experimentos conduzidos, fica evidente também a contribuição do *SWI Gazetteer*, ao disponibilizar seus dados em formato semântico, para incrementar o número de novas localidades com informações na LOD. Grandes repositórios de dados, como DBpedia e Geonames, não contém informações sobre localidades importantes para biodiversidade na Amazônia presentes no *SWI Gazetteer*.

Em suma, a arquitetura apresentada atende aos desafios para a próxima geração de *Gazetteers*, listados na literatura, e fornece um novo processo para lidar com a imprecisão das coordenadas geográficas fornecidas pela contribuição voluntária de usuários, utilizando a Lei de Linus Haklay *et al.* (2010). Desse modo, o *SWI Gazetteer* permite alcançar o objetivo desta dissertação contribuindo para amenizar as lacunas na área de RIG e melhorando os resultados das contribuições geográficas voluntárias, realizadas pelos biólogos e disponibilizadas nos repositórios SpeciesLink e GBIF.

7.1 Produções Científicas

Este projeto de mestrado resultou na publicação de um artigo em periódico internacional, Qualis-A2, e de outro em uma conferência internacional, Qualis-B1, que são listados a seguir:

Cardoso, S.D.; Serique, K.J.; Amanqui, F.K.; Campos Dos Santos, J.L.; Moreira, D.A., *SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data*, **Future Generation Computer Systems**, Available online 16 May 2015, ISSN 0167-739X, <<http://dx.doi.org/10.1016/j.future.2015.05.006>>.

Cardoso, S.D.; Serique, K.J.; Amanqui, F.K.; Campos Dos Santos, J.L.; Moreira, D.A., "A Gazetteer for Biodiversity Data as a Linked Open Data Solution," **WETICE Conference** (WETICE), 2014 IEEE 23rd International , vol., no., pp.435,440, 23-25 June 2014 <[doi:10.1109/WETICE.2014.19](https://doi.org/10.1109/WETICE.2014.19)>

7.2 Dificuldades e Limitações

O presente trabalho apresentou algumas dificuldades e limitações durante seu projeto e implementação:

- A pouca disponibilidade de tempo dos especialistas que possuem conhecimento das regiões contidas no *Gazetteer* para realização dos testes.

- O suporte limitado de algumas funcionalidades nas ferramentas utilizadas para o desenvolvimento, necessitando de várias mudanças na implementação do *Gazetteer* ao decorrer do tempo.

7.3 Trabalhos Futuros

Como trabalhos futuros decorrentes dessa dissertação destacam-se:

- Melhorias na inclusão de novas anotações (sinônimos) nas classes da ontologia utilizada para auxiliar na resolução dos topônimos.
- Melhorias na implementação do algoritmo de mapeamento de dados e busca semântica, utilizando melhor os conceitos descritos por [Zhu et al. \(2012\)](#), para melhorar a representação dos lugares e recuperação de mais informações imprecisas.
- A integração de dados do *SWI Gazetteer* dentro da Web Semântica não é uma tarefa fácil, principalmente devido à complexidade e a dispersão de conteúdo. Dessa forma, as futuras versões do *SWI Gazetteer* deverão abranger o problema de expansão de consultas geográficas, ou seja, após uma consulta ser realizada num *Gazetteer* e nenhum resultado relevante ser encontrado, tal consulta pode ser enviada para outros *Gazetteers*. Esse tipo de abordagem pode ajudar também na manutenção de *Gazetteers*, onde locais de um *Gazetteer* podem ser inseridos em outro, e na resposta para usuários que precisam recuperar resultados de vários repositórios ([KESSLER; JANOWICZ; BISHR, 2009](#)).
- Aprimoramento da técnica de agrupamento de dados para permitir a utilização de *Active Learning*, onde o usuário poderia opinar sobre as localidades agrupadas (agrupamento dos dados semi-supervisionado). Essa opinião poderia incluir novos atributos, como por exemplo, o nome das espécies.
- Um mecanismo para oferecer os registros aprimorados com coordenadas geográficas para instituições, como o INPA, ou repositórios, como o GBIF e SpeciesLink.

REFERÊNCIAS

ALHO, C. The value of biodiversity. **Brazilian Journal of Biology**, scielo, v. 68, p. 1115 – 1118, 11 2008. ISSN 1519-6984. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-69842008000500018&nrm=iso. Citado na página 27.

AMANQUI, F. K.; SERIQUE, K. J.; LAMPING, F.; ALBUQUERQUE, A. C. F.; SANTOS, J. L. C. D.; MOREIRA, D. A. Semantic search architecture for retrieving information in biodiversity repositories. In: BAX, M. P.; ALMEIDA, M. B.; WASSERMANN, R. (Ed.). **ONTOBRAS**. CEUR-WS.org, 2013. (CEUR Workshop Proceedings, v. 1041), p. 83–93. Disponível em: <http://dblp.uni-trier.de/db/conf/ontobras/ontobras2013.html#AmanquiSLASM13>. Citado 6 vezes nas páginas 27, 51, 99, 100, 101 e 107.

AMANQUI, F. K. M. **Uma arquitetura para sistemas de busca semântica para recuperação de informações em repositórios de biodiversidade**. Dissertação (Mestrado) — Universidade de São Paulo, 2014. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-03072014-150009/>. Citado 4 vezes nas páginas 15, 47, 54 e 56.

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. **Estatística Aplicada A Administração e Economia**. 6. ed. [S.l.]: Trilha, 2013. Citado na página 98.

ASLAM, J. A.; PELEKHOV, E.; RUS, D. The star clustering algorithm for static and dynamic information organization. **J. Graph Algorithms Appl.**, v. 8, p. 95–129, 2004. Disponível em: <http://dblp.uni-trier.de/db/journals/jgaa/jgaa8.html#AslamPR04>. Citado 3 vezes nas páginas 77, 93 e 95.

AUER, S.; LEHMANN, J.; HELLMANN, S. Linkedgeodata: Adding a spatial dimension to the web of data. In: **Proceedings of the 8th International Semantic Web Conference**. Berlin, Heidelberg: Springer-Verlag, 2009. (ISWC '09), p. 731–746. ISBN 978-3-642-04929-3. Disponível em: http://dx.doi.org/10.1007/978-3-642-04930-9_46. Citado 2 vezes nas páginas 51 e 78.

BATTLE, R.; KOLAS, D. Enabling the geospatial semantic web with parliament and geosparql. **Semantic Web**, v. 3, n. 4, p. 355–370, 2012. Disponível em: <http://dblp.uni-trier.de/db/journals/semweb/semweb3.html#BattleK12>. Citado 2 vezes nas páginas 56 e 57.

BEARD, K. A semantic web based gazetteer model for vgi. In: **Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information**. New York, NY, USA: ACM, 2012. (GEOCROWD '12), p. 54–61. ISBN 978-1-4503-1694-1. Disponível em: <http://doi.acm.org/10.1145/2442952.2442962>. Citado 4 vezes nas páginas 38, 71, 87 e 107.

BECHHOFFER, S.; HARMELEN, F. van; HENDLER, J.; HORROCKS, I.; MCGUINNESS, D. L.; PATEL-SCHNEIDER, P. F.; STEIN, L. A. **OWL Web Ontology Language Reference**. <http://www.w3.org/TR/owl-ref/>, 2004. Citado 2 vezes nas páginas 53 e 54.

- BERETA, K.; SMEROS, P.; KOUBARAKIS, M. Representation and querying of valid time of triples in linked geospatial data. In: CIMIANO, P.; CORCHO, O.; PRESUTTI, V.; HOLLINK, L.; RUDOLPH, S. (Ed.). **ESWC**. Springer, 2013. (Lecture Notes in Computer Science, v. 7882), p. 259–274. ISBN 978-3-642-38288-8. Disponível em: <<http://dblp.uni-trier.de/db/conf/esws/eswc2013.html#BeretaSK13>>. Citado 5 vezes nas páginas 56, 57, 71, 87 e 107.
- BERNERS-LEE, T.; CHEN, Y.; CHILTON, L.; CONNOLLY, D.; DHANARAJ, R.; HOLLENBACH, J.; LERER, A.; SHEETS, D. Tabulator: Exploring and analyzing linked data on the semantic web. In: **Proceedings of the 3rd International Semantic Web User Interaction**. [s.n.], 2006. Disponível em: <<http://swui.semanticweb.org/swui06/papers/Berners-Lee/Berners-Lee.pdf>>. Citado na página 51.
- BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 34–43, maio 2001. Disponível em: <<http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>>. Citado 3 vezes nas páginas 15, 47 e 48.
- BISBY, F. A. The Quiet Revolution: Biodiversity Informatics and the Internet. **Science**, American Association for the Advancement of Science, v. 289, n. 5488, p. 2309–2312, set. 2000. ISSN 1095-9203. Disponível em: <<http://dx.doi.org/10.1126/science.289.5488.2309>>. Citado na página 27.
- BOLEY, S. T. H.; WAGNER, G. Design rationale of RuleML: A markup language for semantic web rules. In: CRUZ, I. F.; DECKER, S.; EUZENAT, J.; MCGUINNESS, D. L. (Ed.). **Proc. Semantic Web Working Symposium**. Stanford University, California: [s.n.], 2001. p. 381–402. Citado na página 47.
- BORGES, K. A. V.; LAENDER, A. H. F.; MEDEIROS, C. B.; DAVIS, C. A. Discovering geographic locations in web pages using urban addresses. In: PURVES, R.; JONES, C. (Ed.). **GIR**. ACM, 2007. p. 31–36. ISBN 978-1-59593-828-2. Disponível em: <<http://dblp.uni-trier.de/db/conf/gir/gir2007.html#BorgesLMD07>>. Citado na página 40.
- BUYUKOKKTEN, O.; CHO, J.; GARCIA-MOLINA, H.; GRAVANO, L.; SHIVAKUMAR, N. Exploiting geographical location information of web pages. In: **ACM SIGMOD Workshop on The Web and Databases (WebDB'99)**. [s.n.], 1999. Disponível em: <<http://ilpubs.stanford.edu:8090/415/>>. Citado na página 36.
- CARDOSO, D. S.; SERIQUE, J. K.; AMANQUI, K. F.; SANTOS, L. J. Campos dos; MOREIRA, A. D. A gazetteer for biodiversity data as a linked open data solution. In: **WETICE**. [S.l.]: Web2Touch 2014 - Modelling the Collaborative Web Knowledge, 2014. Citado 6 vezes nas páginas 15, 78, 79, 80, 92 e 96.
- CARDOSO, S. D.; AMANQUI, F. K.; SERIQUE, K. J.; SANTOS, J. L. dos; MOREIRA, D. A. **SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data**. 2015. - p. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167739X15001818>>. Citado 8 vezes nas páginas 15, 16, 72, 93, 94, 96, 97 e 99.
- CHIANG, M.-T.; MIRKIN, B. Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. **Journal of Classification**, Springer-Verlag, v. 27, n. 1, p. 3–40, 2010. ISSN 0176-4268. Disponível em: <<http://dx.doi.org/10.1007/s00357-010-9049-5>>. Citado na página 94.

DAVIS, C. A.; FONSECA, F. T.; BORGES, K. A. V. A flexible addressing system for approximate geocoding. In: **GeoInfo**. [s.n.], 2003. Disponível em: <http://dblp.uni-trier.de/db/conf/geoinfo/geoinfo2003.html#DavisFB03>. Citado na página 41.

DENTLER, K.; CORNET, R.; TEIJE, A. ten; KEIZER, N. de. Comparison of reasoners for large ontologies in the owl 2 el profile. **Semant. web**, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 2, n. 2, p. 71–87, abr. 2011. ISSN 1570-0844. Disponível em: <http://dx.doi.org/10.3233/SW-2011-0034>. Citado na página 54.

EGENHOFER, M. J. Toward the semantic geospatial web. In: VOISARD, A.; CHEN, S.-C. (Ed.). **ACM-GIS**. ACM, 2002. p. 1–4. ISBN 1-58113-591-2. Disponível em: <http://dblp.uni-trier.de/db/conf/geoinfo/geoinfo2002.html#Egenhofer02>. Citado 2 vezes nas páginas 40 e 55.

GARBIN, E.; MANI, I. Disambiguating toponyms in news. In: **Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. p. 363–370. Disponível em: <http://dx.doi.org/10.3115/1220575.1220621>. Citado na página 41.

GARBIS, G.; KYZIRAKOS, K.; KOUBARAKIS, M. Geographica: A benchmark for geospatial rdf stores. In: . [s.n.], 2013. abs/1305.5653. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr1305.html#abs-1305-5653>. Citado 3 vezes nas páginas 59, 101 e 104.

GBIF. Data portal of the global biodiversity information facility (gbif). In: **online**. [s.n.], 2014. Disponível em: <http://www.gbif.org>. Citado 3 vezes nas páginas 28, 29 e 89.

GEY, F.; LARSON, R.; SANDERSON, M.; JOHO, H.; CLOUGH, P.; PETRAS, V. Geoclef: The clef 2005 cross-language geographic information retrieval track overview. In: . [s.n.], 2006. p. 908–919. Disponível em: http://dx.doi.org/10.1007/11878773_101. Citado na página 37.

GIL, F. B.; KOZIEVITCH, N. P.; TORRES, R. da S. A geographic annotation service for biodiversity systems. In: BOGORNY, V.; VINHAS, L. (Ed.). **GeoInfo**. MCT/INPE, 2010. p. 33–44. Disponível em: <http://dblp.uni-trier.de/db/conf/geoinfo/geoinfo2010.html#GilKT10>. Citado 5 vezes nas páginas 28, 62, 63, 64 e 65.

GILES, L. C.; BOLLACKER, K.; LAWRENCE, S. Citeseer: An automatic citation indexing system. In: WITTEN, I.; AKSCYN, R.; SHIPMAN, F. M. (Ed.). **Digital Libraries 98 - The Third ACM Conference on Digital Libraries**. Pittsburgh, PA: ACM Press, 1998. p. 89–98. Disponível em: <http://citeseer.ist.psu.edu/108208.html>. Citado na página 36.

GIMENEZ, P. J. A.; TANAKA, A. K.; BAIAO, F. A geo-ontology to support the semantic integration of geoinformation from the national spatial data infrastructure. In: **GeoInfo**. MCT/INPE, 2013. p. 103–114. Disponível em: <http://dblp.uni-trier.de/db/conf/geoinfo/geoinfo2013.html#GimenezTB13>. Citado 2 vezes nas páginas 49 e 50.

GIUNCHIGLIA, F.; DUTTA, B.; MALTESE, V.; FARAZI, F. A facet-based methodology for the construction of a large-scale geospatial ontology. **J. Data Semantics**, v. 1, n. 1, p. 57–73, 2012. Disponível em: <http://dblp.uni-trier.de/db/journals/jodsn/jodsn1.html#GiunchigliaDMF12>. Citado na página 50.

GOODCHILD, M. F.; HILL, L. L. Introduction to digital gazetteer research. **International Journal of Geographical Information Science**, v. 22, n. 10, p. 1039–1044, 2008. Disponível

em: <http://dblp.uni-trier.de/db/journals/gis/gis22.html#GoodchildH08>. Citado 2 vezes nas páginas 33 e 64.

GOUVÊA, C. **Uma Abordagem para o Enriquecimento de Gazetteers a partir de Notícias visando o Georreferenciamento de Textos na Web**. Dissertação (Mestrado) — Universidade Católica de Pelotas, 2009. Disponível em: <http://ppginf.ucpel.tche.br/DM-Arquivos/2009/PPGINF-UCPel-DM-2009-1-001.pdf>. Citado 10 vezes nas páginas 15, 36, 37, 40, 42, 43, 44, 45, 66 e 77.

GRUBER, T. Ontology of folksonomy: a mash-up of apples and oranges. In: **First on-Line conference on Metadata and Semantics Research (MTSR'05)**. [s.n.], 2005. Disponível em: <http://tomgruber.org/writing/ontology-of-folksonomy.htm>. Citado 2 vezes nas páginas 39 e 49.

GUIZZARDI, G. Theoretical foundations and engineering tools for building ontologies as reference conceptual models. **Semantic Web**, v. 1, n. 1-2, p. 3–10, 2010. ISSN 1570-0844. Citado na página 49.

HAKLAY, M. M. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. 2008. Disponível em: http://www.ucl.ac.uk/~ucfamha/OSM%20data%20analysis%20070808_web_orig.pdf. Citado na página 78.

HAKLAY, M. M.; BASIOUKA, S.; ANTONIOU, V.; ATHER, A. How many volunteers does it take to map an area well? the validity of linus law to volunteered geographic information. **Cartographic Journal**, The, Maney Publishing, p. 315–322, nov. 2010. ISSN 0008-7041. Disponível em: <http://dx.doi.org/10.1179/000870410x12911304958827>. Citado 6 vezes nas páginas 15, 67, 68, 69, 76 e 108.

HEATH, T.; BIZER, C. **Linked Data: Evolving the Web into a Global Data Space**. 1st. ed. Morgan & Claypool, 2011. ISBN 9781608454303. Disponível em: <http://linkeddatatbook.com/>. Citado 3 vezes nas páginas 51, 87 e 103.

HILL, L. L. Core elements of digital gazetteers: Placenames, categories, and footprints. In: BORBINHA, J. L.; BAKER, T. (Ed.). **ECDL**. Springer, 2000. (Lecture Notes in Computer Science, v. 1923), p. 280–290. ISBN 3-540-41023-6. Disponível em: <http://dblp.uni-trier.de/db/conf/ercimdl/ecdl2000.html#Hill00>. Citado na página 37.

JAIN, P.; HITZLER, P.; YEH, P. Z.; VERMA, K.; SHETH, A. P. A.p.: Linked data is merely more data. In: **In: AAAI Spring Symposium 'Linked Data Meets Artificial Intelligence'**, AAAI. [S.l.]: Press, 2010. p. 82–86. Citado na página 67.

JONES, C. B.; ALANI, H.; TUDHOPE, D. Geographical information retrieval with ontologies of place. In: MONTELLLO, D. R. (Ed.). **COSIT**. Springer, 2001. (Lecture Notes in Computer Science, v. 2205), p. 322–335. ISBN 3-540-42613-2. Disponível em: <http://dblp.uni-trier.de/db/conf/cosit/cosit2001.html#JonesAT01>. Citado na página 36.

JONES, C. B.; PURVES, R.; RUAS, A.; SANDERSON, M.; SESTER, M.; KREVELD, M. van; WEIBEL, R. Spatial information retrieval and geographical ontologies an overview of the spirit project. In: **Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2002. (SIGIR '02), p. 387–388. ISBN 1-58113-561-0. Disponível em: <http://doi.acm.org/10.1145/564376.564457>. Citado na página 37.

JONES, C. B.; PURVES, R. S. Geographical information retrieval. **International Journal of Geographical Information Science**, v. 22, n. 3, p. 219–228, 2008. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/13658810701626343>>. Citado na página 36.

JR., C. A. D.; VELLOZO, H. de S.; PINHEIRO, M. B. A framework for web and mobile volunteered geographic information applications. In: **GeoInfo**. MCT/INPE, 2013. p. 147–157. Disponível em: <<http://dblp.uni-trier.de/db/conf/geoinfo/geoinfo2013.html#DavisVP13>>. Citado 4 vezes nas páginas 38, 61, 63 e 65.

KESSLER, C.; JANOWICZ, K.; BISHR, M. 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 Systems**. New York, NY, USA: ACM, 2009. (GIS '09), p. 91–100. ISBN 978-1-60558-649-6. Disponível em: <<http://doi.acm.org/10.1145/1653771.1653787>>. Citado 8 vezes nas páginas 29, 37, 55, 66, 71, 87, 107 e 109.

KEßLER, C.; MAUÉ, P.; HEUER, J. T.; BARTOSCHEK, T. Bottom-up gazetteers: Learning from the implicit semantics of geotags. In: JANOWICZ, K.; RAUBAL, M.; LEVASHKIN, S. (Ed.). **GeoS**. Springer, 2009. (Lecture Notes in Computer Science, v. 5892), p. 83–102. ISBN 978-3-642-10435-0. Disponível em: <<http://dblp.uni-trier.de/db/conf/geos/geos2009.html#KesslerMHB09>>. Citado na página 39.

KLINKENBERG, B. **CITIZEN SCIENCE AND VOLUNTEERED GEOGRAPHIC INFORMATION: CAN THESE HELP IN BIODIVERSITY STUDIES?** Biodiversity of British Columbia [www.biodiversity.bc.ca]. Lab for Advanced Spatial Analysis, Department of Geography, University of British Columbia, Vancouver, 2013. Disponível em: <<http://ibis.geog.ubc.ca/biodiversity/VGI--VolunteerGeographicInformation.html>>. Citado 4 vezes nas páginas 33, 34, 35 e 74.

KOLAS, D.; EMMONS, I.; DEAN, M. **Efficient Linked-List RDF Indexing in Parliament**. 2009. Citado na página 104.

KOUBARAKIS, M.; KARPATHIOTAKIS, M.; KYZIRAKOS, K.; NIKOLAOU, C.; SIOUTIS, M. Data models and query languages for linked geospatial data. In: EITER, T.; KRENNWALLNER, T. (Ed.). **Reasoning Web**. Springer, 2012. (Lecture Notes in Computer Science, v. 7487), p. 290–328. ISBN 978-3-642-33157-2. Disponível em: <<http://dblp.uni-trier.de/db/conf/rweb/rweb2012.html#KoubarakisKKNS12>>. Citado 5 vezes nas páginas 15, 57, 58, 60 e 78.

LARSON, R. R. Geographic information retrieval and spatial browsing. **GIS and Libraries: Patrons, Maps and Spatial Information**, University of Illinois, p. 81–124, abr. 1996. Citado na página 37.

LARSON, R. R.; FRONTIERA, P. Geographic information retrieval (gir): searching where and what. In: SANDERSON, M.; JŠRVELIN, K.; ALLAN, J.; BRUZA, P. (Ed.). **SIGIR**. ACM, 2004. p. 600. ISBN 1-58113-881-4. Disponível em: <<http://dblp.uni-trier.de/db/conf/sigir/sigir2004.html#LarsonF04>>. Citado na página 41.

LEIDNER, J. **Towards a reference corpus for automatic toponym resolution evaluation**. 2004. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.7914>>. Citado na página 42.

LEMES, P.; FAMV, F.; TESSAROLO, G.; LOYOLA, R. D. Refinando dados espaciais para a conservação da biodiversidade. **Natureza & Conservação**, v. 9, p. 240–243, 2011. Citado 2 vezes nas páginas 28 e 29.

LEVELING, J.; HARTRUMPF, S. University of Hagen at GeoCLEF 2007: Exploring location indicators for geographic information retrieval. In: **Results of the CLEF 2007 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2007 Workshop**. Budapest, Hungary: [s.n.], 2007. ISBN 2-912335-32-9. Disponível em: <http://www.clef-campaign.org/2007/working_notes/levelingCLEF2007.pdf>. Citado na página 43.

LIU, Y.; LI, R.; CHEN, K.; YUAN, Y.; HUANG, L.; YU, H. Kidxs: A geographical knowledge-informed digital gazetteer service. In: **Geoinformatics, 2009 17th International Conference on**. [S.l.: s.n.], 2009. p. 1–6. Citado 3 vezes nas páginas 62, 64 e 65.

MACHADO, I. M.; ALENCAR, R. O. de; JUNIOR, R. de O. C.; DAVIS, C. A. An ontological gazetteer and its application for place name disambiguation in text. **J. Braz. Comp. Soc.**, v. 17, n. 4, p. 267–279, 2011. Disponível em: <<http://dblp.uni-trier.de/db/journals/jbcs/jbcs17.html#MachadoAJD11>>. Citado 5 vezes nas páginas 61, 62, 63, 65 e 77.

MALLEA, A.; ARENAS, M.; HOGAN, A.; POLLERES, A. On blank nodes. In: **Proceedings of the 10th International Conference on The Semantic Web - Volume Part I**. Berlin, Heidelberg: Springer-Verlag, 2011. (ISWC'11), p. 421–437. ISBN 978-3-642-25072-9. Disponível em: <<http://dl.acm.org/citation.cfm?id=2063016.2063044>>. Citado na página 81.

MANGUINHAS, H.; MARTINS, B.; BORBINHA, J. A geo-temporal web gazetteer integrating data from multiple sources. In: **Digital Information Management, 2009. ICDIM 2009. Third International Conference on**. [S.l.: s.n.], 2009. p. 146–153. Citado 3 vezes nas páginas 62, 64 e 65.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715. Citado na página 100.

MCCONNELL, J. J. Active learning and its use in computer science. **SIGCUE Outlook**, ACM, New York, NY, USA, v. 24, n. 1-3, p. 52–54, jan. 1996. ISSN 0163-5735. Disponível em: <<http://doi.acm.org/10.1145/1013718.237526>>. Citado na página 95.

MCDONALD, D. D. Corpus processing for lexical acquisition. In: BOGURAEV, B.; PUSTEJOVSKY, J. (Ed.). **Hardcover**. Cambridge, MA, USA: MIT Press, 1996. cap. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names, p. 21–39. ISBN 0-262-02392-X. Disponível em: <<http://dl.acm.org/citation.cfm?id=265994.265996>>. Citado na página 43.

MOURA, T. H. V. M.; JR., C. A. D. Expansão do conteúdo de um gazetteer: nomes hidrográficos. In: NAMIKAWA, L. M.; BOGORNY, V. (Ed.). **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), 2012. p. 78–83. ISSN 2179-4820. Disponível em: <<http://urlib.net/sid.inpe.br/mtc-m18/2012/12.04.13.31>>. Acesso em: 25 fev. 2014. Citado 3 vezes nas páginas 61, 63 e 65.

MOURA, T. H. V. M. d.; JR., C. A. D. Linked geospatial data: desafios e oportunidades de pesquisa. In: ANDRADE, P. R.; SANTANCHÈ, A. (Ed.). **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), 2013. p. 6. ISSN 2179-4820. Disponível em:

<<http://urlib.net/sid.inpe.br/mtc-m18/2013/12.10.17.50>>. Acesso em: 31 maio 2014. Citado 5 vezes nas páginas 15, 66, 67, 107 e 108.

NOY, N. F.; Mcguinness, D. L. **Ontology Development 101: A Guide to Creating Your First Ontology**. 2001. Online. Disponível em: <<http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html>>. Citado na página 49.

OGC, O. G. C. **OpenGIS Implementation Standard for Geographic information - Simple feature access**. 1.2.1. ed. [S.l.]: OGC, Open Geospatial Consortium, 2011. Citado na página 59.

_____. Geosparql - a geographic query language for rdf data. In: **OGC 11-052r4 Version: 1.0**. [s.n.], 2012. Disponível em: <https://portal.opengeospatial.org/files/?artifact_id=47664>. Citado 6 vezes nas páginas 15, 56, 57, 71, 87 e 107.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. In: **Proceedings of the 7th International World Wide Web Conference**. Brisbane, Australia: [s.n.], 1998. p. 161–172. Disponível em: <citeseer.nj.nec.com/page98pagerank.html>. Citado na página 36.

PARUNDEKAR, R.; KNOBLOCK, C. A.; AMBITE, J. L. Linking and building ontologies of linked data. In: PATEL-SCHNEIDER, P. F.; PAN, Y.; HITZLER, P.; MIKA, P.; 0007, L. Z.; PAN, J. Z.; HORROCKS, I.; GLIMM, B. (Ed.). **International Semantic Web Conference (1)**. Springer, 2010. (Lecture Notes in Computer Science, v. 6496), p. 598–614. ISBN 978-3-642-17745-3. Disponível em: <<http://dblp.uni-trier.de/db/conf/semweb/iswc2010-1.html#ParundekarKA10>>. Citado 2 vezes nas páginas 51 e 78.

PENG, X.; CHEN, R.; CHENG, C.; YAN, X. A folksonomy-ontology-based digital gazetteer service. In: **Geoinformatics, 2010 18th International Conference on**. [S.l.: s.n.], 2010. p. 1–6. Citado 3 vezes nas páginas 62, 63 e 65.

SALTON, G. **Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0-201-12227-8. Citado na página 37.

Santos, J. L. C. dos. **A biodiversity information system in an open data/metadatabase architecture**. Tese (Doutorado) — University of Twente, Enschede, 2003. Disponível em: <<http://doc.utwente.nl/41436/>>. Citado 2 vezes nas páginas 78 e 90.

SANTOS, J. L. C. dos; NETTO, J. F. de M.; CASTRO, A. N. de; ALBUQUERQUE, A. C. F.; FERNEDA, E.; ALONSO, L.; ROCHA, R. L. da C.; MENDONÇA, D. T. de. Ontologias para interoperabilidade de modelos e sistemas de informação de biodiversidade. In: **Congresso Iberoamericano de Telemática**. Co-located with the 6th Iberoamerican Congress on Telematics, Gramado: Iberoamerican Meeting of Ontological Research, 2011. v. 728. Citado na página 51.

SERIQUE, K. J. A. **Anotação de imagens radiológicas usando a web semântica para colaboração científica e clínica**. Dissertação (Mestrado) — Universidade de São Paulo, 2012. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10092012-155249/>>. Citado 4 vezes nas páginas 15, 48, 49 e 52.

SHAMEEM, M.-U.-S.; FERDOUS, R. An efficient k-means algorithm integrated with jaccard distance measure for document clustering. In: **Internet, 2009. AH-ICI 2009. First Asian Himalayas International Conference on**. [S.l.: s.n.], 2009. p. 1–6. Citado 4 vezes nas páginas 77, 93, 94 e 96.

- SILVA, G. C. da; LIMA, T. de S. **RDF e RDFS na Infra-estrutura de Suporte à Web Semântica**. <http://www2.ic.uff.br/~gsilva/slreic.pdf>, 2004. Disponível em: <http://www2.ic.uff.br/~gsilva/slreic.pdf>. Citado 2 vezes nas páginas 52 e 53.
- SMITH, B.; ASHBURNER, M.; ROSSE, C.; BARD, J.; BUG, W.; CEUSTERS, W.; GOLDBERG, L. J.; EILBECK, K.; IRELAND, A.; MUNGALL, C. J.; LEONTIS, N.; ROCCA-SERRA, P.; RUTTENBERG, A.; SANSONE, S.-A.; SCHEUERMANN, R. H.; SHAH, N.; WHETZEL, P. L.; LEWIS, S. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. **Nat Biotech**, Nature Publishing Group, v. 25, n. 11, p. 1251–1255, nov 2007. ISSN 1087-0156. Disponível em: <http://dx.doi.org/10.1038/nbt1346>. Citado na página 49.
- SMITH, D. A.; MANN, G. S. Bootstrapping toponym classifiers. In: **Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (HLT-NAACL-GEOREF '03), p. 45–49. Disponível em: <http://dx.doi.org/10.3115/1119394.1119401>. Citado na página 42.
- SPECIESLINK. Sistema de informação distribuído para coleções biológicas: a integração do species analyst e do sinbiota (fapesp). In: **online**. [s.n.], 2014. Disponível em: <http://splink.cria.org.br/description?criaLANG=pt>. Citado 3 vezes nas páginas 28, 29 e 89.
- WANG, H.; LI, L.; A, P. chao S. Design of geo-ontology based on concept lattice. In: **XXIst ISPRS Congress, Technical Commission II, v. XXXVII, part B2**. [S.l.: s.n.], 2008. p. 715–720. Citado na página 50.
- YIN, Y.; YASUDA, K. Similarity coefficient methods applied to the cell formation problem: A taxonomy and review. **International Journal of Production Economics**, v. 101, n. 2, p. 329–352, 2006. Disponível em: <http://EconPapers.repec.org/RePEc:eee:proeco:v:101:y:2006:i:2:p:329-352>. Citado na página 78.
- ZHU, X.; CHEN, D.; ZHOU, C.; LI, M.; XIAO, W. Cardinal direction relations query modeling based on geo-ontology. **ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, XXXIX-B2, p. 179–184, 2012. Disponível em: <http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XXXIX-B2/179/2012/>. Citado 3 vezes nas páginas 83, 84 e 109.

CONSULTAS UTILIZADAS PARA TESTAR AS *TRIPLE STORE*

Consultas utilizadas durante o experimento de avaliação das triple store.

Código-fonte 5: Consulta Não Topologica utilizando a função convexHull.

```
1 PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
2 PREFIX datasets: <http://geographica.di.uoa.gr/dataset/>
3 PREFIX geonames: <http://www.geonames.org/ontology#>
4 PREFIX opengis: <http://www.opengis.net/def/uom/OGC/1.0/>
5 SELECT (geof:convexHull(?o1) AS ?ret)
6 WHERE {
7     GRAPH datasets:geonames {?s1 geonames:asWKT ?o1}
8 }
```

Código-fonte 6: Consulta Não Topologica utilizando a função buffer.

```
1 PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
2 PREFIX datasets: <http://geographica.di.uoa.gr/dataset/>
3 PREFIX geonames: <http://www.geonames.org/ontology#>
4 PREFIX opengis: <http://www.opengis.net/def/uom/OGC/1.0/>
5 SELECT (geof:buffer(?o1, 4, opengis:metre) AS ?ret)
6 WHERE {
7     GRAPH datasets:geonames {?s1 geonames:asWKT ?o1}
8 }
```

Código-fonte 7: Consulta para verificar junções espaciais.

```
1 PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
2 PREFIX geonames: <http://www.geonames.org/ontology#>
```

```

3 PREFIX lgd: <http://linkedgeodata.org/ontology/>
4 PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
5 SELECT ?s1 ?o1 ?s2 ?o2
6 WHERE {
7     GRAPH dataset:geonames {?s1 geonames:asWKT ?o1}
8     GRAPH dataset:geonames {?s2 geonames:asWKT ?o2}
9     FILTER(?s1 != ?s2).
10    FILTER(geof:sfWithin(?o1, ?o2)).
11 }

```

Código-fonte 8: Consulta para verificar junções espaciais.

```

1 PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
2 PREFIX geonames: <http://www.geonames.org/ontology#>
3 PREFIX lgd: <http://linkedgeodata.org/ontology/>
4 PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
5 SELECT ?s1 ?o1 ?s2 ?o2
6 WHERE {
7     GRAPH dataset:geonames {?s1 geonames:asWKT ?o1}
8     GRAPH dataset:geonames {?s2 geonames:asWKT ?o2}
9     FILTER(?s1 != ?s2).
10    FILTER(geof:sfIntersects(?o1, ?o2)).
11 }

```

Código-fonte 9: Consulta para verificar seleções espaciais.

```

1 PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
2 PREFIX geo: <http://www.opengis.net/ont/geosparql#>
3 PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
4 PREFIX geonames: <http://www.geonames.org/ontology#>
5 SELECT *
6 WHERE {
7     GRAPH dataset:geonames {?s1 geonames:asWKT ?o1}
8     FILTER(geof:sfWithin(?o1, "Polygon((23.93 33.23, 23.93 36.23,
9         22.63 33.23, 22.63 36.23, 23.93 33.23))"^^geo:wktLiteral))
9 }

```

Código-fonte 10: Consulta para verificar seleções espaciais.

```

1 PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
2 PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
3 PREFIX geo: <http://www.opengis.net/ont/geosparql#>
4 PREFIX opengis: <http://www.opengis.net/def/uom/OGC/1.0/>
5 PREFIX geonames: <http://www.geonames.org/ontology#>
6 SELECT (COUNT(?s1) AS ?NumOfTriples)

```

```
7 WHERE {
8     GRAPH dataset:geonames {?s1 geonames:asWKT ?o1}
9     FILTER(geof:sfWithin(?o1, geof:buffer("Polygon((23.93 33.23,
23.93 36.23, 22.63 33.23, 22.63 36.23, 23.93 33.23))"^^geo:
wktLiteral, 3000, opengis:metre))).
10 }
```