
Construção de um *Gazetteer* Colaborativo usando
tecnologias da *Web Semântica* para auxiliar a
Recuperação de Informações Geográficas sobre
Biodiversidade

Silvio Domingos Cardoso

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 01/08/2014

Assinatura: _____

Construção de um *Gazetteer*
Colaborativo usando tecnologias da Web
Semântica para auxiliar a Recuperação
de Informações Geográficas sobre
Biodiversidade

Silvio Domingos Cardoso

Orientador: *Prof. Dr. Dilvan de Abreu Moreira*

Monografia apresentada ao Instituto de Ciências
Matemáticas e de Computação - ICMC-USP, para
o Exame de Qualificação, como parte dos
requisitos para obtenção do título de Mestre em
Ciências - Ciências de Computação e Matemática
Computacional.

USP – São Carlos
Agosto de 2014

Resumo

O Brasil é considerado o país da megadiversidade por abrigar diversas espécies de flora e fauna. Preservar essa diversidade é extremamente importante, pois a manutenção da vida no planeta depende dos ecossistemas que compõem essa biodiversidade. Por isso, numerosos esforços para realizar essa preservação vêm sendo realizados. Atualmente, vários estudos sobre formas de uso, recuperação e acesso a informações sobre biodiversidade vem sendo discutidos amplamente dentro das agendas acadêmicas de pesquisa, pois pesquisadores e especialistas da área precisam ter acesso a esse conhecimento para conseguirem compartilhar suas novas descobertas sobre seres vivos. Porém, essa tarefa se torna difícil devido ao grande volume de informação gerado, ao fato de que repositórios que disponibilizam informações georreferenciadas sobre coleções biológicas não terem essas informações para uma parcela significativa de seus registros e a falta de interoperabilidade entre as informações de diferentes fontes e formatos. Informações geográficas são extremamente importantes para ações de preservação, como a demarcação de áreas de proteção para espécies ameaçadas de extinção. Com o intuito de amenizar alguns desses problemas, este trabalho tem como objetivo construir um Dicionário Geográfico (*Gazetteer*) Colaborativo, usando tecnologias da Web Semântica, para auxiliar a Recuperação de Informações Geográficas sobre Biodiversidade. Para construir esse *Gazetteer*, este trabalho se baseará em técnicas de recuperação de informação geográfica, para lidar com os dados que contêm algum tipo de significado espacial, e da Web Semântica, para incorporar significado a esses dados. A proposta deste projeto é de que a utilização desse *Gazetteer* possibilitará ao usuário, realizar buscas semânticas para conseguir melhores resultados no que diz respeito à precisão e cobertura de dados, além de permitir a atualização e inserção de informações referente a locais de coleta de espécimes no formato de *Linked Open Data*. Resultados preliminares, publicados em artigos em conferências internacionais, serão discutidos.

Sumário

1. Introdução	1
1.1. Justificativa	2
1.2. Objetivo	4
1.3. Organização	5
2. Recuperação de Informação Geográfica	7
2.1. Utilização de VGI no contexto da biodiversidade	7
2.2. Recuperação de Informação Geográfica	10
2.3. <i>Gazetteers</i>	12
2.4. <i>Volunteered Geographic Information</i>	12
2.5. GeoTAGs	13
2.6. GeoParsing	14
2.7. GeoCoding	15
2.8. Resolução de Topônimos	16
2.8.1. Estratégias para Resolução de Topônimos	18
2.9. Considerações Finais	20
3. Web Semântica	21
3.1. Introdução	21
3.2. Ontologias	23
3.2.1. Geo-Ontologias	23
3.3. <i>Linked Open Data</i>	25
3.4. RDF e RDF-Schema	26
3.5. Web Ontology Language - OWL	27
3.6. <i>Reasoners</i>	28
3.7. Busca Semântica Geoespacial	30
3.8. SPARQL e Geo-SPARQL	30
3.8.1. Well-Known Text (WKT)	32
3.9. Considerações Finais	38

Sumário

4. Trabalhos Relacionados	39
4.1. Introdução	39
4.2. Estado da Arte	44
4.3. Precisão de dados utilizando VIG	46
4.4. Considerações Finais	49
5. Ferramentas	51
5.1. Protégé	51
5.2. Triplestore	52
5.2.1. Parliament	52
5.2.2. Strabon	54
5.2.3. uSeekM	55
5.3. API Jena	56
5.4. GWT	57
5.5. OpenLayer	57
5.6. LinkedGeoData	59
5.7. Repositórios de dados utilizados	59
5.7.1. SpeciesLink	60
5.7.2. Portal de busca do GBIF	61
5.7.3. Geonames	61
5.8. Considerações Finais	62
6. Experimentos	63
6.1. Introdução	63
6.2. Análise dos dados utilizados referente aos repositórios SpeciesLink e GBIF	63
6.3. Método para aprimorar coordenadas geográficas	67
6.3.1. Verificação dos locais agrupados	71
6.4. Mapeamento dos dados e disponibilização do <i>Endpoint</i> GeoSPARQL	74
6.5. Avaliação das <i>triplesotres</i>	79
6.5.1. Comparação de buscas Geoespaciais entre <i>triplesotres</i>	81
6.6. Protótipo para o <i>Gazetteer</i> Colaborativo	83
6.7. Arquitetura proposta para desenvolvimento do trabalho	87
6.8. Considerações Finais	92
7. Plano de Trabalho	93
7.1. Metodologia	93
7.2. Atividades previstas e Cronograma	95

Sumário

7.3.	Atividades concluídas até o momento	96
7.4.	Produções Científicas até o momento	96
7.5.	Dificuldades e Limitações	97
8.	Referências	99
A.	Consultas utilizadas para testar as <i>triplestores</i>.	109

Lista de Figuras

2.1. Exemplos de tipos de ambiguidades. Fonte: Gouvêa (2009)	17
2.2. Ambiguidade presente nas entidades geográficas do TGN. Fonte: Gouvêa (2009)	18
2.3. Expressões de Contexto em Português. Fonte: (Gouvêa, 2009)	19
2.4. Exemplo de tipos de evidências para desambiguação de topônimos. Fonte: (Gouvêa, 2009)	19
3.1. Arquitetura da Web Semântica. Fonte: (Berners-Lee et al., 2001)	22
3.2. Exemplo de triplas em um documento RDF. Fonte: (Seriique, 2012)	27
3.3. Mecanismos de um <i>Reasoner</i> . Fonte:(Amanqui, 2014)	29
3.4. Relacionamentos <i>Simple Feature</i> , GeoSPARQL. Fonte: (OGC, 2012).	32
3.5. Classes das Representações espaciais em WKT. Fonte: Koubarakis et al. (2012)	34
3.6. Exemplos de representação de geometrias em WKT. Fonte: (Koubarakis et al., 2012)	37
4.1. Quantidade de triplas na <i>Web of Data</i> em 2011. Fonte: (Moura und Davir Jr., 2013)	45
4.2. Número de contribuintes e precisão posicional. Fonte: (Haklay et al., 2010)	48
4.3. Número de contribuidores por erro na precisão posicional. Fonte:(Haklay et al., 2010)	49
5.1. Arquitetura do <i>Triple Store</i> Parliament. Fonte: (Kolas et al., 2009)	53
5.2. Estrutura de armazenamento da <i>triplestore</i> Parliament. Fonte: Kolas et al. (2009)	54
5.3. Arquitetura da <i>triplestore</i> Strabon. Fonte: (Bereta et al., 2013)	54
5.4. Visão geral da abordagem de compilação realizada pelo GWT. Fonte: (Cooper und Collins, 2008)	58
5.5. Modelo Relacional da base de dados do LinkedGeoData. Fonte: (Auer et al., 2009a)	60

Lista de Figuras

5.6. Pontos representando coleções no repositório GBIF. Fonte: (Yesson et al., 2007)	61
6.1. Representação das coordenadas extraídas dos dados do SpeciesLink.	65
6.2. Representação das Coordenadas presentes nos dados do GBIF.	66
6.3. Coordenadas referentes a Reserva Florestal Adolpho Ducke contidas nos dados do SpeciesLink Figura(a) e GBIF Figura(b). Pontos vermelhos representam coordenadas geográficas erradas para a reserva. O ponto verde representa a coordenada geográfica correta para a reserva.	66
6.4. Exemplo de vários grupos criados de acordo com o limiar escolhido.	68
6.5. Demonstração do método de sumarização de coordenadas geográficas, onde os valores que ocorrem com mais frequência são identificados e atribuídos para as demais coordenadas..	69
6.6. Resultado das coordenadas geográficas após o método de sumarização. .	70
6.7. Quantidade de coordenadas geográficas recuperadas pelo <i>Gazetteer</i>	71
6.8. Precisão dos locais associados e grupos formados.	72
6.9. Mapeamento das informações geográficas para a especificação <i>as WKT</i> (Representação simplificada).	74
6.10. Representação da união entre um indivíduo referente a uma localidade com outro indivíduo que representa uma geometria usando o <i>object property hasGeometry</i> , disponibilizado pela ontologia GeoSPARQL.	75
6.11. Edição da ontologia utilizada pelo <i>Gazetteer</i> com auxílio da ferramenta Protégé.	76
6.12. Resultados de precisão e revocação após as buscas semânticas realizadas.	77
6.13. Tela referente a funcionalidade de inserção de novos locais no <i>Gazetteer</i> . .	85
6.14. Tela referente a função de buscas para o protótipo do <i>Gazetteer</i>	86
6.15. Tela referente a função de validação das coordenadas do <i>Gazetteer</i>	87
6.16. Tela referente a função de validação dos dados de Linked Open Data. . .	88
6.17. Atual arquitetura do <i>Gazetteer</i>	89
6.18. Arquitetura proposta para o <i>Gazetteer</i> , viabilizando o uso de VGI e padrões de Linked Open Data.	90

Lista de Tabelas

4.1. Comparativo entre os trabalhos relacionados.	43
6.1. Demonstração da qualidade dos dados do SpeciesLink	64
6.2. Demonstração da qualidade dos dados do GBIF	64
6.3. Precisão das coordenadas de acordo com o limiar de agrupamento.	69
6.4. Coordenadas que foram agrupadas de forma imprecisa.	71
6.5. Quantidade de locais presentes no <i>Gazetteer</i>	72
6.6. Vários centroides representando o mesmo local.	73
6.7. Resultados da execução das consultas geoespaciais nas <i>triplestore</i> Strabom, uSeekM e Parliament.	81
7.1. Cronograma.	96

Lista de Algoritmos

3.1.	Representação de um WKT em GeoSPARQL.	36
6.1.	Consulta por fazendas dentro de áreas protegidas.	78
6.2.	Consulta para verificar as cachoeiras que estão a uma distância de 1000 metros de alguma represa.	79
A.1.	Consulta Não Topologica utilizando a função convexHull	109
A.2.	Consulta Não Topologica utilizando a função buffer	109
A.3.	Consulta para verificar junções espaciais	110
A.4.	Consulta para verificar junções espaciais	110
A.5.	Consulta para verificar seleções espaciais	110
A.6.	Consulta para verificar seleções espaciais	111

Lista de Abreviaturas e Siglas

ACR Ambiguidade da Classe do Referente

API Application Programming Interface

ARC Ambiguidade da Referência

ART Ambiguidade do Referente

CSV Comma-Separated Values

GBIF Global Biodiversity Information Facility

GBIF Global Biodiversity Information Facility

GML Geography Markup Language

GPS Global Positioning System

GUI Interface Gráfica do Usuário

GWT Google Web Toolkit

HTML Hypertext Transfer Protocol

INPA Instituto Nacional de Pesquisas da Amazônia

ITN Rede Integrada de Transporte

KB Knowledge Base

LGD LinkedGeoData

OGC Open Geospatial Consortium

OL OpenLayer

OMS Open Street Maps

Lista de Algoritmos

OSM Open Street Map

OWL Web Ontology Language

RDF Resource Description Framework

RDFS Resource Description Framework Schema

REM Reconhecimento de Entidades Mencionadas

RIA Rich Internet Applications

RIG Recuperação de Informação Geográfica

RT Resolução de Topônimos

SIG Sistemas de Informação Geográfica

SPARQL Simple Protocol and RDF Query Language

TGN Getty Thesaurus of Geographic Names

URI Uniform Resources Identifiers

VGI Volunteered Geographic Information

VGI Volunteered Geographic Information

W3C World Wide Web Consortium

WFS Web Feature Service

WGS84 World Geodetic System 1984

WKT Well-Known Text

WKT Well-Known Text

WSD Word Sense Disambiguation

XML Extensible Markup Language

1. 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 vem 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 tem surgido através 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., 2013a).

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 et al., 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

1. Introdução

coleções microbiológicas vêm sendo criados e disponibilizados gratuitamente na internet. 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 lacunas, como, por exemplo, a imprecisão do seu georreferenciamento, problema esse conhecido como déficit Wallaceano (Lemes et al., 2011).

Uma abordagem para lidar com esse problema, é a construção de um *Gazetteer* colaborativo (Dicionário que contém nome de lugares e pode ser atualizado de maneira colaborativa) que visa possibilitar aos biólogos informarem as localidades de ocorrência de espécies.

A proposta deste *Gazetteer* é que, com sua utilização, será possível tornar a busca por informação referente a locais de coleta de espécimes mais precisa. Isso será conseguido através 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 é um sistema distribuído de Informação que integra, em tempo real,

1.1. Justificativa

dados primários de 310 coleções científicas nacionais e estrangeiras. Diariamente, o SpeciesLink trás informações estatísticas sobre seus dados, e, em outubro de 2013, o mesmo registrava 6.151.038 registros *online*. Contudo, apenas 2.572.566 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 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 (SpeciesLink, 2014).

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 a realizar esse 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.

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 et al., 2009)*.

Além disso, a Web Semântica possibilita a expansão de consultas geográficas, ou seja,

1. Introdução

após uma consulta ser realizada num *Gazetteer* e nenhum resultado relevante ser encontrado, tal consulta pode ser enviada para outros *Gazetteers*, por meio de solicitações a *endpoints* para consulta de informações. 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 et al., 2009).

Com o intuito de auxiliar outros trabalhos, que já possuem projetos que abordam problemas envolvidos na área de biodiversidade (Amanqui et al., 2013a), (Lemes et al., 2011), (Farias et al., 2010), (Metzger und Casatti, 2006) e dar continuidade a algumas pesquisas iniciadas pelos mesmos, como, por exemplo, (Amanqui et al., 2013a) e (Farias et al., 2010), este projeto de mestrado, através de uma parceria iniciada entre o grupo do Professor Dilvan Moreira do Laboratório Intermídia do ICMC-USP com o grupo do Professor Laurindo Campos do Laboratório de Interoperabilidade Semântica (LIS) – INPA, propõe a construção de um *Gazetteer* Colaborativo para auxiliar a Recuperação de Informações Geográficas sobre Biodiversidade.

1.2. Objetivo

O objetivo deste projeto de mestrado é construir um *Gazetteer* colaborativo, baseado em tecnologias da Web Semântica, 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 deste *Gazetteer* é promover acesso online a informações geográficas (locais) relevantes a catálogos de dados biológicos (coleções). Ele ainda vai possibilitar que usuários colaborarem com informações para o enriquecimento do *Gazetteer*, utilizem a semântica envolvida nas informações sobre localidades para obter melhores resultados em buscas por informações geográficas e, por fim, recuperem informações sobre coordenadas geográficas ausentes em registros sobre localidades.

Para atingir o objetivo proposto os seguintes objetivos específicos serão realizados:

1. Analise dos dados geográficos (locais referentes a reservas, rios, comunidades indígenas, entre outros) de bancos de dados sobre coleções biológicas (serão usados o SpeciesLink e o GBIF, como exemplos representativos);
2. Desenvolvimento de uma arquitetura e uma ontologia (caso uma existente não possa ser reusada) para que a construção do *Gazetteer* possa ser realizada;

1.3. Organização

3. Disponibilização de um protótipo do *Gazetteer* onde será possível aos usuários buscarem, inserirem e recuperar informações;
4. Analise do protótipo criado, por meio da averiguação das localidades recuperadas, desambiguação de localidades e suporte a prática de *Volunteered Geographic Information*;
5. Disponibilização de um *endpoint GeoSPARQL* para utilização de terceiros e fácil proliferação dos princípios de *Linked Open Data*.

Tendo esses objetivos traçados, acreditamos 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 biológicas. 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 referente 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: Descreve as tecnologias, ferramentas e ontologias que serão utilizadas para o desenvolvimento deste trabalho bem como a aplicação de cada uma no mesmo.

Capítulo 6: Apresenta os experimentos realizados até o momento como, por exemplo, análise dos dados utilizados, mapeamento de registros geográficos para ontologias e protótipos criados.

1. Introdução

Capítulo 7: Este capítulo mostra o cronograma a ser seguido, produções científicas até o momento, as limitações encontradas e o plano de trabalho futuro a ser seguido para construção do *Gazetteer Colaborativo*.

2. 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 a sua versatilidade. Embora o termo, descrito por Goodchild und 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) quanto 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 vi) quanto válido são as informações geográficas voluntárias para pesquisas em

2. Recuperação de Informação Geográfica

biodiversidade (Klinkenberg, 2013).

As respostas para essas questões possuem várias implicações e dependem da precisão das informações fornecidas e quanto útil elas serão no futuro. A precisão das informações é de extrema importância. A capacidade de mapeamento têm se alterado rapidamente com o advento das ferramentas on-line 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 representam os esforços voluntários de centenas de pessoas ao longo dos anos. Embora muitas coleções tenham começado há mais de 100 anos atrás, 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 estão 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.

Nesse contexto, alguns projetos atuais que envolvem VGI enfatizaram a necessidade de precisão, como por exemplo, o projeto eBird, um dos maiores projetos de VGI na área de biodiversidade em curso. Segundo Cornell (2008) um banco de dados somente é tão bom quanto o seu registro mais impreciso. Se alguns registros podem ser considerados questionáveis, todo o conjunto de dados pode ser rotulado como tal.

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).

2.1. Utilização de VGI no contexto da biodiversidade

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, liquens 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.

2. Recuperação de Informação Geográfica

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 VIG viável (Klinkenberg, 2013).

Para auxiliar a prática de VIG 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 et al., 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 (Buyukkokten et al., 1999) tornando difícil a recuperação de informações relevantes para o usuário (Jones et al., 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 et al., 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 (Buyukkokten et al., 1999).

No entanto, a forma semi-estruturada da Web dificulta o acesso a esse tipo de infor-

2.2. Recuperação de Informação Geográfica

mação. Jones und 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 ((Gey et al., 2006) e (Jones et al., 2002)).

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 ii) consultas probabilísticas, onde é possível, realizar buscas por pontos de referência 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. Recuperação de Informação Geográfica

2.3. *Gazetteers*

Um *Gazetteer* é um diretório geográfico que associa nomes de lugares a coordenadas geográficas. Eles são geralmente implementados como diretórios que contém 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 et al., 2009).

Eles oferecem funções para mapear nomes de lugares para entidades (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 et al., 2009).

Gazetteers são comumente construídos a partir da extração de informações de textos. No entanto, atualmente um nova forma de contribuir com informações geográficas vem sendo amplamente difundida, prática essa que se da 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. et al., 2013b).

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. et al., 2013b).

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 usuá-

2.5. GeoTAGs

rios 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 comumente em um *Gazetteer* seja feito por meio de marcações ou GeoTAGs, realizados 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 imediatamente o rótulo (L), vindo de um usuário (U), o vocabulário associado a informação do Item

2. Recuperação de Informação Geográfica

(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 a dimensão do objeto representado quanto a 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, 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 predições e diagnósticos usando um subconjunto conhecido de relações espaciais. Podendo ele 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 et al. (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 precisa 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

2. Recuperação de Informação Geográfica

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 und 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 und 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 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 ambigüidades 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).

		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 2.1.: Exemplos de tipos de ambiguidades. Fonte: 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 2.1 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 und Mann (2003) quantificaram o tipo e o grau de ambiguidade dos topônimos examinando a ontologia geográfica TGN (*Getty Thesaurus of Geographic Names*), conforme ilustrado na Figura 2.2, 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 ambigüidade, 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. Recuperação de Informação Geográfica

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 2.2.: Ambiguidade presente nas entidades geográficas do TGN. Fonte: Gouvêa (2009)

2.8.1. Estratégias para Resolução de Topônimos

A Resolução de Topônimos compreende dois processos a 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 referencias *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 und 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 2.3 (McDonald, 1996):

- 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

2.8. Resolução de Topônimos

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 2.3.: 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 2.4.: Exemplo de tipos de evidências para desambiguação de topônimos. Fonte: (Gouvêa, 2009)

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 2.4.

O problema de se utilizar a abordagem baseada em regras é a necessidade de definirlas 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

2. Recuperação de Informação Geográfica

associados a 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 trabalho é a construção de um *Gazetteer* Colaborativo para um determinado domínio, 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 *Gazetteer* Colaborativo proposto por esse trabalho. Inicialmente foi feita uma breve introdução sobre RIG na Web, evidenciando a forma como as pessoas representam e se referem à conteúdos geográficos. Logo após, o conceito de *Gazetteer* foi explicado e, em seguida, a nova prática de 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 e a utilização da Web Semântica para desenvolver o *Gazetter*.

Estes conceitos serão utilizados no trabalho proposto para desenvolver o *Gazetteer*. Sendo necessário abordar 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.

3. 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 und 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 et al. (2001) propõe uma modelo em camadas para a arquitetura da Web Semântica também conhecida como “bolo de noiva”, a arquitetura lógica é apresentada na Figura 3.1.

A ideia central da arquitetura proposta por Berners-Lee et al. (2001) é de que cada camada seja melhorada e adicionada gradativamente para trazer novas funcionalidades, juntamente com aumento de expressividade para representar os dados e possibilidades de se realizar inferências e autenticações. Essa arquitetura promove o reúso utilizando algo já validado, desse modo minimiza a necessidade de se criar novas camadas a partir do

3. Web Semântica

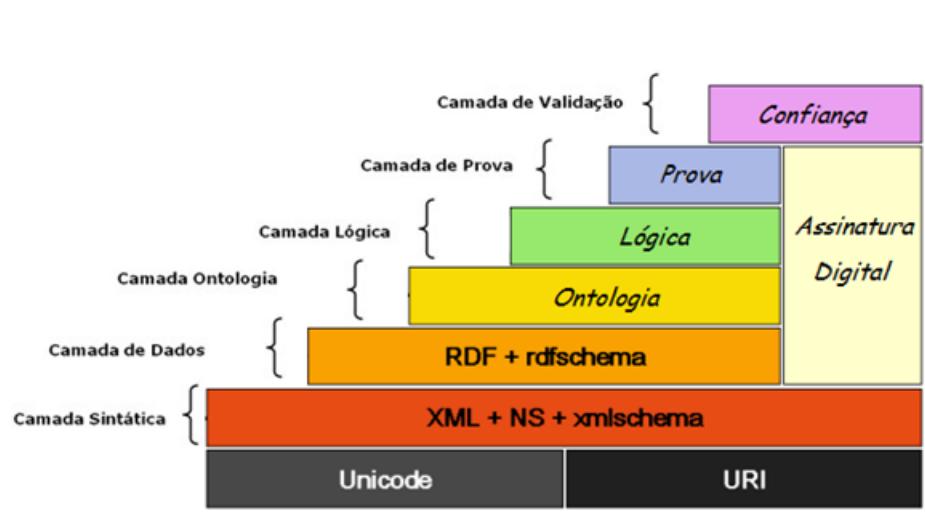


Figura 3.1.: Arquitetura da Web Semântica. Fonte: (Berners-Lee et al., 2001)

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 und McGuinness (2001) relata 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 geocampo e geo-objeto. Quando uma entidade relacionada está focada

3. Web Semântica

em dados espaciais como um conjunto de distribuição contínua ela é denominada como geocampo, ao contrário de um geo-objeto, que trabalha com dados espaciais discretos e identificáveis espalhados pelo mundo (Gimenez et al., 2013).

Ao contrário das ontologias gerais, geo-ontologias devem representar relacionamentos espaciais, fatores espaciais, fatos físicos, coleções de dados, modelos de computação geoespacial (Gimenez et al., 2013). Unificando tais características referentes à geoinformações com conhecimento de especialistas no domínio. Wang et al. (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 fórmula 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).

3.3. Linked Open Data

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 preservam a semântica dos dados para que os mesmos sejam comprehensíveis por computadores e humanos (Heath und 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 und 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 und 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, (Amanqui et al., 2013a); (Parundekar et al., 2010); (Auer et al., 2009b) 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 und 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 consulta-los.

3. Web Semântica

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 und 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 dos Santos et al. (2011) .

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 comprehensível 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 (da Silva und de Souza 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 (da Silva und de Souza 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) (da Silva und de Souza Lima, 2004).

A Figura 3.2 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.

3.5. Web Ontology Language - OWL

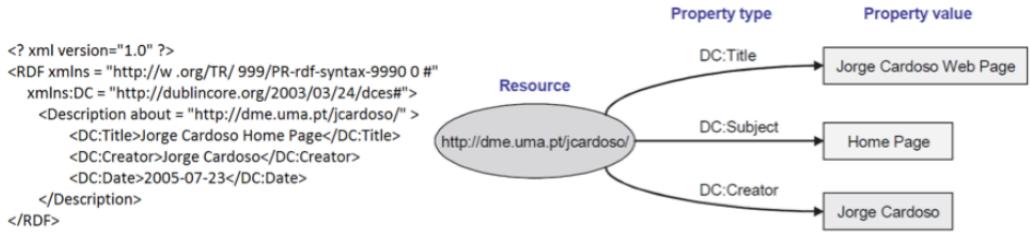


Figura 3.2.: Exemplo de triplas em um documento RDF. Fonte: (Seriique, 2012)

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 (da Silva und de Souza 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 importante 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 (da Silva und de Souza 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 (Bechhofer 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

3. Web Semântica

heranças incorretas e checar as restrições sobre os valores das propriedades e sua cardinalidade (Bechhofer 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 (Bechhofer 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 (Bechhofer 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 (Bechhofer 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 geralmente são mais escaláveis e fáceis de implementar (Bechhofer 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,

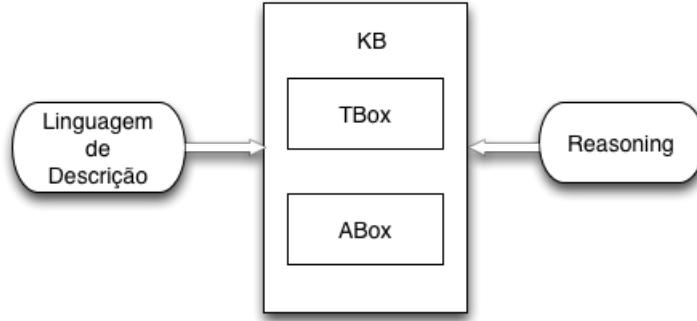


Figura 3.3.: Mecanismos de um *Reasoner*. Fonte:(Amanqui, 2014)

RDF, RDFS (Amanqui, 2014).

Geralmente *reasoners* contém dois mecanismos internos para processar bases de conhecimento (KB, *knowledge base*), são eles Tbox e Abox Dentler et al. (2011). A Figura 3.3 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:

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

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 José é um indivíduo da classe Homem.:

$$\text{Homem} \equiv \text{José}$$

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. Web Semântica

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 et al., 2009).

3.8. SPARQL e Geo-SPARQL

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 *triplesstore*) por meio de um endpoint SPARQL. Um endpoint SPARQL é uma interface que permite a humanos e aplicações acessarem os dados armazenados em uma *triplesstore*. 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.

3.8. SPARQL e Geo-SPARQL

Devido a necessidade de se abordar o relacionamento semântico das entidades geográficas trabalhos como (Battle und Kolas, 2012a; Bereta et al., 2013) têm feito o uso de Geo-SPARQL: 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, se viu necessário o estabelecimento de um padrão (Battle und Kolas, 2012a).

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.

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 3.4. 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 empresas, 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 und Kolas, 2012a) e (Bereta et al., 2013).

3. Web Semântica

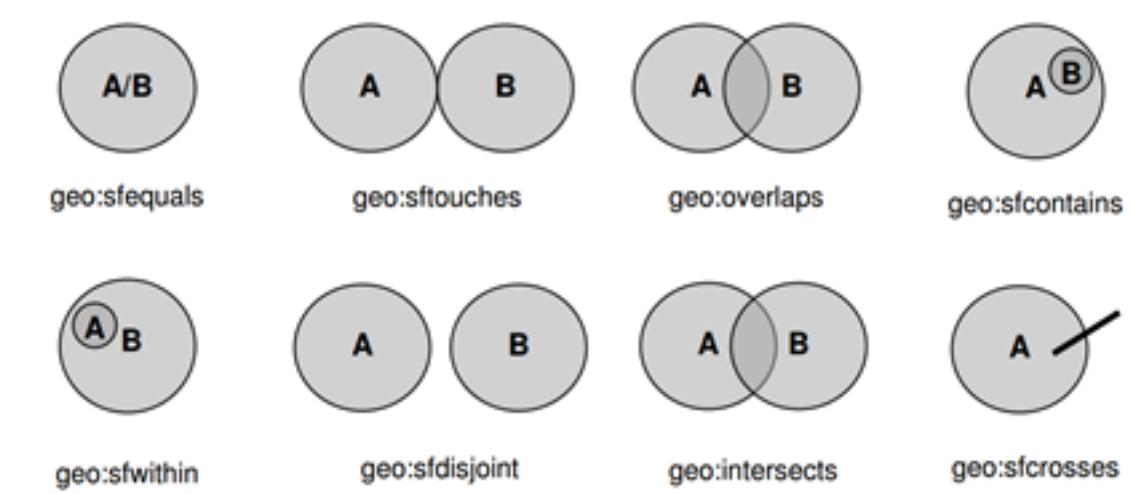


Figura 3.4.: Relacionamentos *Simple Feature*, GeoSPARQL. Fonte: (OGC, 2012).

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

3.8.1. Well-Known Text (WKT)

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 rescritas 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.

Para representar as coordenadas geográficas em geometrias do tipo *Simple Feature* em WKT, o padrão especificado pela OGC descreve a hierarquia de classes da Figura 3.5:

- **Ponto** : (*Point*) representa uma simples localização no espaço, tendo os valores de x e y para representar o espaço geográfico e podendo ter os valores de z e m para representar informações adicionais, dependendo do sistema de referencia para associação de coordenadas.
- **Curva**: (*Curve*) É uma geometria unidimensional. Os subtipos da classe curva definem o tipo de interpolação que é usado entre os pontos.
- **Linha de Pontos**: (*LineString*) É um subtipo da classe curva que usa uma interpolação de linhas entre pontos. Uma linha de pontos é fechada se ela começa com um ponto e termina no mesmo ponto. Uma linha de pontos é simples se ela não possui auto-intersecção.
- **Linha**: (*Line*) É uma linha de pontos (*LineString*) com exatamente dois pontos.
- **Anel Linear**. (*LinearRing*) É uma sequência de linha que é fechada e simples.

3. Web Semântica

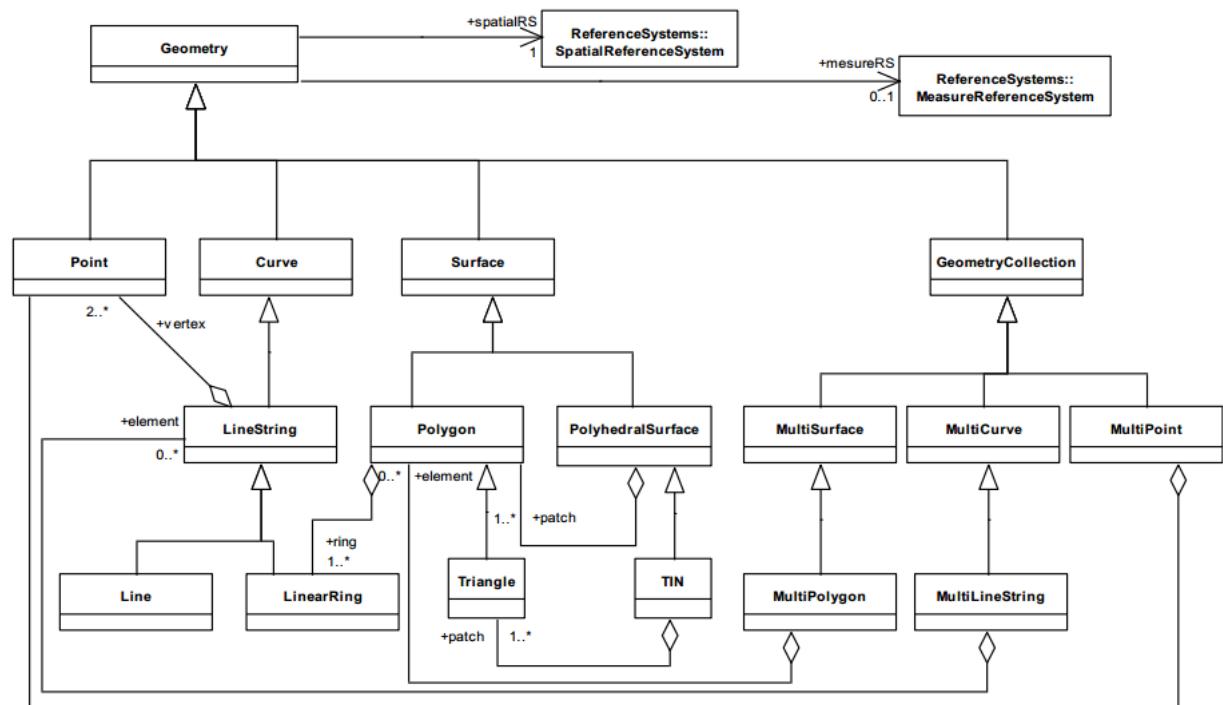


Figura 3.5.: Classes das Representações espaciais em WKT. Fonte: Koubarakis et al. (2012)

- **Superfície:** (*Surface*) É uma geometria bidimensional. Essa classe é abstrata, ou seja, não pode ser instanciada. Uma superfície simples pode consistir de um simples “curso” que tem uma borda exterior e zero ou mais bordas interiores como, por exemplo um polígono com buracos.
- **Polígono.** (*Polygon*) É uma superfície simples e planar que têm exatamente uma borda exterior podendo ter muitas bordas interiores sem interseção. Cada polígono é topologicamente fechado e nenhuma das bordas interiores e exteriores se cruzam. No entanto, duas bordas podem ter um ponto de intersecção em comum, mas somente como uma tangente. O interior de um polígono é um conjunto de pontos conectados, no entanto um polígono com falhas não possui seu exterior conectado.
- **Triângulo:** (*Triangle*) É um polígono com três vértices não colineares sem bordas interiores
- **Superfície Poliédrica:** (*Polyhedral Surface*) É uma coleção continua de polígonos que compartilham segmentos de bordas em comum. Cada par de polígonos que se tocam tem uma borda em comum que é expressa como uma coleção finita de linhas de pontos. Cada linha de pontos é uma parte da borda de dois ou mais polígonos.
- **Rede Irregular Triangulada:** (*Triangulated Irregular Network*) É uma superfície poliédrica que consiste somente de triângulos.
- **Coleção de geometrias:** (*Geometry Collection*) É um conjunto distinto de geometrias.
- **Múltiplos Pontos:** (*MultiPoint*) É uma coleção de geometria da qual elementos são pontos que não estão conectados.
- **Múltiplas Curvas:** (*MultiCurve*) É uma coleção de geometria das quais elementos são curvas
- **Linhas de Pontos Múltiplas:** (*MultiLineString*) É uma coleção de geometrias das quais elementos são linhas de pontos.
- **Superfícies Múltiplas:** (*MultiSurface*) É uma coleção de geometria bidimensional onde elementos são superfícies. O interior de quaisquer duas geometrias não podem se cruzar. As bordas de quaisquer duas geometrias não se cruzam, mas podem se tocar em um número finito de pontos.

3. Web Semântica

Algoritmo 3.1 Representação de um WKT em GeoSPARQL.

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

- **Polígonos Múltiplos:** (*MultiPolygon*) É uma superfície múltipla onde os elementos são polígonos e as bordas de cada polígono não podem se interceptar.

A interpretação das coordenadas de uma geometria depende do sistema de coordenadas de referência que está sendo utilizado em conjunto com a geometria. De acordo com o padrão WKT, o sistema de referência que está associado a uma geometria não é incorporado na representação do objeto, mas é dado separadamente usando a notação adequada de coordenadas, como mostrado na Figura 3.5 (Koubarakis et al., 2012).

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 Geo-SPARQL é exibido no algoritmo 3.1. 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 na listagem 3.1 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))*”.

Alguns exemplos da representação de um WKT são mostrados na Figura 3.6. 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 do *triple store* uSeekM somente trabalhar com esse tipo de representação, conforme descrito por Garbis et al. (2013).

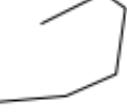
Geometry type	WKT representation	Geometry
Point	Point(5 5)	.
LineString	LineString(5 5,28 7,44 14,47 35,40 40,20 30)	
Polygon	Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5))	
Polygon	Polygon((5 5,28 7,44 14,47 35,40 40,20 30,5 5), (28 29,14.5 11,26.5 12,37.5 20,28 29))	
MultiPoint	MultiPoint((5 5),(28 7),(44 14), (47 35),(40 40),(20 30))	.
Geometry Collection	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)))	

Figura 3.6.: Exemplos de representação de geometrias em WKT. Fonte: (Koubarakis et al., 2012)

3. Web Semântica

3.9. Considerações Finais

Este capítulo apresentou os conceitos necessários para compreender o uso de web semântica para o desenvolvimento do *Gazetteer* Colaborativo. 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, "Quais os rios próximos a Manaus". Além disso, foi descrita a representação de figuras (pontos, linhas, polígonos, etc.) para mapear as representações geométricas 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.

4. Trabalhos Relacionados

4.1. Introdução

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 o 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 analise 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 os 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 *Gazetteer Colaborativo*. A seguir, esses trabalhos são descritos e por fim uma comparação com o trabalho proposto é realizada.

- Jr. et al. (2013a) 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 podem ser facilmente estendidos e reutilizados 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 und 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

4. Trabalhos Relacionados

nomes de rios e 670 bacias.

- O OntoGazetteer, proposto por Machado et al. (2011), utiliza conceitos de ontologias 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 bases 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. (2010a), 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. (2010a) 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 num SGDB.
- O trabalho proposto por Gil et al. (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 SGDB 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 et al. (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 SGDB Apache Derby para armazenar suas informações. Para realizar a extração dos nomes de lugares, Manguinhas et al. (2009) utiliza 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

4.1. Introdução

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 tem 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. et al. (2013a) 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 4.2.
- Já o trabalho proposto por Moura und 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 und Jr. (2012) e Machado et al. (2011) não deixam explicita a ontologia utilizada para desambiguar os nomes de ambos os *Gazetteers*, somente é demonstrado seu modelo.
- No FODGS, proposto por Peng et al. (2010a), 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 contém 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.

4. Trabalhos Relacionados

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 et al. (2010) aborda o contexto biológico, no entanto coordenadas geográficas são manipuladas em um ambiente controlado onde as informações são precisas. O que não ocorre com o trabalho proposto nesta qualificação, 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 et al. (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 et al. (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 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 4.1 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 Section 4.2 descreve o estado da arte e os principais desafios na utilização desses conceitos em *Gazetteers*.

4.1. Introdução

Trabalho	Utiliza Ontologias	Utiliza VGI	Buscas Semânticas	Linked Open Data
Framework proposto por Jr. et al. (2013a)	Não	Apenas descreve um modelo de implementação para VGI	Não	Não
<i>Gazetteer</i> proposto por Moura und Jr. (2012)	Somente para desambiguação	Não	Não	Não
OntoGazetteer Machado et al. (2011)	Somente para desambiguação	Não	Não	Não
FODGS Peng et al. (2010a)	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 et al. (2010)	Não	Não	Não	Não
DIGMAP Manguinhas et al. (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>Gazetteer</i> proposto neste trabalho	Sim	Sim	Sim	Sim

Tabela 4.1.: Comparativo entre os trabalhos relacionados.

4. Trabalhos Relacionados

4.2. Estado da Arte

Sistemas de Informação Geográfica (SIG) são comumente utilizados para manipular dados geográficos, no entanto, nem sempre é possível ou factível extrair seus dados para uso em outras aplicações, ou mesmo em outros SIG. Além disso, a nova forma de coleta de dados, também denominada, Informação Geográfica Voluntária, abre diversas questões a serem discutidas, pois são necessárias fontes de dados confiáveis.

Alguns desses problemas são abordados na literatura e a soluções propostas envolvem recursos como *Gazetteers*, contribuições geográficas voluntárias e *Linked Open Data*. 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 a maioria dos *Gazetteers* atuais são mantidos por pequenos grupos bem definidos, a próxima geração de *Gazetteers* deve também incorporar a utilização de contribuições voluntárias (Goodchild und 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 apresentado na Figura 4.1, 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 und Davir 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 à possibilidade de realizar consultas complexas (Moura und Davir Jr., 2013). Den-

4.2. Estado da Arte

Domínio	Número de datasets	Tripas	%	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 4.1.: Quantidade de tripas na *Web of Data* em 2011. Fonte: (Moura und Davir Jr., 2013)

tro desse contexto, Kessler et al. (2009) relata 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 tripas RDF (Moura und Davir 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 a população, mas não existe uma data indicando a validade da informação,

4. Trabalhos Relacionados

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 und Davir 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 und Davir 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 utilizado a Lei de Linus.

4.3. Precisão de dados utilizando VIG

Tendo em vista a coleta e a natureza distribuída dos dados coletados por voluntários, é extremamente importante validar o quanto bom 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 (2008) descreve 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

4.3. Precisão de dados utilizando VIG

fornecidas por muitos participantes em projetos de VGI como o OSM, são similares ao conjunto de dados mantidos por agencias 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 VIG 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 podem 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 km² 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 4.2, onde mais de 93% dos pontos têm mais de 70% de precisão.

Embora esse valor seja elevado, a qualidade não é mias dependente do número de contribuintes após um determinado número. As coordenadas precisas são editadas por

4. Trabalhos Relacionados

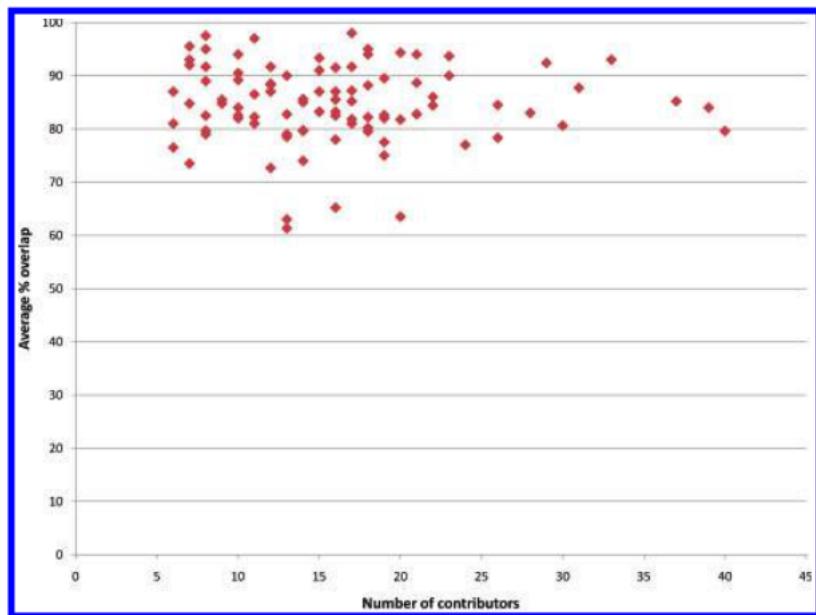


Figura 4.2.: Número de contribuintes e precisão posicional. Fonte: (Haklay et al., 2010)

um número de 5 a 20 colaboradores, sendo que, a partir de 13 contribuintes, a precisão das coordenadas se estabiliza, como apresentado na Figura 4.3.

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 *Gazetteer* proposto neste trabalho utilizará 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 6.3.

4.4. Considerações Finais

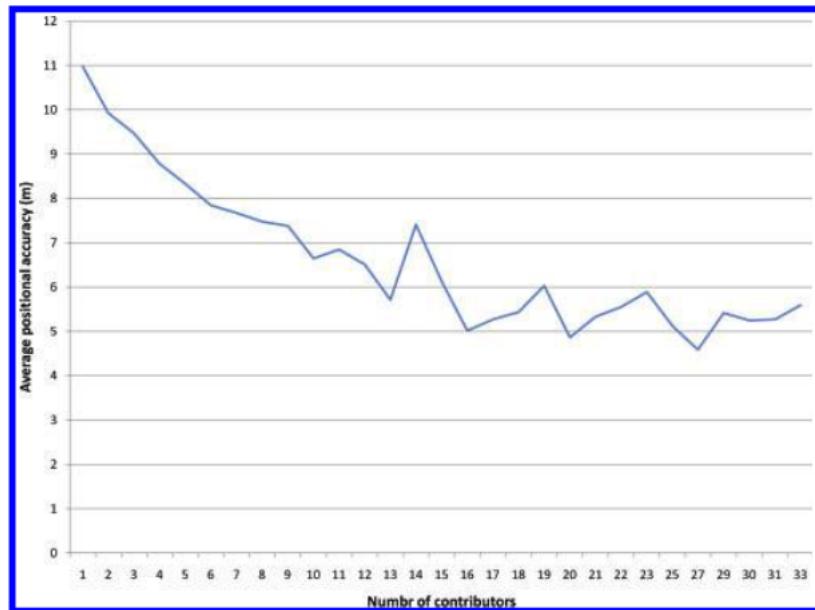


Figura 4.3.: Número de contribuidores por erro na precisão posicional. Fonte:(Haklay et al., 2010)

4.4. 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 trabalho proposto 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 *Gazetteer* proposto 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 VIG, a qualidade, completude e atualização das informações sobre as localidades das coleções biológicas sejam aprimoradas.

5. Ferramentas

Neste capítulo, será descrito o conjunto de ferramentas para o desenvolvimento do *Gazetteer* Colaborativo que possibilitará ao usuário realizar buscas semânticas, inserir e atualizar os dados no formato de *Linked Open Data*. Para representar as localidades utilizaremos a ontologia LinkedGeoData que será editada pela ferramenta Protégé. Os locais utilizados no *Gazetteer* serão extraídos dos repositórios, SpeciesLink, GBIF e Geonames e então serão armazenados em uma *triplesstore* e disponibilizados para o usuário por meio de uma interface em GWT que utilizará mapas digitais feitos com a API OpenLayers.

5.1. Protégé

O Protégé é um ambiente para criação e edição de ontologias desenvolvido pela Universidade de Stanford. Ele permite a construção de ontologias, formulários de entrada de dados customizados e mecanismos de inserção de dados, com suporte para todas as últimas especificações da W3C.(Rubin et al., 2007).

Atualmente, o Protégé possui uma versão Web que permite usuários, compartilhar, procurar e editar ontologias utilizando um navegador Web. O WebProtégé provê um paradigma gráfico similar ao Protégé GUI. A vantagem do WebProtégé é que diversos usuários podem editar a mesma ontologia sem precisar realizar o download para sua máquina (Rubin et al., 2007).

Por ser um sistema open-source, ele permite que desenvolvedores criem novos plugins capazes de aumentar significativamente o conjunto de funcionalidades da ferramenta. As principais características do Protégé, listadas por Rubin et al. (2007), são:

- Utilização um mecanismo de inferência, *reasoner*, para a verificação, de ontologias e sua classificação automática. Esse mecanismo é implementado pelos programas FaCT++ e HermiT, podendo se acoplar outros *reasoners* a ferramenta.
- A versão atual do Protégé, 4.X, é baseada na biblioteca OWL-API (OWL 2), que segue os padrões do propostos pelo W3C, possibilitando o desenvolvimento de ontologias em conformidade com as principais regras descritas pelo W3C.

5. Ferramentas

Neste trabalho, o Protégé será utilizado para criar e editar as classes, propriedades e atributos da ontologia que será utilizada para representar as informações referentes aos locais de coleta das coleções biológicas que compõem o *Gazetteer*. Esses locais serão armazenados em uma *triplesstore* para serem disponibilizados por meio de serviços SPARQL.

5.2. Triplesstore

Uma *triplesstore* é um banco de dados especialmente construído para o armazenamento e recuperação de triplas. Essas triplas podem ser consultadas e representadas em RDF. Existem várias implementações de triplesstores, tanto comerciais como Oracle Database 12c, Franz AllegroGraph RDFStore e StarDog, como *open source*, como o Virtuoso Open Source, BigData e Apache Marmotta. Contudo, só foi possível encontrar três implementações *open source* que permitem trabalhar com representações geoespaciais fornecidas pela linguagem de busca GeoSPARQL Garbis et al. (2013): Parliament, Strabon e UsekM. A seguir, a arquitetura dessas três *triplesstores* são descritas e comparadas.

5.2.1. Parliament

O Parliament é uma *triplesstore* desenvolvida pela BBN e vem sendo usada desde 2001 por um grande número de aplicações que vão desde projetos de pesquisa a projetos finais de produção. Ela permite o gerenciamento de dados compatível com os padrões RDF, RDFS, OWL, SPARQL, e GeoSPARQL.

Em sua implementação, o Parliament incorpora uma série de pacotes de código aberto, como, por exemplo, Jena e ARQ (processador de consultas), Joseki (uma implementação baseada em servlet do protocolo SPARQL), Jetty (um servlet container) e Berkeley DB (usado para armazenar dados). A Figura 5.1 mostra a arquitetura do Parliament segundo Kolas et al. (2009).

A estrutura de armazenamento do Parliament possui três módulos principais: tabela de recursos (*Resource Table*), tabela de declaração (*Statement Table*) e o dicionário de recursos (*Resource Dictionary*). A seguir esses módulos serão descritos de acordo com Kolas et al. (2009):

- *Resource Table*: É um arquivo único com um número fixo de registros, onde cada um representa um recurso ou literal. Os registros são sequencialmente numerados e esse número serve como o ID de correspondência ao recurso. Isso permite acesso direto a um registro fornecendo um ID por meio de uma indexação simples.

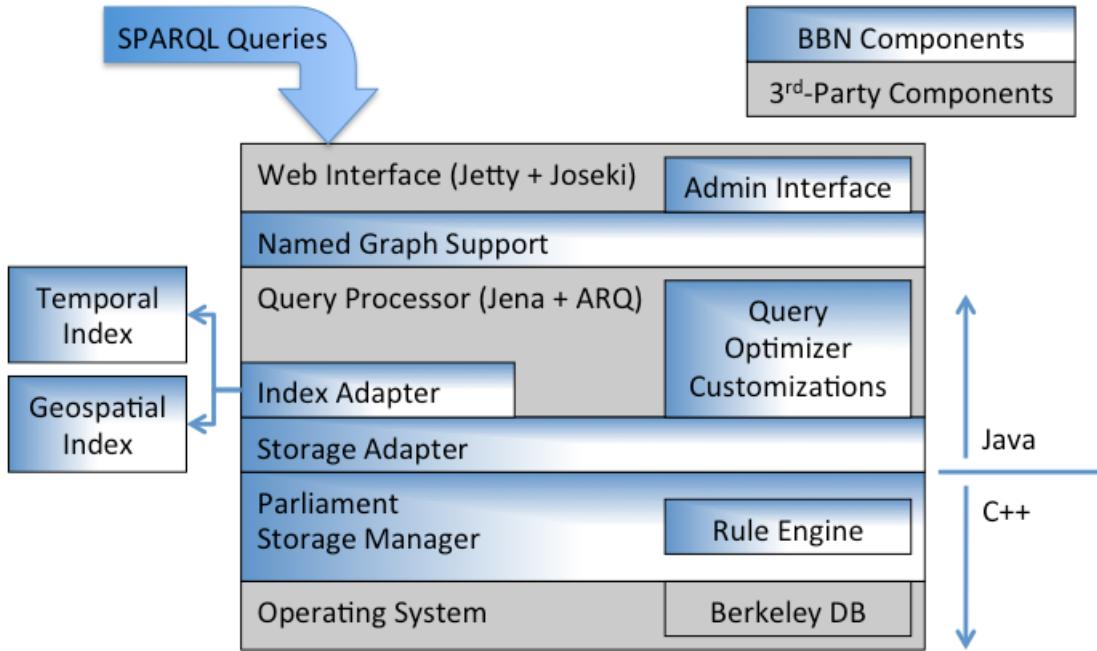


Figura 5.1.: Arquitetura do *Triple Store* Parliament. Fonte: (Kolas et al., 2009)

- *Statement Table*: Este módulo é similar ao *Resource Table*, que faz a utilização de um único arquivo com um número fixo de registros. No entanto, os registros no *Statement Table* são sequencialmente numerados e esse número serve como o ID para um determinado recurso, conforme apresentado na Figure Figure 5.2 on page 54 .
- *Resource Dictionary*: Assim como os outros módulos do Parliament, o *Resource Dictionary* utiliza um dicionário para mapear seus recursos, sendo esse o principal módulo da estrutura de armazenamento do Parliament. O Dicionário de Recursos permite mapear os dados do Parliament utilizando o Berkeley DB implementado a estrutura de dados Árvore B+.

Segundo Kolas et al. (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 $O(n^{2/3})$.

5. Ferramentas

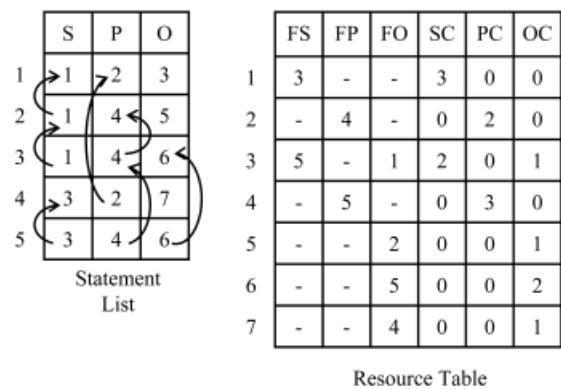


Figura 5.2.: Estrutura de armazenamento da *triplestore* Parliament. Fonte: Kolas et al. (2009)

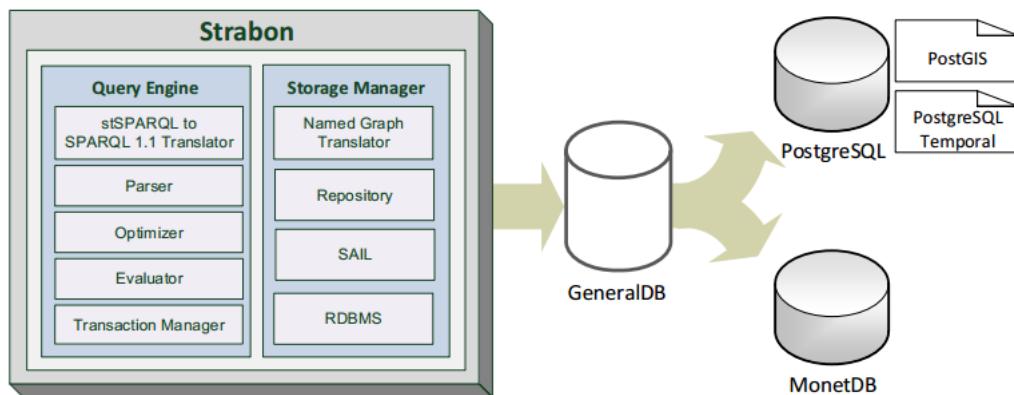


Figura 5.3.: Arquitetura da *triplestore* Strabon. Fonte: (Bereta et al., 2013)

5.2.2. Strabon

O Strabon é uma *triplestore* de código aberto desenvolvido pelo grupo TELEIOS que permite armazenar e consultar dados em stRDF/stSPARQL e GeoSPARQL. Seu desenvolvimento toma como base o *triplestore* Sesame.

A utilização do Sesame como base permite implementar uma vasta gama de funcionalidades e possibilita a utilização de SGBDs, como o PostGIS e MonetDB, como backends para explorar uma variedade de funções espaciais. A implementação do Strabon se dá através de camadas incluídas no pacote de *softwares* do Sesame. Strabon, em sua versão 3.0, usa o Sesame 2.6.3 e comprehende dois módulos básicos, como mostrado na Figura 5.3, o Storage Manager e o Query Engine (Bereta et al., 2013):

- Storage Manager: Carrega as triplas em stRDF usando Sesame e um dicionário de codificação. Seus dados são armazenados em uma Árvore B+ e todos os literais referentes a representações geométricas são armazenados em bancos de dados como PostgreSQL ou MonetDB (Bereta et al., 2013).
- Query engine: Realiza o processamento das queries inseridas no Strabon. O Motor de Busca consiste de um analisador, um otimizador, um avaliador e um sistema de codificação. Primeiramente, o analisador gera uma árvore de sintaxe abstrata, então esta árvore é mapeada para a álgebra interna do Sesame, resultando numa árvore de busca. Essa árvore de busca é processada pelo otimizador que progressivamente a modifica, implementando várias técnicas de otimização feitas pelo Strabon. Logo após, a árvore de busca é repassada para o avaliador para produzir a consulta correspondente em SQL que irá ser avaliada pelo PostgreSQL. Depois que a consulta SQL foi feita, o avaliador recebe os resultados e executa todas as ações de pós-processamento. O passo final envolve formatar esses resultados e retorná-los para visualização (Bereta et al., 2013).

5.2.3. uSeekM

O uSeekM é uma biblioteca que pode ser adicionada a base de dados semânticas que usam as *triplestores* baseadas no Sesame. O uSeekM adiciona a capacidade de indexar e realizar *queries* geoespaciais e provê integração com outras ferramentas e *frameworks*. A maioria das funcionalidades do uSeekM é fornecida através de *Sail wrappers*, ou seja, adaptadores que usam a funcionalidade de *triplestores* Sesame e possuem funcionalidades adicionais uSeekM (2014).

O uSeekM possui os seguintes *Sail wrappers*:

- *IndexingSail*: Estende uma base de dados RDF com a capacidade para armazenamento e realização de consultas SPARQL. Adiciona buscas geoespaciais usando GeoSPARQL. Tecnicamente a *IndexingSail* é um *wrapper Sesame Sail* que pode ser usado em qualquer *triplestore* compatível com o padrão *Sail wrapper* do Sesame.
- *SimpleTypeInferencingSail*: Um componente que infere a transitividade de propriedades *rdfs:subClassof* e *rdf:type*. Ele é útil em casos onde o uso de RDFS completo é muito lento e somente a inferência de subtipo é necessária.
- *SmartSailWrapper*: Utilizado para obter diferentes tipos de conexões a partir de um componente subjacente, caso esse componente ofereça a funcionalidade de conexões.

5. Ferramentas

Isso é útil em casos como, por exemplo, a *triplestore Bigdata* que oferece classes de conexão no padrão Sesame em forma de componentes para se comunicar com a SmartSailWrapper.

As funcionalidades fornecidas pelo uSeekM incluem a implementação de GeoSPARQL permitindo a indexação, busca e computação de geometrias, integradas dentro da linguagem de consulta SPARQL, proporcionando suporte para todas *OpenGIS Simple Features* e relações. Além disso, é possível realizar buscas por texto utilizando a linguagem SPARQL e integrar a *triplestore* com o *framework* Spring para gerenciar a inicialização, escopo de conexão e transações para a *triplestore*.

Para armazenamento das triplas em RDF, o uSeekM utiliza o banco de dados PostgreSQL juntamente com a extensão Postgis. Isso permite ao uSeekM realizar as consultas geoespaciais (uSeekM, 2014).

5.3. API Jena

Uma vez definido a *triplestore* para armazenar e acessar as triplas em RDF é necessário prover um meio de comunicação entre os dados armazenados e a interface do usuário. Neste contexto, a API Jena é utilizada para requisitar dados e criar consultas SPARQL.

Jena é uma API Java desenvolvida com o objetivo de permitir desenvolvedores criarem e manipularem dados e recursos da Web Semântica, em conformidade com as recomendações do W3C. Com a API Jena, é possível ao desenvolvedor de aplicações criar grafos em RDF, realizar *parser* de arquivos RDF/XML, utilizar a API para realizar consultas SPARQL e requisitar dados de Triple Stores, acessar ontologias e processar vocabulários, dentre outros recursos (Carroll et al., 2003).

A API Jena suporta as linguagens RDF/RDFS e OWL, incluindo também um motor de inferência, que pode ser construído a parte e acoplado a API. Além disso, a API permite várias estratégias para armazenamento de dados em triplas RDF na memória ou no disco (Amanqui, 2014).

A Jena foi originalmente desenvolvida pela companhia Hewlett-Packard, em 2000, possuindo uma licença open source para o desenvolvimento de aplicações. Ela é extensivamente utilizada em um vasto número de aplicações da Web Semântica (Amanqui, 2014).

Neste trabalho, utilizaremos a Jena para construir as triplas RDF, armazená-las em disco, utilizando uma *triplestore*, e, por meio do seu motor de busca, realizar consultas para recuperar os dados. Essa API irá se comunicar com a interface do usuário por meio de um aplicação Web desenvolvida em GWT.

5.4. GWT

Dentre as ferramentas para desenvolvimento de aplicações na Web 2.0, temos as Rich Internet Applications (RIAs). Elas são aplicações Web com alto grau de funcionalidade que possuem grande similaridade com interfaces e funcionalidades de um programa desktop. Diversas ferramentas já foram criadas para serem utilizadas com esse propósito, como as applets Java, Adobe Flash, Microsoft Silverlight, entre outras (Shah, 2008). No entanto, essas plataformas para RIAs necessitam de um programa externo ao navegador, como um plugin, para funcionarem corretamente e por isso não estão disponíveis em todos os navegadores e sistemas operacionais (Seriique, 2012).

Dentre as várias ferramentas para desenvolvimento de RIAs, escolhemos o Google Web Toolkit (GWT) para desenvolver a interface gráfica do usuário (GUI) do *Gazetteer* Colaborativo. O GWT é uma ferramenta que permite a desenvolvedores Web criar aplicações com níveis de interatividade próximos a uma aplicação desktop. O desenvolvedor codifica as interfaces em linguagem Java que posteriormente é compilada para Javascript que é então executado no navegador Web. Isso gera um código final otimizado, em Javascript, compatível com os diversos navegadores Web (Seriique, 2012). Isso é possível porque o GWT gera um código Javascript para cada tipo de navegador, tirando essa tarefa laboriosa e nada trivial da responsabilidade dos desenvolvedores. A Figura 5.4, demonstra as fases desse processo de compilação feitas pelo GWT.

Uma das principais vantagens em se utilizar o GWT, em relação as demais tecnologias de desenvolvimento de RIAs, é que o GWT é baseado em Javascript e Ajax, sendo assim requer somente um navegador Web padrão para ser executado. Além disso, o GWT possibilita que os programas desenvolvidos possam ser executados em dispositivos móveis, como tables e smartphones, e não apenas em computadores padrões (Seriique, 2012).

5.5. OpenLayer

Outra forma de acessar os dados do *Gazetteer* é por meio de mapas digitais, onde o usuário pode navegar, marcar pontos e criar polígonos. Para desenvolver essa funcionalidade, a API OpenLayer será utilizada em conjunto com o GWT.

O OpenLayer (OL) é uma API desenvolvida em JavaScript usada para disponibilizar e exibir mapas em navegadores Web. A utilização de OL permite representar geometrias em pontos, polígonos, curvas, linha de pontos, múltiplos pontos, múltipla linha de pontos e múltiplos polígonos (OpenLayers, 2014).

5. Ferramentas

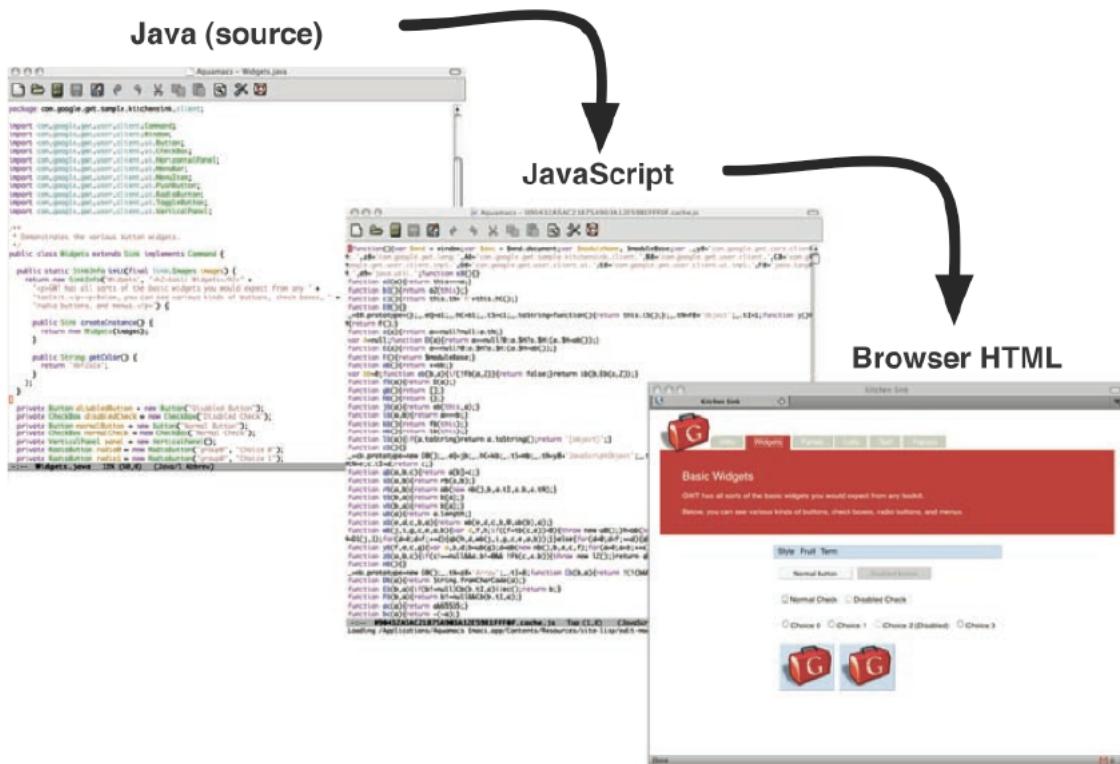


Figura 5.4.: Visão geral da abordagem de compilação realizada pelo GWT. Fonte: (Cooper und Collins, 2008)

A API OpenLayer possui suporte a navegação em mapas com a utilização de mouse e teclado, adição de marcadores e seleção de *layers*. Além disso, é possível obter dados de diversos recursos, tais como *Web Feature Service* (WFS) como, por exemplo, *Google Maps*, *OpenStreetMap*, *Bing Maps*, dentre outros (OpenLayers, 2014).

Por ser desenvolvido em JavaScript, o OL possui APIs em GWT, o GWT-OpenLayers. Facilitando a sua integração e programação para o desenvolvimento de interfaces para manipular mapas. Neste trabalho, a API OpenLayer será utilizada para criar e manipular os mapas que representarão as informações geográficas do *Gazetteer*.

5.6. LinkedGeoData

Uma vez definidas as ferramentas para desenvolver o *Gazetteer* é necessário, representar as informações das entidades geográficas que irão ser acessadas e armazenadas. Como o *Gazetteer* proposto utiliza ontologias para essa tarefa, escolhemos a ontologia LinkedGeoData para representar os dados do *Gazetteer*.

A ontologia LinkedGeoData (LGD) foi desenvolvida com objetivo de mapear os dados do Open Street Map (OSM) para o formato RDF (utilizado pelos dados da Web Semântica), Auer et al. (2009a).

A ontologia LGD é, em parte, derivada de um modelo de dados relacional como apresentado na Figura 5.5 e na inclusão e reuso de classes como *SpatialThing* e geo-WGS84, relações e propriedades como, por exemplo, *locatedNear* e *rdfs:tag*.

No entanto, a sua estrutura principal são as tags do OSM, ou seja, as anotações atributo-valor para os nodos, formas e relações feitas pelos usuários. Todos os atributos são interpretados como classes e os seus valores são representados como subclasses. Dessa forma, *secondary*, *motorway* e *path* são subclasses da classe *highway*. Como resultado desse mapeamento, a ontologia resultante contém cerca de 500 classes, 50 *object properties* e cerca de 15.000 *datatype properties* (Auer et al., 2009a).

5.7. Repositórios de dados utilizados

Neste trabalho, a ontologia LinkedGeoData será utilizada para mapear os dados das coleções biológicas de repositórios de dados, tendo como foco as informações de nomes de lugares e coordenadas geográficas referentes a latitude e longitude.

Nesta seção, serão descritos os repositórios de dados utilizados para o desenvolvimento do *Gazetteer*, as subseções 5.7.1 e 5.7.2 descrevem os repositórios SpeciesLink e GBIF

5. Ferramentas

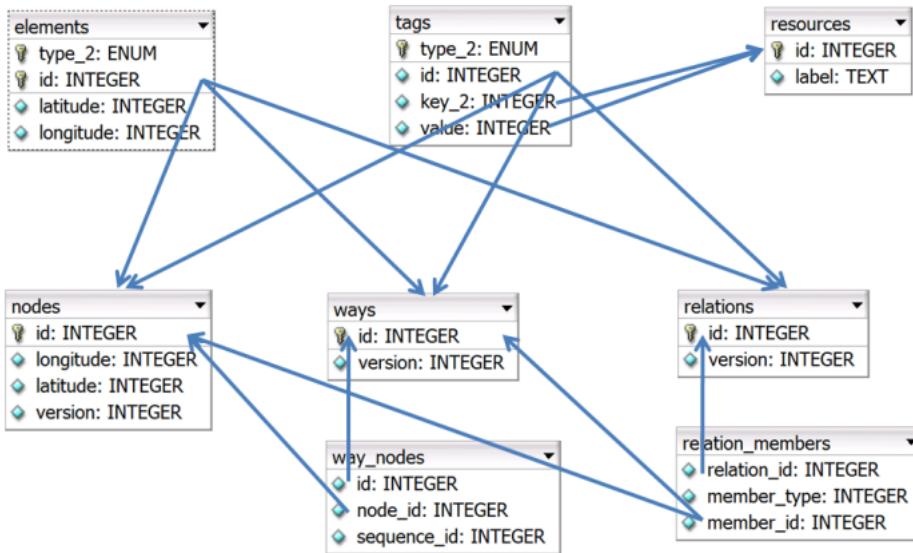


Figura 5.5.: Modelo Relacional da base de dados do LinkedGeoData. Fonte: (Auer et al., 2009a)

respectivamente, esses repositórios foram escolhidos por disponibilizar uma grande quantidade de informações geográficas de forma gratuita e serão utilizados para extrair as informações dos locais das coletas biológicas. Na subseção 5.7.3 é descrito o repositório Geonames que servirá como fonte de dados para recuperar informações ausentes nos dados do SpeciesLink e GBIF. Outra fonte de dados externos, que também é utilizada no *Gazetteer*, são os dados referente aos polígonos das reservas florestas disponibilizados pelo IBGE no link <http://mapas.mma.gov.br/i3geo/datadownload.htm>.

5.7.1. SpeciesLink

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).

Diariamente, o SpeciesLink traz informações estatísticas sobre seus dados, e, em 07 de junho de 2014, o mesmo registrava 6.726.878 registros on-line, sendo que apenas 2.868.386 eram georreferenciados, ou seja, somente 42% de todos seus dados possuíam informações

5.7. Repositórios de dados utilizados

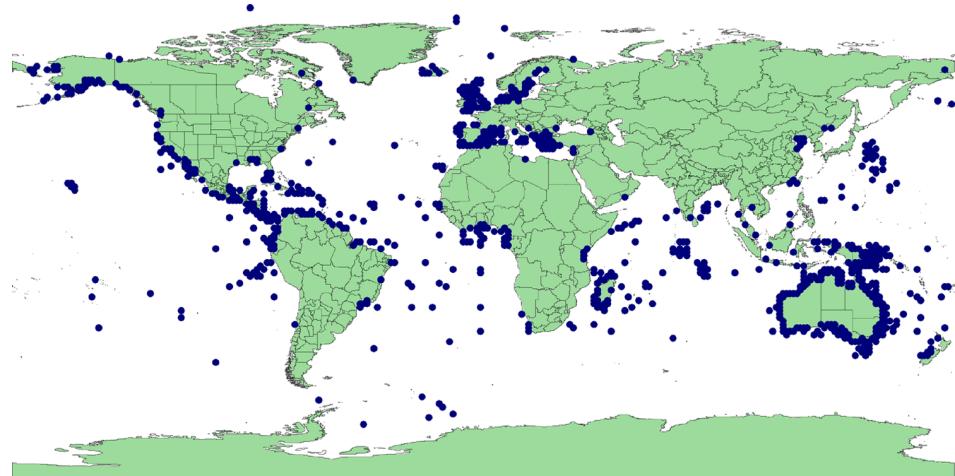


Figura 5.6.: Pontos representando coleções no repositório GBIF. Fonte: (Yesson et al., 2007)

geográficas(SpeciesLink, 2014).

5.7.2. Portal de busca do GBIF

O Global Biodiversity Information Facility (GBIF) é um repositório de dados que fornece acesso livre aos dados científicos sobre biodiversidade mundiais através da Internet utilizando Web Services.

Atualmente, o GBIF conta com cerca de mais de 200 milhões de registros referentes à história natural de espécimes e coletas de campo, que são agregados, indexados e armazenados em mais de 300 fontes.

Segundo Yesson et al. (2007), os registros geográficos contidos no repositório GBIF contém diversas coordenadas geográficas inválidas, como, por exemplo, pontos que referenciam coletas no mar, sendo que esses deveriam estar representados no continente, como exibido na Figura 5.6.

5.7.3. Geonames

Geonames é um *Gazetteer* que contém informações espaciais e temáticas para vários nomes de lugares ao redor do mundo. Seus dados estão disponíveis através de *Web Services* e também publicados em formato RDF (Koubarakis et al., 2012).

Contendo mais de 10 milhões de nomes geográficos e 5,5 milhões de nomes alternativos de lugares, seus dados de conhecimento geográfico são relacionados a bases de dados

5. Ferramentas

públcas dos Estados Unidos e utilizam também a DBpedia e outras fontes de dados (Koubarakis et al., 2012).

Além de nomes de lugares em várias linguagens, seus dados incluem informações como latitude e longitude no padrão *World Geodetic System 1984* (WGS84), informações sobre elevação, população, divisão administrativa e código postal (Koubarakis et al., 2012).

5.8. Considerações Finais

Neste capítulo, as tecnologias para desenvolvimento do *Gazetteer* Colaborativo foram descritas. Inicialmente a ferramenta Protégé para edição de ontologias foi abordada. Essa ferramenta terá como objetivo editar a ontologia LinkedGeoData, descrita na seção 5.6, que será utilizada para mapear os dados referentes aos locais das coletas biológicas. Em seguida, as *triplestore*, que serão analisadas para testes e irão armazenar as informações do *Gazetteer*, foram listadas e logo após, na Seção 5.3, a API Jena utilizada para comunicar com os mesmos foi abordada.

Para criar a GUI do usuário para utilização do *Gazetteer* Colaborativo, as ferramentas GWT e OpenLayer foram descritas nas seções 5.4 e 5.6 respectivamente. Utilizando-se essas APIs, será possível criar funcionalidades de consultas por meio de buscas textuais e mapas digitais que possibilitam a navegação. Por fim, os repositórios SpeciesLink, GBIF e Geonames, que serão utilizados como fonte de dados para o desenvolvimento deste projeto, foram abordados. Os dados do SpeciesLink e GBIF irão compor o *Gazetteer* e os dados do Geonames serão utilizados para aprimorar coordenadas geográficas imprecisas.

6. Experimentos

6.1. Introdução

Neste capítulo serão apresentados alguns resultados que já foram alcançados com o desenvolvimento do *Gazetteer* e discutidos futuros caminhos para aprimorar o projeto. A seção 6.2 descreve a análise dos dados referente aos repositórios SpeciesLink e GBIF, na seção 6.3 é descrito o método para aprimorar as coordenadas geográficas. A seção 6.4 relata como mapear os dados do *Gazetteer* para ontologias e como armazená-los. Na seção 6.5, uma análise entre as principais *triplesstores* para trabalhar com funções GeoSPARQL é relatada. A seção 6.6 descreve o protótipo de telas do *Gazetteer* para suportar a prática de VGI e na seção 6.8 é descrita a arquitetura proposta para o desenvolvimento do trabalho.

6.2. Análise dos dados utilizados referente aos repositórios SpeciesLink e GBIF

Para realizar este trabalho foram utilizados conjuntos de dados, referentes a coletas biológicas do estado do Amazonas, que estão disponíveis nos repositórios biológicos do SpeciesLink (2014) e GBIF (2014). Esses dados estão disponíveis no formato de arquivos “csv” (Comma Separated Values) e “xls” (Excel).

Ao analisar os dados do SpeciesLink e GBIF, Tabela 6.2 e Tabela 6.1, uma métrica de qualidade, descrita na Tabela 6.2, foi criada para avaliar a precisão dos dados geográficos. Nos registros coletados, é possível notar que apenas 24,85% de todos os dados do SpeciesLink e 16,80% do GBIF apresentam informações geográficas precisas, ou seja, contém o nome do local e município que um espécimen foi recolhido e suas coordenadas geográficas (latitude e longitude).

Outro ponto interessante verificado em relação a esses dados é, que entre os registros de coletas marcados com qualidade de informação 2 e 3, se verificou que 31,56% desses registros são referentes a coletas muito antigas, datadas entre 1850 a 1979. Essas coletas não poderiam ser georreferenciadas pela ausência de aparelhos GPS e a dificuldade de

6. Experimentos

se determinar coordenadas no meio da floresta sem eles. Essa afirmação também é comprovada por dos 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 um dos objetivos deste trabalho é tratar os dados que contém informações geográficas imprecisas, a partir de coletas com registros confiáveis, foram utilizados os dados de qualidade 2 a 4 para desenvolver o *Gazetteer*. Ou seja, um total de 80,21% dos registros do SpeciesLink e 88,1% do GBIF. Como os registros remanescentes não apresentam informações de local de coleta, latitude, longitude ou município, eles foram descartados por serem inutilizáveis pois não há como associar a eles informações de localização (além do fato de terem sido coletados no estado do Amazonas). Desta forma, 19,79% dos registros do SpeciesLink e 12,1% do GBIF foram retirados da construção do *Gazetteer*.

Informação Geográfica	Quantidade de Dados	%	Qualidade da informação
Nome do lugar, Latitude, Longitude, e município	60786	24.85%	4
Somente nome do lugar e município	91419	37.38%	3
Somente nome do lugar	43961	17.98%	2
Somente município	41071	16.79%	1
Não contém informações	7310	3%	0

Tabela 6.1.: Demonstração da qualidade dos dados do SpeciesLink

Informação Geográfica	Quantidade de Dados	%	Qualidade da informação
Nome do lugar, Latitude, Longitude, e município	25687	16.80%	4
Somente nome do lugar e município	11287	7.3%	3
Somente nome do lugar	98954	64%	2
Somente município	5915	3.8%	1
Não contém informações	12886	8.3%	0

Tabela 6.2.: Demonstração da qualidade dos dados do GBIF

Ao se analisar os dados do SpeciesLink e GBIF foi possível observar vários registros imprecisos como os exibidos na Figura 6.1 e Figura 6.2. Várias coordenadas apontam

6.2. Análise dos dados utilizados referente aos repositórios SpeciesLink e GBIF

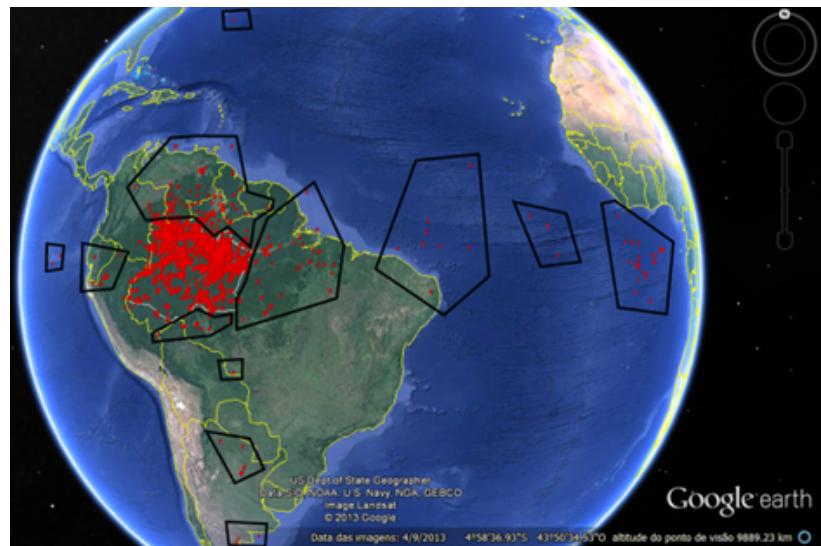


Figura 6.1.: Representação das coordenadas extraídas dos dados do SpeciesLink.

para regiões fora dos limites do estado do Amazonas como, por exemplo, no mar, na Argentina, nos 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 na Figura 6.1. 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 dessa imprecisão, ao se analisar os dados do SpeciesLink e GBIF de forma mais minuciosa, foi possível observar que várias localidades apresentavam informações de latitude e longitude imprecisas como, por exemplo, a situação mostrada na Figura 6.3 on page 66a e Figura 6.3 on page 66b. Nessas figuras, é possível visualizar que a Reserva Florestal Adolpho Ducke, representada pelo ponto verde na sua localização correta, aparece em alguns registros com coordenadas geográficas muito fora dessa posição, marcados em vermelho nas figuras.

Tendo em vista a quantidade de erros presentes e a falta de coordenadas geográficas nos registros do SpeciesLink e GBIF, foi desenvolvido um método para aprimorar essas coordenadas, esse método é descrito na seção seguinte.

6. Experimentos

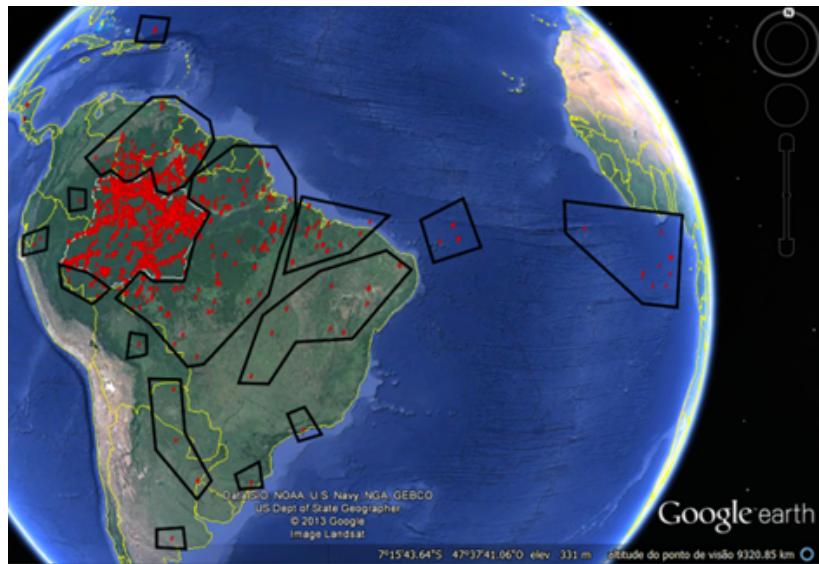
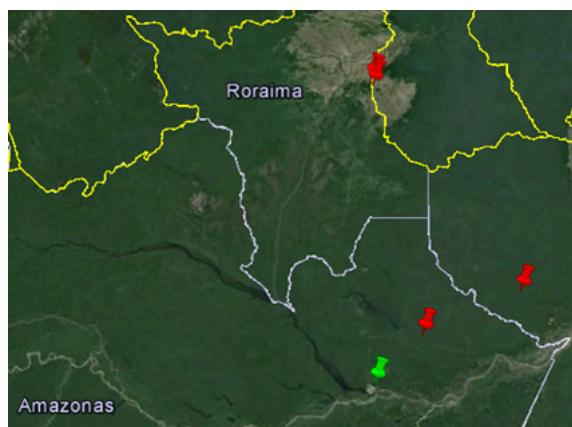
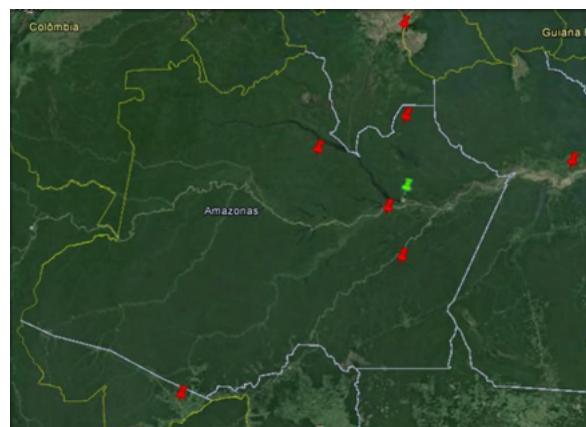


Figura 6.2.: Representação das Coordenadas presentes nos dados do GBIF.



(a) Coordenadas geográficas referentes a Reserva Adolpho Ducke presentes nos dados do SpeciesLink.



(b) Coordenadas geográficas referentes a Reserva Adolpho Ducke presentes nos dados do GBIF.

Figure 6.3.: Coordenadas referentes a Reserva Florestal Adolpho Ducke contidas nos dados do SpeciesLink Figura(a) e GBIF Figura(b). Pontos vermelhos representam coordenadas geográficas erradas para a reserva. O ponto verde representa a coordenada geográfica correta para a reserva.

6.3. Método para aprimorar coordenadas geográficas

A partir dos dados tratados pela etapa anterior, foi construído um *script* para selecionar registros que indicam uma localidade, utilizando-se técnicas presentes na RIG, como resolução de topônimos (expressões usadas para nomear um lugar), onde uma lista de lugares como: rios, reservas, parques nacionais, estradas foi gerada.

Essa lista foi composta por 45 topônimos, onde cada um contém uma expressão regular associada para extrair as informações da base de dados. Por exemplo, foi possível recuperar registros que indicam uma reserva utilizando-se da seguinte expressão regular (*regular expression*) em Java: `(?i)reserva\b(.+?)[,/.;/:]`. Nesse exemplo, todas as tuplas que possuem a palavra reserva no início da palavra são retornadas. Essa técnica também é utilizada por (Gouvea et al., 2008) e (Machado et al., 2011).

Após recuperar os registros de um dado local, os mesmos foram agrupados usando o algoritmo de agrupamento StarAslam et al. (2004a). Nesse algoritmo, dado um objeto, todos os outros objetos similares a ele são identificados e conectados, formando um grafo em forma de estrela, de modo que todos os objetos de um cluster ficam conectados a um centroide.

Em conjunto com o algoritmo de agrupamento, foi utilizada uma lista de *stop words*, que são palavras comuns que não apresentam nenhum significado linguístico. A utilização dessa lista se deve ao fato que vários dados da base escolhida contém termos como, por exemplo, "solo arenoso" ou "terra firme".

Para verificar a similaridade entre as localidades e agrupá-las, foi utilizado o coeficiente de similaridade de Jaccard (Yin und Yasuda, 2006), que observa quão parecida são duas palavras, gerando valores no intervalo de [0,1], onde o valor 0 representa palavras extremamente diferentes e 1 palavras exatamente idênticas. Essa abordagem se deve ao fato que nomes para a mesma localidade podem ter várias formas de escrita para representar o mesmo local, como demonstrado na Figura 6.4 e também relatado por (dos Santos, 2003).

Para utilizar o coeficiente de Jaccard, um limiar de similaridade entre as localidades foi definido manualmente. A partir de testes iniciais verificou-se que, quanto mais próximo de 1 era o valor desse limiar, mais grupos eram gerados. Dessa forma, nomes referentes a uma mesma localidade eram separados em grupos distintos pelo fato da grafia utilizada ser diferente, como demonstrado pela Figura 6.4. Assim, para se obter melhores resultados e verificar qual o melhor limiar a ser utilizado, o experimento limitou-se aos valores 0.4, 0.5 e 0.6.

Para amenizar a imprecisão das coordenadas geográficas mencionadas na seção Section 6.2, foi criado um método para aprimorar as coordenadas geográficas da base de dados

6. Experimentos

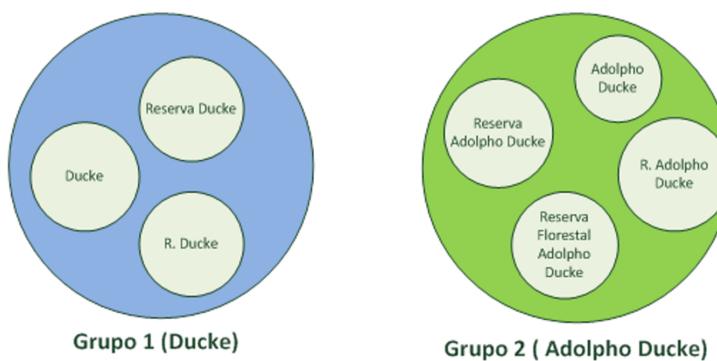


Figura 6.4.: Exemplo de vários grupos criados de acordo com o limiar escolhido.

utilizada. Inicialmente esse método recupera as informações referente as localidades do estado do Amazonas no repositório Geonames por meio de um Web Service. Após obter os dados do Geonames, é feita uma verificação entre os locais da base de dados do SpeciesLink e GBIF que são similares e não possuem coordenadas geográficas, caso algum registro seja similar, a coordenada presente no Geonames é inserida nos dados do SpeciesLink e do GBIF.

A escolha do repositório Geonames se deve ao fato desse *Gazetteer* conter mais de 10 milhões de nomes geográficos e 5,5 milhões de nomes alternativos a lugares. Sendo o mesmo largamente utilizado nos trabalhos em RIG, tais como (Ahlers, 2013) e (KeSSLer et al., 2012a).

Após a atualização das coordenadas, o método de aprimoramento de coordenadas toma como base os diversos grupos existentes e uma sumarização dos valores que se repetem frequentemente é feita, como mostrado na Figura 6.5.

Para satisfazer esse 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 *Gazetteer*, apenas é verificado os valores que aparecem com mais frequência no grupo. Após obter o valor que mais se repete, suas informações geográficas são atribuídas ao centroide indicando que aquela localidade assumirá o novo valor para todos os registros, Figura 6.5.

A escolha desta abordagem, se deve ao fato de que 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). Embora tenhamos

6.3. Método para aprimorar coordenadas geográficas

Limiar	Coordenadas Corretas
0.4	84.3%
0.5	83.8%
0.6	84.1%

Tabela 6.3.: Precisão das coordenadas de acordo com o limiar de agrupamento.

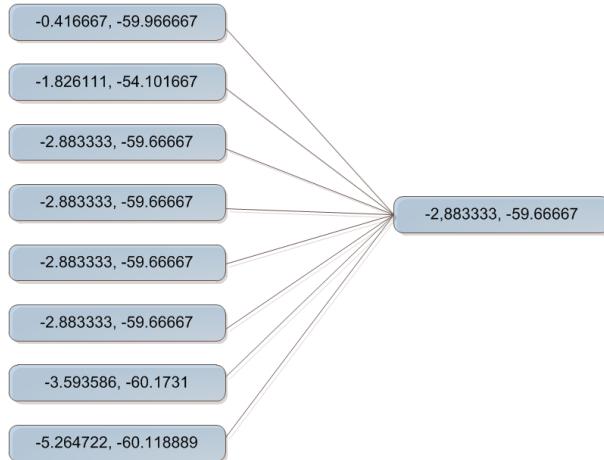


Figura 6.5.: Demonstração do método de sumarização de coordenadas geográficas, onde os valores que ocorrem com mais frequência são identificados e atribuídos para as demais coordenadas..

proposto essa abordagem, garantimos apenas 84% de precisão nessas coordenadas (Tabela 6.3), após realizar a análise das localidades associadas corretamente, apresentada na Figura 6.8.

Como relatado na seção Section 6.2, os dados do SpeciesLink e do GBIF possuem grande imprecisão em suas coordenadas geográficas, sendo possível encontrar informações que apontam para o mar, em países vizinhos ao Brasil na Argentina ou no nordeste do Brasil. Para melhorar a qualidade desses dados, foi utilizado o método de sumarização de coordenadas geográficas descrito neste trabalho, onde, após a execução do mesmo, os dados gerados se mostraram mais precisos, como visualizado na Figura 6.6.

Outra contribuição deste trabalho para melhorar a acurácia dos dados do SpeciesLink e do GBIF, é a atualização os registros que não contém informações geográficas, por meio de registros que possuem e foram validados pela técnica de sumarização.

Os repositórios SpeciesLink e GBIF contêm 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 descrita neste trabalho, o número de registros

6. Experimentos



Figura 6.6.: Resultado das coordenadas geográficas após o método de sumarização.

com coordenadas geográficas do GBIF aumentou para 30,78% (foram inseridos cerca de 20.000 registros) e no SpeciesLink houve um acréscimo de 37,33% (cerca de 30.000 registros foram inseridos), Figura 6.7. Esses números representam um aumento significativo, de cerca de 90%, do número de registos com informação geográfica. Assim, podemos afirmar que o uso do *Gazetteer* (com a técnica de sumarização, proposta neste trabalho) pode levar a um aumento significativo da informação geográfica em dados típicos sobre biodiversidade.

Estes dados evidenciam que, de fato, a construção do *Gazetteer* contribui para amenizar os problemas de acurácia existentes em informações geográficas na área de biodiversidade. Além disso, é importante ressaltar que muitos desses registros são ocorrências antigas, ou seja, quando o uso de equipamentos com GPS não era possível, de valor inestimável pois os habitats onde essas coletas foram feitas não existem mais ou foram alterados pela presença humana.

Além de verificar a quantidade de dados que foram aprimorados, também é necessário verificar se essas localidades agrupadas pertencem aos centroides corretos e contém informações geográficas precisas. Para isso foi realizada uma verificação dos locais agrupados.

6.3. Método para aprimorar coordenadas geográficas

Amount of georeferenced data

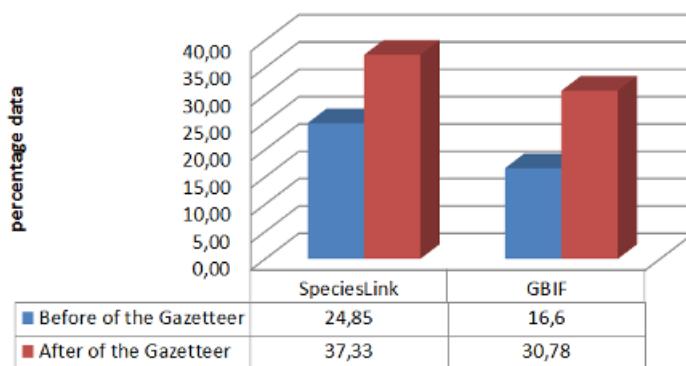


Figura 6.7.: Quantidade de coordenadas geográficas recuperadas pelo *Gazetteer*.

Localidade em um mesmo centroide	
BR-319 sentido Porto Velho a ca 20 Km.	
BR-319 sentido Porto Velho a ca 30 Km.	

Tabela 6.4.: Coordenadas que foram agrupadas de forma imprecisa.

6.3.1. Verificação dos locais agrupados

Para verificar a precisão dos locais do *Gazetteer*, foi feita uma análise com o objetivo de verificar a precisão do agrupamento criado. Para isso, uma amostra de 50 grupos foi selecionada aleatoriamente (as informações dos grupos são representadas pelos círculos brancos, na Figura 6.4). Após essa seleção, cada valor foi verificado manualmente. Nessa verificação, foi analisado se o centroide que a localidade se referia realmente era o correto.

Ao se realizar a análise desses grupos coletados, foi possível verificar que a coerência das localidades associadas foi bem significativa, em média 84% das coordenadas estão associadas corretamente e que nenhum dos limiares escolhidos apresenta grandes variações quanto à precisão dos dados associados, com os valores 84,3% para o limiar 0.4, 83,8% para o limiar 0.5 e 84,1% para o limiar 0.6, como exibido na Figura 6.8, na barra de Localidade associada.

Embora se tenha conseguido valores significativos para esse agrupamento de localidades, após uma análise detalhada desses grupos, foi possível verificar que, embora algumas localidades apresentem maior similaridade entre suas palavras e estejam num mesmo centroide, isso não garante que elas representem um mesmo local como demonstrado na Tabela 6.4.

A ocorrência desse fator levou a redução do valor de localidades associadas correta-

6. Experimentos

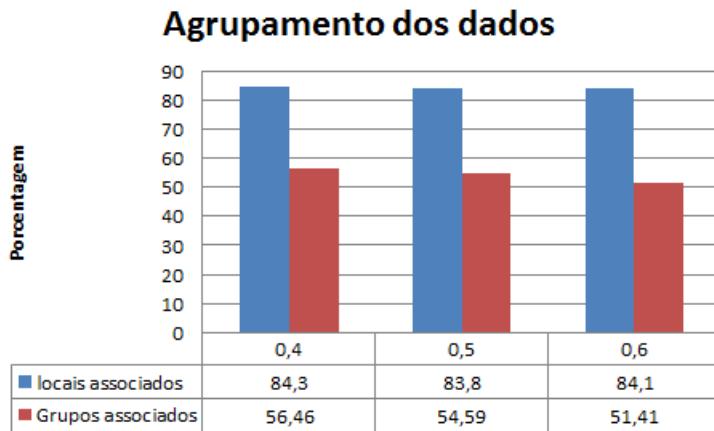


Figura 6.8.: Precisão dos locais associados e grupos formados.

Limiar	Quantidade de locais
0.4	4602
0.5	4782
0.6	4921

Tabela 6.5.: Quantidade de locais presentes no *Gazetteer*.

mente, pois locais distintos eram agrupados em um mesmo grupo. Esse fenômeno é ocasionado pelo fato da técnica de agrupamento por similaridade entre palavras não conseguir determinar a semântica de um local.

Analizando esse fato, pode-se perceber que levar em consideração apenas a similaridade não é uma boa abordagem para referenciar localidades, pois ela despreza a semântica que está sendo passada naquele valor. Por isso, as modificações futuras do *Gazetteer* irão incluir o uso de ontologias para auxiliar na Resolução de Topônimos.

Além da verificação interna entre as localidades agrupadas, também foi realizada uma análise dos nomes representados pelos centroides. Essa análise tem como intuito verificar se os centroides estão corretos e se representam os possíveis grupos presentes na base de dados do SpeciesLink e GBIF.

Inicialmente o *Gazetteer* construído possuía a quantidade de grupos exibidos pela Tabela 6.5 e para realizar a validação desses grupos foi coletada uma amostra de 100 centroides aleatórios (os centroides são representados pelos círculos azul e verde da Figura 6.4), ao se verificar esses dados, foi possível visualizar que vários centroides representavam o mesmo local, como visualizado na Tabela 6.6.

O motivo da ocorrência desse fenômeno é o fato de vários registros terem grafias diferen-

6.3. Método para aprimorar coordenadas geográficas

Valores da amostra de centroides selecionados
Reserva Florestal Adolpho Ducke
Reserva Florestal Ducke (Associação) Ha A3
Reserva Florestal Duckec estr do Acará.
Reserva Florestal Duxke Manaus-Itacoatiara km 26 Área do Acará Floresta de Campinarana

Tabela 6.6.: Vários centroides representando o mesmo local.

tes, que vão desde erros ortográficos, para descrever localidades, a descrições heterogêneas se referindo a locais semelhantes. Gouvea et al. (2008) relatam que o uso de similaridade entre as palavras para agrupar registros pode gerar esse tipo de resultado nos dados, devido a grafia empregada.

Devido à frequente ocorrência desse fato na amostra coletada, a taxa de grupos associados corretamente foi muito penalizada e seus valores ficaram abaixo dos resultados obtidos pelas localidades associadas corretamente, em média 29% menos consistentes que as localidades, como visualizado na Figura 6.8. A diferença de 29% obtida após a verificação dos centroides, informa que foram criados diversos grupos para uma mesma localidade. Em trabalhos futuros, pretendemos verificar a possibilidade de unir esses grupos por suas coordenadas geográficas.

Embora o número de grupos para representar todas as localidades possa ser aprimorado, ainda foi possível contribuir com o aprimoramento das informações geográficas presentes nos repositórios SpeciesLink e GBIF. Além disso, com a construção do *Gazetteer*, especialistas podem extrair informações geográficas de grandes quantidades de dados de forma mais rápida, tarefa essa que era muito laboriosa no quadro inicial deste trabalho, onde os dados estavam armazenados em arquivos “csv”, com várias informações imprecisas e difíceis de serem analisadas.

Vale ressaltar que ao se trabalhar com dados reais, que foram coletados e registrados por humanos, os mesmos estão sujeitos a grandes variações e imprecisões, aumentando a complexidade ao se comparar os resultados deste trabalho com outros *Gazetteers* que utilizam dados sintéticos, extraídos de páginas web (que podem ter vindo de bancos de dados).

Após a construção do *Gazetteer* e tratamento dos dados do SpeciesLink e GBIF, foram utilizadas ontologias geoespaciais para mapear os dados gerados pelo *Gazetteer* para uma *triplesstore*. Esse mapeamento é descrito na próxima seção.

6. Experimentos

```
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 6.9.: Mapeamento das informações geográficas para a especificação *asWKT* (Representação simplificada).

6.4. Mapeamento dos dados e disponibilização do *Endpoint GeoSPARQL*

Inicialmente, os dados geográficos frutos deste experimento foram 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 et al., 2010) e (Auer et al., 2009b). A reutilização de ontologias possibilita seguir os principais padrões já especificados e validados na área, além de reduzir o trabalho para construção do *Gazetteer*.

Essas duas ontologias existem separadamente e para conectá-las foi utilizada ferramenta de edição de ontologias, Protégé 4, Figura 6.11. Por meio da classe *Feature*, presente em ambas ontologias, as mesmas foram unificadas utilizando-se o relacionamento *EquivalentTo* (\equiv) que possibilita dizer que as duas classes são equivalentes, Figura 6.11.

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, conforme exemplificado na Figura 6.9.

Para mapear as informações do *Gazetteer* referentes a localidade, município e data do registro para a ontologia, foram criados os *data properties*: *locality*, *county* e *date*, já que a ontologia LinkedGeoData não continha esses *data properties*. Já as coordenadas geográficas, pontos e polígonos, foram mapeadas para o *data property asWKT*, presente na ontologia GeoSPARQL.

Após representar os dados do *Gazetteer* (locais, municípios, data e pontos) é necessário criar indivíduos para representar as localidades e as geometrias que compõe as coordenadas geográficas. Para realizar essa representação, é necessário criar um indi-

6.4. Mapeamento dos dados e disponibilização do Endpoint GeoSPARQL

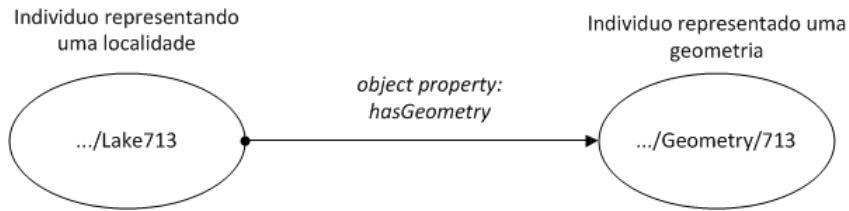


Figura 6.10.: Representação da união entre um indivíduo referente a uma localidade com outro indivíduo que representa uma geometria usando o *object property hasGeometry*, disponibilizado pela ontologia GeoSPARQL.

víduo representando uma localidade e unificá-lo com um indivíduo que representa uma geometria, por meio do *object property hasGeometry* (Figura 6.10). Essa relação define que um indivíduo pode possuir várias Geometrias, permitindo assim à *triplestore* realizar inferências geográficas sobre os dados utilizando as funções, definidas pelo GeoSPARQL.

Após o mapeamento dos dados do *Gazetteer*, usando a ontologia, os registros foram armazenados na *triplestore* Parliament, que tem como função servir de *endpoint* para realização de consultas. O Parliament é um software gratuito que permite o armazenamento de triplas em RDF, manipulação de OWL e disponibilização de *endpoints* GeoSPARQL para realizar consultas. A escolha inicial da *triplestore* Parliament, se deveu a possibilidade dela utilizar OWL e poder disponibilizar *endpoints* GeoSPARQL.

O armazenamento dos dados numa *triplestore* promove a fácil disponibilização dos dados como *Linked Open Data*. Um dos objetivos deste projeto é expor, compartilhar e conectar dados, informação e conhecimento na web semântica sobre dados referentes a localizações geográficas relacionadas a biodiversidade. Linked Open Data basicamente define as melhores técnicas para isso (Heath und Bizer, 2011).

Com intuito de verificar a possibilidade de realizar a expansão de consultas geográficas proposta por Kessler et al. (2009), utilizando-se de *endpoints* para realização de consultas, foram selecionados três repositórios que contém entidades geográficas, Geonames, WikiMapia e Wikipédia. No entanto, verificou-se que nenhum desses repositórios possui implementações de *endpoints* SPARQL para seus dados. Dessa forma, foi necessário encontrar outros repositórios que contivessem *dumps* dos dados do Geonames, WikiMapia e Wikipédia 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 o *parser* de seus dados para esse formato, esse *Gazetteer* foi descartado do experimento. Sendo assim, foram selecionados somente *endpoints* que possuem informações sobre os repositórios Geonames e Wikipédia.

6. Experimentos

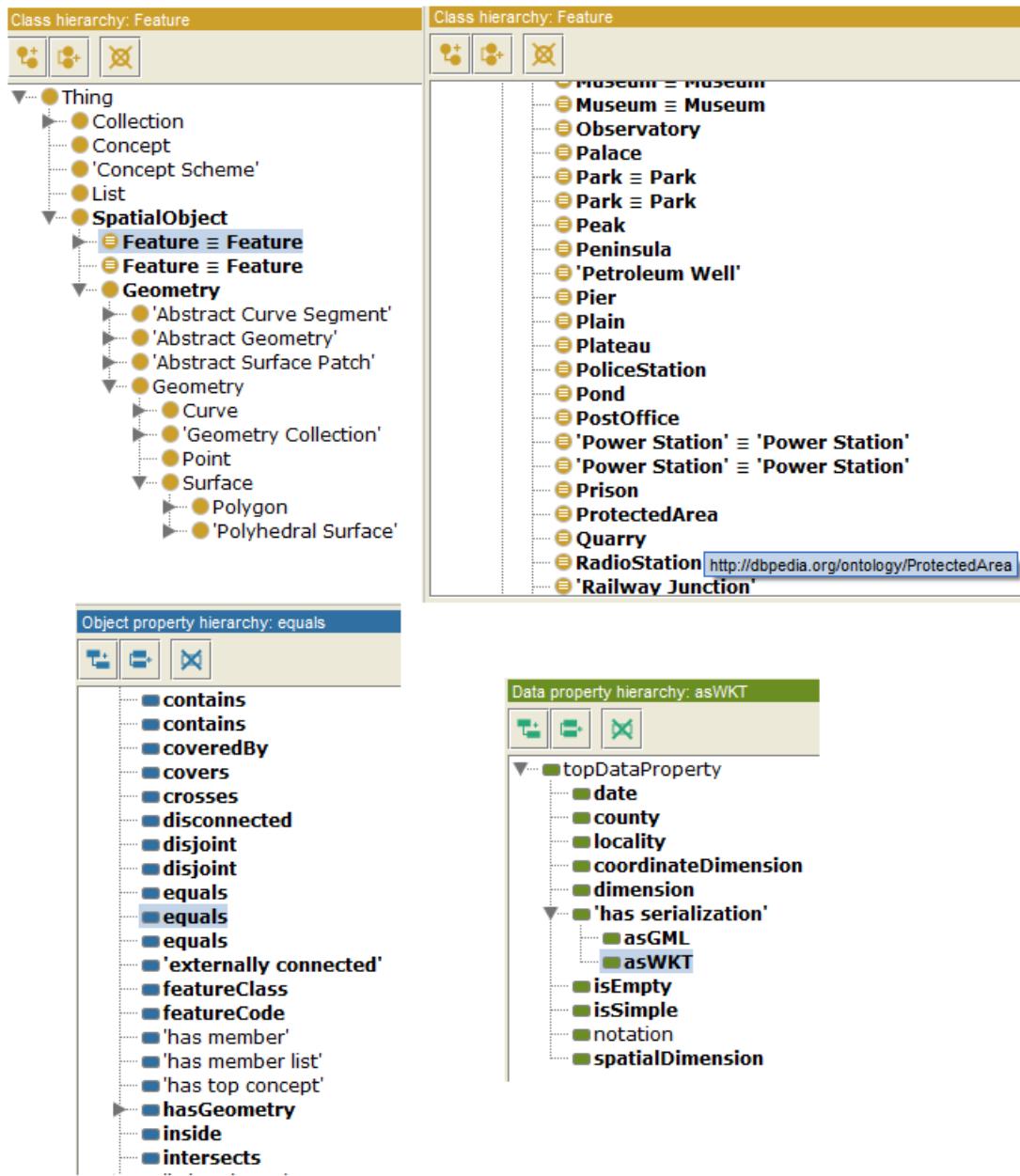


Figura 6.11.: Edição da ontologia utilizada pelo *Gazetteer* com auxílio da ferramenta Protégé.

6.4. Mapeamento dos dados e disponibilização do Endpoint GeoSPARQL

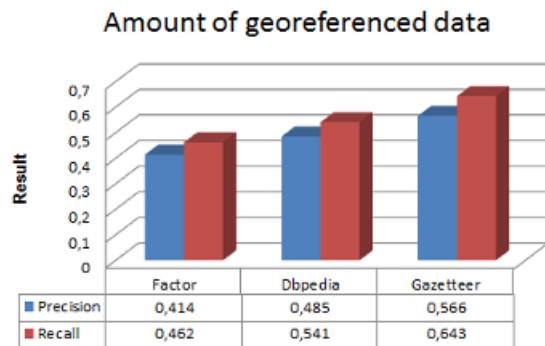


Figura 6.12.: Resultados de precisão e revocação após as buscas semânticas realizadas.

Junto a essa análise, também foi verificado se os *endpoints* disponibilizavam implementações de GeoSPARQL para realizar consultas utilizando as funções geoespaciais, como, 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 *GeoSPARQL*³. 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 notado que o *endpoint GeoSPARQL* contém apenas informações referente aos municípios brasileiros, desprezando locais como reservas naturais, rios, lagos, entre outros. Também foi possível notar que, dentre os três *endpoints*, somente ele permitia realizar funções geoespaciais. No entanto, como seus dados não abordam reservas, lagos, rios, como os outros, ele foi descartado do experimento.

Para realizar as queries, uma amostra de 60 localidades da base de dados inicial do SpeciesLink e GBIF foi selecionada. Dessa amostra, foi criada uma base de dados confiável sobre quais registros são relevantes para uma determinada informação.

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 6.12. Nesses resultados é possível verificar que, com a utilização da busca semântica, é possível obter valores de precisão e revocação próximos, ou seja, no *Gazetteer* implementado foi possível obter 0,566 de precisão e 0,643 de revocação, o que não acontece em sistemas de busca por palavras chaves como relatado por Amanqui et al. (2013a).

No entanto, devido ao fato desses repositórios não conterem todos os locais pesquisados a revocação dos dados foi de, no máximo, 0,54. Demonstrando assim que a criação de

¹<http://dbpedia.org/sparql>

²<http://factforge.net/>

³<http://www.geosparql.org/>

6. Experimentos

Algoritmo 6.1 Consulta por fazendas dentro de áreas protegidas.

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geo:<http://www.opengis.net/ont/geosparql#>
PREFIX geof:<http://www.opengis.net/def/function/geosparql/>
PREFIX dbp:<http://dbpedia.org/ontology/>
PREFIX lgdo:<http://linkedgeodata.org/ontology/>
SELECT ?p ?a ?w1 ?w2
WHERE {
    ?p rdf:type lgdo:Farm .
    ?p geo:hasGeometry ?g2 .
    ?g2 geo:asWKT ?w2 .
    ?a rdf:type dbp:ProtectedArea .
    ?a geo:hasGeometry ?g1 .
    ?g1 geo:asWKT ?w1 .
    FILTER geof:sfWithin(?w2, ?w1)
}
```

outro *endpoint* para disponibilizar dados sobre localidades relacionadas à coletas biológicas é válida. Além disso, ao realizar as consultas foi possível comprovar um dos desafios citados por KeSSLer et al. (2009), em seu artigo sobre a próxima geração de *Gazetteers*, pois, de fato, alguns repositórios, como em especial o DBpedia, possuem algumas localidades que não foram encontradas na *triplestore* utilizado neste trabalho.

Outro motivo que torna válida a criação do *Gazetteer* e a disponibilização de seus dados 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: informações sobre municípios brasileiros puderam ser recuperadas apenas num único *endpoint* que realizava consultas geoespaciais.

A disponibilização de um *endpoint* GeoSPARQL é importante, devido a necessidade de se realizar consultas geoespaciais que contenham significado semântico. Isso fica claro no exemplo da consulta (Algoritmo 6.1) 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. (2013a). 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.

Algoritmo 6.2 Consulta para verificar as cachoeiras que estão a uma distância de 1000 metros de alguma represa.

```

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geo:<http://www.opengis.net/ont/geosparql#>
PREFIX geof:<http://www.opengis.net/def/function/geosparql/>
PREFIX lgdo:<http://linkedgeodata.org/ontology/>
SELECT ?p ?a
WHERE {
    ?p rdf:type lgdo:Dam .
    ?p geo:hasGeometry ?g2 .
    ?g2 geo:asWKT ?w2 .
    ?a rdf:type lgdo:Waterfall .
    ?a geo:hasGeometry ?g1 .
    ?g1 geo:asWKT ?w1 .
    FILTER(geof:distance(?w1,?w2,units:metre)<1000)
}

```

Outro exemplo de consulta que possui significado semântico, é a consulta exibida no Algoritmo 6.2. Nessa consulta, é possível localizar todas as cachoeiras que estão a menos de 1000 metros de alguma represa.

Os resultados obtidos neste trabalho foram publicados, no artigo *A Gazetteer for Biodiversity Data as a Linked Open Data Solution* na conferência Web2Touch 2014 - *Modelling the Collaborative Web Knowledge* (Cardoso et al., 2014).

Durante o processo de desenvolvimento do *Gazetteer* foi possível encontrar outras duas *triplestores* que trabalham com implementações GeoSPARQL, o Strabon e o uSeekM, e, para avaliar-as, foram realizados os testes descritos na próxima seção.

6.5. Avaliação das *triplestores*

Com intuito de verificar a performance das *triplestores* Strabon, uSeekM e Parliament. Além de verificar se a substituição do *triplestore* Parliament era vantajosa, conforme relatado no experimento realizado por Garbis et al. (2013), foram realizados testes utilizando buscas semânticas que continhamém funções GeoSPARQL não topologicas, junções espaciais e seleções espaciais.

Para realizar este teste, uma base de dados contendo 40 mil pontos sobre localidades referentes ao repositório do Geonames foi selecionada, essa base pode ser encontrada no link, <http://geographica.di.uoa.gr/datasets/geonames.tar.gz>. Já as *triplestore* utiliza-

6. Experimentos

dos podem ser acessados por meio dos links, Parliament <http://biomac.icmc.usp.br:8080/parliament/>, Strabon: <http://biomac.icmc.usp.br:8080/strabon/>, UseekM: <http://biomac.icmc.usp.br:8080/useekm-workbench/repositories/SYSTEM/query>.

Com intuito de testar a disponibilidade do serviço GeoSPARQL nessas *triplestores*, um *Web Client* Java usando requisições RESTful (acessando os dados das *triplestores* usando requisições HTTP) foi desenvolvido. Essa cliente permite realizar as solicitações às *triplestores* segundo os padrões de Linked Open Data especificados por Heath und Bizer (2011). O código fonte dessa implementação pode ser acessado no link https://github.com/silviodec/Triple_Store_Test.git, onde é possível obter o projeto para replicar o experimento e contribuir com melhorias para o mesmo.

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 foi instalado o 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 requisições 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 *triplestore*.
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, a partir de um polígono ou ponto informados manualmente verifica-se se algum ponto ou polígono armazenado na *triplestore* está contido ou intercepta o polígono informado manualmente.

Cada uma das implementações, referente as consultas utilizadas, está descrita no Apêndice Chapter A. A seguir, o resultado do experimento com as três *triplestore* é descrito e uma análise das *triplestore* é realizada.

6.5. Avaliação das *triplesstores*

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 6.7.: Resultados da execução das consultas geoespaciais nas *triplesstore* Strabon, uSeekM e Parliament.

6.5.1. Comparação de buscas Geoespaciais entre *triplesotres*

O primeiro conjunto de consultas testadas foram as não topológicas, *queries* 1 e 2 no Apêndice Chapter A, nessas consultas a *tripleStore* uSeekM obteve melhor desempenho em relação às outras duas. Esse resultado é devido 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 et al. (2013).

Nas consultas referentes as junções espaciais, a *triplestore* Strabon obteve os melhores resultados. Esse resultado se deu devido ao fato das *triplesores* 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 6.7.

Por fim, no conjunto de consultas referente a seleções espaciais, as *triplesores* 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 *triplestore* nativa do Sesame e isso acarreta num pouco mais de sobrecarga.

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

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

1. A necessidade de todas as geometrias especificarem a utilização do sistema de referência espacial 4326/WGS84.

6. Experimentos

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 num 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 BNode's 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 *triplestore* 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 *triplestore* Parliament foi possível constatar que:

1. A 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 serem corrompidos e isso acarreta inviabilização de seus dados para uso.

6.6. Protótipo para o Gazetteer Colaborativo

2. Possui suporte para OWL e 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.
5. Sua versão 2.7.6 ainda não é totalmente estável em sistemas UNIX.

Com estes resultados, é possível verificar que todas as *triplestores* necessitam de ajustes. Futuramente será verificada a possibilidade de utilização da *triplestore* uSeekM em conjunto com a *triplestore* BigData: será testado se é possível usar a SAIL API da uSeekM para oferecer as funcionalidades de GeoSPARQL para a *triplestore* BigData (que possui suporte básico para OWL). Caso isso seja possível, ambos serão adotados para o desenvolvimento deste trabalho.

6.6. Protótipo para o Gazetteer Colaborativo

Devido ao fato de iniciativas voluntárias de VGI serem desenvolvidas por um determinado grupo que possui um escopo específico em uma determinada área de interesse, é importante garantir que os sistemas VGI atendam a determinados requisitos. Dentre eles Klinkenberg (2013) descreve que:

1. Um dos aspectos mais importantes para sistemas que oferecem suporte a VIG deve ser sua forma simplificada de receber novas contribuições, pois sistemas complexos, que requerem diversos passos aos usuários para inserirem 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 intuitivo quanto possível, sem a necessidade de tutoriais extensos.
4. Web Sites e interfaces públicas devem disponibilizar suas informações atualizadas, pois, se as informações fornecidas num web site não forem atuais, os usuários e participantes irão se desmotivar com o projeto.

6. Experimentos

Com intuito de atender a esses requisitos, vamos criar um protótipo de telas para ser validado junto aos biólogos. A seguir serão descritas as principais telas, referentes a inserção, buscas, validação de coordenadas e informações de Linked Open Data.

A tela inicial do *Gazetteer Colaborativo*, apresentada na Figura 6.13, contém, na parte superior, um campo de busca que permite aos usuários entrarem com o nome de alguma localidade para ser pesquisada no sistema. Além disso, é possível definir qual a qualidade das informações a serem pesquisadas por meio do campo de escolha de qualidade.

A qualidade escolhida permitirá retornar informações geográficas de acordo com o tipo de usuário que disponibilizou a mesma, como, por exemplo, biólogos do INPA, usuários com grande aceitação de seus registros ou dados do IBGE. Futuramente essa métrica de qualidade será definida em conjunto com os biólogos.

Na Figura 6.13, pode ser observado o mapa de navegação principal que contém as funcionalidades de navegação e marcação das coordenadas geográficas como, por exemplo, pontos, polígonos e linhas. Esse mapa também permite aos usuários visualizarem informações referente as localidades.

Ao lado esquerdo, o usuário encontra um botão para recuperar localidades que não possuem coordenadas geográficas. Quando o sistema recuperar essas informações, as mesmas serão listadas na tabela abaixo desse botão. Vale ressaltar que essa mesma tabela também exibe as informações sobre coordenadas geográficas em conjunto com o mapa de navegação.

Na parte inferior esquerda, 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. As informações necessárias são: nome do local, a GeoTAG referente ao tipo do local, o link referente a alguma fonte de *Linked Open Data* e as informações geográficas que podem ser inseridas manualmente ou através do mapa.

Por fim, no lado direito inferior, são apresentadas as GeoTags, que podem ser utilizadas para facilitar o processo de busca por localidades e as informações estatísticas dos lugares contidas no *Gazetteer*.

A próxima tela, apresentada na Figura 6.14, representa a funcionalidade de busca. Após o usuário digitar alguma localidade e procurar suas informações, por exemplo, "fazendas próximas a Itacoatiara", as informações sobre as localidades que satisfazem essa busca são recuperadas e exibidas no mapa de navegação e na tabela de localidades.

Uma vez que os dados estão na tabela de informações, o usuário pode selecionar alguma localidade. Ao selecionar o registro, as informações referentes a essa fazenda são mostradas no centro do canto inferior da tela. Essas informações são a URI da localidade

6.6. Protótipo para o Gazetteer Colaborativo

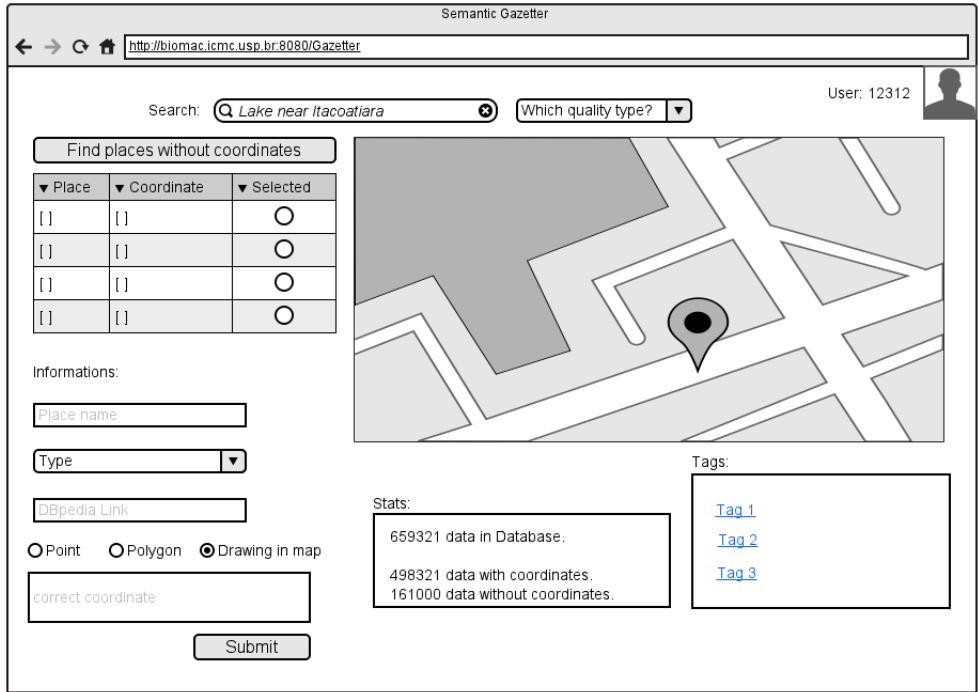


Figura 6.13.: Tela referente a funcionalidade de inserção de novos locais no *Gazetteer*.

pesquisada, o nome da localidade, o número de registros que possuem o mesmo nome de localidade e o número de pessoas que concordam com a informação exibida.

No lado inferior esquerdo, é exibido um gráfico que mostra o número de contribuições ao longo do tempo pela quantidade de pessoas que concordam com aquela informação. Como o registro desse exemplo ainda não possui coordenadas geográficas na *triplesstore*, o gráfico exibido não contém essas informações.

Caso alguma fazenda selecionada possua informações geográficas, a tela apresentada na Figura 6.15 é exibida. No canto inferior esquerdo dessa tela, são exibidas as informações referentes às contribuições de usuários ao longo de um determinado período, indicando se eles concordam ou discordam da informação presente no *Gazetteer*.

Caso a porcentagem de pessoas que concordam seja acima de 70%, os usuários somente podem concordar ou discordar da informação. Se, em algum momento, essa porcentagem estiver abaixo de 70% os usuários podem inserir lugares, que serão armazenados numa lista de candidatos e então um centroide será calculado para essa lista e ele assumirá como nova coordenada geográfica válida.

Para implementar essa funcionalidade, serão atribuídos pesos aos usuários, por exemplo, as sugestões de um voluntário comum recebem peso 1, as de um biólogo do INPA

6. Experimentos

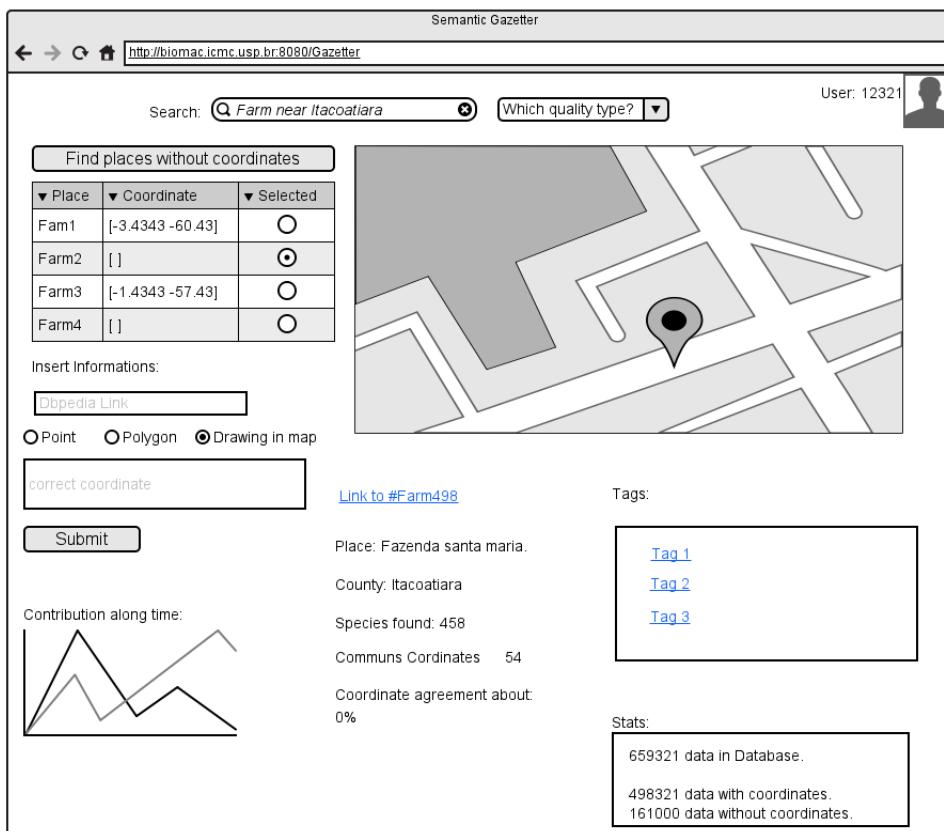


Figura 6.14.: Tela referente a função de buscas para o protótipo do *Gazetteer*.

recebem peso 3 e as de um usuário cartógrafo do IBGE recebem peso 5.

A escolha dos pesos de valor 3 para os biólogos do INPA e 5 para cartógrafos do IBGE, se deve ao fato de que essas informações são mais confiáveis que as informações fornecidas por um usuário comum, assim, para invalidar uma informação inserida por algum biólogo do INPA são necessários mais de 3 usuários comuns. Esses números (1, 3 e 5) ainda não estão definidos e podem mudar a partir de experimentos e opiniões de especialistas. Informações inseridas no Gazetteer de fontes autoritativas, como bancos de dados oficiais do IBGE, só podem ser mudadas manualmente.

A última tela do protótipo, Figura 6.16, tem como objetivo exibir as informações sobre *Linked Open Data*. No exemplo apresentado na Figura 6.16, as informações exibidas, na parte superior à esquerda, representam as informações da localidade e as formas para obter seus dados (RDF, GML, JSON). Já na parte superior à direita, são exibidas as informações referentes a DBpedia, quando alguma informação estiver disponível. Na parte central à esquerda, são mostrados o número de pessoas que contribuíram com

6.7. Arquitetura proposta para desenvolvimento do trabalho

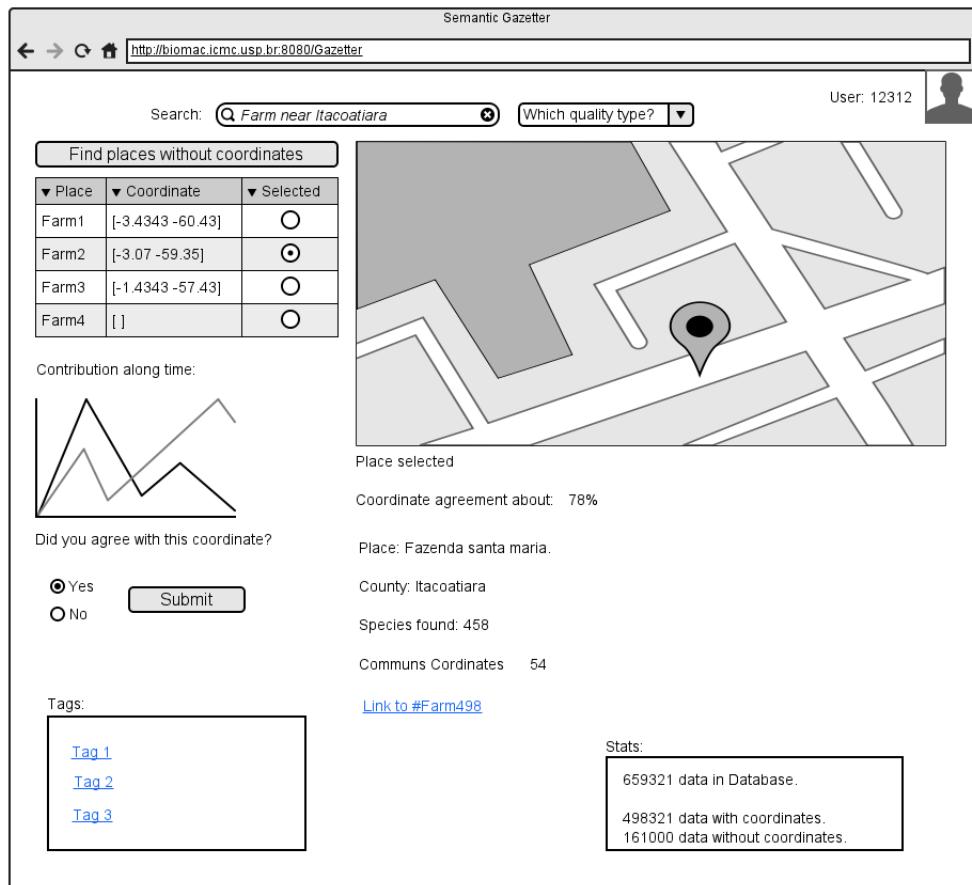


Figura 6.15.: Tela referente a função de validação das coordenadas do *Gazetteer*.

aquela informação e a quantidade de pessoas que concordam com essa conexão para essa fonte de *Linked Open Data*. Por fim, na parte inferior é exibido qual o tipo de geometria de um local.

Essa funcionalidade é importante para representar a qualidade das conexões de *Linked Open Data* que estão sendo disponibilizadas pelo *Gazetteer*, sendo esse um dos desafios listados por Moura und Davir Jr. (2013) para a nova geração de *Gazetteers*.

6.7. Arquitetura proposta para desenvolvimento do trabalho

Ao considerar os trabalhos recentes na área de RIG, que envolvem o desenvolvimento de *Gazetteers*, grande parte deles desenvolvem 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:

6. Experimentos

Semantic Gazetteer
<http://biomac.icmc.usp.br:8080/Gazetteer>

Place:

Place: Fazenda santa maria.
 County: Itacoatiara
 Species found: 458

[RDF](#) | [GML](#) | [JSON](#)

Contributors number:
 12 peoples

agreement about Linked Data: 0%

Contribution along time: Did you agree with this record?



Yes No

Point: Yes
 Point(-3.0711939334869385 -59.35672378540039)

Polygon: Yes
 Polygon not found

Dbpedia Link
[FarmDbpedia](#)
 Wikipedia:


Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus.

Figura 6.16.: Tela referente a função de validação dos dados de Linked Open Data.

apresentação (*Presentation*), negócios (*Businesses*) e dados (*Data*).

Para implementar as funcionalidades presente nessas camadas, grande parte dos *Gazetteers* atuais utilizam tecnologias que não possuem suporte a aplicação de web semântica para abordar os novos desafios da área de RIG e a prática de VIG.

Esta seção tem como objetivo apresentar a arquitetura proposta para o desenvolvimento deste *Gazetteer Colaborativo* (que usa VIG e Web Semântica). Até o momento, a implementação deste *Gazetteer* é a exibida na Figura 6.17. Ela é composta por um mecanismo de aprimoramento de dados que utiliza os dados de acervos biológicos disponibilizados por repositórios como, por exemplo, o SpeciesLink e GBIF. Esse mecanismo permite desambiguar localidades, remover dados de baixa qualidade e aprimorar localidades.

Em conjunto com esse mecanismo, é utilizada uma ontologia que permite mapear os dados para o formato RDF e assim armazená-los numa *triplestore*. Essa *triplestore*, além de armazenar os dados, tem como objetivo disponibilizar um endpoint GeoSPARQL que

6.7. Arquitetura proposta para desenvolvimento do trabalho

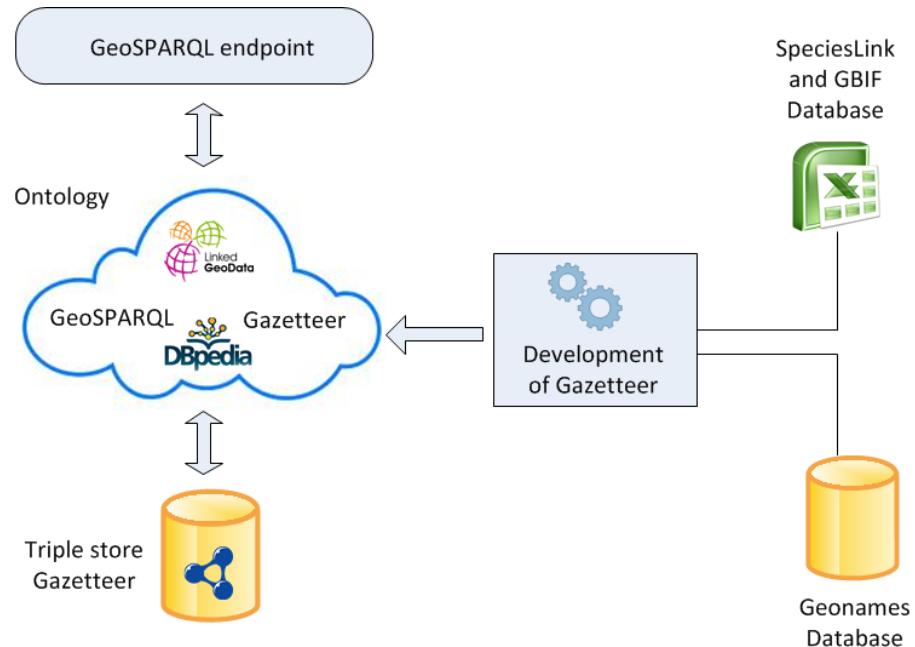


Figura 6.17.: Atual arquitetura do *Gazetteer*.

permite a agentes de software e outros programas realizarem buscas semânticas. No entanto, para dar suporte a prática de VIG é necessário aprimorar a estrutura utilizada.

Verificando essa necessidade e utilizando relatos já listados na literatura por outros trabalhos, a arquitetura do *Gazetteer Colaborativo*, apresentada na Figura 6.18, é sugerida. Essa arquitetura possui três camadas: apresentação, negócios e dados. 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 entre usuários por meio de mapas digitais, descritos na seção Section 6.6, e agentes de software por meio de *endpoints* GeoSPARQL. Desta forma, é possível a usuários e agentes de software incluir ou buscar informações no *Gazetteer*.

Para inserir localidades, informações tais como, tag, localidade, usuário, data e coordenadas geográficas. Essas informações são enviadas a camada de negócios para serem tratadas, validadas e aprimoradas. Isso também acontece quando alguma busca é enviada ao *Gazetteer*, as informações referentes a *strings* de busca são enviadas para o componente de reformulação de consultas na camada de negócios.

A camada de negócios, ao receber tais dados, é responsável por diversas tarefas que são realizadas por seus principais componentes, explicados a seguir:

6. Experimentos

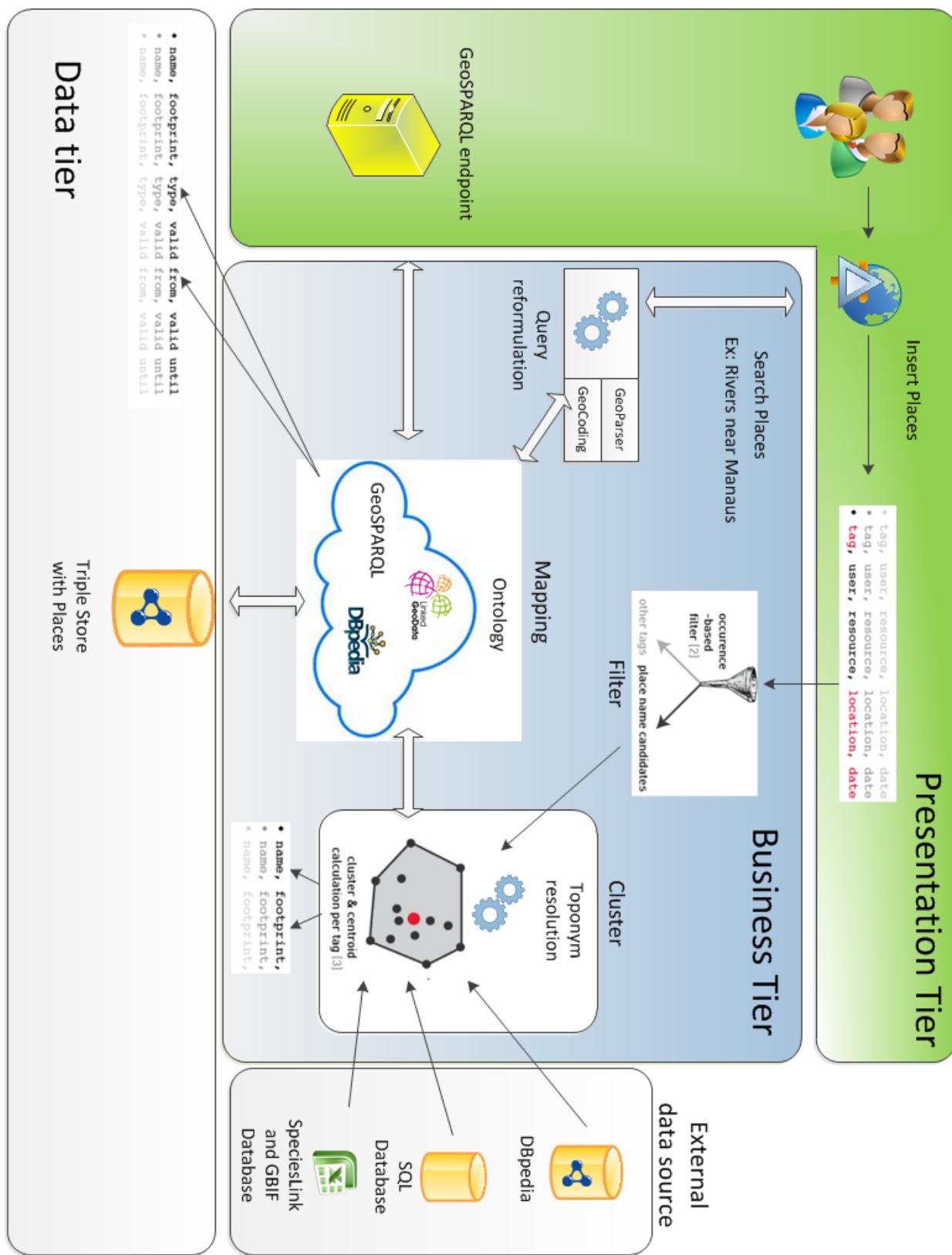


Figura 6.18.: Arquitetura proposta para o *Gazetteer*, viabilizando o uso de VGI e padrões de Linked Open Data.

6.7. Arquitetura proposta para desenvolvimento do trabalho

- i. **Módulo para filtrar informações:** Este módulo é responsável por verificar se as informações inseridas pelo usuário são válidas, por exemplo, se a localidade informada se encontra dentro dos limites administrativos do estado do Amazonas. Após realizar a filtragem dos dados, aqueles que são válidos são repassados para o módulo de agrupamento e aprimoramento de registros
- ii. **Módulo para agrupar e aprimorar os registros:** Ao se obter os dados já validados, este módulo é responsável por realizar o agrupamento dos dados, verificando se existem outras localidades que podem ser aprimoradas com a informação inserida pelo usuário. Além disso, esse modulo tem a tarefa de resolução de topônimos e leitura de dados de fontes externas, onde os mesmos processos de aprimoramento e agrupamento são realizados com os dados já disponíveis no *Gazetteer*. Após todo esse processo, os dados são enviados para o módulo de mapeamento de dados em RDF.
- iii. **Módulo para mapear os dados em RDF:** Este módulo tem como objetivo transformar os dados que foram agrupados e aprimorados pelo módulo anterior em dados no formato RDF. Para isso, as informações referentes a ontologia são consultadas e os registros sobre coordenadas geográficas, localidades, data de do registro, entre outros, são estruturados em um formato semântico para assim serem enviados para a camada de dados.
- iv. **Módulo de reformulação de consultas:** Este módulo tem como função receber as informações referentes as consultas do usuário que foram transmitidas pela camada de apresentação e reformular as mesmas num 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 modulo se comunica com a ontologia com objetivo de mapear as informações para uma consulta semântica e então enviar a query para a *triplesstore* na camada de dados.

A camada de dados, por sua vez, tem como função armazenar os dados do *Gazetteer* em uma *triplesstore*. Além disso, é também possível fornecer a camada de negócio dados adicionais referentes a modificações temporais, isto é, informações sobre localidades que tiveram seus nomes ou coordenadas geográficas alteradas ao longo do tempo, como, por exemplo, um município que foi dividido.

6. Experimentos

Informações geográficas referentes aos locais “candidatos,” que forem informados pelos usuários para alterar algum registro no *Gazetteer*, ou seja, os locais que tiveram uma baixa porcentagem de aceitação descritos na seção Section 6.6, serão armazenadas num outro grafo RDF auxiliar. Desse modo, quando um número suficiente de registros forem inseridos por usuários comuns ou biólogos, uma lista de candidatos será gerada e então armazenada no grafo principal da *triplestore*.

A implementação da arquitetura proposta neste trabalho demanda muito esforço. Diversas implementações poderiam ser realizadas de maneira coletiva e simultaneamente, e o resultado de tais implementações, ou seja, os módulos, poderiam ser conectados futuramente. Por isso, o presente trabalho irá se limitar a um determinado escopo de forma a tornar factível a verificação de alguns resultados num tempo hábil com o cronograma descrito no capítulo Chapter 7.

6.8. Considerações Finais

Neste capítulo, foram descritos os resultados obtidos com o desenvolvimento deste trabalho até o momento. Inicialmente, uma análise dos dados dos repositórios SpeciesLink e GBIF foi realizada, nessa análise foi demonstrado como os dados geográficos sobre biodiversidade são imprecisos. Além disso, foi descrito o método para aprimorar essas coordenadas imprecisas, por meio de um agrupamento de dados e sumarização de coordenadas. Os resultados obtidos por essa técnica se mostraram bem promissores e podem auxiliar na recuperação de registros biológicos.

Outra contribuição do trabalho, até o momento, é a disponibilização dos dados do *Gazetteer* na Web of Data. Isso foi possível graças ao mapeamento de seus dados para formatos utilizados na web semântica, como o RDF, e a disponibilização de serviços SPARQL utilizando uma *triplestore*. Além disso, foi realizada uma análise das principais *triplestore* open source para se trabalhar com funções geográficas, por meio de GeoSPARQL. Foram evidenciados a performance e os principais pontos a serem abordados para o aprimoramento desses sistemas.

Com intuito de aprimorar o *Gazetteer* deste trabalho, um mokup de interface web foi proposto para que usuários possam inserir e consultar dados, contribuindo assim para com a prática de VGI. Futuramente, essa interface será avaliada em conjunto com biólogos do INPA e desenvolvida para se integrar ao *Gazetteer*.

Por fim, foi descrita uma proposta de arquitetura para o desenvolvimento da versão final deste *Gazetteer*, evidenciando os módulos e funcionalidades que devem ser abordadas para utilização de VGI e *Linked Open Data*.

7. Plano de Trabalho

7.1. Metodologia

Nesta seção será descrita a metodologia a ser seguida para o desenvolvimento deste trabalho e construção do *Gazetteer Colaborativo* proposto que será baseado nas seguintes etapas:

1. Recuperação dos dados dos repositórios SpeciesLink e GBIF referente aos locais que foram realizadas coletas de espécimes
 - a) Análise dos dados de baixa qualidade, ou seja, dados que não tem informações importantes, por exemplo, localidade e município
 - b) Verificação do número de dados imprecisos e exibição dos mesmos em um mapa
 - c) Verificação do número de dados que contém informações sobre local, latitude, longitude
 - d) Verificação da quantidade de dados que possuem informações de latitude e longitude antes e depois do uso de GPS por biólogos em coletas
2. Utilização de dados de fontes externas como Geonames, Wikimapia, DBpedia
3. Implementação de um método para agrupar todos os dados dos repositórios SpeciesLink e GBIF e realizar a resolução de topônimos utilizando técnicas de Recuperação de Informação Geográfica e ontologias
4. Aprimoramento das informações geográficas ausentes nos dados do SpeciesLink e GBIF
 - a) Utilização dos dados referente aos repositórios externos para melhorar os registros das localidades dos repositórios SpeciesLink e GBIF
 - b) Criação um método para sumarizar as coordenadas geográficas de acordo com a abordagem da Lei de Linus

7. Plano de Trabalho

5. Verificação da quantidade de dados que tiveram suas informações aprimoradas
 - a) Contagem dos dados que não contém informações sobre latitude, longitude e foram recuperados
 - b) Contagem dos dados que possui informações geográficas recuperadas e eram muito velhos, ou seja, antes do uso de GPS por biólogos em coletas
 - c) Análise dos resultados dos passos a) e b) anteriores utilizando o teste t de Student, para verificar o quanto boa foi a abordagem utilizada
6. Mapeamento dos dados para uma *triples store* utilizando ontologias
 - a) Definição das triplas em RDF, que deveram possuir um sujeito, predicado e objeto para cada localidade.
 - b) Mapeamento de coordenadas geográficas para GeoSPARQL
7. Construção de uma base de teste com informações sobre qual consulta representa uma localidade
 - a) Construção de consultas semânticas e verificação de resultados utilizando as medidas de precisão, revocação e F1
 - b) Realização da mesma abordagem do passo a) para os repositórios DBpedia e Geonames, com intuito de verificar a viabilidade para expansão de consultas
8. Desenvolvimento de uma interface que permita aos biólogos inserir dados no *Gazetteer* por meio de mapas
 - a) Desenvolvimento de um método que permita aos biólogos inserir lugares usando GeoTAGS
 - b) Agrupamento dos dados inseridos e aprimoramento dos dados referentes a localidades que são similares.
 - c) Desenvolvimento de um método que permita aos biólogos inserirem links do DBpedia, quando os mesmos existirem, para auxiliar no crescimento da *Web of Data*.
 - d) Desenvolvimento de um método que permita aos biólogos verificar a acurácia dos links sobre a DBpedia inseridos no *Gazetteer*
9. Verificação do número de lugares inseridos pelos usuários e qualidade dos dados

7.2. Atividades previstas e Cronograma

- a) Verificação da média de usuários que concordam com as coordenadas geográficas
- b) Verificação da média de usuários que concordam com as informações de *Linked Open Data* presentes no *Gazetteer*
- c) Verificação da quantidade de dados que tiveram suas informações aprimoradas

7.2. Atividades previstas e Cronograma

As seguintes atividades estão previstas, com início em Agosto de 2013 e duração de 24 meses, conforme o cronograma a seguir:

A1 - Obtenção de créditos referente as disciplinas do programa de mestrado.

A2 - Levantamento bibliográfico sobre a área de pesquisa.

A3 - Exame de proficiência na língua inglesa.

A4 - Análise das informações do Banco de dados do SpeciesLink e GBIF (etapa 1)

A5 - Estudo das técnicas envolvidas para criação do *Gazetteer*. (etapa 3)

A6 - Estudos aprofundados, definição detalhada da proposta, protótipos e testes preliminares. (etapas 1, 2, 3, 4, 5 e 6)

A7 - Qualificação: redação da monografia de qualificação, incluindo detalhamento do projeto.

A8 - Exame de Qualificação (apresentação da proposta à Comissão Examinadora).

A9 - Implementação do protótipo do *Gazetteer*. (etapas 5, 6, 7, 8 e 9)

A10 - Testes preliminares, refinamento das funções do *Gazetteer*, ajustes. (etapas 1, 2, 3, 4, 5, 6, 7, 8 e 9)

A11 - Testes e validação: estudos de caso e refinamento. (etapas 5, 7, 8 e 9)

A12 - Redação da dissertação.

A13 - Redação e submissão de artigos com os resultados obtidos.

7. Plano de Trabalho

A14 - Defesa.

A Tabela Table 7.1 on page 96 apresenta o cronograma de execução deste projeto de mestrado.

Ativ.	2013		2014				2015		
	3 Tri	4 Tri	1 Tri	2 Tri	3 Tri	4 Tri	1 Tri	2 Tri	3 Tri
A1					
A2					
A3		.							
A4			..						
A5				..					
A6							
A7							
A8						.			
A9							
A10							..		
A11							.	..	
A12							.	..	
A13	
A14									.

Tabela 7.1.: Cronograma.

7.3. Atividades concluídas até o momento

Quanto a metodologia proposta para desenvolvimento do Gazetteer, os passos 1 ao 7 já foram concluídos, necessitando apenas alguns ajustes e integração das novas funcionalidades que serão implementadas nos passos 8 e 9. No cronograma, todas as atividades de A1 a A6 foram concluídas. Além disso, a redação e submissão de artigos com os resultados obtidos, já vem sendo realizada desde o início do projeto.

7.4. Produções Científicas até o momento

1. CARDOSO S. ; SERIQUE, K. J. ; AMANQUI, F. M. ; SANTOS, J. L. ; MOREIRA, D. A. A Gazetteer for Biodiversity Data as a Linked Open Data Solution. In: Web2Touch 2014 - Modelling the Collaborative Web Knowledge, Italy June 2014. *Qualis CAPES B1*.

7.5. Dificuldades e Limitações

2. AMANQUI, F. M. ; SERIQUE, K. J. ; CARDOSO S. ; ALBUQUERQUE, A. ; SANTOS, J. L. ; MOREIRA, D. A. Improving Biodiversity Data Retrieval through Semantic Search and Ontologies. In: The 2014 IEEE/WIC/ACM International Conference on Web Intelligence, Poland 2014 (Artigo Aceito). *Qualis CAPES A2.*

7.5. Dificuldades e Limitações

Até o presente momento, foi evidenciado como dificuldade para desenvolvimento do projeto, a busca por uma *triplestore* que ofereça a manipulação tanto de funções GeoS-PARQL, para realizar buscas semânticas geográficas, quanto das funcionalidades relacionadas a OWL. A *triplestore* Parliament realiza ambas funções, mas, devido a problemas de performance, o tempo demasiado longo para se realizar buscas com ela impede a sua utilização no projeto.

8. Referências

- Ahlers, Dirk (2013): Assessment of the Accuracy of GeoNames Gazetteer Data. In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*. ACM, New York, NY, USA, GIR '13, S. 74–81.
- Alho, CJR. (11 2008): The value of biodiversity. *Brazilian Journal of Biology*, 68:1115 – 1118.
- Amanqui, Flor K.; Kleberson J. Serique; Franco Lamping; Andrea C. F. Albuquerque; José L. C. Dos Santos und Dilvan A. Moreira (2013a): Semantic Search Architecture for Retrieving Information in Biodiversity Repositories. In: *ONTOBRAS*, Hg. Marcello Peixoto Bax; Mauricio Barcellos Almeida und Renata Wassermann. CEUR-WS.org, Bd. 1041 von *CEUR Workshop Proceedings*, S. 83–93.
- Amanqui, Flor Karina Mamani (2014): *Uma arquitetura para sistemas de busca semântica para recuperação de informações em repositórios de biodiversidade*. Mestrado.
- Aslam, Javed A.; Ekaterina Pelekhou und Daniela Rus (2004a): The Star Clustering Algorithm for Static and Dynamic Information Organization. *J. Graph Algorithms Appl.*, 8:95–129.
- Auer, Sören; Jens Lehmann und Sebastian Hellmann (2009a): LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: *Proceedings of the 8th International Semantic Web Conference*. Springer-Verlag, Berlin, Heidelberg, ISWC '09, S. 731–746.
- Auer, Sören; Jens Lehmann und Sebastian Hellmann (2009b): LinkedGeoData - Adding a Spatial Dimension to the Web of Data. In: *Proc. of 7th International Semantic Web Conference (ISWC)*.
- Battle, Robert und Dave Kolas (2012a): Enabling the geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web*, 3(4):355–370.
- Beard, Kate (2012): A Semantic Web Based Gazetteer Model for VGI. In: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. ACM, New York, NY, USA, GEOCROWD '12, S. 54–61.

8. Referências

- Bechhofer, Sean; Frank van Harmelen; Jim Hendler; Ian Horrocks; Deborah L. McGuinness; Peter F. Patel-Schneider und Lynn Andrea Stein (February 2004): *OWL Web Ontology Language Reference*. Techn. Ber., W3C, <http://www.w3.org/TR/owl-ref/>.
- Bereta, Konstantina; Panayiotis Smeros und Manolis Koubarakis (2013): Representation and Querying of Valid Time of Triples in Linked Geospatial Data. In: *ESWC*, Hg. Philipp Cimiano; Óscar Corcho; Valentina Presutti; Laura Hollink und Sebastian Rudolph. Springer, Bd. 7882 von *Lecture Notes in Computer Science*, S. 259–274.
- Berners-Lee, Tim; Yuhsin Chen; Lydia Chilton; Dan Connolly; Ruth Dhanaraj; James Hollenbach; Adam Lerer und David Sheets (2006): Tabulator: Exploring and Analyzing linked data on the Semantic Web. In: *Proceedings of the 3rd International Semantic Web User Interaction*.
- Berners-Lee, Tim; James Hendler und Ora Lassila (Mai 2001): The Semantic Web. *Scientific American*, 284(5):34–43.
- Bisby, Frank A. (Sept. 2000): The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*, 289(5488):2309–2312.
- Boley, S. Tabet H. und G. Wagner (2001): Design Rationale of RuleML: A Markup Language for Semantic Web Rules. In: *Proc. Semantic Web Working Symposium*, Hg. I. F. Cruz; S. Decker; J. Euzenat und D. L. McGuinness. Stanford University, California, S. 381–402.
- Borges, Karla A. V.; Alberto H. F. Laender; Claudia Bauzer Medeiros und Clodoveu A. Davis (2007): Discovering geographic locations in web pages using urban addresses. In: *GIR*, Hg. Ross Purves und Chris Jones. ACM, S. 31–36.
- Buyukkokten, O.; J. Cho; H. Garcia-Molina; L. Gravano und N. Shivakumar (1999): Exploiting Geographical Location Information of Web Pages. In: *ACM SIGMOD Workshop on The Web and Databases (WebDB'99)*.
- Cardoso, D. S.; J. K. Serique; K. F. Amanqui; L. J. Campos dos Santos und A. D. Moreira (Jun 2014): A Gazetteer for Biodiversity Data as a Linked Open Data Solution. Web2Touch 2014 - Modelling the Collaborative Web Knowledge.
- Carroll, J.; I. Dickinson; C. Dollin; D. Reynolds; A. Seaborne und K. Wilkinson (2003): Jena: Implementing the semantic web recommendations.
- Cooper, Robert und Charles Collins (2008): *GWT in Practice*. Manning Publications.

- Cornell (2008): eBird. online Lab of Ornithology and National Audubon Society.
- Davis, Clodoveu A.; Frederico T. Fonseca und Karla A. V. Borges (2003): A Flexible Addressing System for Approximate Geocoding. In: *GeoInfo*.
- Dentler, Kathrin; Ronald Cornet; Annette ten Teije und Nicolette de Keizer (Apr. 2011): Comparison of Reasoners for Large Ontologies in the OWL 2 EL Profile. *Semant. web*, 2(2):71–87.
- Egenhofer, Max J. (2002): Toward the semantic geospatial web. In: *ACM-GIS*, Hg. Agnès Voisard und Shu-Ching Chen. ACM, S. 1–4.
- Farias, E. M. B.; Campos dos Santos J. L. und M. T. b. Geller (2010): Proposta para Integração de Geo-dados Biológicos um Gazetteer Colaborativo. In: *II Jornada Científica de Computação (JCC 2010)*. Anais da II Jornada Científica de Computação.
- Garbin, Eric und Inderjeet Mani (2005): Disambiguating toponyms in news. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, S. 363–370.
- Garbis, George; Kostis Kyzirakos und Manolis Koubarakis (2013): Geographica: A Benchmark for Geospatial RDF Stores. *CoRR*, abs/1305.5653.
- GBIF (2014): Data Portal of the Global Biodiversity Information Facility (GBIF). In: *online*.
- Gey, Fredric; Ray Larson; Mark Sanderson; Hideo Joho; Paul Clough und Vivien Petras (2006): GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. S. 908–919.
- Gil, Fabiana B.; Nádia P. Koziavitch und Ricardo da Silva Torres (2010): A Geographic Annotation Service for Biodiversity Systems. In: *GeoInfo*, Hg. Vania Bogorny und Lúbia Vinhas. MCT/INPE, S. 33–44.
- Giles, Lee C.; Kurt Bollacker und Steve Lawrence (June FebruaryMarch–FebruaryJune 1998): CiteSeer: An Automatic Citation Indexing System. In: *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, Hg. Ian Witten; Rob Akscyn und Frank M. Shipman. ACM Press, Pittsburgh, PA, S. 89–98.

8. Referências

- Gimenez, Paulo J. A.; Asterio K. Tanaka und Fernanda Baiao (2013): A geo-ontology to support the semantic integration of geoinformation from the National Spatial Data Infrastructure. In: *GeoInfo*. MCT/INPE, S. 103–114.
- Giunchiglia, Fausto; Biswanath Dutta; Vincenzo Maltese und Feroz Farazi (2012): A Facet-Based Methodology for the Construction of a Large-Scale Geospatial Ontology. *J. Data Semantics*, 1(1):57–73.
- Goodchild, Michael F. und Linda L. Hill (2008): Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044.
- Gouvea, Cleber; Stanley Loh; Luís Fernando Fortes Garcia; Evandro Brasil da Fonseca und Igor Wendt (2008): Discovering Location Indicators of Toponyms from News to Improve Gazetteer-Based Geo-Referencing. In: *GeoInfo*, Hg. Marcelo Tílio Monteiro de Carvalho; Marco A. Casanova; Marcelo Gattass und Lúbia Vinhas. INPE, S. 51–62.
- Gouvêa, Cleber (2009): *Uma Abordagem para o Enriquecimento de Gazetteers a partir de Notícias visando o Georreferenciamento de Textos na Web*. Mestrado.
- Gruber, Tom (2005): Ontology of Folksonomy: A Mash-up of Apples and Oranges.
- Guizzardi, Giancarlo (2010): Theoretical Foundations and Engineering Tools for Building Ontologies as Reference Conceptual Models. *Semantic Web*, 1(1-2):3–10.
- Haklay, Mordechai M. (2008): How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England.
- Haklay, Mordechai M.; Sofia Basiouka; Vyron Antoniou und Aamer Ather (Nov. 2010): How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus Law to Volunteered Geographic Information. *Cartographic Journal, The*, S. 315–322.
- Heath, Tom und Christian Bizer (2011): *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1. Aufl.
- Hill, Linda L. (2000): Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: *ECDL*, Hg. José Luis Borbinha und Thomas Baker. Springer, Bd. 1923 von *Lecture Notes in Computer Science*, S. 280–290.
- Jain, Prateek; Pascal Hitzler; Peter Z. Yeh; Kunal Verma und Amit P. Sheth (2010): A.P.: Linked Data is Merely More Data. In: *In: AAAI Spring Symposium ŠLinked Data Meets Artificial IntelligenceŠ, AAAI*. Press, S. 82–86.

- Jones, Christopher B.; Harith Alani und Douglas Tudhope (2001): Geographical Information Retrieval with Ontologies of Place. In: *COSIT*, Hg. Daniel R. Montello. Springer, Bd. 2205 von *Lecture Notes in Computer Science*, S. 322–335.
- Jones, Christopher B.; R. Purves; A. Ruas; M. Sanderson; M. Sester; M. van Kreveld und R. Weibel (2002): 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*. ACM, New York, NY, USA, SIGIR '02, S. 387–388.
- Jones, Christopher B. und Ross S. Purves (2008): Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.
- Jr., Clodoveu A. Davis; Hugo de Souza Vellozo und Michele Brito Pinheiro (2013a): A Framework for Web and Mobile Volunteered Geographic Information Applications. In: *GeoInfo*. MCT/INPE, S. 147–157.
- Jr., Clodoveu A. Davis; Hugo de Souza Vellozo und Michele Brito Pinheiro (2013b): A Framework for Web and Mobile Volunteered Geographic Information Applications. In: *GeoInfo*. MCT/INPE, S. 147–157.
- Kessler, Carsten; Krzysztof Janowicz und Mohamed Bishr (2009): An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, USA, GIS '09, S. 91–100.
- KeSSLer, Carsten; Krzysztof Janowicz und Tomi Kauppinen (2012a): spatial@linkedscience - Exploring the Research Field of GIScience with Linked Data. In: *GIScience*, Hg. Ningchuan Xiao; Mei-Po Kwan; Michael F. Goodchild und Shashi Shekhar. Springer, Bd. 7478 von *Lecture Notes in Computer Science*, S. 102–115.
- KeSSLer, Carsten; Patrick Maué; Jan Torben Heuer und Thomas Bartoschek (2009): Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. In: *GeoS*, Hg. Krzysztof Janowicz; Martin Raubal und Sergei Levashkin. Springer, Bd. 5892 von *Lecture Notes in Computer Science*, S. 83–102.
- Klinkenberg, Brian (2013): 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.

8. Referências

- Kolas, Dave; Ian Emmons und Mike Dean (2009): Efficient Linked-List RDF Indexing in Parliament.
- Koubarakis, Manolis; Manos Karpathiotakis; Kostis Kyzirakos; Charalampos Nikolaou und Michael Sioutis (2012): Data Models and Query Languages for Linked Geospatial Data. In: *Reasoning Web*, Hg. Thomas Eiter und Thomas Krennwallner. Springer, Bd. 7487 von *Lecture Notes in Computer Science*, S. 290–328.
- Larson, Ray R. (Apr. 1996): Geographic Information Retrieval and Spatial Browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, S. 81–124.
- Larson, Ray R. und Patricia Frontiera (2004): Geographic information retrieval (GIR): searching where and what. In: *SIGIR*, Hg. Mark Sanderson; Kalervo Järvelin; James Allan und Peter Bruza. ACM, S. 600.
- Leidner, J. (2004): Towards a reference corpus for automatic toponym resolution evaluation.
- Lemes, P; Faleiro Famv; G Tessarolo und R D Loyola (2011): Refinando dados espaciais para a conservação da biodiversidade. *Natureza & Conservação*, 9:240–243.
- Leveling, Johannes und Sven Hartrumpf (2007): 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.
- Liu, Yu; Runqiang Li; Kaichen Chen; Yihong Yuan; Lingli Huang und Hao Yu (Aug 2009): KIDGS: A geographical knowledge-informed digital gazetteer service. In: *Geoinformatics, 2009 17th International Conference on*. S. 1–6.
- Machado, Ivre Marjorie; Rafael Odon de Alencar; Roberto de Oliveira Campos Junior und Clodoveu A. Davis (2011): An ontological gazetteer and its application for place name disambiguation in text. *J. Braz. Comp. Soc.*, 17(4):267–279.
- Manguinhas, H.; B. Martins und J. Borbinha (Nov 2009): A geo-temporal Web gazetteer integrating data from multiple sources. In: *Digital Information Management, 2009. ICDIM 2009. Third International Conference on*. S. 146–153.
- McDonald, David D. (1996): Corpus Processing for Lexical Acquisition. MIT Press, Cambridge, MA, USA, Kap. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names, S. 21–39.

- Metzger, Jean-paul und Lilian Casatti (2006): Do diagnóstico à conservação da biodiversidade: o estado da arte do programa BIOTA/FAPESP. *Biota Neotropica*, 6(2):1–23.
- Moura, Tiago Henrique V. M. und Clodoveu A. Davis Jr. (2012): Expansão do conteúdo de um gazetteer: nomes hidrográficos. In: *Anais...*, Hg. Laércio Massaru Namikawa und Vania Bogorny. Simpósio Brasileiro de Geoinformática, 13. (GEOINFO), Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, S. 78–83.
- Moura, Tiago Henrique V. M. de und Clodoveu A. Davir Jr. (2013): Linked Geospatial Data: desafios e oportunidades de pesquisa. In: *Anais...*, Hg. Pedro Ribeiro Andrade und André Santanchè. Simpósio Brasileiro de Geoinformática, 14. (GEOINFO), Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, S. 6.
- Noy, Natalya F. und Deborah L. McGuinness (2001): Ontology Development 101: A Guide to Creating Your First Ontology. Online.
- OGC, Open Geospatial Consortium (2011): *OpenGIS Implementation Standard for Geographic information - Simple feature access*. 1. Aufl.
- OGC, Open Geospatial Consortium (2012): GeoSPARQL - A Geographic Query Language for RDF Data. In: *OGC 11-052r4 Version: 1.0*.
- OpenLayers (2014): OpenLayers: Free Maps for the Web.
- Page, L.; S. Brin; R. Motwani und T. Winograd (1998): The PageRank citation ranking: Bringing order to the Web. In: *Proceedings of the 7th International World Wide Web Conference*. Brisbane, Australia, S. 161–172.
- Parundekar, Rahul; Craig A. Knoblock und José Luis Ambite (2010): Linking and Building Ontologies of Linked Data. In: *International Semantic Web Conference (1)*, Hg. Peter F. Patel-Schneider; Yue Pan; Pascal Hitzler; Peter Mika; Lei Zhang 0007; Jeff Z. Pan; Ian Horrocks und Birte Glimm. Springer, Bd. 6496 von *Lecture Notes in Computer Science*, S. 598–614.
- Peng, Xiaobo; Rongguo Chen; Changxiu Cheng und Xun Yan (June 2010a): A folksonomy-ontology-based digital gazetteer service. In: *Geoinformatics, 2010 18th International Conference on*. S. 1–6.
- Rubin, Daniel L.; Natalya F. Noy und Mark A. Musen (2007): Protégé: A Tool for Managing and Using Terminology in Radiology Applications.

8. Referências

- Salton, Gerard (1989): *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- dos Santos, José Laurindo Campos (2003): *A biodiversity information system in an open data/metadatabase architecture*. Dissertation, Enschede.
- dos Santos, Jose Laurindo Campos; Jose F. de Magalhaes Netto; Alberto Nogueira de Castro; Andrea C. F. Albuquerque; Edilson Ferneda; Luiza Alonso; Ricardo L. da C. Rocha und Daniel T. de Mendonca (2011): Ontologias para Interoperabilidade de Modelos e Sistemas de Informação de Biodiversidade. Iberoamerican Meeting of Ontological Research, Co-located with the 6th Iberoamerican Congress on Telematics, Gramado, Bd. 728.
- Serique, Kleberson Junio Amaral (2012): *Anotação de imagens radiológicas usando a web semântica para colaboração científica e clínica*. Mestrado.
- Shah, Shreeraj (2008): *Web 2.0 security : defending Ajax, RIA, and SOA*. Charles River Media Internet series. Charles River Media, Boston. Index.
- da Silva, Geiza Cristina und Tarcísio de Souza Lima (2004): *RDF e RDFS na Infra-estrutura de Suporte à Web Semântica*. Techn. Ber., UFJF, <http://www2.ic.uff.br/gsilva/slreic.pdf>.
- Smith, Barry; Michael Ashburner; Cornelius Rosse; Jonathan Bard; William Bug; Werner Ceusters; Louis J Goldberg; Karen Eilbeck; Amelia Ireland; Christopher J Mungall; Neocles Leontis; Philippe Rocca-Serra; Alan Ruttenberg; Susanna-Assunta Sansone; Richard H Scheuermann; Nigam Shah; Patricia L Whetzel und Suzanna Lewis (nov 2007): The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255.
- Smith, David A. und Gideon S. Mann (2003): Bootstrapping Toponym Classifiers. In: *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-NAACL-GEOREF '03, S. 45–49.
- SpeciesLink (2014): Sistema de Informação Distribuído para Coleções Biológicas: a Integração do Species Analyst e do SinBiota (FAPESP). In: *online*.
- uSeekM (2014): uSeekM Extensions for Semantic Databases.

Wang, Hong; Lin Li und Ping chao Song A (2008): DESIGN OF GEO-ONTOLOGY BASED ON CONCEPT LATTICE. In: *XXIst ISPRS Congress, Technical Commission II, v. XXXVII, part B2*. S. 715–720.

Yesson, Chris; Peter W. Brewer; Tim Sutton; Neil Caithness; Jaspreet S. Pahwa; Mikhaila Burgess; W. Alec Gray; Richard J. White; Andrew C. Jones; Frank A. Bisby und Alastair Culham (Nov. 2007): How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, 2(11):e1124+.

Yin, Yong und Kazuhiko Yasuda (2006): Similarity coefficient methods applied to the cell formation problem: A taxonomy and review. *International Journal of Production Economics*, 101(2):329–352.

A. Consultas utilizadas para testar as *triples* *stores*.

Algoritmo A.1 Consulta Não Topologica utilizando a função convexHull

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

Algoritmo A.2 Consulta Não Topologica utilizando a função buffer

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

A. Consultas utilizadas para testar as triplestores.

Algoritmo A.3 Consulta para verificar junções espaciais

```
PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
PREFIX geonames: <http://www.geonames.org/ontology#>
PREFIX lgd: <http://linkedgeodata.org/ontology/>
PREFIX geof:<http://www.opengis.net/def/function/geosparql/>
SELECT ?s1 ?o1 ?s2 ?o2
WHERE {
  GRAPH dataset:geonames {?s1 geonames:asWKT ?o1}
  GRAPH dataset:geonames {?s2 geonames:asWKT ?o2}
  FILTER(?s1 != ?s2).
  FILTER(geof:sfWithin(?o1, ?o2)).
}
```

Algoritmo A.4 Consulta para verificar junções espaciais

```
PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
PREFIX geonames: <http://www.geonames.org/ontology#>
PREFIX lgd: <http://linkedgeodata.org/ontology/>
PREFIX geof:<http://www.opengis.net/def/function/geosparql/>
SELECT ?s1 ?o1 ?s2 ?o2
WHERE {
  GRAPH dataset:geonames {?s1 geonames:asWKT ?o1}
  GRAPH dataset:geonames {?s2 geonames:asWKT ?o2}
  FILTER(?s1 != ?s2).
  FILTER(geof:sfIntersects(?o1, ?o2)).
}
```

Algoritmo A.5 Consulta para verificar seleções espaciais

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

Algoritmo A.6 Consulta para verificar seleções espaciais

```
PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX opengis: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX geonames: <http://www.geonames.org/ontology#>
SELECT (COUNT(?s1) AS ?NumOfTriples)
WHERE {
    GRAPH dataset:geonames {?s1 geonames:asWKT ?o1}
    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))).
```
