Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

QualiOSM: Uma Arquitetura para Melhorar a Qualidade de Dados no OpenStreetMap

Gabriel Franklin Braz de Medeiros

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientadora
Prof.ª Dr.ª Maristela Terto de Holanda

Brasília
2020

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

QualiOSM: Uma Arquitetura para Melhorar a Qualidade de Dados no OpenStreetMap

Gabriel Franklin Braz de Medeiros

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof.ª Dr.ª Maristela Terto de Holanda (Orientadora)
CIC/UnB

Prof. Dr. Angelo Roncalli Alencar Brayner
Universidade Federal do Ceará (UFC)

Prof.ª Dr.ª Lívia Castro Degrossi
Fundação Getúlio Vargas (FGV)

Prof. Dr. Bruno Luiggi Macchiavello Espinoza
Coordenador do Programa de Pós-graduação em Informática

Brasília, 03 de dezembro de 2020

Dedicatória

Dedico este trabalho a todos os meus familiares e amigos que estiveram presentes ao longo da realização deste trabalho de mestrado. Em especial, dedico àqueles que estiveram ao meu lado ao longo dos períodos mais difíceis de minha caminhada acadêmica.

Agradecimentos

Agradeço primeiramente a Deus, por ter me dado forças para chegar até aqui. Agradeço aos meus pais, por estarem sempre presentes em minha vida, me ajudando em todos os momentos. Agradeço ao Departamento de Ciência da Computação da Universidade de Brasília, por ter fornecido a infraestrutura necessária para a realização deste trabalho, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro. Agradeço também às professoras Maristela Terto de Holanda e Lívia Castro Degrossi, por me ajudarem e terem feito parte deste projeto.

Resumo

Em Sistemas de Informações Geográficas (SIG) e, mais especificamente, em Sistemas de Informações Geográficas Voluntárias (SIGV), os usuários participam ativamente dos processos de inclusão, alteração e exclusão de informações. Nesse contexto, a questão da qualidade dos dados nesses tipos de sistemas tornou-se um tópico central, uma vez que os usuários possuem diferentes níveis de conhecimento e experiência, sendo fundamental garantir que alguns parâmetros, denominados de dimensões da qualidade, sejam minimamente satisfeitos, tais como a acurácia, a consistência lógica, e a completude dos dados. Dessa maneira, esta dissertação apresenta a arquitetura QualiOSM, com o objetivo de contribuir para a melhoria da completude das informações de endereço associadas aos objetos dentro da ferramenta de mapeamento colaborativo OpenStreetMap (OSM). Para a realização deste trabalho, uma Revisão Sistemática da Literatura (RSL) foi elaborada de forma a fazer um levantamento dos principais artigos acadêmicos relacionados com o tema. Em seguida, procedeu-se à realização da parte prática do trabalho, com a elaboração da arquitetura QualiOSM, a qual apresenta o desenvolvimento de um adicionador automático de *tags* com o propósito de acrescentar informações de endereço faltantes aos objetos da plataforma OSM, utilizando-se para isso das ferramentas de geocodificação reversa Nominatim, CEP Aberto e a base de dados dos Correios. A ferramenta QualiOSM apresentou bons resultados para a melhoria da completude das informações de cidade, bairro e logradouro em objetos do OpenStreetMap, especialmente em cenários de grandes centros urbanos, onde o nível de mapeamento costuma ser melhor comparado a cenários em ambientes rurais ou periféricos. Em relação à *tag* de código postal, os melhores resultados foram obtidos ao se utilizar a ferramenta QualiOSM em conjunto com a base de dados dos Correios.

Palavras-chave: Dados Geográficos, Sistemas de Informações Geográficas, Dimensões da Qualidade, Mapeamento, OpenStreetMap

Abstract

In Geographic Information Systems (GIS) and, more specifically, in Volunteered Geographic Information (VGI), users actively participate in the processes of inclusion, editing and exclusion of information. In this context, the issue of data quality in these types of systems has become a central topic, since users have different levels of knowledge and experience, and it is essential to ensure that some parameters, called quality dimensions, are being minimally satisfied, such as accuracy, logical consistency, and completeness of information. Thus, this dissertation presents the QualiOSM architecture, with the objective of contributing to the improvement of the completeness of the address information associated with the objects within the collaborative mapping tool OpenStreetMap (OSM). For the accomplishment of this work, a Systematic Literature Review (RSL) was elaborated in order to make a survey of the main academic articles related to the theme. Then, the practical part of the work was carried out, with the development of the QualiOSM architecture, which presents the elaboration of an automatic adder of tags with the purpose to add missing address information to the objects of the OSM platform, using the reverse geocoding tools Nominatim, CEP Aberto and the database from Correios. The QualiOSM tool showed good results for improving the completeness of city, neighborhood and street information in OpenStreetMap objects, especially in scenarios of large urban centers, where the level of mapping is usually better compared to scenarios in rural or peripheral environments. Regarding the postal code information, the best results were obtained when using the QualiOSM tool in conjunction with the Correios database.

Keywords: Geographic Data, Geographic Information Systems, Quality Dimensions, Mapping, OpenStreetMap

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Um Sistema de Informações Geográficas (SIG) consiste em uma agregação de fatores, tais como hardware, software, dados, pessoas e instituições, que realizam a coleta, o armazenamento, a análise e a disseminação de informações relativas às áreas da superfície terrestre [19]. A partir dos anos 2000, com o advento da chamada Web 2.0, surgiram sistemas capazes de proporcionar uma interação cada vez maior com seus usuários, uma vez que permitiram que as informações geográficas pudessem ser adicionadas, editadas ou excluídas via formulários preenchidos dinamicamente ou diretamente dentro de um mapa. Esses tipos de sistema ficaram conhecidos como Sistemas de Informações Geográficas Voluntárias (SIGV) [28].

O termo SIGV foi desenvolvido com o propósito de descrever sistemas computacionais em que um grande número de usuários colaboradores estão engajados na criação ou edição das informações geográficas. Nas últimas décadas, os SIGV foram utilizados para inúmeras finalidades, seja para o monitoramento ambiental ou para o gerenciamento de desastres naturais, além de auxiliarem no mapeamento de regiões mais remotas do planeta, onde é mais díficil o acesso para o mapeamento [62].

Um exemplo bem sucedido de SIGV é a ferramenta de mapeamento colaborativo OpenStreetMap (OSM), utilizada neste trabalho como estudo de caso. Criada no ano de 2004 pelo estudante de computação Steve Coast, da University College London (UCL), a ideia inicial do projeto era realizar o mapeamento apenas da região do Reino Unido, mas logo despertou o interesse de pesquisadores de outros países [57]. Entretanto, os dados fornecidos por voluntários requerem uma atenção especial em relação à questão da qualidade. Uma das principais razões para tal fato acontecer é a grande heterogeneidade observada por parte dos usuários, uma vez que estes utilizam ferramentas e tecnologias distintas, e possuem diferentes níveis de detalhamento ou precisão [62].

Um dos grandes desafios de ferramentas colaborativas é que elas necessitam da participação dos usuários para a inclusão de novas informações à base de dados. Nesse contexto,

[54] apresenta a Regra do 1%, também chamada de Regra 90-9-1, responsável por descrever a chamada desigualdade de participação presente nesse tipo de ferramenta. Segundo essa regra, 90% dos usuários desses tipos de sistemas apenas consomem o serviço sem qualquer tipo de colaboração ativa, 9% colaboram de forma esporádica e somente 1% corresponde aos membros que efetivamente colaboram com o projeto, participando da criação efetiva de conteúdo. Assim, a manutenção da qualidade dos dados em sistemas de participação voluntária é um grande desafio, uma vez que uma pequena parcela de usuários é responsável pela criação de uma grande quantidade de dados. Dessa forma, nem sempre as informações são inclusas de forma correta ou completa à base de dados.

Uma vez que dentro da ferramenta OpenStreetMap os usuários participam ativamente dos processos de inclusão, alteração e exclusão dos dados, a questão da qualidade das informações geográficas está sempre presente, visto que é necessário verificar se os dados estão sendo inseridos ou modificados da forma correta. Dessa maneira, o objetivo deste trabalho é realizar o desenvolvimento da arquitetura QualiOSM, de forma a auxiliar na melhoria da completude das informações de endereço dentro da plataforma do OSM. A arquitetura QualiOSM foi aplicada na forma de um *plugin* para o editor de dados Java Open Street Map Editor (JOSM), editor de dados responsável pelo maior número de edições dos dados do OpenStreetMap[1].

1.1 Descrição do Problema

A ferramenta OpenStreetMap é um projeto de mapeamento colaborativo criado com o objetivo de gerar um mapa livre e editável do mundo, construído por voluntários e lançado com uma licença de conteúdo aberto[2]. Devido ao seu caráter colaborativo, é possível que um usuário acrescente informações erradas à base de dados ou efetue a exclusão de objetos indevidamente. Um outro fator que pode levar ao acréscimo de informações incorretas é a falta de conhecimento dos voluntários tanto em relação à ferramenta quanto ao território. Por causa da constante preocupação com a melhoria da qualidade dos dados nesses tipos de ferramenta, este trabalho desenvolve a arquitetura QualiOSM com a finalidade de melhorar a dimensão da qualidade da completude dentro da ferramenta OpenStreetMap, uma vez que ainda existem muitas informações incompletas dentro da plataforma do OSM.

Analisando estatísticas presentes no site TagInfo[3], sistema criado com o objetivo de encontrar e agregar informações sobre as *tags* do OSM, foi observado que dentre as cinco

[1]https://wiki.openstreetmap.org/wiki/Editor_usage_statsTables_and_figures [Acesso em janeiro de 2020].

[2]https://wiki.openstreetmap.org/wiki/Main_Page [Acesso em fevereiro de 2019].

[3]https://taginfo.openstreetmap.org/ [Acesso em outubro de 2020].

tags mais utilizadas para os pontos do OpenStreetMap, quatro são *tags* de endereço ("addr: house-number", "addr: street", "addr: city" e "addr: postcode"). Também foi possível observar que essas quatro tags estão entre as dez mais utilizadas tanto para linhas quanto para objetos do OpenStreetMap em geral. Além disso, a etiqueta de endereço mais utilizada, "addr: house-number", estava associada a mais de 51 milhões de pontos em 1º de março de 2020, correspondendo a mais de um terço do total de pontos contidos na plataforma OSM. Dessa forma, percebeu-se a importância de melhorar a completude das informações de endereço dentro da plataforma do OpenStreetMap.

1.2 Objetivos

O objetivo geral deste trabalho é a elaboração da arquitetura QualiOSM com o propósito de melhorar a dimensão da completude dos dados dentro da ferramenta OpenStreetMap, por meio da implementação de um adicionador de *tags* de endereço aos objetos da plataforma do OSM.

1.3 Objetivos Específicos

Os objetivos específicos são:

- Desenvolvimento da arquitetura QualiOSM na forma de um *plugin* para o editor de dados JOSM;
- Análise da completude de dados de endereço dentro da ferramenta OpenStreetMap, aplicando a arquitetura desenvolvida em diferentes cenários de teste;
- Comparação da acurácia e completude das ferramentas de geocodificação reversa Nominatim e CEP Aberto;
- Validação da ferramenta QualiOSM por meio de comparação com a base de dados dos Correios.

1.4 Estrutura do Trabalho

O restante deste documento está estruturado da seguinte maneira:

- O Capítulo 2 apresenta a fundamentação teórica necessária para o desenvolvimento desta pesquisa, tais como os conceitos de Bancos de Dados Espaciais, Sistemas de Informações Geográficas, Dimensões da Qualidade, além de trazer uma introdução sobre a ferramenta OpenStreetMap;

- O Capítulo 3 apresenta o estado da arte, sob a forma de uma Revisão Sistemática da Literatura;
- O Capítulo 4 apresenta o desenvolvimento da ferramenta QualiOSM, demonstrando a metodologia e a arquitetura utilizadas para a implementação;
- O Capítulo 5 apresenta os resultados obtidos após o desenvolvimento da ferramenta;
- O Capítulo 6 apresenta a conclusão e a apresentação dos trabalhos futuros.

Capítulo 2

Referencial Teórico

Este capítulo dedica-se a apresentar os principais conceitos teóricos envolvidos para a elaboração deste trabalho. O capítulo está dividido nas seguintes seções: A Seção 2.1 traz a definição de bancos de dados geoespaciais; A Seção 2.2 aborda o conceito de Sistemas de Informações Geográficas (SIG); A Seção 2.3 realiza uma introdução sobre a ferramenta OpenStreetMap; Por fim, a Seção 2.4 aborda o conceito das dimensões da qualidade.

2.1 Bancos de Dados Geoespaciais

Dados geoespaciais, também chamados de dados geograficamente referenciados ou dados geográficos, são dados que podem ser exibidos, manipulados e analisados por meio de um atributo espacial, que denota um local na superfície terrestre. Esse atributo espacial é normalmente fornecido na forma de pares de coordenadas geográficas, que permitem que a posição e o formato de um objeto sejam medidos e representados graficamente. Dados geoespaciais possuem duas importantes propriedades [69]:

- Fazem referência a um espaço geográfico, o que significa que os dados são registrados em um sistema de coordenadas geográficas abrangendo alguma área da superfície terrestre.
- Podem ser representados em uma variedade de escalas geográficas e quando representados em escalas muito pequenas podem ser generalizados ou sofrer aproximações.

Para a representação de objetos em um Banco de Dados Geográficos, frequentemente é utilizado o Espaço Euclidiano, ou seja, um espaço vetorial real de dimensão finita [29]. Isso significa que um ponto no plano será dado por um par de números reais. Por outro lado, os sistemas computacionais existentes trabalham com aproximações finitas e limitadas, o que dificulta o processo de modelagem desses tipos de bancos de dados e pode trazer

alguns problemas na representação. Por exemplo, o ponto de intersecção entre duas linhas pode ser arredondado para o ponto mais próximo representável, não correspondendo necessariamente às coordenadas geográficas do mundo real [42]. As subseções 2.1.1 a 2.1.3 detalham um pouco mais de que forma é realizada a modelagem de dados geográficos.

2.1.1 Modelagem de Dados Geoespaciais

Dados geográficos necessitam de uma modelagem própria para serem representados. Nesse sentido, existem dois diferentes aspectos que precisam estar contemplados em uma modelagem para bancos de dados geográficos [30]:

1. Objetos no espaço: Há o interesse de se representar entidades distintas organizadas no espaço, cada uma das quais com sua própria descrição geométrica;
2. Espaço: É necessário que o espaço propriamente dito também seja inserido na representação, colocando-se em um mapa cada ponto desse espaço.

O primeiro aspecto permite modelar objetos como cidades, florestas ou rios. Já o segundo aspecto é utilizado para a construção de mapas temáticos descrevendo o uso da terra ou a partição de uma cidade em bairros ou distritos. Em um Banco de Dados Geoespaciais, esses dois aspectos são conciliados a fim de proporcionar a modelagem de objetos individuais, e coleções de objetos relacionados espacialmente entre si. Dados geográficos podem ser representados em diferentes formatos, sendo que as principais representações são o formato vetorial e o formato matricial. Essas duas representações serão detalhadas nas subseções 2.1.2 e 2.1.3.

2.1.2 Representação Vetorial

Em geral, dados no formato vetorial são baseados em coordenadas geográficas, e isso quer dizer que cada objeto será representado por meio de um conjunto de coordenadas *(x,y)* ou *(x,y,z)* . Os objetos podem conter também outras informações adicionais por meio de seus diferentes atributos, armazenados no formato de *tags*.

Para a modelagem dos objetos individuais nesse tipo de representação, existe a noção de três abstrações fundamentais: o ponto, a linha e o polígono. Dessa forma, tem-se que [49]:

- Um ponto simboliza um objeto cuja localização é relevante, mas cuja área é desconsiderada na representação;
- Uma linha (neste contexto sempre entendida como uma curva no espaço, geralmente representada por uma sequência de segmentos de linha) é a abstração básica para ca-

minhos entre dois pontos. Geralmente, estradas, rios, cabos de telefonia e eletricidade são objetos espaciais modelados como linhas em um Banco de Dados Geoespaciais;

- Um polígono representa um objeto cuja área é relevante na representação, demarcando uma região bem definida. Essa região inclusive pode conter espaços vazios ou vários pedaços disjuntos.

Na representação vetorial, um polígono é definido por um conjunto de coordenadas geográficas, sendo que cada coordenada corresponderá a um vértice desse polígono em um espaço bidimensional ou tridimensional. Dessa forma, o polígono definido por P = <[4,4],[6,1],[3,0],[0,2],[2,2] > terá sua representação no espaço *xy* ilustrada pela Figura 2.1.

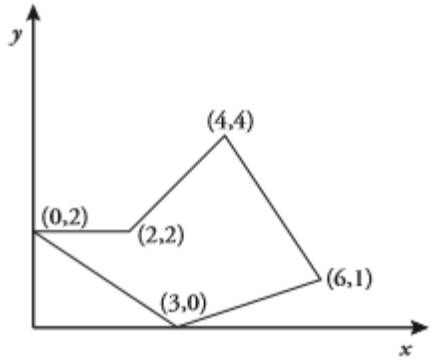


Figura 2.1: Representação Vetorial do Polígono P [58].

2.1.3 Representação Matricial

Dados no formato matricial (ou *raster*) são representados no formato de um *grid*, ou seja, um conjunto de linhas horizontais e verticais formando uma espécie de matriz. Geralmente, são dados provenientes de sensoriamento remoto ou imagens digitalizadas, produzidos por dispositivos eletrônicos ou satélites. Diferentemente dos dados vetoriais, os dados *raster* não contêm um registro em um banco de dados associado com cada célula, uma vez que os dados são geocodificados para cada *pixel* da resolução da imagem. Dessa maneira, é possível fazer a representação de paisagens com grandes variações de cores ou texturas, como ocorre em regiões de plantações, por exemplo [49]. Nesse caso, um polígono é definido não por um conjunto de coordenadas, mas sim por um conjunto de células do *grid*. Dessa forma, o polígono definido por P = < 5,12,13,14,17,18,19,20,21, 22,26,27,28, 29,30,31,35,36,37,38 > terá a sua representação no espaço *xy* ilustrada pela Figura 2.2.

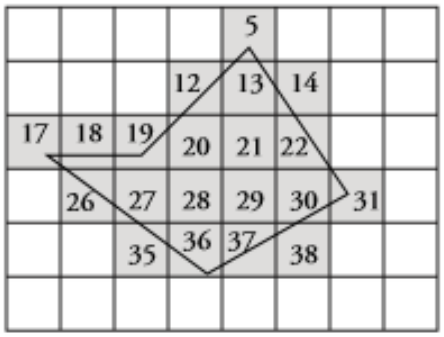


Figura 2.2: Representação Matricial do Polígono P [58].

2.2 Sistemas de Informações Geográficas

O termo Sistema de Informações Geográficas refere-se a um conjunto de ferramentas utilizadas para a captura, preparação, armazenamento, análise e apresentação de dados geográficos. Isso implica que um usuário de um SIG pode esperar que esse tipo de sistema acesse dados georeferenciados com a finalidade de que sejam analisados e visualizados na forma de um mapa [35]. Em geral, um SIG possui os seguintes componentes [30]:

- Interface com o usuário - Define como o sistema será operado e controlado;
- Entrada e integração de dados - Um SIG deve possuir mecanismos básicos de processamento de dados, o que envolve a parte de coleta e manipulação desses dados;
- Funções de processamento de gráficos e imagens - Uma vez que um SIG trabalha com elementos gráficos, é fundamental que seja capaz de manipular corretamente esses tipos de funções;
- Visualização e plotagem de dados em um mapa - A produção de mapas é uma atividade bastante recorrente em um SIG, uma vez que a representação visual dos dados é muito importante nesses tipos de sistemas;
- Armazenamento e recuperação dos dados por meio de um Sistema Gerenciador de Banco de Dados (SGBD) - Uma boa integração entre o SIG e um SGBD é fundamental a fim de facilitar o compartilhamento dos dados, recuperação através de mecanismos de *backup*, além de aumentar a disponibilidade e a integridade dos dados.

Os componentes de um Sistema de Informações Geográficas se relacionam entre si de acordo com uma arquitetura própria, conforme pode ser observado na Figura 2.3. Conforme pode ser observado, a arquitetura de um SIG é estruturada de uma forma hierárquica, apresentando três diferentes camadas:

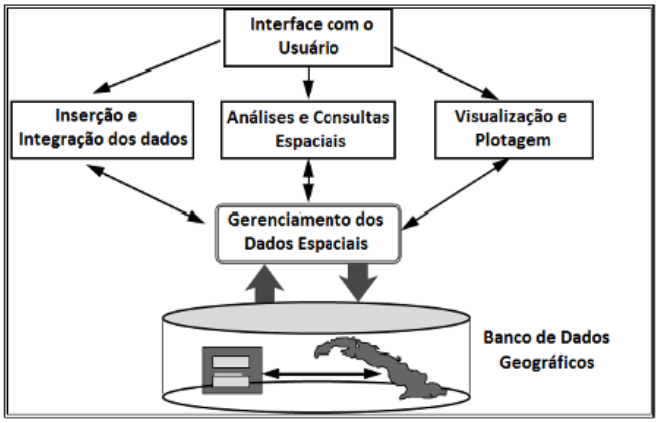


Figura 2.3: Arquitetura Padrão de um SIG [9].

1. Camada de Apresentação – Implementa a interface com o usuário do sistema, definindo como este será operado e controlado. Geralmente, os usuários acessam os dados de um SIG por meio de instruções específicas e os resultados podem ser visualizados através de uma ferramenta própria ou um módulo de visualização dentro do próprio SIG;
2. Camada de Aplicação – Provê a maior parte das funcionalidades de um SIG, incluindo as partes de inserção e integração dos dados, análises e consultas espaciais, além da visualização e plotagem dos dados na forma de mapas. Ela realiza a ligação entre as camadas de apresentação e de dados;
3. Camada de Dados – Realiza o gerenciamento dos dados espaciais através de um SGBD, o que envolve, portanto, operações de armazenamento e recuperação dos dados.

Portanto, a arquitetura de um SIG também pode ser descrita de forma mais simplificada por meio da Figura 2.4.

2.2.1 Sistemas de Informações Geográficas Voluntárias

O fenônemo dos Sistemas de Informações Geográficas Voluntárias faz parte de uma profunda transformação no modo como as informações geográficas estão sendo produzidas e distribuídas no mundo atualmente. As últimas décadas testemunharam uma profunda mudança na forma como os dados geográficos, informações e, mais amplamente, o conhecimento estão sendo produzidos e disseminados devido ao crescimento fenomenal de uma infinidade de tecnologias relacionadas, as quais ficaram popularmente conhecidas como Web 2.0. Embora diferentes conceitos tenham surgido para descrever essa nova tendência, a ideia geral se traduz no uso de ferramentas para criar, compartilhar e analisar infor-

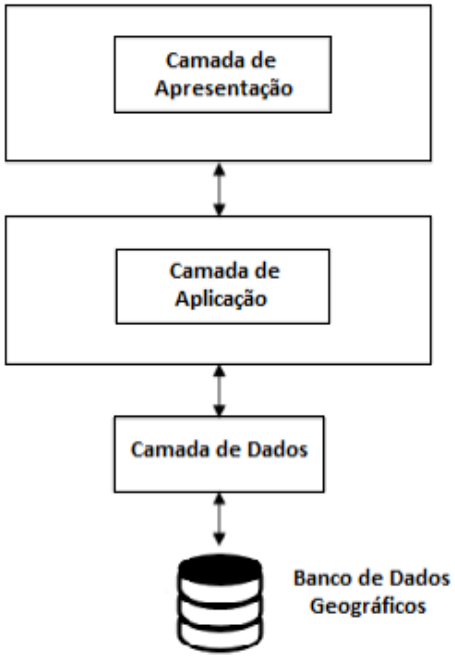


Figura 2.4: Camadas da Arquitetura de um SIG [38].

mações geográficas através de múltiplos usuários e múltiplos dispositivos/plataformas de computação [64].

Para entender o funcionamento de um Sistema de Informações Geográficas Voluntárias, é importante destacar as principais ferramentas computacionais que tornaram esse tipo de tecnologia possível. Dessa maneira, pode-se citar as seguintes tecnologias [28]:

- ***Web* 2.0:** No começo dos anos 2000, novos protocolos de *Internet* foram desenvolvidos, os quais permitiram com que usuários pudessem ter acesso a bancos de dados armazenados em servidores *web*, podendo inclusive fazer a alteração de registros através de formulários. Tal fato permitiu a existência de *sites* cujo conteúdo é quase totalmente inserido por meio de seus usuários, com pouca moderação ou controle, e quase nenhuma restrição em relação à natureza do conteúdo inserido;
- **Georeferenciamento:** Para possibilitar a criação de dados geográficos pelo público em geral, é necessário dispor de um conjunto de ferramentas facilmente disponíveis a fim de identificar coordenadas geográficas (latitude e longitude) na superfície terrestre. Várias ferramentas utilizam-se de imagens capturadas por satélites, para assim criarem registros digitais de ruas inteiras e, em seguida, essas imagens são carregadas e compiladas na forma de mapas digitais;
- ***Geotags* (Etiquetas Geográficas):** Uma *geotag* ou etiqueta geográfica é um código normalizado que pode ser inserido em uma informação geográfica a fim de que

possa ser melhor observada a sua localização. Muitas *geotags* foram utilizadas a fim de relacionar artigos da Wikipedia a objetos geográficos associados, por exemplo;

- ***Global Positioning System (GPS):*** O Sistema de Posicionamento Global é sem dúvida o primeiro sistema na história humana a permitir uma medição direta da posição de objetos na superfície terrestre. Os receptores GPS são fáceis de serem utilizados e fornecem uma estimativa instantânea da posição de um objeto, com acurácia muitas vezes melhor do que a precisão de 10 metros. Incorporado a sistemas de navegação, o GPS permite a localização de determinado veículo comparando-se ao conteúdo de um mapa digital. O GPS tem suscitado uma série de atividades envolvendo Sistemas de Informações Geográficas, destacando-se a criação de mapas digitais para pedestres, ciclistas ou veículos;
- **Recursos Gráficos:** Gráficos de alta qualidade constituem uma inovação relativamente recente na história da computação. A visualização dinâmica de objetos tridimensionais, como ocorre com o Google Earth[1] por exemplo, exige componentes de hardware com poderosos recursos gráficos, o que se tornou viável apenas no começo dos anos 2000;
- **Comunicação Banda Larga:** O desenvolvimento dos SIGV não poderia se tornar possível sem um bom acesso à *Internet*, preferencialmente via uma conexão de alta capacidade. Atualmente, existe um grande número de pessoas com *Internet* via conexão banda larga, cuja infraestrutura pode fazer uso de satélites, cabos ou ainda aproveitar a estrutura utilizada para telefonia.

2.2.2 Crowdsourcing

O termo *crowdsourcing* é utilizado para descrever um conjunto de técnicas e métodos computacionais que dependem de uma grande quantidade de usuários para a resolução de tarefas específicas [18]. A adaptabilidade das ferramentas de *crowdsourcing* permite que esta seja uma prática eficaz e poderosa, mas dificulta a sua definição. Dessa forma, [20] forneceu uma definição ampla de modo a compreender a maioria (se não todos) os processos de *crowdsourcing* existentes. Através da análise de várias definições, foram encontradas algumas características comuns a qualquer iniciativa de *crowdsourcing*, destacando-se os seguintes elementos: presença de um grande número de usuários (*crowd* = multidão); presença de tarefas a serem completadas; presença de mecanismos de recompensa aos usuários e a presença de um usuário iniciador da atividade de *crowdsourcing*, denominado *crowdsourcer*.

[1] https://earth.google.com/web/ [Acesso em fevereiro de 2018].

O recrutamento de usuários é um dos mais importantes desafios dos sistemas de *crowdsourcing*. Segundo [18], existem basicamente cinco possíveis soluções para esse problema. Primeiro, pode-se exigir que os usuários façam contribuições por meio do uso da autoridade (por exemplo, um gerente pode exigir que 100 funcionários ajudem a criar um sistema colaborativo para toda a empresa). Em segundo lugar, pode-se contratar usuários pagos para a realização de tarefas específicas. A ferramenta Mechanical Turk[2], por exemplo, fornece uma maneira de pagar usuários na Web para que estes realizem tarefas colaborativas. Terceiro, pode-se utilizar a ajuda de voluntários. Esta solução é gratuita e fácil de executar e, portanto, é a mais popular. A maioria dos sistemas de *crowdsourcing* atuais utilizam essa solução, incluindo a Wikipedia e a ferramenta OpenStreetMap. O lado negativo do voluntariado é que é difícil prever quantos usuários é possível recrutar para uma aplicação específica.

A quarta solução é fazer com que os usuários paguem pelo serviço. A ideia básica é exigir que os usuários de um sistema A paguem pelo uso de A, contribuindo para um sistema de *crowdsourcing* B. A quinta solução é utilizar os rastros deixados pelos usuários em outro sistema bem estabelecido (como construir um sistema de correção ortográfica explorando os rastros deixados por um usuário em um mecanismo de busca).

As ferramentas de *crowdsourcing* têm ganhado cada vez mais notoriedade nos últimos anos, e por esse motivo, espera-se que muitas técnicas sejam desenvolvidas para engajar uma gama cada vez maior de usuários em *crowdsourcing*. Dessa forma, usuários inexperientes tornam-se capazes de realizar contribuições cada vez mais complexas, inserindo informações em Sistemas de Informações Geográficas sem necessariamente conhecer alguma linguagem de consulta estruturada [18].

Uma categoria cada vez mais popular de *crowdsourcing* é o chamado *crowdsourcing* espacial, onde as tarefas devem ser concluídas em um local e horário específicos. O *crowdsourcing* espacial estimulou uma série de sucessos industriais recentes, incluindo serviços urbanos de economia compartilhada (Uber e Gigwalk) e coleta de dados espaço-temporais (OpenStreetMap e Waze)[65]. Para a realização deste trabalho, foi escolhida a ferramenta OpenStreetMap para estudo de caso, devido ao seu crescimento no cenário brasileiro e à sua licença de conteúdo aberto, o que facilita a coleta e utilização dos dados.

2.3 OpenStreetMap

A ferramenta OpenStreetMap é um projeto de *crowdsourcing* espacial criado para construir um banco de dados geográfico livre do planeta, tendo como objetivo obter um registro de todas as feições geográficas existentes. No início, a ferramenta era utilizada

[2]https://www.mturk.com/ [Acesso em fevereiro de 2019].

basicamente para o mapeamento de ruas, porém atualmente provê a inclusão de trilhas, prédios, bosques, praias, entre outros objetos. Juntamente com as características geográficas, o projeto também inclui fronteiras administrativas, detalhes do uso da terra, rotas de ônibus e outras informações que podem ser adicionadas por meio do uso de *tags*[7].

Pode-se dizer que existem dois fatores primordiais que tornaram o OpenStreetMap uma ferramenta de sucesso: o primeiro foi a flexibilização das restrições em relação ao uso e acesso às informações geográficas do mundo inteiro; O segundo foi o barateamento dos dispositivos portáteis de navegação por satélite.

Além disso, os dados do OSM possuem uma grande vantagem por serem completamente livres, uma vez que estão dentro de uma licença de conteúdo aberto. Os dados na ferramenta são constantemente atualizados através de seus milhares de usuários cadastrados, os quais são capazes de inserir pontos relevantes ao mapa. O OSM possui ainda um significativo potencial para atrair voluntários do mundo todo, incluindo as regiões menos desenvolvidas do planeta, em que a obtenção de dados pode ser mais difícil para a maioria das empresas de mapeamento comercial

A principal saída cartográfica do OpenStreetMap é apresentada na página web da ferramenta.[3] A página utiliza-se da biblioteca AJAX[4] para que o mapa possa ser atualizado em tempo real, e assim permitir uma interatividade maior com os usuários. Dessa forma, à medida que os usuários clicam em diferentes pontos do mapa, novos quadros são solicitados em segundo plano, sem precisar recarregar toda a página HTML. A Figura 2.5 representa a página padrão do OpenStreetMap, mostrando parte da área central de Londres, destacando-se em vermelho as principais ruas e avenidas situadas nessa parte da cidade.

2.3.1 Tipos de Objetos no OSM

O OpenStreetMap trabalha com três tipos de objetos fundamentais: nós, caminhos e relações. Cada um desses objetos é descrito a seguir [55]:

- **Nós:** Um nó representa um ponto na superfície terrestre, definido por sua latitude e longitude. Cada nó deve conter no mínimo um número de identificação (id) e um par de coordenadas. Nós podem ser utilizados para representar objetos como um banco de uma praça ou uma parada de ônibus, por exemplo. Os nós também podem ser utilizados como pontos de referência ao longo de caminhos, como para a representação de placas de trânsito ao longo de rodovias;

[3]https://www.openstreetmap.org/ [Acesso em fevereiro de 2019].
[4]http://ajaxian.com/by/topic/ajax [Acesso em em fevereiro de 2019].

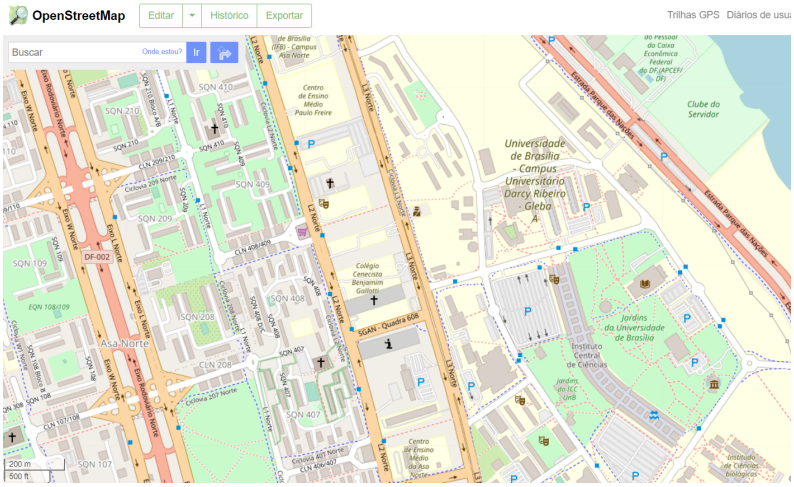


Figura 2.5: Página Padrão da Ferramenta OpenStreetMap.

- **Caminhos:** Um caminho representa uma lista ordenada contendo de 2 a 2000 nós, definindo dessa forma o tipo "polilinha". Caminhos representam estruturas lineares tais como rios e rodovias. Caminhos também são utilizados para definir limites de regiões fechadas como construções ou florestas e, nesse caso, o primeiro nó deve coincidir com o último nó. Áreas que contêm buracos, ou cuja fronteira supera 2000 nós, não podem ser representadas por um simples caminho, devendo ser representadas através de relações;
- **Relações:** Uma relação é uma estrutura de dados que documenta o relacionamento entre dois ou mais elementos (nós, caminhos ou outras relações). Alguns possíveis exemplos são:
  a. Relação de rota – Lista os caminhos que delimitam uma avenida, ciclovia ou rota de ônibus;
  b. Relação de restrição – Informa os locais em que o acesso de veículos, ciclistas ou pedestres é restrito.

O significado de muitas relações da ferramenta OSM é definido por meio de *tags* (etiquetas) específicas. Uma *tag* consiste em um par "chave-valor", sendo que cada um desses campos contém uma *string Unicode* de até 255 caracteres.

2.3.2 Histórico da Ferramenta

O projeto OpenStreetMap começou em agosto de 2004, quando o programador britânico Steve Coast quis experimentar um receptor GPS USB que ele havia comprado. Ele usou um software chamado GPSDrive, que pegou mapas do Microsoft MapPoint, quebrando as condições da licença. Não querendo violar direitos autorais nesses mapas, ele

procurou uma alternativa. Coast descobriu que não havia fontes de dados de mapeamento disponíveis que ele pudesse incorporar ao software de código aberto sem quebrar as condições de licenciamento ou pagar enormes quantias. Coast percebeu então que poderia desenhar seu próprio mapa, com a ajuda de usuários voluntários [7].

A ideia de cartógrafos amadores construir mapas foi recebida com algum ceticismo a princípio. Alguns disseram que os receptores de GPS padrão eram muito imprecisos para fazer mapas, pois um erro de 10 metros significaria que as estradas estariam no lugar errado. Outros alegaram que era necessária uma infraestrutura complexa para um projeto tão grande.

Em março de 2006, o primeiro aplicativo de edição para o OpenStreetMap, foi lançado: o Java OpenStreetMap (JOSM). Escrito na linguagem de programação Java, o JOSM permitiu o mapeamento offline pela primeira vez. Logo depois, o primeiro mapa colorido foi criado usando um renderizador escrito especificamente para o OpenStreetMap, denominado Osmarender [7].

O crescimento do OpenStreetMap aconteceu de forma bastante acelerada, atualmente ultrapassando a barreira dos sete milhões de usuários registrados, conforme pode ser observado na Figura 2.6. No mês de janeiro de 2010, a ferramenta possuía aproximadamente 200.000 usuários registrados. Esse número foi multiplicado em cerca de dez vezes já no começo do ano de 2015, quando a ferramenta já contava com mais de dois milhões de usuários registrados.

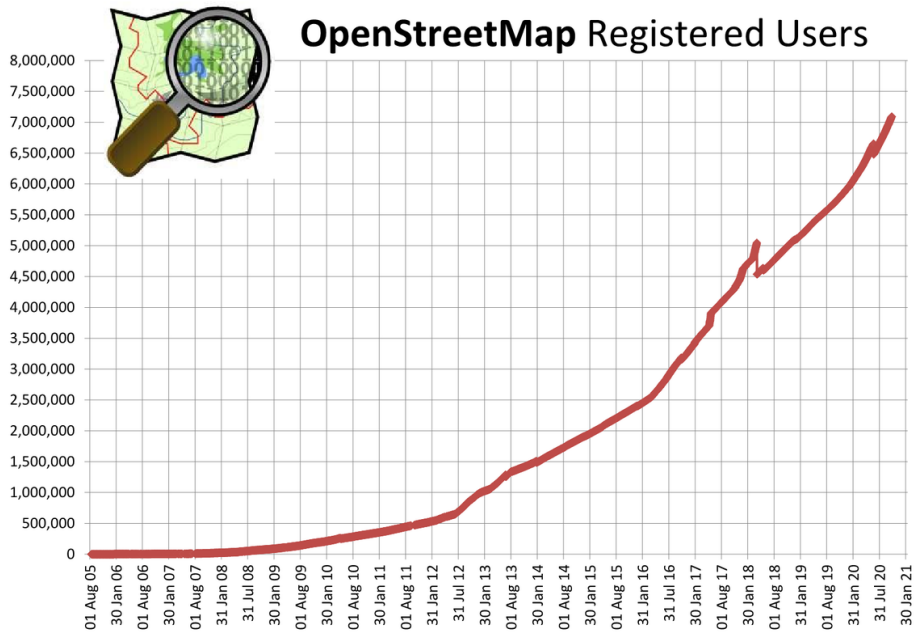


Figura 2.6: Usuários Registrados no OpenStreetMap.[5]

2.4 Dimensões da Qualidade

A qualidade é um componente chave de qualquer conjunto de dados. Especificamente, em um conjunto de dados geoespaciais, as tomadas de decisões para uma determinada finalidade são fortemente baseadas em medidas de qualidade, tais como acurácia, completude e consitência lógica [24]. Isso se aplica de forma mais intensa aos SIGV, visto que os usuários participam ativamente dos processos de inclusão, edição e exclusão das informações [60].

Um conjunto de elementos é especificado na norma ISO 19157 para qualidade de dados geoespaciais [36]. Essa estrutura atende de forma adequada as agências de mapeamento, que contam com profissionais que seguem protocolos rigorosos e múltiplos processos de controle de qualidade, de modo a produzir produtos de alta qualidade com especificações mínimas aceitáveis. Entretando, a qualidade dos dados em relação aos SIGV traz novos desafios para o campo da avaliação da qualidade, devido à natureza heterogênea desses sistemas [21].

Muitas vezes, existe um viés sociocultural nas informações coletadas em ferramentas colaborativas, com mais dados coletados em áreas urbanas do que rurais [53] e mesmo dentro de uma área urbana, as áreas mais populares e turísticas costumam receber mais atenção e, portanto, mais dados com maior detalhamento, do que áreas urbanas desconhecidas ou periféricas [5]. Essas tendências podem ser influenciadas ainda mais pelo acesso e conhecimento de recursos digitais, diferenças culturais e quanto tempo os usuários têm efetivamente para participar do processo de mapeamento [34].

Além disso, um dos desafios que os profissionais e pesquisadores têm em discutir, avaliar e medir a qualidade é que não existe uma única definição ou padrão de qualidade. Dessa forma, o conceito de qualidade pode ser dividido em diferentes aspectos, os quais são denominados na literatura de dimensões da qualidade. As dimensões da qualidade da informação podem ser agrupadas em clusters de acordo com [6]. Assim, as principais dimensões da qualidade estão listadas a seguir [23]:

1. Acurácia, correção e precisão têm como foco a adesão a uma dada realidade de interesse;
2. Completude, pertinência e relevância referem-se à capacidade de representar todos e apenas os aspectos relevantes da realidade de interesse;
3. Consistência, coesão e coerência referem-se à capacidade das informações para cumprir sem contradições a todas as propriedades da realidade de interesse, conforme especificado em termos de restrições de integridade, regras de negócios, regras de modelagem e outros formalismos;

[5]https://wiki.openstreetmap.org/wiki/Stats [Acesso em outubro de 2020].

4. Redundância, minimalidade e concisão referem-se à capacidade de representar os aspectos da realidade de interesse com o uso mínimo de informações;
5. Legibilidade, inteligibilidade, clareza e simplicidade referem-se à facilidade de compreensão de informações pelos usuários;
6. Confiança, credibilidade, confiabilidade e reputação concentra-se em quanto a informação deriva de uma fonte confiável.

Este trabalho tem como foco a dimensão da qualidade da completude, mais especificamente em relação às informações de endereço associadas aos objetos da ferramenta OpenStreetMap. Dessa forma, a aplicação QualiOSM foi desenvolvida com o objetivo de melhorar a completude dos objetos dentro da plataforma do OSM.

Capítulo 3

Estado da Arte

A revisão do estado da arte foi implementada por meio de uma Revisão Sistemática da Literatura (RSL), que consiste em um meio de avaliar e interpretar toda a pesquisa disponível relevante para um campo do conhecimento específico, área temática ou fenômeno de interesse. As revisões sistemáticas têm como objetivo apresentar uma avaliação justa de um tópico de pesquisa usando uma metodologia confiável, rigorosa e auditável [42]. As próximas seções deste capítulo descrevem o processo realizado durante a revisão, explicitando o protocolo desenvolvido e quais questões de pesquisa foram respondidas por meio deste trabalho. O processo de revisão sistemática foi realizado entre os dias 1º de abril e 30 de abril de 2019, sendo atualizado entre os dias 1º de setembro e 30 de setembro de 2020.

3.1 Protocolo de Pesquisa

Para se realizar a revisão sistemática na área de qualidade dos dados em relação aos SIG e SIGV, foi utilizado o mesmo processo abordado em [42], em que os autores dividiram o processo da revisão em três fases distintas: planejamento, execução e documentação da revisão.

Dessa forma, a fase de planejamento é responsável por realizar a identificação da necessidade da revisão, especificação das questões de pesquisa a serem respondidas ao longo da fase de execução, além de realizar o desenvolvimento e avaliação do protocolo de pesquisa, instrumento que servirá de apoio aos pesquisadores envolvidos durante todo o processo de revisão.

A fase de execução é responsável pela identificação do trabalho de pesquisa, seleção dos trabalhos relevantes, análise de qualidade, extração e monitoramento dos dados. Já a fase de documentação da revisão é responsável pela consolidação da pesquisa desenvolvida no formato de relatório, o qual deverá ser devidamente formatado e avaliado.

Segundo [46], o estabelecimento de um protocolo de pesquisa é necessário para assegurar que a revisão seja sistemática e para que sejam minimizados quaisquer aspectos de natureza subjetiva por parte do pesquisador. Dessa maneira, o protocolo de pesquisa possui o intuito de responder, de forma objetiva, um conjunto de questões pré-estabelecidas pelo pesquisador, as quais neste trabalho estão descritas na subseção 3.1.1. Além disso, a fase da execução da pesquisa foi dividida em duas etapas (seleção e extração dos dados), descritas respectivamente nas subseções 3.1.2 e 3.1.3.

3.1.1 Questões de Pesquisa

O objetivo desta revisão sistemática foi responder às seguintes Questões de Pesquisa (QP):

- QP1: Quais dimensões da qualidade estão sendo exploradas em relação aos Sistemas de Informações Geográficas?
- QP2: Que tipos de soluções estão sendo propostas em relação à qualidade dos dados em Sistemas de Informações Geográficas?
- QP3: Que tipos de soluções estão sendo propostas em relação à qualidade dos dados especificamente em Sistemas de Informações Geográficas Voluntárias, tais como a ferramenta OpenStreetMap?

O protocolo de pesquisa desenvolvido foi implementado por meio de duas etapas: uma etapa para a seleção dos dados e outra etapa para a extração dos dados. Essas duas etapas encontram-se descritas nas subseções a seguir.

3.1.2 Seleção dos Dados

A pesquisa foi desenvolvida utilizando-se de três bases científicas distintas, escolhidas levando-se em conta a relevância para a área dos Sistemas de Informações. Dessa forma, as bases bibliográficas escolhidas foram:

1. The ACM Digital Library[1];
2. IEEE Xplore[2];
3. Scopus (Elsevier)[3].

Para realizar a pesquisa nas fontes científicas, duas *strings* de pesquisa foram elaboradas, considerando a qualidade dos dados em Sistemas de Informações Geográficas em

[1]http://dl.acm.org/ [Acesso em fevereiro de 2019].
[2]http://ieeexplore.ieee.org [Acesso em fevereiro de 2019].
[3]https://www.scopus.com/ [Acesso em fevereiro de 2019].

geral (a fim de responder QP1 e QP2), e considerando a qualidade dos dados em Sistemas de Informações Geográficas Voluntárias tais como na ferramenta OpenStreetMap. Dessa forma, as *strings* de pesquisa elaboradas foram as seguintes:

(i) ("Geographic Information System") AND (QUALITY) AND (PARAMETERS OR ASSESSMENT OR INDICATORS OR METRICS OR STUDY OR ANALYSIS)

(ii) ((VGI OR OPENSTREETMAP OR "Volunteered Geographic Information") AND (QUALITY) AND (PARAMETERS OR ASSESSMENT OR INDICATORS OR METRICS OR STUDY OR ANALYSIS))

A *string* de busca é utilizada para a procura de termos-chave de forma a responder as questões de pesquisa elaboradas pelo pesquisador. Dessa forma, de acordo com as necessidades da pesquisa, novos termos poderiam ser adicionados de forma a abranger outros termos relacionados com os SIG e os SIGV, tais como *crowdsourcing* ou *collaborative mapping*, por exemplo. Além disso, para que um artigo fosse selecionado na revisão sistemática, foram considerados os seguintes critérios de inclusão/exclusão:

- O artigo deveria estar disponível para *download* na base de dados pesquisada;
- Foi dada preferência a artigos publicados a partir do ano de 2007. Porém, artigos tidos como clássicos, associados a conceitos básicos ou definições, também foram considerados;
- Artigos que continham pelo menos 3 citações foram priorizados;
- O artigo deveria mencionar de forma explícita a questão da qualidade em Sistemas de Informações Geográficas.

Por outro lado, os seguintes critérios de exclusão foram aplicados:

- Artigos publicados antes de 2007 que não trouxessem definições clássicas não foram considerados;
- Trabalhos duplicados, publicados em mais de uma das bases de dados pesquisadas, não foram considerados;
- Artigos com menos de 3 citações não foram considerados;
- Artigos identificados como fora do escopo ou tema da pesquisa não foram considerados.

3.1.3 Extração dos Dados

Aplicando as duas *strings* de pesquisa, 1.192 documentos foram extraídos das bibliotecas digitais. Após a aplicação dos critérios de inclusão/exclusão, 1.146 artigos foram rejeitados e apenas 51 artigos foram considerados relevantes. A Figura 3.1 mostra a distribuição de artigos de acordo com a fonte pesquisada. De acordo com a figura, 12% dos

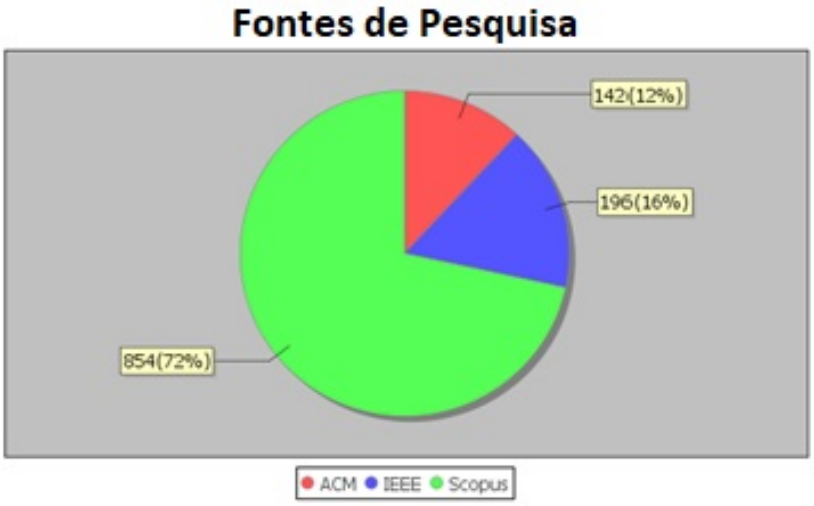


Figura 3.1: Distribuição de Artigos por Fonte de Pesquisa.

artigos foram encontrados na base de dados da ACM, 16% na base de dados da IEEE e 72% na base de dados da Scopus.

Durante a fase de extração de dados, foram lidos o resumo e a introdução dos artigos selecionados a fim de extrair suas informações mais importantes, dando prioridade em responder às questões de pesquisa pré-estabelecidas. Depois de ler o resumo e a introdução de todos os 51 artigos escolhidos na fase de seleção, ainda foram identificados 16 artigos fora do escopo da pesquisa e que por esse motivo foram descartados. Dessa forma, ao final da Revisão Sistemática realizada, foi obtido um total de 35 artigos considerados relevantes. Para realizar o processo de classificação dos artigos, foi utilizada a ferramenta StArt[4], desenvolvida pelo Laboratório de Pesquisa em Engenharia de Software (LaPES), na Universidade Federal de São Carlos (UFSCar).

3.2 Resultados da Revisão

Respondendo à questão de pesquisa QP1, observou-se que a dimensão da qualidade mais explorada pelos pesquisadores nos últimos anos tem sido a acurácia, sendo abordada em todos os 35 artigos considerados relevantes. A dimensão da completude foi explorada em 18 artigos (51,43% do total) e a dimensão da consistência lógica foi explorada em 12 artigos (34,29% do total). Alguns artigos abordaram mais de uma dimensão e por esse motivo, o total excede 100%. O grande número de trabalhos explorando a dimensão da acurácia pode ser facilmente explicado, uma vez que esta dimensão é responsável por descrever a exatidão ou incorreção dos objetos no banco de dados, funcionando como um

[4]http://lapes.dc.ufscar.br/tools/start_tool [Acesso em fevereiro de 2019].

aspecto essencial para a qualidade dos dados. Estes resultados podem ser observados por meio da Figura 3.2.

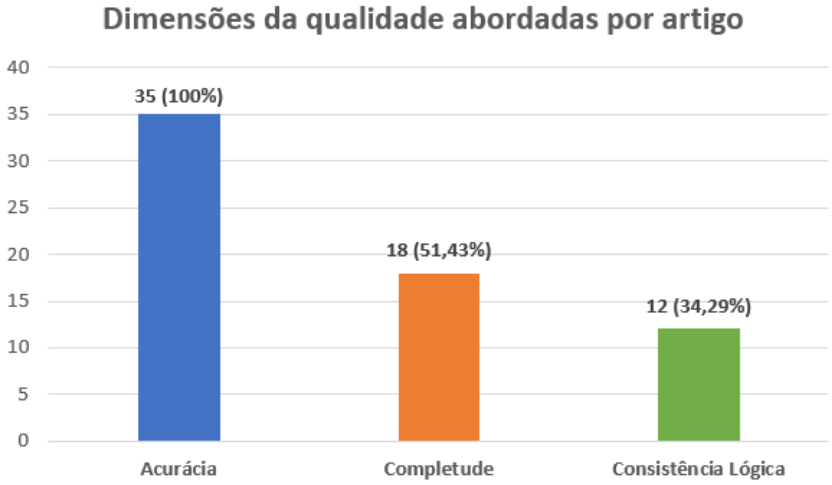


Figura 3.2: Dimensões da Qualidade Abordadas por Artigo Acadêmico.

Em relação à questão de pesquisa QP2, foi observado o seguinte resultado: 21 trabalhos propuseram a elaboração de um framework ou algoritmo para a melhoria da qualidade dos dados; 19 artigos utilizaram serviços baseados em localização; 11 artigos fizeram uma comparação com imagens de satélite ou utilizaram técnicas de processamento de imagens; 9 artigos estruturaram os seus trabalhos sob a forma de um relatório, sem qualquer tipo de implementação; 8 artigos realizaram análises baseadas em redes viárias; 7 artigos fizeram uso de técnicas de aprendizado de máquina; 6 artigos fizeram uso de metadados; 4 artigos utilizaram técnicas de interpolação; 2 trabalhos realizaram uma espécie de ranking da confiabilidade dos usuários; 2 artigos fizeram uso de técnicas de computação humana. Estes resultados podem ser vistos por meio da Tabela 3.1, reparando também que alguns artigos utilizaram mais de uma solução.

Finalmente, respondendo à questão de pesquisa QP3, os seguintes resultados foram obtidos: 12 artigos utilizaram serviços baseados em localização; 11 trabalhos propuseram a elaboração de um framework ou algoritmo para a melhoria da qualidade dos dados; 11 artigos fizeram comparações com imagens de satélite ou utilizaram técnicas de processamento de imagens; 8 artigos construíram seu trabalho na forma de um relatório, sem nenhum tipo de implementação; 6 artigos fizeram uso de metadados; 4 trabalhos realizaram análises baseadas em redes rodoviárias; 4 artigos fizeram uso de técnicas de *machine learning*; 2 artigos realizaram uma espécie de ranking de confiabilidade dos usuários; 2 artigos fizeram uso de técnicas de computação humana; nenhum artigo foi encontrado

lidando com técnicas de interpolação. Esses resultados podem ser observados na Tabela 3.2.

Tabela 3.1: Soluções Propostas para a Questão da Qualidade dos Dados em SIG.

Tipo de Solução Proposta	Artigos	Total
Implementação de algoritmo / framework	[2, 3, 8, 10, 32, 33, 39, 70, 68, 71, 31] [37, 11, 72, 12, 63, 16, 27, 66, 40, 67]	21
Serviços baseados em localização	[31, 1, 22, 25, 11, 12, 8, 33, 39, 2] [51, 61, 70, 40, 67, 68, 72, 71, 44]	19
Comparação com imagens de satélite / Técnicas de processamento de imagem	[1, 25, 13, 52, 61, 8, 32, 33, 39, 70, 63]	11
Relatório	[1, 22, 25, 13, 17, 52, 51, 61, 44]	9
Análise de redes viárias	[61, 8, 33, 39, 2, 16, 66, 37]	8
Técnicas de *machine learning*	[61, 32, 16, 71, 3, 31, 37]	7
Uso de metadados	[22, 52, 51, 11, 10, 44]	6
Melhoria em *tags*	[14, 41, 4, 15, 50]	5
Técnicas de interpolação	[13, 32, 70, 27]	4
Ranking de confiabilidade dos usuários	[10, 40]	2
Técnicas de computação humana	[11, 40]	2

Tabela 3.2: Soluções Propostas para a Questão da Qualidade dos Dados em SIGV.

Tipo de Solução Proposta	Artigos	Total
Serviços baseados em localização	[1, 22, 25, 11, 8, 40, 44] [51, 61, 68, 72, 71]	12
Implementação de algoritmo / framework	[3, 8, 10, 68, 71, 37, 11, 72, 63, 66, 40]	11
Comparação com imagens de satélite / Técnicas de processamento de imagem	[1, 25, 13, 52, 61, 8, 32, 33, 39, 70, 63]	11
Relatório	[1, 22, 25, 17, 52, 51, 61, 44]	8
Uso de metadados	[22, 52, 51, 11, 10, 44]	6
Melhoria em *tags*	[14, 41, 4, 15, 50]	5
Análise de redes viárias	[61, 8, 66, 37]	4
Técnicas de *machine learning*	[61, 71, 3, 37]	4
Ranking de confiabilidade dos usuários	[10, 40]	2
Técnicas de computação humana	[11, 40]	2
Técnicas de interpolação	-	0

3.3 Análise da Revisão

Por meio da realização da revisão sistemática, foram obtidos 35 artigos relevantes a partir de um total de 1.192 artigos, coletados de diferentes fontes de pesquisa. Pode-se observar que a área de qualidade de dados em relação aos SIG e SIGV ainda está em expansão, possuindo muitos campos a serem explorados. Além disso, percebe-se uma grande quantidade de trabalhos escritos envolvendo alguma forma de implementação, seja de um framework ou de um algoritmo. Este fato evidencia a importância do trabalho prático envolvendo a questão da qualidade dos dados sobretudo em relação aos SIGV.

Após esta revisão sistemática, outros resultados importantes foram obtidos. Em muitos dos trabalhos acadêmicos lidos, destaca-se a heterogeneidade presente na qualidade dos dados, causada especialmente por conta da discrepância no grau de conhecimento entre os usuários desses tipos de sistema. Além disso, foi observado um considerável número de artigos que utilizaram como solução técnicas de processamento de imagens de satélite ou técnicas de *machine learning*.

3.4 Trabalhos Relacionados

Dentro dos trabalhos revisados, foram encontrados cinco estudos relevantes na literatura que exploraram o processo de adição de *tags* em ferramentas colaborativas. [4] exploraram a motivação para atribuir etiquetas em imagens no Flickr, concluindo que a maioria dos usuários realizam esse tipo de marcação para tornar as informações mais acessíveis ao público em geral. Além disso, [41] avaliaram o desempenho de classificadores treinados com fotos do Flickr e suas *tags* associadas, demonstrando que as *tags* fornecidas pelos usuários contêm muitas informações incorretas.

Em relação às ferramentas de mapeamento colaborativo, [14] organizaram uma ontologia com o objetivo de padronizar e facilitar a hierarquia de *tags* dentro da ferramenta OpenStreetMap, mas concluíram que o uso de uma ontologia só é eficiente se o usuário mantiver as *tags* constantemente atualizadas dentro da plataforma do OSM.

Ainda dentro da ferramenta OpenStreetMap, [50] realizaram a análise de mais de 25.000 objetos no banco de dados da Irlanda, Reino Unido, Alemanha e Áustria. Os resultados indicaram que existem alguns problemas decorrentes da forma como os usuários atribuem tags aos objetos no OSM. O estudo também mostrou que esses problemas identificados são uma combinação da flexibilidade do processo de etiquetagem e a falta de um mecanismo mais rígido para verificar a aderência à ontologia OpenStreetMap em relação às etiquetas adicionadas pelos seus usuários.

[15] usaram as recomendações fornecidas na página "Map Features" da Wiki do projeto OpenStreetMap e analisaram a base de dados do OSM em quarenta cidades ao redor do mundo para ver se os usuários colaboradores nessas áreas urbanas estavam usando as diretrizes em suas práticas de marcação. O estudo concluiu que o cumprimento das sugestões e orientações geralmente é médio ou ruim, uma vez que os usuários dessas áreas nem sempre possuem o mesmo nível de conhecimento.

Diferentemente dos trabalhos citados, este trabalho propõe a implementação da ferramenta QualiOSM com o objetivo de melhorar a qualidade das informações geográficas dentro do OpenStreetMap, principalmente no que diz respeito ao processo de atribuição de *tags* de endereço aos objetos. Dessa forma, a intenção da ferramenta é contribuir para a completude das informações de endereços de objetos na plataforma OSM, auxiliando na automatização da inserção dessas informações dentro plataforma do OSM.

Capítulo 4

Desenvolvimento da Arquitetura QualiOSM

Este capítulo apresenta o desenvolvimento da arquitetura QualiOSM, elaborada com o objetivo de melhorar a completude das informações de endereço em objetos da ferramenta OpenStreetMap. Dessa forma, o capítulo descreve as fases de desenvolvimento deste trabalho, compreendidas pelas fases de projeto, implementação e utilização da arquitetura QualiOSM, a qual foi concebida na forma de uma extensão (*plugin*) para o editor de dados Java OpenStreetMap (JOSM), responsável pelo maior número de edições dentro da plataforma OSM.

4.1 Projeto

A arquitetura proposta para a melhoria da qualidade dos dados dentro da ferramenta OpenStreetMap baseia-se em um modelo de três camadas, na qual a lógica do negócio, o acesso aos dados e a interface com o usuário são desenvolvidas e mantidas como módulos independentes [26]. Conforme pode ser observado a partir da Figura 4.1, a arquitetura QualiOSM foi dividida em três camadas distintas: a camada superior é a Camada de Apresentação, responsável por fornecer a interface entre o usuário e o editor de dados JOSM, além de fornecer o carregamento das imagens aéreas; o plugin QualiOSM e a funcionalidade do adicionador de *tags* foram desenvolvidos dentro da Camada de Aplicação, na qual também é possível ver a interação com a API da ferramenta OpenStreetMap; por fim, a Camada de Dados é responsável por prover o gerenciamento dos dados na base da plataforma do OSM e interagir com as ferramentas Nominatim, CEP Aberto e a base de dados dos Correios. As ferramentas Nominatim e CEP Aberto foram utilizadas para a inclusão das tags gerais de endereçamento (*addr:city, addr:building, addr:suburb e addr:neighbourhood*) e também para a inclusão da *tag* de código postal (*addr:postcode*),

enquanto a base de dados dos Correios foi utilizada para a inclusão da *tag* de código postal e para fazer a validação das demais *tags*.

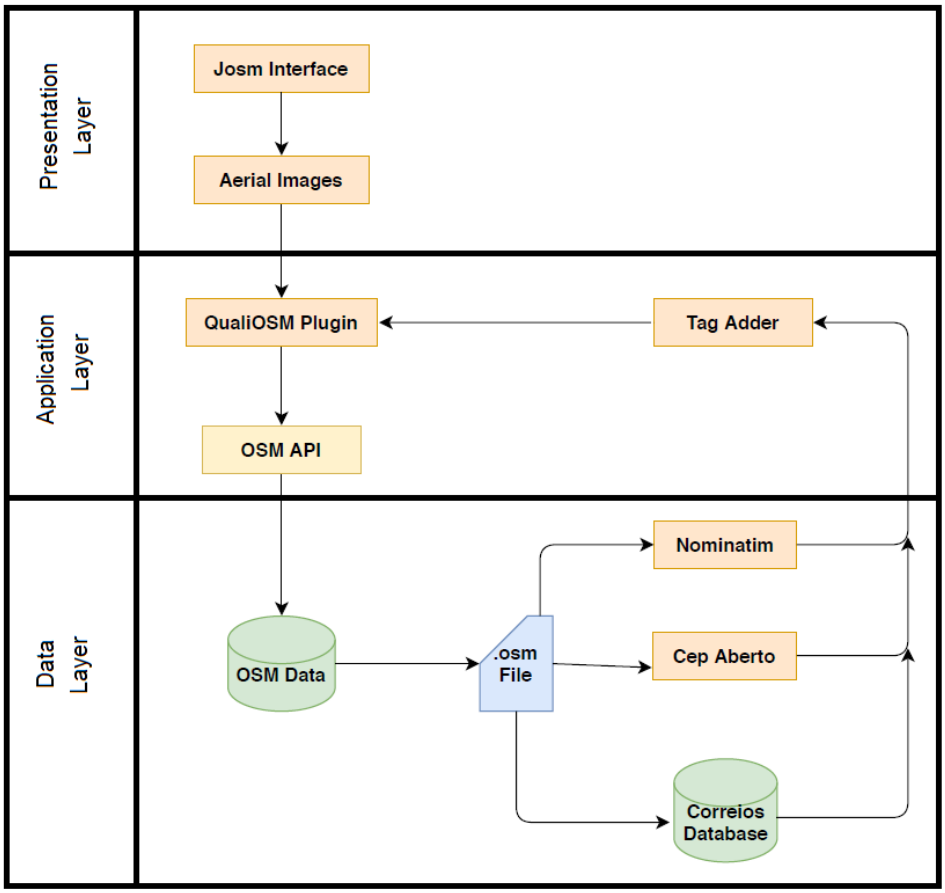


Figura 4.1: Arquitetura para Implementação da Ferramenta QualiOSM.

4.1.1 Camada de Apresentação

A Camada de Apresentação é a camada mais externa da arquitetura desenvolvida e é responsável por prover a interface de interação com o usuário. É nesta camada em que o usuário será capaz de realizar o carregamento das imagens aéreas e visualizar o carregamento dos dados exportados a partir da ferramenta OpenStreetMap em seu formato padrão na extensão *.osm*.

A Figura 4.2 apresenta os principais componentes da interface do editor de dados JOSM. Dessa forma, na parte superior da tela encontra-se o menu principal, o qual em sua configuração padrão terá as seguintes opções de escolha para o usuário: Arquivo, Editar, Visualizar, Ferramentas, Seleção, Predefinições, Camadas, Janelas, Áudio e Ajuda. Abaixo do menu principal, encontra-se a barra de ferramentas principal, representada na forma de botões ilustrados. Tanto o menu principal quanto a barra de ferramentas são

customizáveis, sendo possível adicionar e remover itens desses menus de acordo com a finalidade do *plugin* implementado.

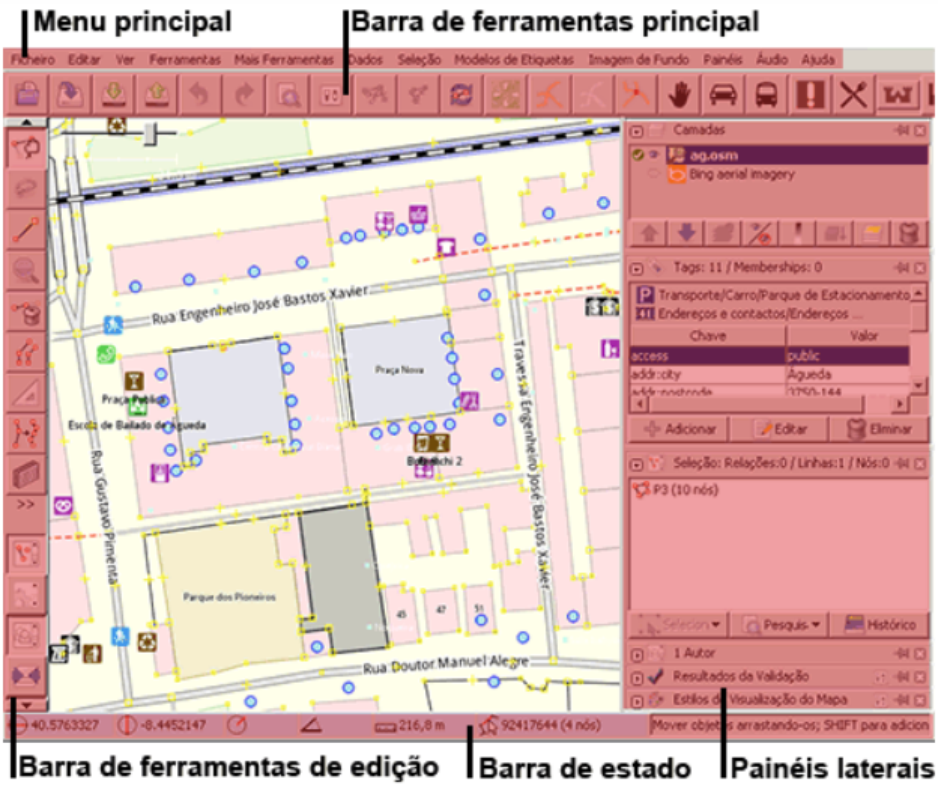


Figura 4.2: Interface do Editor de Dados JOSM.[1]

Na parte central da Figura 4.2, encontra-se a janela de visualização do mapa, local onde o usuário poderá carregar camadas de imagens aéreas, bem como visualizar os dados exportados a partir da ferramenta OpenStreetMap no formato *.osm*. Na parte lateral esquerda da Figura 4.2, encontra-se a barra de ferramentas de edição, por meio da qual o usuário poderá, por exemplo, realizar a inclusão ou seleção de objetos dentro da plataforma do OSM. Na parte inferior da figura, encontra-se a barra de estado, local onde é possível visualizar as coordenadas geográficas dos objetos selecionados no mapa. Por fim, na parte lateral direita é possível visualizar os painéis laterais, em que o usuário poderá enxergar as camadas carregadas e as *tags* associadas aos objetos selecionados.

4.1.2 Camada de Aplicação

A Camada de Aplicação é responsável por conter a lógica do negócio e realizar a interação com a API da ferramenta OpenStreetMap. A Figura 4.3 apresenta o diagrama de classes do *plugin* desenvolvido neste projeto o qual também foi denominado de QualiOSM.

[1]https://josm.openstreetmap.de/wiki/Pt:Help [Acesso em outubro de 2020].

Conforme pode ser observado nesse diagrama, a classe principal da aplicação está representada no centro pela classe QualiOSM_Plugin, a qual será responsável por fazer a chamada dos métodos AddTags_Correios(), AddTags_Nominatim(), AddTags_Cepaberto() e Clean_Tags(). A classe QualiOSM_Plugin é derivada da classe Plugin, a qual por sua vez é responsável por encapsular as informações gerais do *plugin*. Dessa maneira, o método Plugin(PluginInformation info) cria um objeto de informações do *plugin*, realizando a leitura do arquivo de manifesto contido no arquivo QualiOSM.jar, gerado após a compilação do projeto.

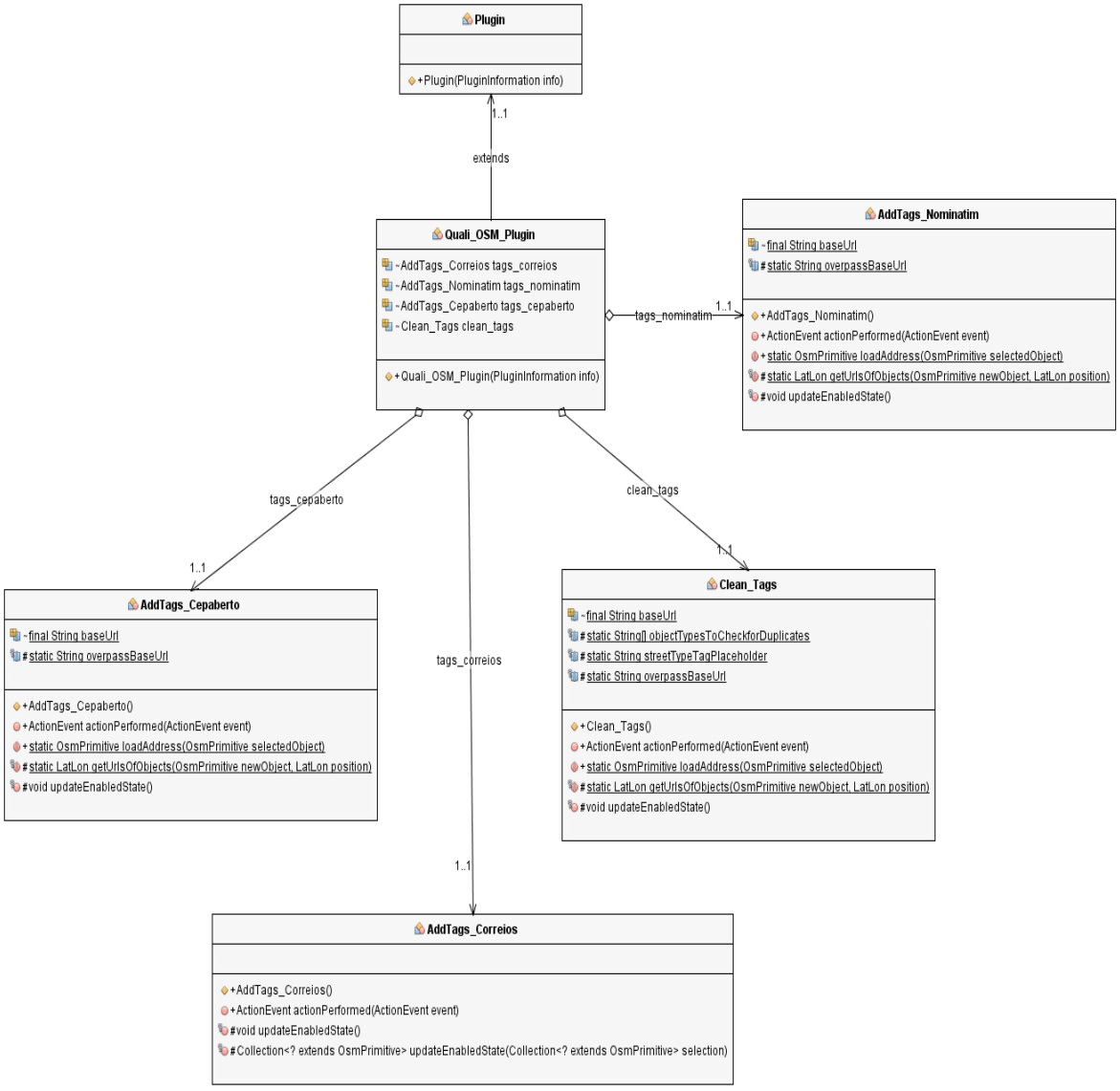


Figura 4.3: Diagrama de Classes da Ferramenta QualiOSM.

A Camada de Aplicação é a camada responsável também por fazer a chamada aos principais métodos que compreendem o adicionador de *tags*, ou seja, os métodos AddTags_Correios() para a adição de *tags* de endereço utilizando a base de dados dos Cor-

reios, AddTags_Nominatim() para a adição de *tags* de endereço utilizando a ferramenta Nominatim, AddTags_Cepaberto(), para a adição de *tags* de endereço utilizando a ferramenta CEP Aberto, além do método Clean_Tags() para fazer a exclusão de *tags* de endereço.

Neste projeto, as ferramentas Nominatim e CEP Aberto foram utilizadas para a inclusão das *tags addr:city* (cidade) , *addr:neighbourhood*(logradouro), *addr:suburb* (bairro) e *addr:postcode* (código postal) . Já a base de dados dos Correios foi utilizada para a inclusão da *tag addr:postcode* e para fazer a validação das demais *tags* de endereço.

4.1.3 Camada de Dados

A Camada de Dados é responsável por fazer a interação com as ferramentas Nominatim, CEP Aberto e a base de dados dos Correios. Conforme pode ser observado no diagrama de classes da Figura 4.3, as classes AddTags_Nominatim e AddTags_Cepaberto irão agir de forma semelhante, pois as duas ferramentas de geocodificação reversa irão buscar as informações de endereço associadas aos objetos selecionados utilizando suas respectivas plataformas na Web, a partir de um endereço *url*, os quais serão retornados utilizando o método *loadAddress*. Para a ferramenta Nominatim, foi utilizada a base *url* https://nominatim.openstreetmap.org/reverse?format=json e para a ferramenta CEP Aberto foi utilizada a base *url* https://www.cepaberto.com/api/v3/nearest?. Essas *urls* foram então concatenadas com as latitudes e longitudes dos objetos selecionados.

A base de dados dos Correios, por sua vez, é composta por um arquivo *.sql*, com as coordenadas de cada objeto já associadas. Assim, as informações do código postal foram inseridas com base na coordenada mais próxima do centro do objeto selecionado no JOSM. A distância foi calculada de acordo com a fórmula da menor distância entre dois pontos, expressa na Equação 4.1.

$$distância = \sqrt{(lat2 - lat1)^2 + (lon2 - lon1)^2} \tag{4.1}$$

Onde (lat1, lon1) corresponde às coordenadas do centro do objeto selecionado e (lat2, lon2) corresponde às coordenadas do objeto na base de dados dos Correios. O algoritmo encontra o código postal do objeto selecionado quando a distância calculada é inferior a 10^{-4}.

A base de dados dos Correios também foi utilizada para verificar a acurácia das informações de endereço ao realizar as simulações de inserção de *tags* utilizando as ferramentas Nominatim e CEP Aberto. A base de dados foi baixada por meio da *url* https://www.qualocep.com/base-de-cep/. Os dados foram então carregados no SGBD PostreSQL e processados com o objetivo de capturar as colunas contendo as informações

de CEP, logradouro, latitude e longitude. Em seguida, os dados foram transformados em um arquivo *.json* para as informações serem lidas pelo adicionador de *tags*.

4.2 Implementação

A ferramenta QualiOSM foi desenvolvida com o objetivo de melhorar a completude das informações de endereço associadas aos objetos da plataforma OpenStreetMap. O aplicativo foi escrito na linguagem de programação Java e implementado como extensão (*plugin*) dentro do editor de dados Java OpenStreetMap (JOSM). A aplicação possui código aberto e encontra-se disponível para *download* em repositório do *site* GitHub[2].

Para a implementação do adicionador de *tags* dentro da ferramenta QualiOSM, foi utilizada a técnica de geocodificação reversa, na qual a extração de informações textuais, tais como nome ou endereço, é realizada a partir de um par de coordenadas geográficas (latitude e longitude). Esta técnica é comum em muitos cenários de aplicação geográfica, como por exemplo, serviços de mapeamento online gratuitos [43]. Neste trabalho, foram utilizadas as ferramentas de geocodificação reversa Nominatim[3] e CEP Aberto[4]. Além disso, a lista de códigos postais do Brasil, que se encontra na base de dados oficiais dos Correios, foi baixada na forma de um arquivo *.sql* com o objetivo de inserir informações mais precisas em relação ao Código de Endereçamento Postal (CEP) associado a objetos dentro da plataforma do OSM.

A escolha de implementar o aplicativo QualiOSM dentro do editor de dados JOSM se deu por várias razões: (i) é o editor de dados mais amplamente utilizado pelos usuários do OSM; (ii) é multiplataforma, sendo escrito na linguagem de programação Java; (iii) oferece um mecanismo de *plugin* para estender sua funcionalidade principal. Com uma interface de usuário facilmente compreensível, a ferramenta proposta pode fazer com que qualquer colaborador do OpenStreetMap seja capaz de enriquecer o mapa com as informações de endereço, uma vez que nenhum conhecimento específico em linguagens de Web semântica ou formalismos subjacentes são necessários [59].

4.2.1 *Tags* de endereço no OSM - Brasil

A fim de realizar algumas análises em relação às *tags* de endereço em todo o território brasileiro, os dados do OpenStreetMap - Brasil foram baixados seguindo uma arquitetura em duas diferentes camadas: uma camada para a coleta de dados e outra para a

[2]https://github.com/gmedeiros93/josm/tree/master/josm/plugins/Quali_OSM [Acesso em novembro de 2020].

[3]https://nominatim.openstreetmap.org/ [Acesso em março de 2019].

[4]https://cepaberto.com/ [Acesso em outubro de 2020].

visualização e a análise dos dados. Conforme pode ser observado a partir da Figura 4.4, dois arquivos foram utilizados para a coleta dos dados geográficos do Brasil: o arquivo FullHistory.osm[5], contendo o histórico dos dados da ferramenta OpenStreetMap correspondentes a todo o planeta até 31 de outubro de 2019; e o arquivo Brazil.poly, contendo o contorno da região do Brasil, disponibilizado no site do projeto Geofabrik. Em seguida, esses dois arquivos foram processados com a ferramenta osmconvert para a criação do arquivoBrazilHistory.osm, contendo o histórico dos dados do OpenStreetMap no Brasil. A seguir, o arquivoBrasilHistory.osm foi processado na ferramenta osm2pgsql com o objetivo de importar os dados para o banco de dados PostgreSQL. Além disso, a extensão Postgis foi utilizada para o tratamento dos dados espaciais e a extensão Hstore foi utilizada para a captura de *tags* dos objetos do OpenStreetMap. O software QGIS foi utilizado para a visualização dos dados.

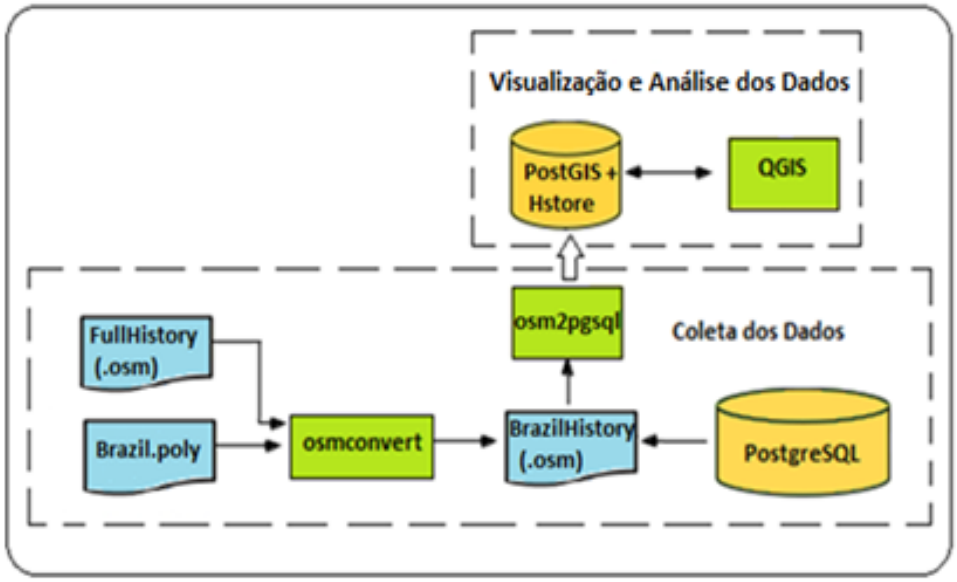


Figura 4.4: Arquitetura para Coleta e Visualização de Dados no OSM - Brasil.

Dentro da ferramenta OpenStreetMap, os edifícios são objetos que geralmente precisam de informações de endereço associadas, uma vez que os usuários desejam acrescentar dados sobre a localização de pontos de interesse, tais como o código postal, o bairro ou nome do edifício. Desta forma, foi realizada uma análise sobre o número de edifícios que atualmente possuem etiquetas de endereço associadas no Brasil e como essas inclusões foram feitas ao longo do tempo.

A Figura 4.5 apresenta uma evolução da inclusão de *tags* de endereço nos edifícios do OpenStreetMap no Brasil entre os anos de 2009 e 2019. Nesta figura, observa-se que a inclusão deste tipo de etiqueta tem crescido desde 2015, mas ainda há um pequeno número

[5]

de edifícios com etiquetas de endereço associadas (em 2017, havia mais de 860.000 edifícios mapeados, porém apenas pouco mais de 100.000 edifícios tinham etiquetas de endereço associadas). Ainda na Figura 4.5 destaca-se o pico de inclusão desses tipos de *tags* no ano de 2017, principalmente em relação à *tag* "addr: street", correspondente aos nomes das ruas. A predominância desta *tag* justifica-se porque a ferramenta OpenStreetMap é especializada na questão do mapeamento de estradas e rodovias, sendo que informações sobre nomes de estradas próximas aos edifícios podem facilitar mecanismos de roteamento.

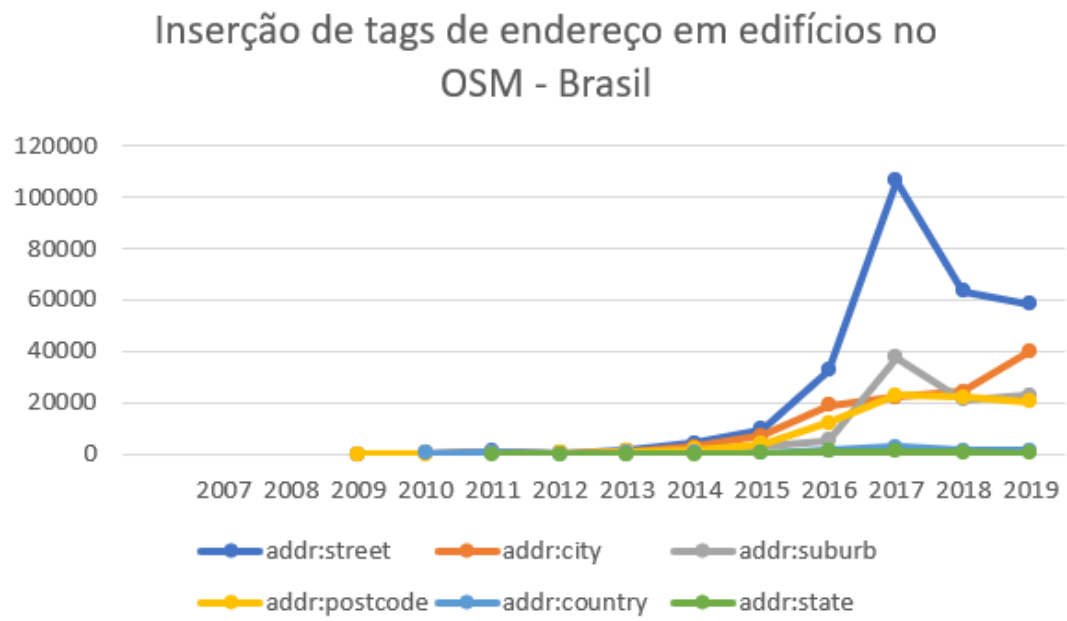


Figura 4.5: Inserção de *Tags* de Endereço em Edifícios no OSM - Brasil.

4.2.2 Inserção de *tags* no OpenStreetMap

Os recursos geográficos do mundo real são representados no OSM como pontos, linhas e polígonos. Atributos temáticos para esses recursos são armazenados na forma de etiquetas geográficas, denominadas de *tags*. No OSM, não há um limite superior para o número de *tags* associadas a qualquer objeto. Embora seja desencorajado, e o software de edição identificará o problema, os objetos não precisam necessariamente receber tags [50].

O uso de *tags* (e seu conjunto de valores recomendados) aumenta a probabilidade de que os dados espaciais sejam entendidos por vários mecanismos de renderização cartográfica que criam visualizações de mapa a partir dos dados do OSM. A etiquetagem adequada e exaustiva de todos os objetos é trabalhosa e demorada, mas é importante para a qualidade geral da coleção de objetos. Dessa forma, a etiquetagem incorreta pode levar a resultados não satisfatórios dos dados geográficos [45].

No OpenStreetMap, as *tags* são apresentadas como um par do tipo chave = valor. A chave é usada para descrever um tópico, categoria ou tipo de recurso (por exemplo, *highway*, *name*). As chaves podem ainda ser qualificadas com prefixos, infixos ou sufixos (geralmente separados por dois pontos :), formando super ou subcategorias. O valor detalha a forma do recurso especificado pela chave. Comumente, valores são texto de forma livre (por exemplo, name = "Eixo Rodoviário Norte"), um conjunto de valores distintos (uma enumeração; por exemplo, highway = primary), vários valores de uma enumeração (separados por um ponto-e-vírgula) ou um número (inteiro ou decimal), como uma distância [56].

4.2.3 Ferramenta Nominatim

O Nominatim é uma ferramenta para pesquisar dados do OpenStreetMap por nome de objetos com o objetivo de capturar informações de endereços de pontos do OSM por meio da técnica da geocodificação reversa, definida como a extração de informações textuais, como um nome ou um endereço, a partir de coordenadas geográficas. Essa técnica é comum em muitos cenários de aplicativos geográficos, por exemplo, em serviços de mapeamento online gratuitos disponíveis [43].

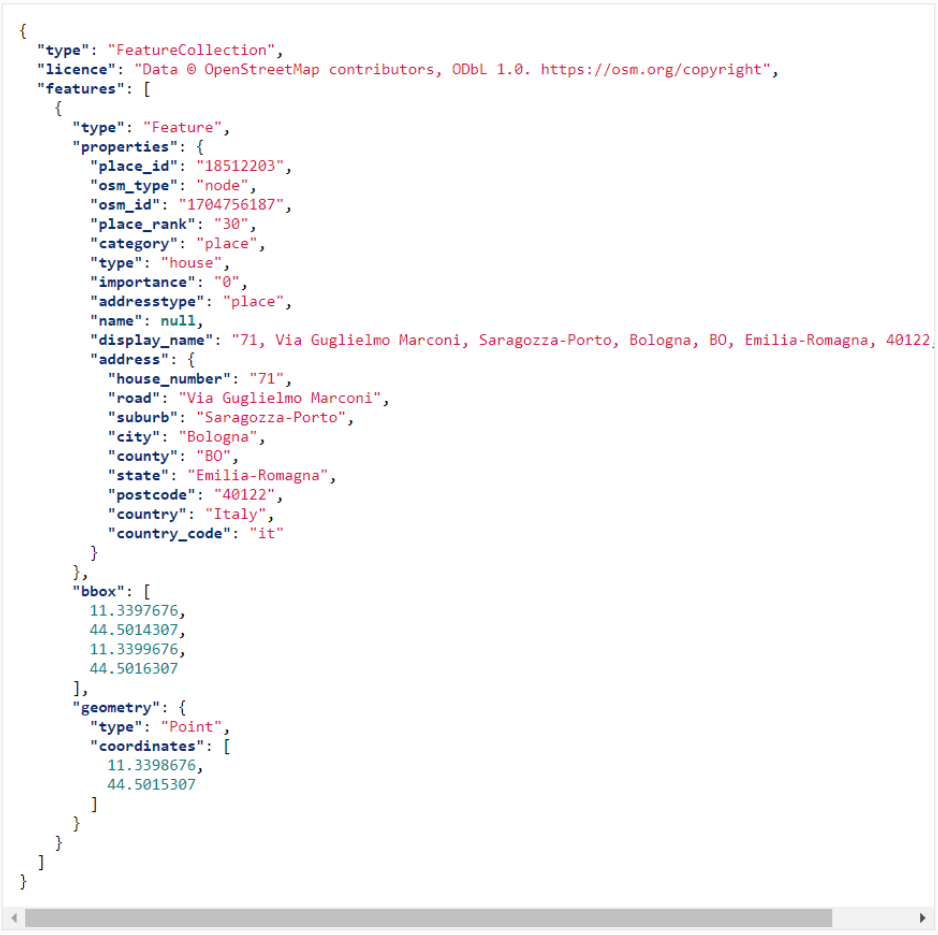
Assim, a partir de um par de coordenadas geográficas selecionado, os dados são gerados no formato *Extensible Markup Language* (*.xml*) ou *JavaScript Object Notation* (*.json*), sendo muito úteis no desenvolvimento de sistemas de informações geográficas para obter dados do local cujos dados são disponibilizados pelo OpenStreetMap com dados abertos, sob a licença *Open Data Commons Open Database License (ODbL)* pela Fundação OpenStreetMap.

A API de geocodificação reversa implementada dentro da ferramenta Nominatim não calcula exatamente o endereço da coordenada que recebe, mas encontra o objeto do OpenStreetMap mais próximo da coordenada solicitada e retorna suas informações de endereço. Isso pode ocasionalmente levar ao acréscimo de informações incorretas à base do OpenStreetMap[6]. O formato padrão para a API de geocodificação reversa da ferramenta Nominatim pode ser traduzido em um endereço *url* do tipo `https://nominatim.openstreetmap.org/reverse?lat=<value>&lon=<value>&<params>`, em que *lat* e *lon* são, respectivamente, a latitude e a longitude de um par de coordenadas na projeção WGS84. A API retorna exatamente um resultado ou um erro quando a coordenada está em uma área sem cobertura de dados do OpenStreetMap.

A Figura 4.6 ilustra o arquivo JSON retornado quando é selecionado deteminado objeto na latitude igual a 44.50155 e longitude igual a 11.33989, correspondendo a um edifício na cidade de Bolonha, na Itália. Conforme pode ser observado, a ferramenta Nominatim

[6]https://nominatim.org/release-docs/develop/api/Reverse/ [Acesso em novembro de 2020.]

traz as informações de endereço por meio do *array address*, o qual contém as seguintes informações: *house_number* (número da casa), *road* (rua), *suburb* (bairro), *city* (cidade), *county* (condado), *state* (estado), *postcode* (código postal), *country* (país) e *country_code* (código do país).



```
{
  "type": "FeatureCollection",
  "licence": "Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright",
  "features": [
    {
      "type": "Feature",
      "properties": {
        "place_id": "18512203",
        "osm_type": "node",
        "osm_id": "1704756187",
        "place_rank": "30",
        "category": "place",
        "type": "house",
        "importance": "0",
        "addresstype": "place",
        "name": null,
        "display_name": "71, Via Guglielmo Marconi, Saragozza-Porto, Bologna, BO, Emilia-Romagna, 40122,
        "address": {
          "house_number": "71",
          "road": "Via Guglielmo Marconi",
          "suburb": "Saragozza-Porto",
          "city": "Bologna",
          "county": "BO",
          "state": "Emilia-Romagna",
          "postcode": "40122",
          "country": "Italy",
          "country_code": "it"
        }
      },
      "bbox": [
        11.3397676,
        44.5014307,
        11.3399676,
        44.5016307
      ],
      "geometry": {
        "type": "Point",
        "coordinates": [
          11.3398676,
          44.5015307
        ]
      }
    }
  ]
}
```


Figura 4.6: Arquivo *.json* Retornado Após Utilizar a Ferramenta Nominatim.

4.2.4 Ferramenta CEP Aberto

A ferramenta CEP Aberto consiste em um projeto que visa prover acesso gratuito e construir de maneira colaborativa uma base de dados com os Códigos de Endereçamento Postal (CEP) geolocalizados de todo o Brasil. Assim como a ferramenta Nominatim, a ferramenta CEP Aberto consegue capturar atributos tais como o endereço e o CEP de objetos do OpenStreetMap. Em seguida, essas informações podem ser utilizadas para serem inseridas no formato de *tags*.

O CEP Aberto possui informação de 1.137.139 CEPs distribuídos em 10.660 cidades e municípios[7]. A URL base da API segue o modelo `https://www.cepaberto.com/api/v3/metodo?parametro=valor[&parametro2=valor2]`, em que o método especifica a busca a ser feita e pode ser igual aos seguintes valores: cep (busca das informações de endereço de um objeto pelo número do CEP); nearest (busca das informações de endereço pela latitude e longitude mais próxima), address (busca das informações de endereço por estado, cidade, bairro ou logradouro), cities (busca das informações de endereço pelo nome da cidade ou município), ou update (método para realizar atualização de CEPs na base da ferramenta CEP Aberto).

Para a utilização da técnica de geocodificação reversa dentro da ferramenta CEP Aberto, é necessária a utilização da API com o método *nearest*. Dessa forma, dado um par latitude e longitude, retorna-se o CEP mais próximo do ponto correspondente a estas coordenadas. A busca limita-se a um raio de 10km a partir do ponto referente às coordenadas passadas como parâmetro. A Figura 4.7 ilustra a busca de informações de endereço para um objeto situado na latitude igual a -20.55 e longitude = -43.63, utilizando a API da ferramenta CEP Aberto em um programa escrito em Python.

```python
import requests

url = "https://www.cepaberto.com/api/v3/nearest"
# O seu token está visível apenas pra você
headers = {'Authorization': 'Token token=66acd61c3fce9cdf2cfde02ccdb71a2e'}
params = {'lat': -20.55, 'lng': -43.63}
response = requests.get(url, headers=headers, params=params)

print(response.json())
```

Figura 4.7: Busca de informações de endereço utilizando a API da ferramenta CEP Aberto.

A Figura 4.8 apresenta o arquivo *.json* retornado após consulta utilizando a ferramenta Cep Aberto para a latitude igual a -20.55 e longitude = -43.63. Conforme pode ser observado, foram extraídas as informações de cidade, ddd, estado, altitude, latitude, longitude, bairro, complemento, CEP e logradouro.


```
{"cidade": {"ibge": "3550308", "nome": "São Paulo", "ddd":
11}, "estado": {"sigla": "SP"}, "altitude": 760.0,
"longitude": "-46.636", "bairro": "Sé", "complemento": "-
lado ímpar", "cep": "01001000", "logradouro": "Praça da Sé",
"latitude": "-23.5479099981"}
```


Figura 4.8: Arquivo *.json* Retornado Após Utilizar a Ferramenta CEP Aberto.

[7]https://cepaberto.com/ [acesso em novembro de 2020]

4.2.5 Base de Dados dos Correios

O Código de Endereçamento Postal (CEP), com estrutura de 5 dígitos, foi criado pela empresa Brasileira de Correios e Telégrafos, em maio de 1971. Sua divulgação ao público em geral ocorreu com a publicação do Guia Postal Brasileiro, na edição de 1971. Em maio de 1992, sua estrutura foi alterada para 8 dígitos e oficializada junto ao público em geral, com a publicação do Guia Postal Brasileiro, edição de 1992[8].

Dessa maneira, o CEP é um conjunto numérico constituído de oito algarismos, cujo objetivo principal é orientar e acelerar o encaminhamento, o tratamento e a distribuição de objetos de correspondências, por meio da sua atribuição a localidades, logradouros, serviços, órgãos públicos, empresas e edifícios. A finalidade do CEP é racionalizar os métodos de separação da correspondência por meio da simplificação das fases dos processos de triagem, encaminhamento e distribuição, permitindo o tratamento mecanizado com a utilização de equipamentos eletrônicos de triagem.

Neste trabalho, a base de dados dos Correios foi utilizada com o objetivo de fazer a inclusão da *tag* de código postal e validação das demais *tags*. A base é composta por mais de 1 milhão de CEPs, sendo atualizada conforme o último censo do IBGE. Os CEPs estão estruturados segundo o sistema decimal, sendo composto de Região, Sub-região, Setor, Subsetor, Divisor de Subsetor e Identificadores de Distribuição, conforme demonstrado na Figura 4.9[9].

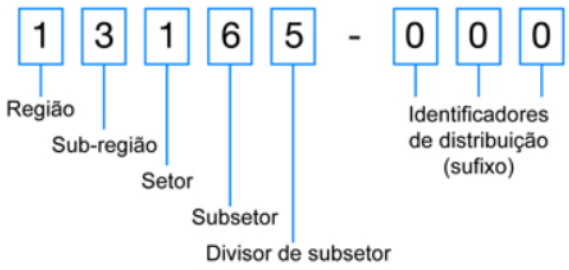


Figura 4.9: Estrutura do Código de Enderaçamento Postal (CEP) no Brasil.

Nesse modelo dos Correios, o Brasil foi dividido em dez regiões para fins de codificação postal, utilizando como parâmetros o desenvolvimento sócio-econômico e fatores de crescimento demográfico de cada Unidade da Federação. A distribuição do CEP foi feita no sentido anti-horário a partir do estado de São Paulo e pode ser observada por meio da Figura 4.10.

[8]https://www.correios.com.br/enviar-e-receber/ferramentas/cep/o-que-e-cep [acesso em dezembro de 2020.]

[9]https://www.correios.com.br/enviar-e-receber/ferramentas/cep/estrutura-do-cep [Acesso em novembro de 2020.]

Figura 4.10: Distribuição dos CEPS no Brasil por Região Postal.

4.3 Utilização da QualiOSM

No primeiro acesso à ferramenta QualiOSM, o usuário deverá realizar a instalação do *plugin* dentro do editor JOSM. Para isso, ele deverá selecionar no menu principal a opção 'Editar' e em seguida deverá abrir a caixa de diálogos de 'Preferências'. Feito isso, o usuário deverá selecionar a opção 'Configurar plugins disponíveis', representada pelo símbolo de uma tomada no menu lateral esquerdo, conforme apresentado na Figura 4.11. A seguir, o usuário deverá buscar o plugin QualiOSM na caixa de pesquisa e selecioná-lo.

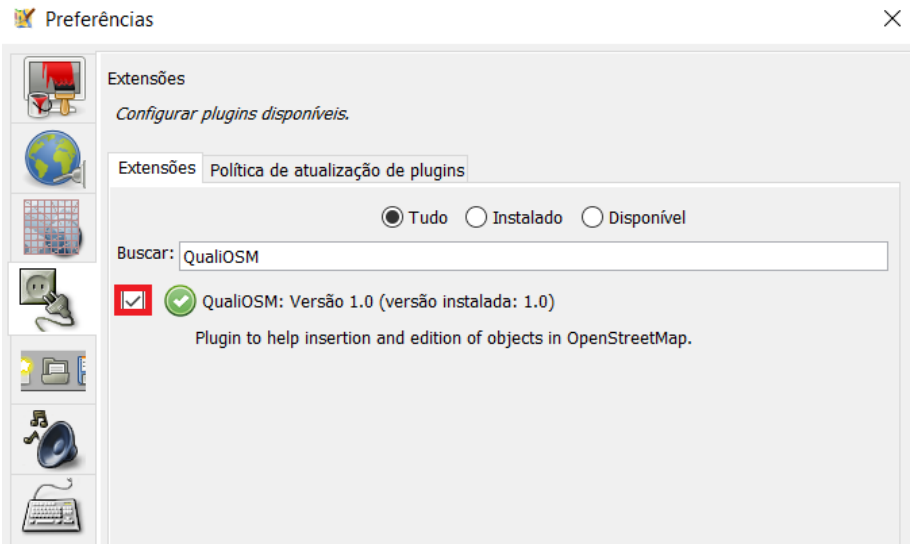


Figura 4.11: Instalação do *Plugin* QualiOSM.

Feita a instalação do *plugin*, o usuário será capaz de utilizar a funcionalidade do

adicionador de *tags* de endereço para os objetos da ferramenta OpenStreetMap. Para isso, o usuário deverá realizar a carga dos dados da ferramenta OpenStreetMap utilizando-se da interface do JOSM e selecionar os objetos em que deseja realizar a inserção das *tags* de endereço. O usuário poderá realizar tal inserção utilizando a ferramenta Nominatim, CEP Aberto ou a base de dados dos Correios.

Após a instalação do *plugin* QualiOSM, ficará disponível para o usuário selecionar a opção "QualiOSM" dentro do menu principal. Conforme pode ser observado a partir da Figura 4.12, os botões para adicionar as *tags* de endereço com a ferramenta Nominatim ("Add address tags - Nominatim"), com a ferramenta CEP Aberto ("Add address tags - CEP Aberto") ou com a base de dados dos Correios ("Add address tags - Correios") só ficam ativos quando existem objetos selecionados no mapa, o que foi possível por meio da implementação do método *updateEnabledState()*. Além disso, também foi acrescentado um botão para que o usuário apague as *tags* de endereço de objetos selecionados ("Clean address tags") para o caso em que o usuário queira fazer a exclusão desse tipo de informação quando encontrar *tags* de endereço atribuídas erroneamente a objetos.

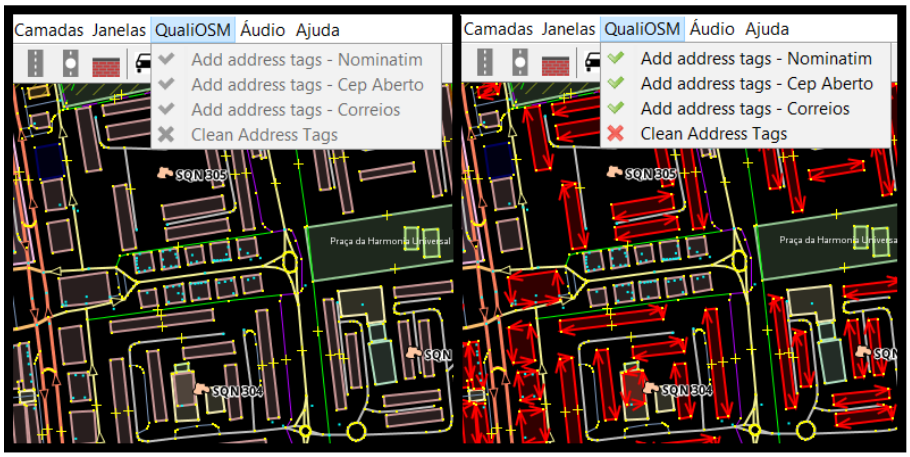


Figura 4.12: Mudança de Estado dos Botões no QualiOSM Após Seleção de Objetos.

Ao utilizar o *plugin* QualiOSM, o usuário pode optar por fazer o carregamento de uma camada de imagem aérea ou não, a fim de melhorar a visibilidade dos dados. Além disso, é possível que o usuário faça a seleção de múltiplos objetos utilizando-se do painel "Predefinições", contido na barra de ferramentas. Neste trabalho, os testes foram realizados selecionando a predefinição de "Edifícios Residenciais", correspondente aos casos em que a *tag building* está associada aos seguintes valores: "yes" (sim), "house" (casa), "apartment" (apartamento) ou "residential" (residencial). Os resultados dos testes realizados encontram-se descritos no Capítulo 5: Resultados.

Capítulo 5

Resultados

Este capítulo apresenta os resultados obtidos a partir da execução da ferramenta QualiOSM nos diferentes cenários de teste. Dessa forma, o capítulo está estruturado da seguinte forma: a Seção 5.1 apresenta a definição dos cenários de teste; a Seção 5.2 apresenta os resultados obtidos para o Cenário I; a Seção 5.3 apresenta os resultados obtidos para o Cenário II; a Seção 5.4 apresenta os resultados obtidos para o Cenário III; a Seção 5.5 apresenta os resultados obtidos para o Cenário IV. Além disso, a Seção 5.1 apresenta a discussão dos resultados obtidos e a Seção 5.8 apresenta os resultados acadêmicos obtidos a partir da realização deste mestrado.

5.1 Definição dos Cenários de Teste

A ferramenta QualiOSM foi testada em diferentes cenários de teste realizando a extração dos dados no formato padrão do OpenStreetMap em diferentes áreas de interesse, descritas a seguir:

- Cenário I: corresponde à parte da região administrativa do Plano Piloto, na cidade de Brasília. O centro da capital do Brasil é conhecido por ser uma cidade planejada, em que os prédios são dispostos de forma organizada e não muito próximos uns dos outros. Os dados foram coletados dentro da seguinte caixa delimitadora: latitude mínima = -15,7929; latitude máxima = -15,7322; longitude mínima = -47,9093; longitude máxima = -47,8561.
- Cenário II: corresponde à parte da cidade do Rio Branco, no estado do Acre (AC). Esta região foi escolhida com base no projeto "Mapping Flood Prone Urban Areas in Brazil"(Mapeamento de Áreas Urbanas Propensas a Inundações no Brasil), disponível na ferramenta Hot Tasking Manager[1] Conforme pode ser observado, neste

[1]https://tasks.hotosm.org/projects/6124 [Acesso em outubro de 2020].

cenário as casas estão dispostas muito mais próximas umas das outras, tornando a tarefa de mapear as edificações mais desafiadora. Os dados foram coletados dentro da seguinte caixa delimitadora: latitude mínima = -9,9903; latitude máxima = -9,9733; longitude mínima = -67,8242; longitude máxima = -67,8021.

- Cenário III: Considera a parte periferica das cidades interioranas de Mogi das Cruzes e Suzano, no estado de São Paulo. Os dados foram coletados dentro da seguinte caixa delimitadora: latitude mínima = -23.6901; latitude máxima = -23.6253; longitude mínima = -46.3922; longitude máxima = -46.2971.
- Cenário IV: Considera a area pertencente à comunidade da Rocinha, no estado do Rio de Janeiro. Nesse ambiente os edifícios encontram-se muito próximos uns dos outros, com a presença inclusive de edifícios sobrepostos. Os dados foram coletados dentro da seguinte caixa delimitadora: latitude mínima = -22.9896; latitude máxima = -22.9866; longitude mínima = -43.2498; longitude máxima = -43.2397.

Os dados de cada cenário foram baixados respeitando as caixas delimitadoras de cada área. Depois, foram realizadas simulações de como seriam as inclusões das *tags* de endereço *addr:city, addr:building e addr:suburb* utilizando as ferramentas de geocodificação reversa Nominatim e CEP Aberto. Em seguida, a base de dados de CEPS dos Correios foi utilizada a fim de verificar a acurácia das informações inseridas em relação à *tag addr:postcode*, relativa à informação de código postal. Para realizar as simulações, foram selecionados os objetos caracterizados como "edifícios residenciais" em cada cenário observado. Dentro da ferramenta OpenStreeMap, um objeto será do tipo edifício residencial quando a *tag building* estiver associada aos seguintes valores: "*yes*" (sim), "*residential*" (residencial), "*house*" (casa) ou "*apartment*" (apartamento).

5.2 Cenário I:

O adicionador de *tags* foi aplicado de forma a simular uma inclusão de atributos de endereço utilizando as ferramentas de geocodificação reversa Nominatim e CEP Aberto. Para o Cenário I: Ambiente Urbano I, foi escolhida parte da Região Administrativa do Plano Piloto em Brasília. O resultado das simulações pode ser observado a partir da Tabela 5.1, onde pode-se perceber que a ferramenta Nominatim provou ser eficiente para a inclusão das *tags addr:city* (cidade) e *addr:suburb* (bairro), atingindo todos os 210 edifícios encontrados nesse cenário. As *tags addr:building* (nome do edifício) e *addr:neighbourhood* (utilizada para logradouro) também obtiveram melhoria, com o acréscimo de informações em 67,62% e 91,43%, respectivamente.

A ferramenta CEP Aberto obteve um desempenho um pouco inferior em relação à ferramenta Nominatim, promovendo um acréscimo de 98,09% de edifícios associados com

a tag addr:city e 94,76% de acréscimo de objetos associados com as *tags addr:suburb* e *addr:neighbourhood*. A ferramenta CEP Aberto não continha os dados referentes aos nomes de edifícios separados de seu respectivo logradouro e por esse motivo, não houve alteração para a simulação da inclusão da *tag addr:suburb*.

Tabela 5.1: Simulação da Inserção de *Tags* de Endereço no Cenário I.

	Antes	**Nominatim**	**CEP Aberto**
addr:city	0,48%	100%	98,57%
addr:building	0%	67,62%	0%
addr:suburb	0%	100%	94,76%
addr:neighbourhood	0%	91,43%	94,76%

Também foi realizada uma análise em relação à adição da *tag addr:postcode*, uma vez que durante as simulações, foi detectada a inclusão de códigos postais errados ao se utilizar as ferramentas Nominatim e CEP Aberto. Ao se utilizar a base de dados dos Correios, apesar de reduzir o percentual de objetos associados com a *tag* de código postal, a ferramenta garante que os CEPS inseridos estarão corretos. Conforme pode ser observado a partir da Figura 5.1, após a simulação da adição da *tag addr:postcode* em edifícios para o Cenário I, foi verificado que a ferramentas Nominatim e CEP Aberto estavam acrescentando mais informações erradas do que informações corretas. Utilizando a ferramenta Nominatim, 96,15% dos edifícios foram associados com uma informação de código postal errada e 3,85% foram associados com a *tag* de forma correta. No caso da ferramenta CEP Aberto, apenas 17,31% dos edifícios foram corretamente associados com a *tag* de código postal, enquanto 26,92% foram associados de forma errada. Já utilizando a base de dados dos Correios, 67,31% dos 210 edifícios foram associados corretamente à *tag* de código postal e não houve acréscimo de informações incorretas. Além disso, 32,69% dos edifícios permaneceram inalterados.

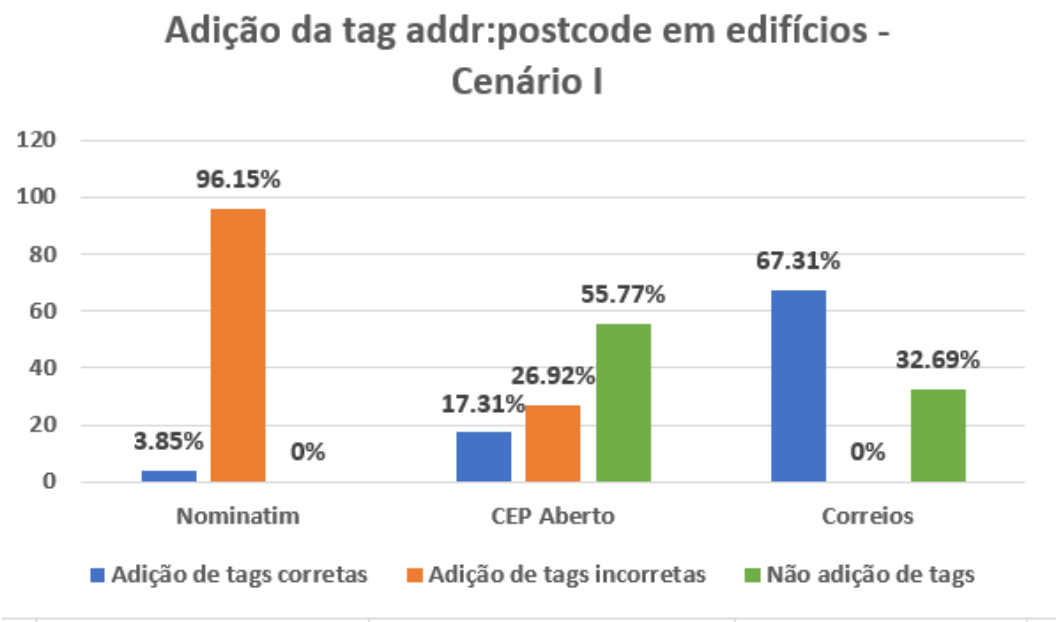


Figura 5.1: Gráfico da Simulação para a Adição da *Tag* de CEP - Cenário I.

A Figura 5.1 ilustra a simulação da adição da *tag* de código postal utilizando a base de dados dos Correios. Antes da execução da ferramenta, havia 0,95% de edifícios associados com a *tag addr:postcode*, enquanto 67,31% ficaram associados com a *tag* após a execução da ferramenta. Os edifícios associados com a *tag* de código postal estão marcados em vermelho.

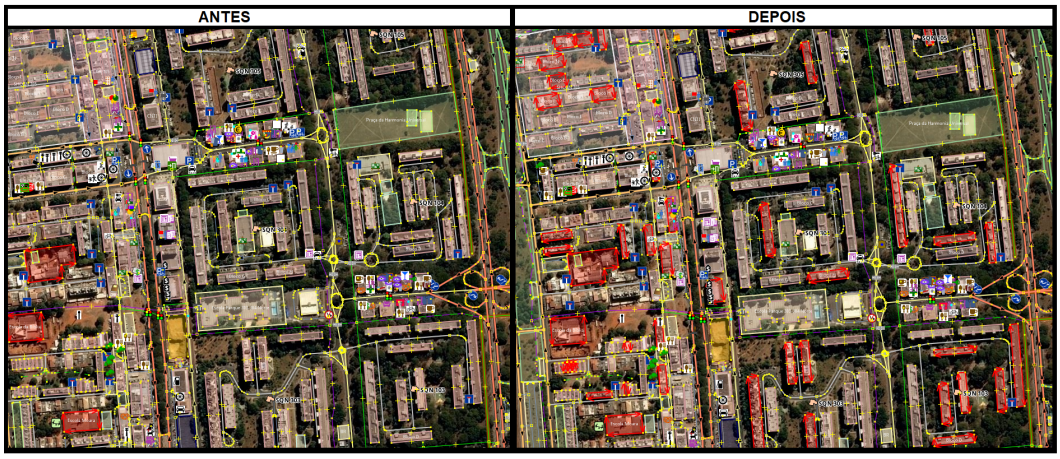


Figura 5.2: Simulação para a Adição da *Tag addr:postcode* - Cenário I.

5.3 Cenário II:

Para o Cenário II, foi escolhida parte da cidade do Rio Branco, no estado do Acre. O resultado das simulações pode ser observado a partir da Tabela 5.2 e, conforme pode ser observado, a ferramenta Nominatim provou ser eficiente apenas para a inclusão da *tag addr:city*, a qual foi acrescentada em 100% dos 4.049 edifícios residenciais encontrados no cenário. A *tag addr:suburb* foi associada a 0,02% e a *tag addr:neighbourhood* foi associada a 4,57% dos edifícios. A ferramenta CEP Aberto também não obteve bom resultado, acrescentando as *tags addr:city, addr:neighbourhood* e *addr:suburb* em aproximadamente 12% dos edifícios.

Tabela 5.2: Simulação da Inserção de *Tags* de Endereço no Cenário II.

	Antes	**Nominatim**	**CEP Aberto**
addr:city	0,1%	100%	12,15%
addr:building	0%	0%	0%
addr:suburb	0,02%	0,02%	12,08%
addr:neighbourhood	0%	4,57%	12,05%

Realizando a análise em relação à adição da *tag addr:postcode*, utilizando a base de dados dos Correios, obtém-se os resultados ilustrados a partir da Figura 5.3. Após a simulação da adição da *tag addr:postcode* em edifícios para o Cenário II, foi verificado que a ferramentas Nominatim e CEP Aberto estavam acrescentando mais informações erradas do que informações corretas. Com a ferramenta Nominatim 96,39% dos edifícios foram associados com uma informação de código postal errada e apenas 3,61% dos edifícios foram associados com a informação correta. No caso da ferramenta CEP Aberto, 0,2% dos edifícios foram corretamente associados com a *tag* de código postal, 0,1% dos edíficios foram associados com a *tag* incorreta e 99,7% dos edifícios permaneceram inalterados. Já utilizando a base de dados dos Correios, 39,34% dos edifícios foram associados corretamente à *tag* de código postal, 60,66% dos edifícios permaneceram inalterados e não houve acréscimo de informações incorretas.

A Figura 5.4 ilustra a simulação da adição da *tag* de código postal utilizando a base de dados dos Correios no Cenário II. Antes da execução da ferramenta, havia apenas 2 edifícios associados com a *tag addr:postcode*, enquanto 39,34% ficaram associados com a *tag* após a execução da ferramenta. Os edifícios associados com a *tag* de código postal estão marcados em vermelho. Conforme pode ser observado a partir da Figura 5.3, enquanto a ferramenta Nominatim possui uma baixa acurácia para as informações de código postal, a ferramenta CEP Aberto possui uma baixa completude, uma vez que não conseguiu completar as informações de código postal para 99,7% dos edifícios. Portanto, a inclusão

da *tag addr:postcode* apresentou os melhores resultados ao utilizar a base de dados dos Correios.

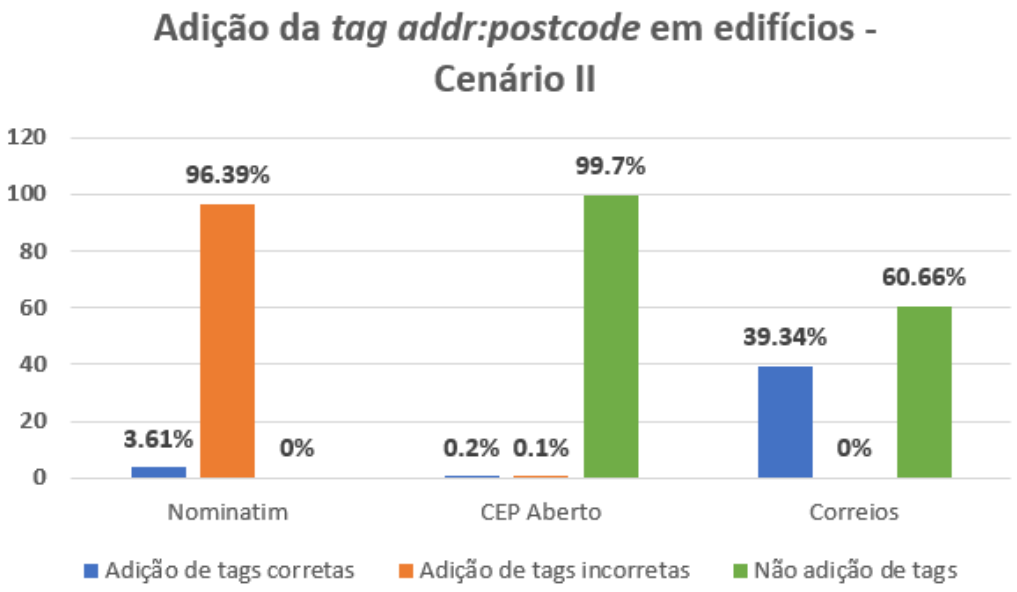


Figura 5.3: Gráfico da Simulação para a Adição da *Tag* de CEP - Cenário II.

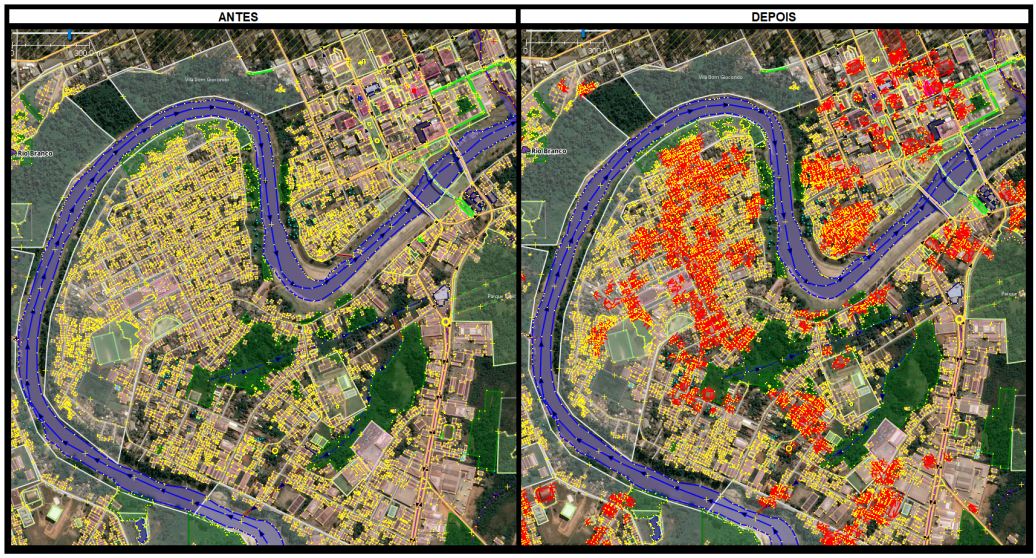


Figura 5.4: Simulação para a Adição da *Tag addr:postcode* - Cenário II.

5.4 Cenário III

Para o Cenário III, foi escolhida parte da zona rural correspondente entre as cidades interioranas de Mogi das Cruzes e Suzano, no estado de São Paulo. O resultado das simulações pode ser observado a partir da Tabela 5.3 e, conforme pode ser observado, a ferramenta Nominatim provou ser eficiente para a inclusão da *tag addr:city*, que foi associada a 100% dos 47 edifícios residenciais encontrados no cenário. A *tag addr:suburb* foi associada a 70,21% dos edifícios e a *tag addr:neighbourhood* foi associada a apenas 2,13% dos edifícios. Já a ferramenta CEP Aberto teve um desempenho insatisfatório, associando as *tags* de endereço a apenas 4 edifícios, correspondentes a 8,51% dos edifícios encontrados no cenário. Não ocorreram mudanças em relação à *tag addr:building* devido à falta dessa informação tanto na ferramenta Nominatim quanto na ferramenta CEP Aberto em relação ao cenário observado.

Tabela 5.3: Simulação da Inserção de *Tags* de Endereço no Cenário III.

	Antes	**Nominatim**	**CEP Aberto**
addr:city	0%	100%	8,51%
addr:building	0%	0%	0%
addr:suburb	0%	70,21%	8,51%
addr:neighbourhood	0%	2,13%	8,51%

Realizando a análise em relação à adição da *tag addr:postcode* e utilizando a base de dados dos Correios, obtém-se os resultados ilustrados na Figura 5.5. Conforme pode ser observado, após a simulação da adição da *tag addr:postcode* em edifícios para o Cenário III, foi verificado novamente que a ferramenta Nominatim estava acrescentando mais informações erradas do que informações corretas. Com a ferramenta Nominatim, 70,21% dos edifícios foram associados com uma informação de código postal errada e 29,79% foram associados com a informação correta. No caso da ferramenta CEP Aberto, 14,89% dos edifícios foram corretamente associados com a *tag* de código postal, 85,11% dos edifícios permaneceram inalterados e não ocorreu acréscimo de informações incorretas. Já utilizando a base de dados dos Correios, 40,82% dos 47 edifícios foram associados corretamente à *tag* de código postal, 59,18% dos edifícios permaneceram inalterados e também não ocorreu acréscimo de informações incorretas.

A Figura 5.6 ilustra a simulação da adição da *tag* de código postal utilizando a base de dados dos Correios no Cenário III. Antes da execução da ferramenta, não havia nenhum edifício associado com a *tag addr:postcode*, enquanto 40,82% ficaram corretamente associados com a *tag* após a execução da ferramenta. Neste cenário, nota-se a presença de poucos edifícios mapeados em uma área extensa, confirmando a hipótese de que áreas

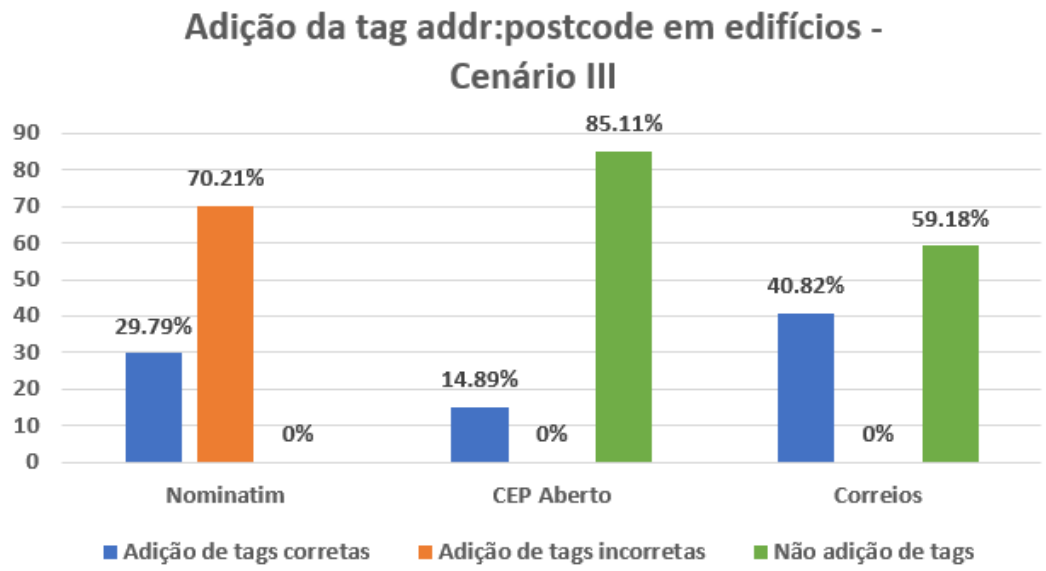


Figura 5.5: Gráfico da Simulação para a Adição da *Tag* de CEP - Cenário III.

rurais costumam possuir menor detalhamento do que áreas urbanas dentro de ferramentas colaborativas. Ainda assim, a ferramenta conseguiu atribuir uma boa porcentagem de *tags* de forma correta à base de dados do OpenStreetMap.

Figura 5.6: Simulação para a Adição da *Tag addr:postcode* - Cenário III.

5.5 Cenário IV

Para o Cenário IV, foi escolhida a região correspondente à comunidade da Rocinha, no estado do Rio de Janeiro. Conforme pode ser observado a partir da Tabela 5.4, a ferramenta Nominatim conseguiu contribuir em 100% para a inclusão da *tag addr:city* e *addr:suburb*, atingindo todos os 280 edifícios encontrados nesse cenário. A *tag addr:building* também obteve uma boa melhoria, com o acréscimo de *tags* em 67,62% dos edifícios associados.

Em relação à ferramenta CEP Aberto, ocorreu um acréscimo de 25,36% de edifícios associados com a *tag addr:suburb* e 26,78% de edifícios associados com a *tag addr:neighbourhood*. Dessa forma, verifica-se mais uma vez que a ferramenta Nominatim provou ser mais eficiente para a inclusão de *tags* de endereço em comparação com a ferramenta CEP Aberto.

Tabela 5.4: Simulação da Inserção de *Tags* de Endereço no Cenário IV.

	Antes	**Nominatim**	**CEP Aberto**
addr:city	0.71%	100%	0.71%
addr:building	0%	67,62%	0%
addr:suburb	0.71%	100%	26.07%
addr:neighbourhood	0%	0%	26.78%

Realizando a análise em relação à adição da *tag addr:postcode* e utilizando a base de dados dos Correios, foram obtidos os resultados presentes na Figura 5.7. Conforme pode ser observado, após a simulação da adição da *tag addr:postcode* em edifícios para o Cenário IV, foi verificado que a ferramenta Nominatim apenas inseriu informações de código postal incorretas. Após a simulação da inserção de *tags* utilizando a ferramenta CEP Aberto, apenas 0.36% dos edifícios tiveram suas *tags* associadas de forma correta; 26,07% dos edifícios foram associados com *tags* incorretas e 73,57% dos edifícios permaneceram inalterados. Utilizando a base de dados dos Correios, 9,29% dos edifícios foram associados à *tags* de forma correta, 90,71% dos edifícios permaneceram inalterados e não ocorreu acréscimo de informações incorretas.

A Figura 5.8 ilustra a simulação da adição da *tag* de código postal utilizando a base de dados dos Correios no Cenário IV. Antes da execução da ferramenta, havia apenas um único edifício associado com a *tag addr:postcode*, enquanto 9,29% dos edifícios foram corretamente associados com a *tag* após a execução da ferramenta. Os edifícios associados com a *tag* de código postal estão marcados em vermelho.

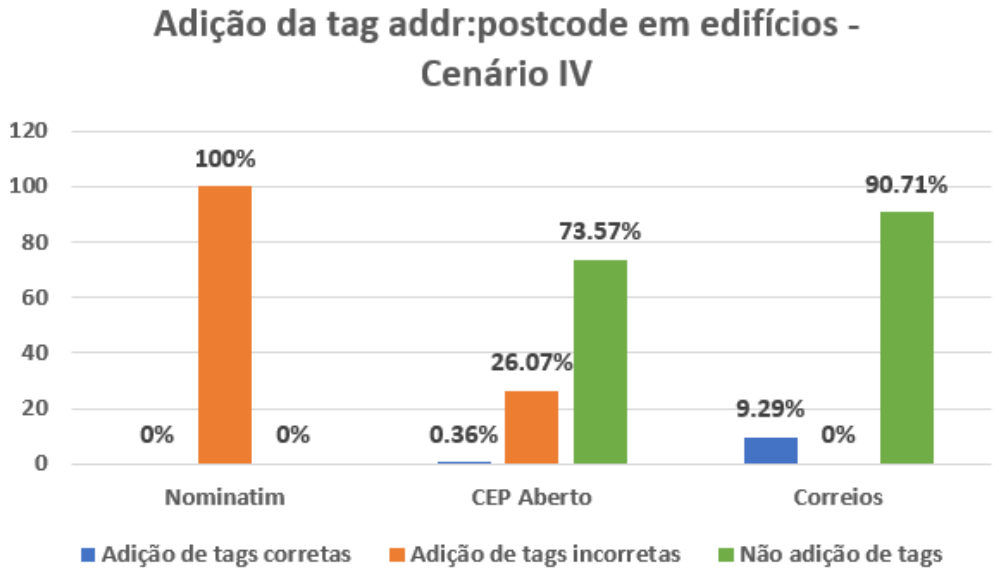


Figura 5.7: Gráfico da Simulação para a Adição da *Tag* de CEP - Cenário IV.

5.6 Discussão dos Resultados

Para a melhoria da qualidade dos dados dentro da ferramenta OpenStreetMap, fez-se necessário recorrer a outras ferramentas, e dessa maneira, a arquitetura QualiOSM tornou-se dependente das dimensões da qualidade existentes nessas ferramentas. A ferramenta Nominatim, apesar de ter uma boa completude e tentar preencher completamente as informações de código postal, possui uma baixa acurácia, o que faz com que sejam acrescentadas informações erradas à base de dados do OSM. A ferramenta CEP Aberto, por sua vez, não possui uma boa completude quando comparada à ferramenta Nominatim e nota-se que há muitas informações faltantes na ferramenta, especialmente em relação às informações de código postal.

Entre as três abordagens utilizadas para a inclusão da *tag* de código postal, a que se mostrou mais eficiente foi a utilização da base de dados dos Correios, uma vez que a base é acurada e assim diminui as chances do acréscimo de informações erradas dentro do OpenStreetMap. Todavia, uma das dificuldades existentes para a aplicação dessa abordagem é o fato de que nem todos os logradouros situados no Brasil encontram-se georreferenciados com a disposição de suas respectivas coordenadas geográficas, o que compromete a completude da base de dados. Dessa forma, uma base de dados completa e acurada de edifícios seria de fundamental importância para a melhoria da qualidade dos dados geográficos colaborativos no Brasil.

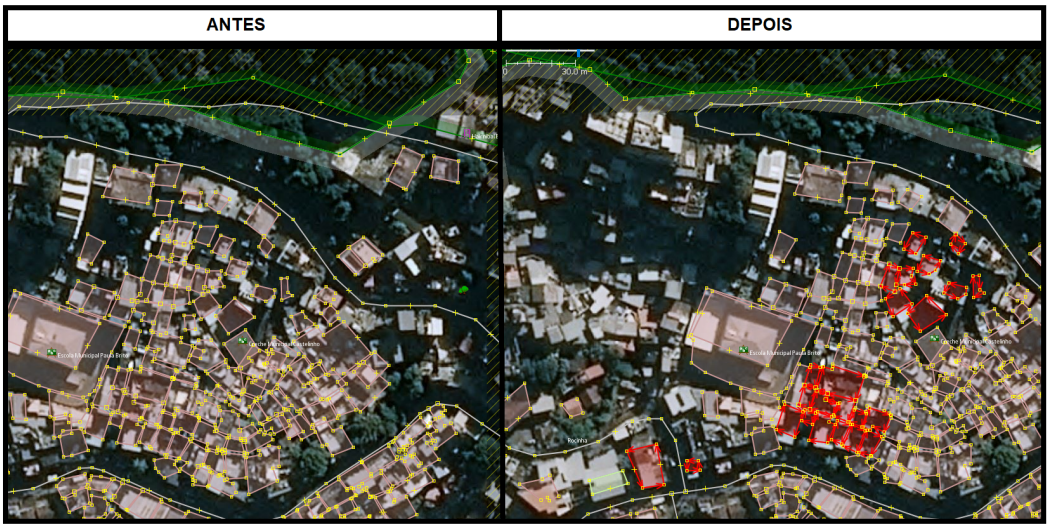


Figura 5.8: Simulação para a Adição da *Tag addr:postcode* - Cenário IV.

5.7 Limitações da Arquitetura QualiOSM

Foram observadas algumas limitações durante a utilização da arquitetura QualiOSM. A ferramenta apresentou os melhores resultados para o Cenário I, pois a área central da cidade de Brasília corresponde a uma região totalmente planejada, em que os edifícios estão corretamente associados a um CEP bem definido dentro da base de dados dos Correios. Entretanto, em relação ao Cenário IV, foi observado que a base de dados dos Correios não é completa, pois trata-se de uma área caracterizada pela existência de moradias precárias e poucos edifícios formais.

Em relação ao Cenário III, correspondente a uma região rural, a ferramenta QualiOSM também obteve limitações, uma vez que para áreas rurais, os Correios geralmente utiliza um CEP geral e assim, muitos edifícios não estão associados a um código postal exclusivo. Como o CEP para área rural tem a finalidade exclusiva para ser usado no cadastramento de endereços rurais, ele foi associado a toda extensão rural de cada distrito (localidade) integrante da codificação postal do município.[2].

Além disso, durante a inclusão das *tags* de endereço foram priorizados os edificícios residenciais, associados com as *tags residential, yes, house e apartment*, o que pode ter limitado a atuação da ferramenta, pois muitos usuários da ferramenta OpenStreetMap realizam o mapeamento de edifícios sem utilização dessas *tags*.

A ferramenta CEP Aberto também possui outra limitação em relação à ferramenta Nominatim, pois enquanto a segunda pode ser utilizada em qualquer país, a primeira é de uso exclusivo do Brasil. Por tanto, para realizar a extensão da arquitetura QualiOSM a fim

[2]https://www.correios.com.br/enviar-e-receber/ferramentas/cep/cep-para-areas-rurais [Acesso em dezembro de 2020].

de atender outras localidades fora do Brasil, seria necessário realizar algumas adaptações e utilizar outras bases de códigos postais para fazer as devidas validações.

5.8 Resultados Acadêmicos

Ao longo deste projeto de mestrado, foram aceitos três artigos em diferentes conferências abrangendo as áreas de Banco de Dados e GeoInformática. O trabalho [47] foi publicado na 7th World Conference on Information Systems and Technologies - WorldCist 2019 e apresentou a Revisão Sistemática da Literatura realizada em relação à questão da qualidade dos dados nos SIG e SIGV. O artigo [48] foi publicado no 35º Simpósio Brasileiro de Banco de Dados (SBBD 2020) e apresentou a ferramenta desenvolvida QualiOSM, realizando uma comparação entre os diferentes cenários de teste. Por fim, o artigo *QualiOSM: Improving Data Quality in the Collaborative Mapping Tool OpenStreetMap*, ainda não publicado, foi aceito no 21º Simpósio Brasileiro de GeoInformática (GEOINFO 2020) e detalhou um pouco mais o funcionamento do adicionador de *tags* desenvolvido e realizou uma comparação entre os cenários urbanos da cidade de Brasília e Rio Branco.

Capítulo 6

Conclusão e Trabalhos Futuros

A arquitetura QualiOSM foi desenvolvida com o objetivo de melhorar a completude das informações de endereço dentro da ferramenta OpenStreetMap. A ferramenta implementada na forma de *plugin* para o editor de dados JOSM apresentou bons resultados em relação à inclusão de novas informações referentes ao nome da cidade, bairro e logradouro em objetos da plataforma do OSM, contribuindo assim para uma melhoria no percentual de objetos com etiquetas de endereço associadas. Esse fato foi observado especialmente em relação aos dois cenários urbanos utilizados de teste, onde o nível de mapeamento é melhor quando comparado ao nível de mapeamento em regiões rurais ou periféricas.

Além disso, foi verificado que entre as três abordagens utilizadas para a inclusão da informação de código postal, a que se mostrou mais eficiente foi a abordagem utilizando a base de dados dos Correios, uma vez que a ferramenta Nominatim mostrou-se ter baixa acurácia assim contribuído para a inclusão de informações incorretas ao OpenStreetMap. A ferramenta CEP Aberto, por sua vez, demonstrou possuir uma baixa completude, uma vez que possui uma alta porcentagem de informações faltantes que pudessem contribuir com a inclusão da *tag* de código postal.

Como trabalho futuro, pretende-se explorar outras *tags* além das *tags* de endereço utilizadas neste trabalho, bem como utilizar outras ferramentas além da Nominatim e CEP Aberto para obtenção de novas informações de objetos do OSM. Pretende-se também testar a ferramenta em outros cenários e avaliar outras dimensões de qualidade em sistemas colaborativos, bem como testar a aplicação da arquitetura implementada em outros editores fora do JOSM. Além disso, a utilização da base de dados dos Correios pode ser utilizada para outros propósitos, como por exemplo, para a identificação e mapeamento de edifícios dentro da ferramenta OpenStreetMap.

Referências

[1] Ahlers, D.: *Assessment of the accuracy of geonames gazetteer data.* Proceedings of the 7th Workshop on Geographic Information Retrieval (GIR '13). ACM, New York, NY, USA, páginas 74–81, 2013. 23

[2] Ahmed, M., B. T. Fasy e C. Wenk: *Local persistent homology based distance between maps.* Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, páginas 43–52, 2014. 23

[3] Ali, A. L., F. Schmid, R. Al-Salman e T. Kauppinen: *Ambiguity and plausibility: Managing classification quality in volunteered geographic information.* Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, páginas 143–152, 2014. 23

[4] Ames, M. e M. Naaman: *Why we tag: Motivations for annotation in mobile and online media.* ACM SIGCHI Conf. Human Factors in Computing Systems, página 971–980, 2007. 23, 24

[5] Antoniou, V. e C. Schlieder: *Participation patterns, VGI and gamification.* Proceedings of AGILE 2014. Presented at the AGILE, páginas 3–6, 2014. 16

[6] Batini, C., M. Palmonari e G. Viscusi: *The many faces of information and their impact on information quality.* In: Proceedings of the 17th international conference on information quality, 2012. 16

[7] Bennett, J.: *OpenStreetMap: Be your own cartographer.* Packt Publishing, 2010. 13, 15

[8] Biagioni, J. e J. Eriksson: *Map inference in the face of noise and disparity.* Proceedings of the 20th International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, páginas 79–88, 2012. 23

[9] Câmara, G., M.A. Casanova e G.C. Magalhães: *Anatomia de sistemas de informação geográfica.* 1996. ix, 9

[10] Camara, J. H. S., L. F. M. Vegi, R. O. Pereira, Z .A. Geocze e J.: Lisboa-Filho: *Click-onmap: A platform for development of volunteered geographic information systems.* 12th Iberian Conference on Information Systems and Technologies (CISTI), páginas 1–6, 2017. 23

[11] Celino, I.: *Human computation VGI provenance: Semantic web-based representation and publishing.* IEEE Transactions on Geoscience and Remote Sensing, 51(11):5137–5144, 2013. 23

[12] Che, Y., K. Chiew, X. Hong e Q. He: *Sals: semantics-aware location sharing based on cloaking zone in mobile social networks.* Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems. ACM, New York, NY, USA, páginas 49–56, 2012. 23

[13] Chiang, Y., S. Leyk, N. H. Nazari, S. Moghaddam e T. X.: Tan: *Assessing the impact of graphical quality on automatic text recognition in digital maps.* Computers Geosciences, 93:21–35, 2016. 23

[14] Codescu, M., G. Horsinka, O. Kutz, T. Mossakowski e R. Rau: *Osmonto-an ontology of OpenStreetMap tags.* State of the map Europe (SOTM-EU), 2011, 2011. 23, 24

[15] Davidovic, N., P. Mooney, L. Stoimenov e M. Minghini: *Tagging in volunteered geographic information: an analysis of tagging practices for cities and urban regions in OpenStreetMap.* ISPRS International Journal of Geo-Information, 5(12):232, 2016. 23, 25

[16] Delling, D., A.V. Goldberg, M. Goldszmidt, J. Krumm, K. Talwar e R.F. Werneck: *Navigation made personal: inferring driving preferences from gps traces.* Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, 2015. 23

[17] Demetriou, D.: *Uncertainty of OpenStreetMap data for the road network in cyprus.* Fourth International Conference on Remote Sensing and Geoinformation of the Environment, Paphos, Cyprus, 2016. 23

[18] Doan, A., R. Ramakrishnan e A. Y. Halevy: *Crowdsourcing systems on the world-wide web.* Communications of the ACM, 54(4):86–96, 2011. 11, 12

[19] Dueker, K. J. e D. Kjerne: *Multipurpose Cadastre: Terms and Definitions.* Falls Church VA, American Society for Photogrammetry and Remote Sensing and American Congress on Surveying and Mapping, 1989. 1

[20] Estellés-Arolas, E. e F. González-Ladrón-de Guevara: *Towards an integrated crowdsourcing definition.* Journal of Information Science, 38(2):189–200, 2012. 11

[21] Estima, J., C.C. Fonte e M. Painho: *Comparative study of land use/cover classification using flickr photos, satellite imagery and corine land cover database.* 2014. 16

[22] Estima, J. e M. Painho: *Exploratory analysis of OpenStreetMap for land use classification.* Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD'13). ACM, New York, NY, USA, páginas 39–46, 2013. 23

[23] Firmani, D., M. Mecella, M. Scannapieco e C. Batini: *On the meaningfulness of "Big Data quality" (invited paper).* Data Science and Engineering, 1(1):6–20, 2016. 16

[24] Fonte, C. C., V. Antoniou, L. Bastin, J. Estima, J. J. Arsanjani, J.C.L. Bayas, L. See e R. Vatseva: *Assessing VGI data quality.* Mapping and the citizen sensor, páginas 137–163, 2017. 16

[25] Foody, G. M., Fellow, IEEE e D.S.: Boyd: *Using volunteered data in land cover map validation: Mapping west african forests.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,, 6(3), 2013. 23

[26] Fowler, M.: *Patterns of enterprise application architecture*. Addison-Wesley Longman Publishing Co., Inc., 2002. 26

[27] Galarus, D. E. e R. A Angryk: *A smart approach to quality assessment of site-based spatio-temporal data*. IEEE International Conference on Big Data, páginas 2636–2645, 2016. 23

[28] Goodchild, M. F.: *Citizens as sensors: The world of volunteered geography*. GeoJournal, 69(4):211 – 221, 2007. 1, 10

[29] Gray, A.: *Euclidean Spaces*. Modern Differential Geometry of Curves and Surfaces with Mathematica, Chapman and Hall/CRC, 3rd edition, 2006. 5

[30] Güting, R. H.: *An introduction to spatial database systems*. The VLDB Journal, 3(4):357–399, 1994. 6, 8

[31] Hassan, A., R. Jones e F. Diaz: *A case study of using geographic cues to predict query news intent*. Proceedings of the 17th ACM Sigspatial International Conference on Advances in Geographic Information Systems, páginas 33–41, 2009. 23

[32] Helmholz, P., U. Buschenfeldc, T .and Breitkopf e Rottensteiner F. Muller, S and: *Multitemporal quality assessment of grassland and cropland objects of a topographic dataset*. The International Archives of the Photogrammetry, páginas 67–72, 2012. 23

[33] Hoffman, S. e C. Brenner: *Quality assessment of automatically generated feature maps for future driver assistance systems*. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, páginas 500–503, 2009. 23

[34] Holloway, T., M. Bozicevic e K. Börner: *Analyzing and visualizing the semantic coverage of wikipedia and its authors*. Complexity, 12(3):30–40, 2007. 16

[35] Huisman, O. e R. A. By: *Principles of Geographic Information Systems - An Introductory Textbook*. ITC Educational Textbook Series, 2009. 8

[36] ISO, ISO: *19157: 2013: Geographic information—data quality*. International Organization for Standardization: Geneva, Switzerland, 2013. 16

[37] Jilani, M., P. Corcoran e M. Bertolotto: *Automated highway tag assessment of OpenStreetMap road networks*. Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, páginas 449–452, 2014. 23

[38] Kanapaka, R. R. N., Neelisete R. K: *A survey of tools for visualizing geo spatial data*. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), páginas 22–27, 2015. ix, 10

[39] Karagiorgou, S. e D. Pfoser: *On vehicle tracking data-based road network generation*. Proceedings of the 17th ACM Sigspatial International Conference on Advances in Geographic Information Systems, páginas 89–98, 2012. 23

[40] Karam, R. e M. Melchiori: *A crowdsourcing-based framework for improving geospatial open data*. IEEE International Conference on Systems, Man, and Cybernetics, 2013. 23

[41] Kennedy, L., S. F. Chang e I. Kozintsev: *To search or to label? predicting the performance of search-based automatic image classifiers*. ACM Workshop Multimedia Information Retrieval, página 249–258, 2006. 23, 24

[42] Kitchenham, B. e S. Charters: *Guidelines for performing systematic literature reviews in software engineering*. EBSE 2007-001. Keele University and Durham University Joint Report, 2007. 6, 18

[43] Kounadi, O., T. J. Lampoltshammer, M. L. e T. Heistracher: *Accuracy and privacy aspects in free online reverse geocoding services*. Cartography and Geographic Information Science, 40(2):140–153, 2013. 31, 34

[44] Langley, S. A., J. P. Messina e N. Moore: *Using meta-quality to assess the utility of volunteered geographic information for science*. International Journal of Health Geographics, 2017. 23

[45] Liu, D., M. Wang, X S Hua e H J Zhang: *Semi-automatic tagging of photo albums via exemplar selection and tag inference.* IEEE Transactions on Multimedia, 13:82–91, 2011. 33

[46] Manikas, K. e K. M. Hansen: *Software ecosystems – a systematic literature review*. Journal of Systems and Software, 86(5):1294 – 1306, 2013. 19

[47] Medeiros, G. e M. Holanda: *Solutions for data quality in GIS and VGI: a systematic literature review*. Em *World Conference on Information Systems and Technologies*, páginas 645–654. Springer, 2019. 51

[48] Medeiros, G.F.B. de, L. C Degrossi e M. Holanda: *QualiOSM: Melhorando a qualidade dos dados na ferramenta de mapeamento colaborativo OpenStreetMap*. Simpósio Brasileiro de Banco de Dados (SBBD). 51

[49] Monteiro, A. M., G. Camara, S.D. Fucks e M.S Carvalho: *Spatial analysis and GIS: A primer*. National Institute for Space Research, 2001. 6, 7

[50] Mooney, P. e P. Corcoran: *The annotation process in OpenStreetMap*. Transactions in GIS, 16:561–579, 2012. 23, 24, 33

[51] Mooney, P., P. Corcoran, H. Sun e L. Yan: *Citizen generated spatial data and information: Risks and opportunities*. 2012 International Conference on Industrial Control and Electronics Engineering, 2012. 23

[52] Mooney, P., P. Corcoran e A. C. Winstanley: *Towards quality metrics for OpenStreetMap*. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, páginas 514–517, 2010. 23

[53] Neis, P. e D. Zielstra: *Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap*. Future Internet, 6(1):76–106, 2014. 16

[54] Nielsen, J.: *The 90-9-1 rule for participation inequality in social media and online communities*, 2006. 2

[55] OpenStreetMap_Elements. `https://wiki.openstreetmap.org/wiki/Elements`. Acessado em fevereiro de 2019. 13

[56] OpenStreetMap_Tags. `https://wiki.openstreetmap.org/wiki/Tags`. Acessado em fevereiro de 2019. 34

[57] Ramm, F., J. Topf e S. Chilton: *OpenStreetMap: Using and enhancing the free map of the world*. UIT Cambridge, 2010. 1

[58] Rigaux, P., M. Scholl e A. Voisard: *Spatial Databases with Application to GIS*. Morgan Kaufmann Publishers, 2002. ix, 7, 8

[59] Ruta, M., Scioscia F. Ieva S. Loseto G. e E. Di Sciascio: *Semantic annotation of OpenStreetMap points of interest for mobile discovery and navigation*. IEEE First International Conference on Mobile Services, páginas 33–39, 2012. 31

[60] See, L., J. Estima, A. Pődör, J.J. Arsanjani, J.C.L Bayas e R. Vatseva: *Sources of VGI for mapping*. Citizen Sensor, página 13, 2017. 16

[61] Sehra, S. S., J. Singh e H. S. Rai: *A systematic study of OpenStreetMap data quality assessment*. 11th International Conference on Information Technology: New Generations, 2014. 23

[62] Senaratne, H., A. Mobasheri, A. L. Ali, C. Capineri e M. Haklay: *A review of volunteered geographic information quality assessment methods*. International Journal of Geographical Information Science, 31(1):139 – 167, 2017. 1

[63] Subbiah, G., A. Alam, L. Khan e B. Thuraisingham: *Geospatial data qualities as web services performance metrics*. Proceedings of the 15th International Symposium on Advances in Geographic Information Systems, 2007. 23

[64] Sui, D., S. Elwood e M. Goodchild: *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer, 2013. 10

[65] Tong, Y., Z. Zhou, Y. Zeng, L. Chen e C. Shahabi: *Spatial crowdsourcing: a survey*. The VLDB Journal, 29(1):217–250, 2020. 12

[66] Wang, M., Q. Li, Q. Hu e M. Zhou: *Quality analysis of open street map data. international archives of the photogrammetry*. Remote Sensing and Spatial Information Sciences, 8th International Symposium on Spatial Data Quality, Hong Kong, 2013. 23

[67] Xu, T. e Y. Cai: *Location anonymity in continuous location-based services*. Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems. ACM, New York, NY, USA, 2007. 23

[68] Ye, M., K. Janowicz e Lee W. Mulligann, C and: *What you are is when you are: the temporal dimension of feature types in location-based social networks*. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, páginas 102–111, 2011. 23

[69] Yeung, A. K. W e G. B. Hall: *Spatial Database Systems: Design, Implementation and Project Management*. Springer, 2007. 5

[70] Zhang, C. e C. S. Fraser: *Automated registration of high-resolution satellite images*. The Photogrammetric Record, 22(117):75–87, 2007. 23

[71] Zheng, Y., X. Fen, X. Xie, S. Peng e J. Fu: *Detecting nearly duplicated records in location datasets*. Proceedings of the 18th ACM Sigspatial International Conference on Advances in Geographic Information Systems, páginas 137–143, 2010. 23

[72] Zhou, M., Q. Hu e M. Wang: *A quality analysis and uncertainty modeling approach for crowd-sourcing location check-in data*. 8th International Symposium on Spatial Data Quality, 2013. 23