Métricas de qualidade para VGI para o projeto Pauliceia

Rodrigo M. Mariano[1], **Karine R. Ferreira**[1], **Luis Ferla**[2]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP – Brasil

[2]Universidade Federal de São Paulo (UNIFESP)
Guarulhos – SP – Brasil

{rodrigo.mariano, karine.ferreira}@inpe.br, ferla@unifesp.br

Abstract. *Pauliceia is a project whose aim is to develop a computational platform for historical data manipulation collaboratively. Researchers will contribute with this data and assist in quality control. Apart from the insertion of historical data, there will be the opportunity for the volunteers to make the mapping of old maps and it is intended that the system be in charge of verifying a consensus in the geometries obtained. Quality questions and metrics were evaluated for data collection by VGI, in order to improve data quality. With it, a protocol for the Pauliceia's project was developed, because the lack of it can generate contributions with poor quality.*

Resumo. *Pauliceia é um projeto cujo propósito é desenvolver uma plataforma computacional para manipulação de dados históricos colaborativamente. Os pesquisadores contribuirão com esses dados e auxiliarão no controle de qualidade. Fora a inserção de dados históricos, haverá a oportunidade dos voluntários fazerem a vetorização de mapas antigos e pretende-se que o sistema se encarregue de verificar um consenso nas geometrias obtidas. Questões e métricas de qualidade foram avaliadas, para dados coletados por VGI, com o intuito de melhorar a qualidade dos dados. Com isso, desenvolvendo-se um protocolo para o projeto Pauliceia, porque a falta dele pode gerar contribuições com má qualidade.*

1. Introdução

O projeto Pauliceia tem como objetivo desenvolver uma plataforma computacional online para gerenciamento de dados históricos com uma localização geográfica (dados espaço-temporais) colaborativamente. Os estudiosos poderão produzir mapas e visualizações de suas próprias pesquisas e ao mesmo tempo, contribuir para os dados dentro do sistema. Uma atividade online com participação da comunidade em um propósito, é conhecido como crowdsourcing e quando esses voluntários contribuem com informações geográficas, é conhecido como VGI. O projeto irá enriquecer a compreensão da história de São Paulo (SP) durante o período de 1870 a 1940, além de oferecer um modelo inovador de pesquisa para as Humanidades Digitais, que promova o trabalho colaborativo e o fluxo de conhecimento gratuito. O recorte histórico de 1870 a 1940 foi escolhido, pois foi uma época em que a cidade de São Paulo cresceu muito, saindo de aproximadamente 30.000 habitantes para 1.300.000 em torno de 70 anos [Secretaria Municipal de Urbanismo e Licenciamento 2017]. Os principais motivos para

este acontecimento, se dá no aumento da produção de café, vinda de imigrantes, criação da ferrovia e desenvolvimento da indústria [Mota et al. 2007] [Carvalho 2017].

[Estellés-Arolas and González-Ladrón-de Guevara 2012] estudaram e produziram uma única definição de crowdsourcing, que seria: *"a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken"*. Enquanto a coleta de dados em projetos de ciência cidadã pode ser feitos, por exemplo, com formulários em papel, por definição o crowdsourcing tem natureza online. Isto torna o crowdsourcing mais restrito. Esta categoria de participação não precisa ser abertos a todos, podendo ser restringidos a certos grupos [See et al. 2016].

Originalmente o termo Volunteered Geographic Information (VGI) aparece primeiramente por Goodchild [Goodchild 2007] como *"the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals"* [See et al. 2016] e quando o termo é definido, ele diz: *"the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate. But collectively, they represent a dramatic innovation that will have profound impacts on geographic information systems (GIS) and on the discipline of geography and its relationship to the general public. VGI is a special case of the more general Web phenomenon of user generated content"*, em outro trabalho, ele com Li [Goodchild and Li 2012] resumem VGI como sendo *"a version of crowd-sourcing in which members of the general public create and contribute georeferenced facts about the Earth's surface and near-surface to websites where the facts are synthesized into databases"*. Em suma, é um fenômeno recente que oferece um mecanismo alternativo para a aquisição e compilação de informações geográficas de voluntários [Goodchild 2007]. Essas contribuições contêm uma localização geográfica e uma descrição, com vários atributos, recorrentes dessa localização [Goodchild and Li 2012]. Os usuários são, frequentemente, não treinados, e apesar de seus conhecimentos e antecedentes, criam informações geográficas em plataformas web, como por exemplo: OpenStreetMap (OSM), Google MyMaps, Flickr ou Wikimapia [Senaratne et al. 2017]. A representação do dado pode ser feita por um ponto, uma linha ou polígono. Mesmo com alto potencial, por adquirir informações geográficas de maneira rápida, detalhada e com baixo custo, o VGI por padrão não oferece garantia de qualidade em seus dados [Goodchild and Li 2012].

O OpenStreetMap (OSM) é o mais conhecido sistema de VGI. Ele permite trabalhar com dados geográficos gratuitos, tendo uma licença de conteúdo aberto [OpenStreetMap 2017a]. Vários estudos medem a precisão dos dados de VGI com base no OSM, por exemplo: [Haklay 2010], [Girres and Touya 2010] ou

[Ciepłuch et al. 2010]. O trabalho de [Senaratne et al. 2017] faz uma revisão detalhada sobre os métodos para avaliar a qualidade dos dados de VGI. Todos os estudos fornecem conhecimentos proveitosos sobre a precisão do VGI, ajudando a garantir e melhorar a qualidade.

Relacionado a tipologia do VGI, pode-se classificá-lo com base no tipo explícito ou implícito, relacionado a categoria de voluntariado (explícito ou implícito). No VGI explícito, o foco são nas atividades de mapeamento, anotando explicitamente os dados com conteúdos geográficos, como por exemplo: as geometrias em OSM ou Wikimapia. No VGI implícito os dados são associados a uma localização geográfica específica, podendo ser qualquer tipo de mídia, como texto, imagem ou vídeo. Por exemplo, microblogs geoetiquetados (e.g. Tweets) ou imagens geoetiquetadas do Flicker. Para cada VGI, implícito ou explícito, existem diferentes abordagens para avaliar a qualidade [Antoniou et al. 2010] [Craglia et al. 2012][Senaratne et al. 2017]. Por conta do crescimento do uso do VGI, torna-se cada vez mais importante estar atento a qualidade de seus dados, para obter informações e, consequentemente, decisões precisas. Por conta da falta de padronização, a qualidade no VGI tem mostrado inúmeras fontes de dados heterogêneas, como textos, imagens e etc [Liu et al. 2008] [Jacob et al. 2009] [Chunara et al. 2012] [Fuchs et al. 2013] [Senaratne et al. 2017].

Esforços foram criados para estudar a qualidade da informação geográfica, gerando discussões sobre a possibilidade de um padrão de dados geoespaciais. Isto levou a um concordância de cinco dimensões fundamentais: precisão de posicionamento, precisão de atributos, consistência lógica, integridade (ou completude) e linhagem. Posteriormente surgiu outros elementos, como precisão temporal e precisão semântica [Goodchild and Li 2012] [Guptill and Morrison 2013]

A imprecisão dos dados de VGI é explicado pelo fato de que os seres humanos expressam as regiões geográficas e suas relações imprecisamente, através de conceitos vagos. A imprecisão na conceptualização humana da localização deve-se, não apenas ao fato de que as entidades geográficas são de natureza contínua, mas também por conta da qualidade e limitações do conhecimento espacial [Montello et al. 2003] [Hollenstein and Purves 2010].

Fornecer serviços confiáveis ou retirar informações úteis desses dados, requer das contribuições pelo menos um padrão de qualidade. Informações imprecisas, sejam elas maliciosas ou não, podem ser minimizados por indicadores de qualidade apropriados e medidas para essas várias contribuições VGI. Segundo Goodchild e Li [Goodchild and Li 2012] há três abordagens para assegurar a qualidade do VGI:

- crowdsourcing (geração de informação por várias pessoas): é o envolvimento de um grupo para validar e corrigir erros cometidos por um colaborador individual;
- abordagens sociais: são indivíduos confiáveis que tem uma boa reputação com suas contribuições para o VGI, podendo atuar como *gatekeepers* (porteiros) para manter e controlar a qualidade de outras contribuições de VGI;
- abordagens geográficas: é o uso de leis e conhecimento da geografia, como a primeira lei da geografia[1], para avaliar a qualidade.

[1]"Everything is related to everything else but near things are more related than distant things" [Tobler 2004]

As Agências Nacionais de Mapeamento (NMAs) e as Empresas de Topografia Comercial (CSCs) utilizam protocolos robustos e padronizados que regem e orientam a coleta de dados geográficos, no entanto os projetos VGI muitas vezes carecem de padrões, ou apenas fornecem diretrizes e sugestões soltas, em vez de especificações rigorosas. Embora o VGI possa teoricamente atingir altos padrões de qualidade sem padrões rigorosos, sua ausência é muitas vezes uma fonte importante de erros nos dados, representando uma barreira à sua maior difusão e reutilização [Mooney et al. 2016].

A necessidade de estabelecer padrões e protocolos para projetos VGI não é uma novidade. Girres e Touya [Girres and Touya 2010] citam que a falta de especificação é um dos pontos chave que pode causar a má qualidade nos dados. Que a heterogeneidade das contribuições podem ser causadas pela falta de especificação na coleta dos dados. De acordo com Mooney et al. [Mooney et al. 2016] alguns pesquisadores alertaram sobre as ameaças para a comunidade e a sociedade, colocadas pela falta de protocolos e mencionaram a relevância deles para projetos de VGI, sugerindo a definição de protocolos para garantir alta qualidade de dados. Os protocolos são importantes para facilitar e ampliar a reutilização dos dados do VGI, para fins e aplicativos diferentes do que originalmente foram coletados.

Mesmo que os projetos de VGI forneçam instruções aos seus contribuintes sobre como manipular os dados geográficos, essas instruções são geralmente flexíveis e podem faltar rigor de pesquisa geográfica profissional. Os voluntários só são incentivados a usar essas instruções, e muitas vezes acontece que eles coletam dados sem estudar as recomendações do projeto VGI. A falta de adoção e implementação de rigorosas estratégias de coleta e pesquisa no VGI preocupa os usuários e potenciais contribuidores do VGI [Mooney et al. 2016].

Segundo Bonney et al. [Bonney et al. 2009], para assegurar aos cidadãos coletar e enviar dados precisos dependem de três coisas: protocolos claros de coleta de dados, formulários de dados simples e lógicos, e apoio aos participantes para entender como seguir os protocolos e enviar suas informações. A maioria dos voluntários está disposta a seguir os protocolos, mesmo que complexos, a fim de coletar dados de forma recomendada e padronizada para garantir que sua contribuição tenha valor [Mooney et al. 2016]. De acordo com Pilz et al. [Pilz et al. 2006] na maioria dos casos, a maior recompensa para os participantes é ver que os resultados de seus esforços voluntários de coleta de dados são avaliados.

De acordo com a pesquisa feita por Schmidt et al. [Schmidt et al. 2013], sobre contribuição no OSM, 29.7% dos participantes tem medo de fazer alguma coisa errada, 65.1% dizem que o tempo de contribuição é muito grande e 40.1% deles diz que a edição é muito complexa. Mooney el al. [Mooney et al. 2016] descrevem que em vários projetos do VGI, a forma desestruturada de contribuição dos dados de vários contribuintes criou-se mais problemas, do que aqueles que estava-se tentando resolver. Por conta disso, há a necessidade de uma moderação na criação ou integração de dados de múltiplas fontes.

Muitas obras desenvolveram métodos para avaliar a qualidade do VGI, mesmo sendo perceptível que não haja somente um único método que resolva todos os problemas, afinal as contribuições de VGI tem perfil heterogêneo [Senaratne et al. 2017]. A criação de um protocolo bem definido, pode diminuir as barreiras descritas, por conta

da padronização ou formalização criada, provendo uma melhora na coleta de dados. Por conta disso, o presente trabalho fará uma revisão das técnicas de qualidade de VGI existentes e apresentará as que melhores se aplicam ao contexto do projeto Pauliceia, desenvolvendo um protocolo para ele.

1.1. Trabalho Relacionado

Nesta seção, será apresentado projetos que possuem características semelhantes ao projeto Pauliceia.

A plataforma ATLMaps[2] é uma colaboração entre a Georgia State University e a Emory University. Nele é possível combinar mapas, visualização de dados geoespaciais e pontos de localização, contribuídos por voluntários para aumentar o conhecimento sobre Atlanta. Os contribuidores podem criar seus próprios projetos em cima das camadas disponíveis, adicionando anotações, áudios, imagens e etc. Os usuários podem superpor mapas de várias épocas; podendo, por exemplo, comparar os limites históricos da cidade nos tempos [White and Gilbert 2016].

O Building Inspector[3] é um projeto dos Laboratórios da Biblioteca Pública de Nova York em colaboração com Lionel Pincus e Princess Firyal Map Division da Biblioteca Pública de Nova York, feito para gerenciar dados de mapas históricos. Os computadores são treinados para fazer o levantamento pesado e distribui-se as demais tarefas de controle de qualidade para os cidadãos, produzindo um diretório abrangente da antiga Nova York. Com essa informação é possível explorar o passado da cidade. O projeto permite vincular documentos históricos (arquivos, jornais antigos, fotografias e etc.) aos lugares, dando oportunidade de novas maneiras de pesquisar, aprender e descobrir o passado. É usado um algoritmo proposto por Budig et al. [Budig et al. 2016] para a extração da melhor representação poligonal, dado vários polígonos, de construções dos atlas do século XIX e início do século XX. Dado um conjunto de polígonos, que descrevem o mesmo objeto (e.g. um edifício), o algoritmo retorna o melhor consenso.

O Digital Harlem é um projeto de pesquisa colaborativa sobre Harlem, entre 1915 e 1930, realizado por historiadores do Departamento de História da Universidade de Sydney, na Austrália. Ele se concentra na vida de nova-iorquinos africanos comuns, capturando atividades, lugares e relacionamentos que compunham a vida cotidiana, através de registros legais e jornais [Digital Harlem Blog 2017]. O Digital Harlem não foi desenvolvido como o intuito de ser um projeto de história pública ou um recurso de ensino [Robertson 2016]. O site permite a busca de eventos e locais, gerar mapas interativos, localizar pessoas em Harlem para descobrir lugares que elas costumavam ir e os resultados podem ser mostrados em um mapa com inúmeras camadas.

OpenHistoricalMap[4] (OHM) é um projeto que usa a infraestrutura OSM para criação de um mapa universal e detalhado da história mundo, de forma colaborativa. Pode-se inserir dados de margens, fronteiras políticas, edifícios e caminhos, semelhantes ao OSM. Ele tem política de dados aberto, podendo usá-lo para qualquer finalidade, creditando ao OHM e seus contribuidores.

[2]https://atlmaps.org/

[3]http://buildinginspector.nypl.org/

[4]http://www.openhistoricalmap.org

O HistOSM[5] é um aplicativo para exploração de objetos históricos do OpenStreetMap. Pode-se rastrear as características históricas regionais ao ampliar e filtrar o mapa, mostrando os objetos de interesse. Se clicar nos objetos de maneira individual, será mostrado as informações detalhadas deles, como tags e links (e.g. imagens associadas ou sites), podendo também ir para o site do OpenStreetMap para editar os objetos e se necessário, adicionar ou atualizar informações.

O projeto Pauliceia tem características parecidas com esses trabalhos, também sendo influenciados por alguns. Enquanto o OpenHistoricalMap e o HistOSM trabalham atualmente só com objetos históricos (e.g. monumentos), no Pauliceia também haverão eventos históricos (e.g. assaltos de uma época). Há um projeto chamado Historic Event[6], cujo objetivo era adicionar eventos históricos no OSM, mas ele não foi continuado. Um dos principais focos do projeto Pauliceia é o crowdsourcing, pois quem alimentará a plataforma serão os usuários finais, que são pessoas com dados históricos interessadas em colaborar. O VGI será utilizado para motivar os cidadãos a participar na coleta de informações espaciais com qualidade, atuando como sensores. Os usuários providenciarão feedbacks (revisões, comentários, votações e etc.) para melhora dos dados.

1.2. Objetivo

Fazer uma revisão das métricas e estratégias utilizadas em VGI para avaliar a qualidade de dados colaborativos. Analisar quais métricas seriam viáveis para os dados históricos do projeto Pauliceia e como incorporar isto no serviço web para VGI que está sendo desenvolvido.

2. Informação Geográfica Voluntária

Com o VGI, se tornou possível que as pessoas poderiam criar sua própria informação geográfica digital, através de mapas on-line, sem custo. Isto se dá por conta que as coordenadas podem ser obtidas com o GPS ou utilizar imagens disponibilizadas por terceiros como, por exemplo, o Google Earth. Conhecimento em cartografia já não é necessário, pois há softwares open-source para construção de mapas com alta qualidade. Um usuário pode desenvolver mapas de sua área local, podendo ser mais eficaz, do que um especialista em mapeamento, devido seu próprio conhecimento local [Goodchild and Li 2012]. Por conta disso é interessante estudar a qualidade e melhoramento dos dados gerados por VGI, que será descrito nas próximas seções.

2.1. Medidas e indicadores para a qualidade do VGI

A qualidade das contribuições do VGI podem ser descritas por: medidas de qualidade e indicadores de qualidade. As medidas de qualidade utilizam os dados autoritários (gerados por agências confiáveis, e.g. NMAs ou CSCs), como um conjunto de dados de referência, para avaliar os dados gerados pelo VGI, comparando-os. Isto é feito, pois acredita-se que os dados autoritários têm sempre alta qualidade. Os indicadores de qualidade são utilizados quando os dados autoritários não estão disponíveis, o que pode ser frequente, dependendo da categoria dos dados. Além de que, Over et al. [Over et al. 2010] observam que a qualidade dos dados do OSM difere de dados autoritários. A natureza do

[5]http://histosm.org/

[6]http://wiki.openstreetmap.org/wiki/Proposed_features/historic_event

voluntariado pode apresentar tendências nas contribuições. Isto ocorre por vários fatores, desde capacidade técnicas até diferenças culturais das pessoas. Conforme os dados de VGI ficam mais detalhados ao longo do tempo e em determinadas áreas, torna-se menos proveitoso a utilização de dados autoritários para avaliar a qualidade dos dados gerados por VGI. Isto ocorre, pois os dados de VGI são muitas vezes, mais completos e precisos, do que os conjuntos de dados autoritários existentes [Antoniou and Skopeliti 2015].

Em relação a medidas de qualidade para VGI, a Organização Internacional de Normalização (ISO) definiu a qualidade da informação geográfica como: *"a totalidade das características de um produto que tem sua capacidade de satisfazer as necessidades declaradas e implícitas"*[7]. Existe um conjunto de padrões explicados pela ISO, que define as medidas de qualidade da informação geográfica, que são [TC 2009] [Jakobsson and Giversen 2007]:

- integridade: descreve a presença e ausência de dados, seus atributos e relacionamentos entre objetos. Para ser avaliado pode-se utilizar: comparação de comprimento baseado em grade em relação a dados autoritários, comparação do número de características, comparação do comprimento total ou da área total, medida de integridade e índice de integridade;
- consistência lógica: nível de coerência às regras lógicas das estrutura de dados (como conceitual, lógica ou física), atribuição e relacionamentos. Para ser avaliado pode-se utilizar: similaridade espacial na multi-representação, semelhança semântica entre as etiquetas, identificação de entidades com classificação inadequada ou sistema de recomendação de etiquetas (algoritmo que sugere etiquetas relevantes);
- precisão do posicionamento: precisão da posição dos recursos. É a proximidade entre uma medida de uma quantidade e o valor verdadeiro aceito dela. Para ser avaliado pode-se utilizar: a distância entre cruzamentos correspondentes de uma rede, a distância euclidiana dos atributos pontuais, distância de erro x e y, distâncias entre polígonos centroides e similaridade espacial em multi-representação;
- precisão temática: correção de classificação, correção de atributos não quantitativos e precisão de atributos quantitativos. Para ser avaliado pode-se utilizar: medição da porcentagem (%) da classificação correta, o número de características com atributos específicos, matriz de confusão e análise padrão do índice kappa;
- precisão temporal: precisão de uma medida de tempo, consistência temporal e validade dos dados em relação ao tempo. Para ser avaliado pode-se estudar a evolução dos dados VGI.

No que diz respeito aos indicadores de qualidade para VGI, eles são qualitativos (expressam a qualidade dos dados), como o propósito, uso e linhagem. O propósito expõe o objetivo dos dados. O uso indica quais as funcionalidades dos dados. Linhagem se refere ao histórico dos dados, desde sua coleta, aquisição e formação para o seu uso final. Os seus indicadores podem ser descritos como [Senaratne et al. 2017]:

- confiabilidade: um julgamento baseado em características subjetivas, como confiança. Sendo adquiridas por boas classificações das contribuições ou maior frequência de uso delas [Flanagin and Metzger 2008];

[7] https://www.iso.org/obp/ui/#iso:std:iso:19109:ed-1:v1:en

- credibilidade: definida como a credibilidade de uma fonte ou mensagem, na qual têm duas dimensões: confiabilidade e experiência. Para avaliá-lo, usa-se a fonte da informação como base, porém não é direto, pois como os dados de VGI não são autoritários, portanto a fonte talvez não seja disponível É necessário considerar fatores de confiabilidade e experiência, para conseguir fazer esta categoria de avaliação. Os metadados sobre a origem do VGI podem fornecer uma base para isso;
- qualidade do conteúdo de texto: é a qualidade dos dados de texto baseada no uso das características do texto, como: comprimento dele, estrutura, legibilidade, histórico de revisões, uso de termos específicos e entre outros. Normalmente aplicável em VGI baseado em texto;
- imprecisão: ambiguidade na captura dos dados, como por exemplo: imprecisão por baixa resolução;
- conhecimento local: conhecimento dos usuários em relação ao ambiente geográfico que ele está mapeando;
- experiência: experiência do usuário na plataforma VGI. Pode ser capturado quando o voluntário se registra no portal, pela quantidade de contribuições feitas (seja adicionadas ou editadas) ou número de vezes que o usuário utilizou os fóruns online para discussão dos dados;
- reconhecimento: dar recompensas ao contribuidor pela utilização da plataforma, como prêmios virtuais, conhecido como Gamificação, e oportunidade de revisão de suas contribuições por outros usuários;
- reputação: é a capacidade de avaliar, marcar, discutir e anotar as contribuições, afetando na reputação do usuário. Avaliação do histórico das interações do voluntário entre os outros colaboradores.

2.2. Qualidade em relação ao VGI baseado em mapa, imagem e texto

O uso do VGI está totalmente ligado à qualidade dos dados. Isto depende desde o tipo do VGI, da forma como os dados são coletados e do contexto do uso. O VGI pode ser descrito com base em três formas: mapa, imagem e texto; de acordo com os métodos de coleta dos dados.

O VGI baseado em mapa é quando a fonte dos dados incluem as geometrias básicas como pontos, linhas e polígonos (e.g. OSM ou Wikimapia). O OSM é o sistema mais conhecido. O seu objetivo é desenvolver um mapa gratuito do mundo. Sua comunidade ativa é muito grande, com milhões de contribuintes registrados. Ele fornece mecanismos flexíveis para trabalhar com os dados, utilizados para geo-visualização, pesquisa de pontos de interesse (POI) e etc. Os dados do OSM são feitos na forma de nós, linhas ou polígonos, referenciados por uma latitude e longitude, e os atributos são feitos por etiquetas (pares chave-valor), descrevendo as características de uma entidade geográfica. Não há restrições ao uso dessas etiquetas, porém o OSM fornece padrões, que precisam ser seguidos. Esta padronização é interessante para evitar classificação errada ou conflitos, que podem levar uma redução na qualidade dos dados, pois este tipo de VGI é normalmente usado para navegação ou pesquisa de POI. Por conta disso, é importante ter uma precisão de posição, precisão dos atributos e a consistência topológica das entidades. Este cuidado é bom para garantir o fornecimento de serviços confiáveis, porque há uma falta de dados autoritários para fazer uma comparação [Senaratne et al. 2017].

O VGI baseado em imagem é produzido de maneira implícita em plataformas como Flickr ou Instagram. Os usuários tiram fotos de um objeto ou local, com o uso de câmeras ou smartphones, e anexam uma localização geográfica para ele, podendo referenciá-la depois. Este tipo de VGI têm vários usos, como por exemplo: monitoramento ambiental, navegação de pedestres ou análise de trajetórias. [Fuchs et al. 2013] [Robinson et al. 2012] [Andrienko et al. 2009]. A adição de metadados pode ser feita através de marcação de etiquetas, que descreve aquela contribuição. As coordenadas geográficas são metadados que indicam a localização da imagem, conhecidas como geoetiquetas. Essas geoetiquetas podem ser obtidas por um dispositivo GPS (atualmente vem incluso nas câmeras e smartphones) ou pelo posicionamento de uma foto em uma interface de mapa [Golder and Huberman 2006] [Valli and Hannay 2010]. Erros de precisão do GPS ou usuários que indicam incorretamente uma geoetiqueta nas fotos (não indicam a localização da origem dela), são fatores que afetam a qualidade dos dados, causando problemas na precisão de posição [Keßler et al. 2009]. Esta é uma dificuldade quando se quer utilizar as imagens para fazer alguma análise, como análise de trajetória humana, gerenciamento de desastres ou monitoramento ambiental [Senaratne et al. 2017].

O VGI baseado em texto é produzido de maneira implícita em plataformas como Twitter ou Blogs. Os voluntários fornecem textos através de dispositivos (e.g. computadores ou smartphones), dando uma localização geográfica para ele. Essas informações podem ser utilizadas para busca de informações (como no Twitter), detecção de propagação de doenças ou eventos, análise de trajetórias de pessoas e etc [MacEachren et al. 2011] [Chunara et al. 2012] [Bosch et al. 2013] [Huberman et al. 2008] [Senaratne et al. 2017]. O metadado espacial pode uma geoetiqueta da origem do texto ou estar dentro do texto (e.g. “A banda Rosa de Saron está fazendo um show em São Paulo hoje”). A análise da qualidade desses dados é importante para filtrar as informações úteis, pois pode haver erros nos dados. Estes problemas podem ser causados por falhas nos dispositivos de GPS, localização incorretamente especificada, localização com baixa resolução, informações inúteis dadas (e.g. spam) ou informações não precisas (e.g. um usuário geoetiquetar um texto sobre um evento a quilômetros de distância dele) [Senaratne et al. 2017].

2.3. Métodos para avaliar a qualidade do VGI

Senaratnea et al. [Senaratne et al. 2017] revisaram métodos para avaliar várias medidas de qualidade e indicadores de VGI. Um método é um procedimento seguido para avaliar as medidas de qualidade e os indicadores de qualidade (e.g. comparar imagens de satélite para avaliar a precisão de posição). Algumas dessas medidas serão citadas posteriormente.

Como o projeto Pauliceia por enquanto só trabalhará com VGI baseado em mapa, será abordado somente métodos de avaliação de qualidade neste tipo. Caso o leitor queira estudar sobre VGI baseado em imagem ou texto, o estudo de Senaratne et al. [Senaratne et al. 2017] faz uma revisão sobre eles.

2.3.1. Avaliação de qualidade em VGI baseado em mapa

Em relação a precisão de posicionamento, Haklay et al. [Haklay et al. 2010] estudou a aplicação da Lei de Linus[8] e descobriu que quanto mais voluntários existirem, maior probabilidade de se identificar os objetos corretamente, aumentando a qualidade. Muito dos trabalhos citados por ele usam dados autoritários para comparação com os gerados por voluntários. Isto o que levou a concluir que o uso da Lei de Linus é aplicável, portanto o uso de dados de referência pode não ser a única maneira de avaliar a qualidade dos dados gerados por usuários. De Tré et al. [De Tré et al. 2010] apresentam uma técnica para a detecção (semi-)automática e a fusão de POIs correferentes em um único POI consistente. Para isto é utilizando um valor de verdade possibilista (possibilistic truth value, PTV) para determinar se dois POIs são correferentes ou não. Um ponto de interesse (point of interest, POI) indica um local geográfico que pode ser de interesse para algum usuário (e.g. edifícios históricos). Os POIs correferentes são POIs que indicam a mesma localização geográfica física, causados por virem de inúmeras fontes ou usuários, gerando imperfeições. Portanto, eles devem ser evitados, pois podem inserir incertezas e inconsistências nos dados.

A consistência topológica nos dados do OSM são qualificados por verificações de seus atributos, com o intuito de diminuir os problemas relacionados a sobreposição de características. Alguns desses problemas são conhecidos como "dangles", que é quando a digitalização não é bem feita, ocasionando que, por exemplo, duas linhas que deveriam se encontrar, não se encontram. Pode ser desenvolvido e aplicado: regras de integridade de topologia como medida, análise geométrica para avaliar a coerência topológica dos dados e utilizar a Matriz de Nove Interseções Dimensionalmente Estendida para calcular a relação espacial entre os objetos [Senaratne et al. 2017].

Em relação a precisão temática e semântica, a maioria dos erros são causados por inserção incorreta das características dos dados, feita algumas vezes por parte dos contribuintes, causando imprecisão. As razões disso são a falta de padronização na classificação, a falta de atenção que os contribuintes têm ao inserirem etiquetas e valores que não estão presentes na especificação OSM, a falta de regulamentação na nomenclatura e etc. Necessitando de especificações padronizadas para melhorar a precisão semântica e de atributos dos dados [Senaratne et al. 2017]. A heterogeneidade semântica dos objetos coletados por VGI é um problema, pois há vários usos que poderiam ser feitos com essa informação. Por conta disso Vandecasteele e a Devillers [Vandecasteele and Devillers 2015], mostram uma abordagem para melhorar a qualidade semântica dos dados, reduzindo a heterogeneidade semântica deles. Esta abordagem é feita usando um sistema de recomendação de etiquetas, chamado OSMantic. O OSMantic recomenda etiquetas significativas aos voluntários durante a coleta de dados. Com isso, os usuários podem encontrar etiquetais mais adequadas para uma contribuição, diminuindo a heterogeneidade semântica.

Para avaliar a integridade Koukoletsos et al. [Koukoletsos et al. 2012] propõem uma abordagem de correspondência automatizada para comparar dados com um conjunto de dados de referência. Isto é feito em vários estágios, combinando restrições geométricas e atributos. Eles aplicam essa abordagem nos dados do OSM comparando-os com os

[8] "Given enough eyeballs, all bugs are shallow" [Raymond 2017]

dados oficiais do Ordnance Survey. Girres e Touya [Girres and Touya 2010] utilizaram diversos métodos de amostragem e áreas de estudo para avaliar a qualidade dos dados espaciais. Em relação a integridade, foi feito uma análise quantitativa das amostras de dados, verificando a correlação de integridade e o número de contribuintes em uma área. Isto mostrou que a quantidade de contribuições em uma área, cresce em relação ao número de contribuidores nela, contudo de forma não linear.

Girres e Touya [Girres and Touya 2010] analisaram a precisão temporal através da correlação entre o número de contribuintes com a data média da coleta, e entre o número de usuários e a versão média do objeto coletado, avaliando a quantidade de objetos atualizados. Isto serve para avaliar quantos objetos são atualizados. Neste estudo foi observado que quanto mais voluntários, mais as contribuições eram recentes.

Em relação a linhagem dos dados, ela pode ser avaliada considerando a maneira como a coleta de dados é feita, de acordo com a fonte dos dados. Uma maneira adequada de resolver isto, seria criar uma forma de permitir o gerenciamento de contribuições confiáveis, impedindo as menos confiáveis e dar oportunidade de correção para os dados menos corretos. Isto pode ser feito com a inserção de voluntários moderadores [Girres and Touya 2010]. Pode ser feito também uma avaliação seguindo uma abordagem orientada a dados, com essência na origem dos itens dos dados específicos. A utilização de um vocabulário de origem mostra a linhagem dos recursos dos dados online [Keßler et al. 2011].

3. Protocolo para projetos VGI

Mooney el al. [Mooney et al. 2016] propõem alguns tópicos do protocolo, desenvolvido por eles, para atingir os objetivos, padronização e por fim a qualidade de projetos baseados em VGI. Esses tópicos são a Modelo de Dados, Métodos de Coleta de Dados e Características de Dados Vetoriais, com isso, gerando o protocolo propriamente dito.

No Modelo de Dados deve-se apresentar o projeto VGI em detalhes, explicando a motivação e os objetivos. Isto facilita o contribuinte a entender o porquê e como coletar os dados. Propondo uma lista de camadas temáticas e mantendo a possibilidade dos contribuintes de criarem novas. Em relação ao dado contribuído, deve-se definir quais os tipos de geometrias que serão utilizadas; seus atributos (fixos ou dinâmicos) e regras para garantir a homogeneidade (como um objeto do mundo real deve ser mapeado). Deve-se incluir exemplos de casos de uso as camadas temáticas. Isto fará com que os colaboradores sejam encorajados a se familiarizar com o serviço antes de serem matriculados na coleta de dados. Sendo os contribuintes incitados a fornecer comentários e observações [Mooney et al. 2016].

No Métodos de Coleta de Dados descreve como a coleta será feita. A vetorização manual é a aquisição de dados vetoriais de mapas, imagens aéreas ou de satélite. Na tela, traçasse um mouse sobre os recursos exibidos em uma tela do computador, é o método mais popular para a aquisição de dados. A pesquisa de campo é a coleta de dados vetoriais usando equipamentos, como dispositivos Global Navigation Satellite System (GNSS), smartphones e etc. A importação em massa é a integração de dados vetoriais existentes no projeto VGI. Podem ser considerados dados espaciais de outras fontes de dados [Mooney et al. 2016].

Na Características de Dados Vetoriais são as características que são importantes

e devem ser levadas em consideração, como o Sistemas de Referência de Coordenadas (CRS); topologia e regras topológicas; nível de detalhe/escala; metadados (fonte dos dados, resolução, data/hora de digitalização, comentários sobre a qualidade das imagens, licença e etc) e qualidade dos dados [Mooney et al. 2016]

Um passo-a-passo, de como será a coleta de dados por parte dos usuários é proposto por Mooney el al. [Mooney et al. 2016], que será utilizado no projeto Pauliceia. A descrição dele é a seguinte:

- inicialização: o usuário deve se compreender as especificações do projeto, investindo algum tempo em um projeto de exemplo, para sanar dúvidas antes do início da contribuição real. O contribuinte deve verificar se os dados coletados são adequados para o projeto em relação ao conteúdo e qualidade.
- coleta de dados: o usuário deve planejar o processo de coleta de dados, separando uma parte do seu tempo para a coleta. Ele deve ter acesso as especificações do projeto durante a coleta de dados para consulta e fazê-la de acordo com as instruções. É aconselhável o voluntário anotar qualquer problema técnico do sistema ou situação problemática que encontre.
- controle de qualidade: o contribuinte deve revisar os dados coletados antes de serem enviados ao servidor. Deve-se verificar se os dados estão adequados em relação ao conteúdo geométrico, de metadados e de acordo com as especificações. Se encontrar erros, deve-se consertá-los editando.
- envio dos dados: após as revisões necessárias, o usuário enviará os dados coletados.
- verificação de envio de dados: é incentivado o voluntário a fazer uma verificação final aos dados que acabou de enviar, diferentemente do item anterior. Será analisado a qualidade dos dados em relação de coerência do projeto. Sendo feito em relação aos seus dados e de outros contribuintes.
- feedback para a comunidade: o projeto deve disponibilizar canais de discussão (listas de discussão, redes sociais e etc.) para que o usuário possa expressar seus comentários e observações.

4. Protocolo VGI para o projeto Pauliceia

Mooney el al. [Mooney et al. 2016] propõem um protocolo genérico, para ser aplicado a diversos projetos de VGI. Este protocolo visa criar uma padronização para os projetos que trabalham com dados geográficos colaborativos. Esta padronização pode acarretar em uma melhora na qualidade das contribuições feitas pelos usuários. Por conta disso, foi-se capturado ideias deste artigo para serem aplicadas ao projeto Pauliceia.

O cenário é aplicado ao projeto Pauliceia, descrito em detalhes no começo deste trabalho. Em suma é um projeto que visa construir uma plataforma computacional para pesquisa histórica, feita de maneira colaborativa. Esta plataforma permitirá que os pesquisadores de história possam manipular dados históricos (dados espaço-temporais) da cidade de SP. O período histórico se dá entre 1870 a 1940, pois foi uma época em que a cidade de SP cresceu muito e a área piloto é o centro da cidade.

Nesta seção é descrito o protocolo aplicado ao projeto Pauliceia.

4.1. Inicialização

Os pesquisadores de história, alunos ou quaisquer cidadãos entrarão no portal Pauliceia utilizando um navegador, pelo computador. Para fazer o gerenciamento dos dados, é necessário que eles façam um cadastro ou acessem com um login social. Antes de começar a contribuição, o usuário deverá aceitar um termo de Política de Uso do projeto. Este termo descreve, que o sistema não se responsabiliza pelas contribuições e que os dados do projeto são públicos, portanto não se pode inserir dados com direitos autorais, logo o usuário se responsabiliza disso.

Os voluntários são encorajados a preencherem os campos opcionais em seu perfil, como instituição que faz parte, nível (graduação, mestrado ou etc), entre outros. Isto é feito, pois como o projeto é voltado para academia, é conveniente que se tenha esse tipo de informação para determinar a confiabilidade das contribuições de um usuário, baseado no seu perfil acadêmico. Por exemplo, supõe-se que os dados contribuídos por um usuário com pós-doutorado seja mais confiável, do que o de um aluno de graduação. O procedimento que o voluntário deve seguir antes de iniciar a coleta de dados é o seguinte:

- ler sobre o projeto Pauliceia, suas especificações e objetivos, que estarão disponíveis no site. Isto fará que o usuário se familiarize mais com o sistema, fazendo-o entender o porquê e como coletar os dados. Os usuários são encorajados a fazer comentários e dar sugestões da plataforma, enviando um e-mail para `support_pauliceia@googlegroups.com`.
- simular a coleta de dados em um ambiente "sandbox", que é um ambiente que contêm exemplos de como utilizar o sistema. Neste ambiente o usuário poderá sanar suas principais dúvidas, ter um treinamento prévio da ferramenta e se acostumar com os processos da coleta de dados, fazendo-o se sentir confiante para inserir os dados reais.
- ter certeza que os dados são de domínio público. Os dados não podem ter alguma licença privada.
- criar um documento, de pelo menos uma página, no formato PDF, explicando e detalhando sobre quais dados serão trabalhados, a licença deles, quem são os responsáveis e qualquer outra questão que acredite ser necessária. Este documento deve ser anexado no projeto que o usuário criar.

4.2. Modelo de Dados

Serão usadas diversas de camadas temáticas que agrupam os dados de acordo com o tipo (por exemplo, ocorrências de crime em uma região ou pontos de incêndios em determinado ano). Estas camadas estarão associadas a um projeto. As camadas-base são mapas de plano de fundo no navegador para auxiliar na coleta de dados vetoriais. Um mapa atual da cidade de São Paulo será provido pelo OpenStreetMap (OSM) e haverá um conjunto de cartas antigas da área central de São Paulo, como a do Sara Brasil [Gestão Urbana SP 2017], para auxiliar no processo de vetorização desses dados históricos.

Um projeto é um conjunto de contribuições feitas por um ou mais usuários, indicando uma camada no sistema. Por exemplo: um pesquisador trabalha com dados de crimes em 1900, então este projeto será uma camada dos crimes que ocorreram naquela data. O historiador poderá criar um projeto relacionado a essa sua pesquisa, inserindo seus

dados no sistema. Esse pesquisador, dono do projeto, poderá adicionar outros usuários a ele (como seus alunos ou colaboradores) que podem manipular os dados históricos desse conjunto. Um voluntário só poderá modificar dados de projetos a qual ele pertença. Se algum contribuinte, fora do projeto, perceber que há dados com má qualidade, ele não poderá editá-los manualmente, porém poderá solicitar uma revisão aos dados na página "Revisões" do projeto.

O Sistema de Referência de Coordenadas (CRS) escolhido para a vetorização dos mapas e inserção dos dados é o WGS84 (EPSG:4326). Este CRS foi definido por cobrir o mundo inteiro [Spatial Reference 2017], deixando o sistema mais genérico e possivelmente internacionalizável, para caso o projeto ganhe repercussão em outros países.

As contribuições poderão ser compartilhadas nas mídias sociais, como Facebook e Google+.

Como o projeto Pauliceia é voltado para pesquisa histórica urbana da cidade de São Paulo de 1870 a 1940, os dados a serem coletados devem se restringir a esse espaço e período de tempo. Mais precisamente, os dados devem abranger a área central de São Paulo, que constitui a área piloto definida. Os tipos de dado que podem ser coletados e inseridos na plataforma serão:

- Dados de vetoriais: dados históricos com uma localização geográfica, por exemplo: ocorrência de um assalto em 1930, na Rua São Bento, n. 3 (um ponto); uma área de epidemia de uma doença ou edifício (um polígono); vetorização de uma carta antiga (uma linha) e etc. Os dados históricos contêm atributos fixos (como nome, data início e fim, fontes), todavia o usuário pode inserir atributos dinamicamente. O atributo fonte descreve a origem dos dados, se ele é de um artigo, livro, dissertação ou outros, nele conterá a referência bibliográfica do documento original do dado. A referência pode ser descrita seguindo as normas da ABNT ou em bibtex (dica: no Scholar[9] do Google, quando se pesquisa por algum trabalho, tem a opção "citar" onde têm as bibliografias prontas). Pode-se inserir mais de uma fonte.
- Contribuições textuais, que serão georreferenciadas usando o endereço fornecido. O usuário insere, por exemplo, a ocorrência de um roubo no ano de 1900, na rua São Pio X, n. 42; nisso o sistema se encarregará, dado o endereço, de geolocalizar no mapa o lugar. O dado será inserido como um ponto. Esta tipo de dado será usado principalmente para importação de dados em grande volume, para evitar a geolocalização manual de grandes quantidades de dados.
- Dados de mídia (em anexo a um dado vetorial ou textual ou um dado de mídia georreferenciado) como fotos históricas, vídeos, depoimentos em áudio e etc. Esses arquivos devem ser armazenados em um repositório, cujo link será adicionado no projeto. Por exemplo: os vídeos são colocados no Youtube, as imagens no Google Photos e documentos são adicionados no Google Drive (ou Dropbox); na contribuição será adicionada a URLs pública do anexo. É aconselhável ao usuário criar uma conta do projeto no repositório que for usar, para que seja evitado excluir acidentalmente, se estiver na conta pessoal. Em relação as fotos históricas, será possível fazer o upload no sistema, caso seja necessário.

[9] https://scholar.google.com.br/

- Dados de pesquisas, como artigos e outros trabalhos. Se o documento for o trabalho de origem do dado, é aconselhável que coloque a URL do site no atributo fonte. Caso contrário, o documento pode ser adicionado como um dado de mídia, como descrito acima.

Um atributo importante que os dados coletados devem possuir é uma data, que indica para qual período aquele dado é válido. A data de início representa quando a contribuição ocorreu ou começou a existir, e data final representa quando a contribuição terminou de acontecer ou deixou de existir. O formato da data pode ser com os três campos completos, com mês e ano ou apenas ano. Isto faz com que temos os seguintes casos:

- caso exista data inicio e final, o usuário deve colocá-los, pois representam a existência temporal dele neste período. Por exemplo: a contribuição tem data início igual à 13/05/1917 e final 13/10/1917, logo a contribuição existirá neste período.
- quando não houver certeza de início e fim do período, a data será considerada discreta, se encaixando em uma das seguintes categorias, onde é sabido a:
  - data início, mas não a final;
  - data final, mas não a de início;
  - data pontual.

  Caso exista a data início, mas não a final, ou vice-versa, o usuário deve preencher somente a data que possuir certeza e a data em aberto (não preenchida) ficará disponível como uma data não certa, com uma abrangência máxima de dois anos. Por exemplo: a contribuição tem data início igual à 13/05/1917, mas a data final em aberto. Isto fará com que a contribuição "exista" entre 13/05/1917 e 13/05/1919, pois a data final está em aberto.
  A data pontual é quando o usuário sabe somente uma aproximação da existência de uma contribuição, não sabendo uma data exata. Se isto ocorrer, o contribuinte deve informar o período que sabe da existência da dado. Por exemplo: o voluntário sabe que um teatro existiu em 1930, mas não sabe quando foi construído ou demolido, então ele deve colocar a data início e final iguais à 1930.
  No caso de data discreta haverá um checkbox para o usuário indicar se há ou não este fato.

4.3. Métodos de Coleta de Dados

Certas recomendações são descritas para que seja este processo seja realizado com sucesso, como: (1) o usuário deve reservar um tempo suficiente para o sucesso dos processos do projeto; evitando distrações e se possível, concentrar o tempo em uma única quantidade, em vez de vários estágios curtos, porque melhora a qualidade dos dados; (2) ter acesso as especificações do projeto, para fazer consulta ou sanar dúvidas e (3) fazer a coleta de dados de acordo com as especificações do projeto. O contribuinte deve anotar qualquer problema técnico do sistema ou situação problemática que encontre, excluindo-se os por mal funcionamento do dispositivo.

A coleta de dados vetoriais pode acontecer em dois cenários diferentes: HistMapathons ou contribuições individuais. Um HistMapathon (Maratona de Mapeamento Histórico) será uma reunião organizada de pessoas com o objetivo de vetorizar mapas

antigos de São Paulo de 1870 a 1940, semelhantes aos realizados pelo Google Maps [Tech2 2017] e OpenStreetMap [OpenStreetMap 2017b]. Esses encontros serão promovidos pelos historiadores e seus alunos. O objetivo principal desses eventos é promover a contribuição em massa de dados vetoriais históricos para a plataforma Pauliceia. Um colaborador também pode vetorizar individualmente fora de um HistMapathon, se desejado.

A Figura 1 ilustra o processo de coleta de dados, onde um usuário poderá fazer a vetorização de mapas antigos ou inserir dados históricos. As contribuições de dados históricos poderão ser feitos manualmente ou por importação em massa, onde o próprio usuário poderá fazer uma auto-avaliação deles. Em relação a vetorização de mapas, vários usuários podem vetorizar o mesmo objeto (e.g. edifício), por conta disso, depois da coleta um algoritmo achará o melhor consenso entre as geometrias inseridas. Todas contribuições serão salvas em um banco de dados:

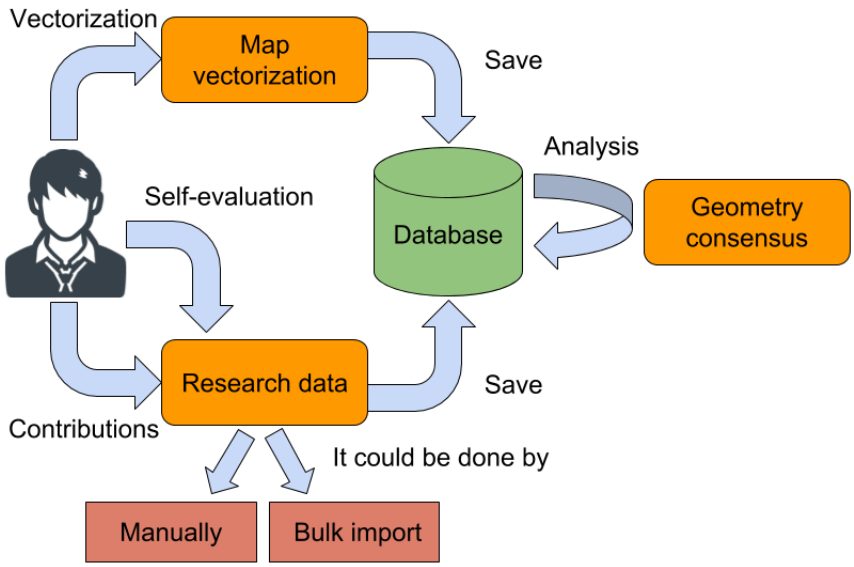

Figura 1. Processo de coleta de dados

Para que as contribuições feitas via vetorização manual obtenha o melhor resultado possível, são recomendadas algumas práticas durante o processo [Mooney et al. 2016]: (1) uso de um mouse ao invés de um touchpad, pois auxilia na precisão do posicionamento dos dados; (2) verificar o nível de detalhe (zoom) da imagem do mapa no navegador para a melhor coleta dos dados e (3) deve ser realizado de acordo com as especificações indicadas no modelo de dados e nas características dos dados.

A pesquisa de campo se refere à coleta de dados usando equipamentos como smartphones e equipamentos tecnologia GNSS. Como as contribuições a serem inseridas são de dados históricos, a pesquisa de campo não se qualifica como um método aplicável a esse projeto.

A importação em massa se refere à integração de dados na plataforma, em grandes quantidades. Nesse projeto, a importação em massa será permitida para dados tex-

tuais, onde (1) os dados necessitam estar em formato CSV. Se eles não estiverem, é necessário convertê-los; (2) o usuário deve enviar um e-mail para `import_pauliceia@googlegroups.com` para discutir com os administradores sobre o plano dos dados. Enviar o nome do projeto em questão e uma amostra dos dados (pelo menos duzentos registros, se tiver) junto. Não será possível carregar os dados, sem antes os administradores fazerem uma revisão do documento e dos dados e (3) se os administradores concordarem e for dada a permissão para importação em massa de dados, o contribuinte deve seguir com o plano de trabalho, fornecendo atualização frequente no mesmo e-mail.

4.4. Características dos dados

É necessário que o usuário forneça metadados para as contribuições que faz. Alguns tipos de metadados como conversões de Sistemas de Coordenadas realizados, fonte dos dados (que pode ser feita como uma referência bibliográfica ao documento original), explicação e detalhamento dos dados a serem trabalhados e a licença que possuem, devem ser inseridos nos campos correspondentes de metadados que serão fornecidos ou podem ser anexados ao projeto, como um documento. Recomenda-se que o colaborador faça comentários a respeito da qualidade do dado ou de possíveis problemas ocorridos durante sua coleta. Já outros metadados, como qual usuário que inseriu o dado, data de inserção, resolução do mapa e modificações feitas naquela contribuição serão coletadas automaticamente e armazenada em arquivos de log.

4.5. Tipos de usuários

Existem 3 tipos de usuário no projeto, que tem acesso às seguintes funcionalidades (de acordo com tipo):

1. administrador: usuário com habilidade de gerenciar os usuários e o sistema, bem como contribuir com dados para a plataforma. Este tipo pode dar privilégios de moderador a um usuário autenticado.
2. autenticado: usuário com cadastro. Ele pode:
    - fazer download dos dados do projeto, como ruas, eventos, edificações, mapas históricos, anexos e etc.
    - adicionar comentários relacionados as contribuições existentes. Haverá uma página do projeto, chamada Revisões, onde o voluntário poderá solicitar revisões sobres os dados de projetos dos quais eles não tem acesso direto a correção.
    - criar um projeto relacionado a uma pesquisa (como crimes em 1890, teatros em 1910 e etc.). Este projeto é um conjunto de contribuições feitas. Se pesquisador o projeto será relacionado a sua pesquisa profissional; se aluno, terá relação com uma pesquisa de, por exemplo, mestrado ou referente a um artigo; se outro, alguma pesquisa pessoal de interesse. Este projeto será um conjunto de dados históricos que podem ser editados colaborativamente por outros membros do projeto.
    - se tiver privilégios de moderador, pode dar permissão de adição de dados em grande volume para outros usuários autenticados.
3. sem senha: usuários que não contêm um login. Poderão acessar o portal apenas para consulta e visualização simples dos dados.

Por questões de segurança, as ações principais do usuário (como manipular dados históricos) serão salvos em históricos (arquivos de log). Isto poderá ajudar a rastrear, com mais facilidade, usuários mal intencionados que queiram danificar o sistema. Isto ajuda a definir a confiabilidade e reputação dos contribuintes [Davis Jr et al. 2013].

4.6. Controle de qualidade

É necessário que os contribuintes leiam este protocolo, antes de começarem a coleta de dados. Isto colabora para a sanação de dúvidas, evitando que problemas possam acontecer, aumentando a qualidade dos dados inseridos.

Os usuários devem fazer uma auto-avaliação dos dados que eles inseriram, uma revisão dos dados antes de serem enviados ao servidor. Esta revisão será principalmente em relação aos seus dados que acabou de contribuir. Eles revisarão se os dados estão coerentes, se tem qualidade adequada, verificando se a geometria e seus atributos estão corretos, de acordo com as especificações e baseados no seu conhecimento pessoal da área de pesquisa.

Esta auto-avaliação é manual, o usuário verificará se a geometria e os atributos estão corretos. Este tipo de avaliação para o projeto Pauliceia é importante, pois público alvo do portal são os historiadores, que tem um conhecimento prévio do assunto, logo eles estão capacitados a corrigir eventuais problemas. Como diria a Lei de Linus: "Dado uma quantidade suficiente de olhos, todos os bugs são superficiais" [Raymond 2017]. Caso o contribuinte encontre erros nos dados, ele deve consertá-los editando. Lembrando que o usuário só pode editar as contribuições suas e de outros que estão no mesmo projeto. Se o voluntário necessitar modificar dados de outros projetos, ele deverá solicitar revisão deles na página de Revisões.

Tem-se a intenção de utilizar um algoritmo (como o proposto por Budig et al. [Budig et al. 2016]) para a recuperação da melhor representação poligonal ou linear (melhor consenso), dado várias geometrias de um mesmo objeto. Isto será utilizado principalmente nos dados recebidos nos HistMapathons.

Baseado nas recomendações de Cechanowicz et al. [Cechanowicz et al. 2013] e Hamari et al. [Hamari et al. 2014], no contexto do projeto Pauliceia, utilizaremos algumas técnicas de Gamificação para incentivar a participação dos usuários, instigando-os na qualidade dos dados.

Gamificação é o uso de elementos de jogo em contextos que não sejam jogos. É uma forma de melhorar o engajamento e a motivação dos usuários, aumenta a participação e os proporciona uma melhor experiência. Isto pode ser utilizado para obter mais de um contribuinte: mais dados, dados de maior qualidade, aumento da frequência e duração da participação [Cechanowicz et al. 2013]. É um processo de melhoramento dos serviços com recursos estimulantes, utilizando as experiências de jogos para obter resultados comportamentais adicionais. É usado para apoiar o empenho dos usuários e melhorar o uso de serviços. Por exemplo: aumentar a atividade do usuário, interação social, qualidade e produtividade de suas ações [Hamari et al. 2014].

Uma das técnicas de Gamificação que será utilizado no projeto é o de Ranking. O voluntário começará com uma quantidade de "pontos" igual à zero, podendo ganhar ou perder esses pontos ao decorrer da utilização do sistema. Isto instigará o usuário a con-

tribuir mais e com mais qualidade, pois se contribuir com frequência, fazer comentários e receber boas votações em seus dados, seu Ranking aumentará. Se as contribuições forem de má qualidade, será recebido votações negativas, portanto o Ranking do voluntário diminuirá. A aquisição de pontos será feita através do esquema de votação nos dados e obtenção de emblemas.

A reputação dos contribuintes será dada pelo seu ranking em relação aos dos demais.

4.7. Envio dos dados

Depois de todas as verificações, o usuário clicará no botão "salvar", enviando suas mudanças para o servidor do projeto Pauliceia.

4.8. Última verificação da contribuição

A partir desse momento as contribuições estão disponíveis para a comunidade do projeto. O usuário deve fazer uma verificação final nos dados que acabou de enviar. Analisar a qualidade dos dados de acordo com a coerência do projeto e aos demais dados dos outros contribuidores. Caso os erros sejam detectados o usuário deve editá-los. Esta operação aplica-se tanto aos dados que o próprio usuário coletou quanto aos coletados por outros.

4.9. Feedback para a comunidade

Um projeto colaborativo melhora à medida que mais usuários contribuem para isso, portanto é interessante o usuário fornecer um feedback sobre sua experiência. O contribuinte pode expressar seus comentários, opiniões e observações nos canais disponíveis do projeto, como: lista de discussão[10], grupo no Facebook[11] e página no Facebook[12].

O contribuinte é encorajado a descrever se o processo de coleta de dados foi fácil ou difícil, e o porquê. Escrevendo os problemas ou situações inesperadas que encontrou, sugerindo melhorias ou mudanças. Explicando precisamente o que aconteceu, para facilitar a compreensão dos administradores e aplicar uma correção adequada.

É aconselhável que o usuário divulgue o projeto para pessoas que conheça, com intenção de atrair novos contribuintes. Isto fará com que a plataforma cresça e melhore com frequência. Quanto mais houver participantes, mais o sistema se tornará rico em seus dados, consequentemente melhorando a qualidade deles.

5. Conclusão

VGI gera uma grande quantidade de dados com características heterogêneas, tornando a comparações tradicionais já não tão viáveis. Por conta disso, Goodchild e Li [Goodchild and Li 2012] propõem três abordagens para garantir a qualidade do VGI: crowd-sourced (multidões), sociais e geográficas. Senaratnea et al [Senaratne et al. 2017] ainda descreve uma abordagem extra que é a mineração de dados, que ajuda a avaliar a qualidade, descobrindo padrões e aprendendo com o suporte dos dados. A mineração de dados pode ser usada como uma abordagem autônoma, independente do conhecimento

[10] pauliceia@googlegroups.com
[11] https://www.facebook.com/groups/pauliceia2.0/
[12] fb.me/pauliceia2.0

em geografia, abordagens sociais ou de pessoas para avaliar a qualidade das contribuições. Por exemplo, pode-se utilizar métodos de detecção de outlier, análise de cluster, análise de regressão, correlação e etc, para avaliar a qualidade dos dados descobrindo e aprendendo padrões de dados.

Gamificação é uma abordagem que está sendo utilizada para envolver pessoas a contribuir com dados espaciais (por exemplo: Foursquare e Ingres). Tais maneiras de gamificação aumentam a participação, bem como a cobertura espacial. Os voluntários ganham incentivos através desta abordagem de coleta de dados, como aumentando o seu ranking, ganhando emblemas e etc. Isto pode ser utilizado para controlar o processo de coleta de dados, tornando-os mais precisos, aumentando a qualidade deles [Antoniou et al. 2010] [Antoniou and Schlieder 2014] [Yanenko and Schlieder 2014].

Estes mecanismos encorajam e motivam as pessoas a contribuir, mas não só, deve-se utilizar métodos para conscientização das pessoas em relação a qualidade do conteúdo dos dados. Uma maneira que pode ser utilizada é dar um ranking para o voluntário, baseado na qualidade dos dados que ele inserir. Revelar os ranking para a comunidade incentivará os usuários a terem mais cuidado (atenção) à qualidade de suas contribuições. A utilização desses recursos para medir a qualidade de VGI baseado em mapas é possível, como descrito acima, porém não é tão utilizado em VGI baseado em imagem e texto. Isto ocorre em especial, por conta da complexidade dos dados de imagem e texto, pois a análise sistemática desses recursos não é simples. É visível que não existe um único método que possa resolver todos os problemas relacionados à qualidade das contribuições de VGI. Isto se dá essencialmente pela característica heterogênea que são esses dados. Por conta disso, temos que ter em mente que existem várias técnicas para resolver diversos tipos de problemas, havendo também suas limitações. Trabalhar em cima dessas limitações, pode, por exemplo, melhorar os métodos já existentes, aumentando a contribuição científica em estudos sobre VGI [Senaratne et al. 2017].

A especificação de um protocolo bem estruturado para o projeto Pauliceia dá a oportunidade de ajudar a melhorar a qualidade dos dados que serão coletados na plataforma final. Sem contar que, contribui para a definição das funcionalidades do Serviço Web para VGI.

As questões principais discutidas neste trabalho estão sendo incorporas no serviço web para VGI do projeto Pauliceia na forma de funções bem definidas. As funcionalidades já existentes são de autenticação de usuário e gerenciamento de geometrias. Fica-se então para trabalhos futuros: detalhar mais o protocolo do projeto, com o intuito de melhorá-lo; a implementação da Gamificação e construção do algoritmo para achar o melhor consenso, dado várias geometrias.

6. Agradecimento

Agradeço à FAPESP pela concessão da bolsa, referente ao processo nº 2017/03852-9, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Agradeço também ao Dr. Antônio M. V. Monteiro e ao Dr. Sergio Rosim pela oportunidade de fazer este trabalho na matéria e supervisão geral.

Referências

Andrienko, G., Andrienko, N., Bak, P., Kisilevich, S., and Keim, D. (2009). Analysis of community-contributed space-and time-referenced data (example of flickr and panoramio photos). In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 213–214. IEEE.

Antoniou, V., Morley, J., and Haklay, M. (2010). Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1):99–110.

Antoniou, V. and Schlieder, C. (2014). Participation patterns, vgi and gamification. In *Proceedings of the 17th AGILE Conference on Geographic Information Science, Castellón, Spain*, pages 3–6.

Antoniou, V. and Skopeliti, A. (2015). Measures and indicators of vgi quality: An overview. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., and Shirk, J. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11):977–984.

Bosch, H., Thom, D., Heimerl, F., Püttmann, E., Koch, S., Krüger, R., Wörner, M., and Ertl, T. (2013). Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031.

Budig, B., van Dijk, T. C., Feitsch, F., and Arteaga, M. G. (2016). Polygon consensus: smart crowdsourcing for extracting building footprints from historical maps. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 66. ACM.

Carvalho, D. F. (2017). Café, ferrovias e crescimento populacional: o florescimento da região noroeste paulista. `http://www.historica.arquivoestado.sp.gov.br/materias/anteriores/edicao27/materia02/`. Accessed on 04/04/2017.

Cechanowicz, J., Gutwin, C., Brownell, B., and Goodfellow, L. (2013). Effects of gamification on participation and data quality in a real-world market research domain. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, pages 58–65. ACM.

Chunara, R., Andrews, J. R., and Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American journal of tropical medicine and hygiene*, 86(1):39–45.

Ciepłuch, B., Jacob, R., Mooney, P., and Winstanley, A. C. (2010). Comparison of the accuracy of openstreetmap for ireland with google maps and bing maps. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, page 337. University of Leicester.

Craglia, M., Ostermann, F., and Spinsanti, L. (2012). Digital earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth*, 5(5):398–416.

Davis Jr, C. A., de Souza Vellozo, H., and Pinheiro, M. B. (2013). A framework for web and mobile volunteered geographic information applications. In *GeoInfo*, pages 147–157.

De Tré, G., Bronselaer, A., Matthé, T., Van de Weghe, N., and De Maeyer, P. (2010). Consistently handling geographical user data. *Information processing and management of uncertainty in knowledgebased systems, applications*, 28:85–94.

Digital Harlem Blog (2017). The project. `http://wiki.openstreetmap.org/wiki/Mapathon`. Accessed on 18/09/2017.

Estellés-Arolas, E. and González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200.

Flanagin, A. J. and Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72(3-4):137–148.

Fuchs, G., Andrienko, N., Andrienko, G., Bothe, S., and Stange, H. (2013). Tracing the german centennial flood in the stream of tweets: first lessons learned. In *Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, pages 31–38. ACM.

Gestão Urbana SP (2017). Prefeitura disponibiliza mapa histórico de 1930 no geosampa. `http://gestaourbana.prefeitura.sp.gov.br/noticias/prefeitura-disponibiliza-mapa-historico-de-1930-no-geosampa/`. Accessed on 04/04/2017.

Girres, J.-F. and Touya, G. (2010). Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14(4):435–459.

Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of information science*, 32(2):198–208.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Goodchild, M. F. and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, 1:110–120.

Guptill, S. C. and Morrison, J. L. (2013). *Elements of spatial data quality*. Elsevier.

Haklay, M. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B: Planning and design*, 37(4):682–703.

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How many volunteers does it take to map an area well? the validity of linus' law to volunteered geographic information. *The Cartographic Journal*, 47(4):315–322.

Hamari, J., Koivisto, J., and Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 3025–3034. IEEE.

Hollenstein, L. and Purves, R. (2010). Exploring place through user-generated content: Using flickr tags to describe city cores. *Journal of Spatial Information Science*, 2010(1):21–48.

Huberman, B. A., Romero, D. M., and Wu, F. (2008). Social networks that matter: Twitter under the microscope.

Jacob, R., Zheng, J., Ciepłuch, B., Mooney, P., and Winstanley, A. C. (2009). Campus guidance system for international conferences based on openstreetmap. In *International Symposium on Web and Wireless Geographical Information Systems*, pages 187–198. Springer.

Jakobsson, A. and Giversen, J. (2007). Guidelines for implementing the iso 19100 geographic information quality standards in national mapping and cadastral agencies. *Eurogeographics Expert Group on Quality*.

Keßler, C., Maué, P., Heuer, J., and Bartoschek, T. (2009). Bottom-up gazetteers: Learning from the implicit semantics of geotags. *GeoSpatial semantics*, pages 83–102.

Keßler, C., Trame, J., and Kauppinen, T. (2011). Tracking editing processes in volunteered geographic information: The case of openstreetmap. In *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory*, volume 12.

Koukoletsos, T., Haklay, M., and Ellul, C. (2012). Assessing data completeness of vgi through an automated matching procedure for linear data. *Transactions in GIS*, 16(4):477–498.

Liu, S. B., Palen, L., Sutton, J., Hughes, A., and Vieweg, S. (2008). In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster. In *Proceedings of the information systems for crisis response and management conference (ISCRAM)*, pages 969–980.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J. (2011). Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 181–190. IEEE.

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3):185–204.

Mooney, P., Minghini, M., Laakso, M., Antoniou, V., Olteanu-Raimond, A.-M., and Skopeliti, A. (2016). Towards a protocol for the collection of vgi vector data. *ISPRS International Journal of Geo-Information*, 5(11):217.

Mota, P. d. B. et al. (2007). A cidade de são paulo de 1870 a 1930: café, imigrantes, ferrovia, indústria.

OpenStreetMap (2017a). About openstreetmap. `http://wiki.openstreetmap.org/wiki/About_OpenStreetMap`. Accessed on 01/09/2017.

OpenStreetMap (2017b). Mapathon. `http://wiki.openstreetmap.org/wiki/Mapathon`. Accessed on 18/09/2017.

Over, M., Schilling, A., Neubauer, S., and Zipf, A. (2010). Generating web-based 3d city models from openstreetmap: The current situation in germany. *Computers, Environment and Urban Systems*, 34(6):496–507.

Pilz, D., Ballard, H. L., and Jones, E. T. (2006). Broadening participation in biological monitoring: handbook for scientists and managers.

Raymond, E. S. (2017). Release early, release often. http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/ar01s04.html. Accessed on 01/09/2017.

Robertson, S. (2016). Digital mapping as a research tool: Digital harlem: Everyday life, 1915–1930. *The American Historical Review*, 121(1):156–166.

Robinson, S., Jones, M., Williamson, J., Murray-Smith, R., Eslambolchilar, P., and Lindborg, M. (2012). Navigation your way: from spontaneous independent exploration to dynamic social journeys. *Personal and Ubiquitous Computing*, 16(8):973–985.

Schmidt, M., Klettner, S., and Steinmann, R. (2013). Barriers for contributing to vgi projects. In *Proc. ICC*, volume 13.

Secretaria Municipal de Urbanismo e Licenciamento (2017). História demográfica do município de são paulo. http://smul.prefeitura.sp.gov.br/historico_demografico/tabelas/pop_brasil.php. Accessed on 04/04/2017.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., et al. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5):55.

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

Spatial Reference (2017). Epsg:4326. http://spatialreference.org/ref/epsg/wgs-84/. Accessed on 04/08/2017.

TC, I. (2009). Iso/tc 211 geographic information/geomatics.

Tech2 (2017). Why is google's mapathon in hot waters in india? all you need to know. http://www.firstpost.com/tech/news-analysis/why-is-googles-mapathon-in-hot-waters-in-india-all-you-need-to-know-
html. Accessed on 18/09/2017.

Tobler, W. (2004). On the first law of geography: A reply. *Annals of the Association of American Geographers*, 94(2):304–310.

Valli, C. and Hannay, P. (2010). Geotagging where cyberspace comes to your place. In *Security and Management*, pages 627–632.

Vandecasteele, A. and Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system: the case of openstreetmap. In *OpenStreetMap in GIScience*, pages 59–80. Springer.

White, J. W. and Gilbert, H. (2016). Laying the foundation.

Yanenko, O. and Schlieder, C. (2014). Game principles for enhancing the quality of user-generated data collections. In *Proc. AGILE, workshop geogames geoplay*, pages 1–5.