ISSN 1828-5961

Clodoveu Augusto Davis Júnior
Civil engineer, Master's and PhD in Computer Science from Federal University of Minas Gerais. Professor at the Computer Science Department of UFMG, teaching graduate and undergraduate courses. Main research subjects: spatial databases, spatial data infrastructures, geographic data modeling, geocoding and urban GIS applications.

Challenges in Crowdsourcing Geospatial Data to Replace or Enhance Official Sources

Desafios para Substituir ou Aprimorar Fontes Oficiais de Dados Geoespaciais através de Crowdsourcing

Urban dwellers demand geospatial data on various aspects of contemporary life. Governmental sources account for several important data categories, but not necessarily up-to-date, or even covering all aspects. Technologies such as smartphones allow citizens to become geospatial data producers and to contribute with valuable georeferenced data in near real time. Crowdsourced often has problems in aspects such as coverage, reliability and positional accuracy. This paper focuses on challenges that exist in the integration of geospatial data provided by governmental or corporate sources to crowdsourced data, provided dynamically and voluntarily by concerned citizens, or unconsciously through online tools and social networks.

Moradores urbanos demandam dados geoespaciais sobre muitos aspectos da vida contemporânea. As fontes governamentais disponibilizam categorias de dados importantes, mas não necessariamente atualizadas, ou mesmo sobre todos os aspectos. Tecnologias como smartphones permitem que os cidadãos sejam produtores de dados e contribuam com dados georreferenciados valiosos, em tempo quase real. Dados por crowdsourcing têm problemas relativos a cobertura, confiabilidade e precisão locacional. O artigo aborda os desafios da integração de dados geoespaciais de fontes governamentais ou corporativas aos dados de crowdsourcing, fornecidos dinamicamente e voluntariamente por cidadãos interessados, ou inconscientemente por ferramentas on-line e redes

Keywords:
Crowdsourcing; urban computing; urban GIS; Volunteered Geographic Information
Palavras-chave:
Crowdsourcing; computação urbana; SIG urbano; informação geográfica voluntária

ISSN 1828-5961

1. INTRODUCTION

Traditionally, data that comprise urban geographic information systems (GIS) are produced by public or private organizations whose responsibilities include tax-oriented cadastral surveys, infrastructure management, or distributed public services, as in the case of utility companies. Such traditional data sources, however, are increasingly unable to create and maintain detailed datasets on various aspects of urban life. Today's citizens, equipped with Internet-connected smartphones and their array of sensors, services and applications, demand more geographic information for their daily activities.

As a result, a wide array of new information providers supplies valuable information that is not connected or derived from the traditional sources. Many of the features in OpenStreetMap [1] (OSM), and the contents of location-based services such as Google Places [2] , Foursquare [3] , Yelp [4] and others go beyond basic cartography and official data sources, and complement them with data on mobility, commercial activities, points of interest and many other information categories.

Much of the content found in these novel sources, however, is provided by volunteers, through crowdsourcing platforms such as OSM, or by business owners themselves, based on their interest to show up in widely used apps. There is no guarantee, for these alternative sources, of broad or complete coverage, of uniform quality, or of unbiased content generation (Goodchild, 2007a). If there are more volunteers who know a region and are able to contribute, data density and quality in that region tends to be higher (Flanagin and Metzger 2008, Haklay 2010, Goodchild and Li 2012). Businesses that do not seek the attraction of passersby or of online users, such as travelers, are not attracted by these platforms. Examples include law firms and business-to-business companies (Wenceslau, Davis Jr. et al. 2017).

As a result, we observe that, nowadays, useful and relevant geographic data are not always cartographic in nature, not always obtained from official sources, and, as a consequence, not always complete or homogeneous. Nevertheless, we expect such data to complement, enhance or even partially replace official sources. In the next sections, we will explore ideas and limitations in that direction

2. COLLABORATIVE DATA COLLECTION

The quantity and variety of spatial data available online to the common citizen have increased rapidly. Since the introduction of Google Earth, in 2004, and of Google Maps, in 2006, there is a growing interest on tools and resources that allow people to spatially locate points of their interest and that offer location-based services, such as address location and vehicle routing. Such a wide interest expanded the range of possibilities to other directions. Following the ideas behind the Web 2.0 movement, common individuals, equipped with low-cost and widely available hardware and software, have become important actors in the creation, dissemination and maintenance of geographic information (Goodchild 2007b, Kuhn 2007). Many recent initiatives show that people are willing to contribute, voluntarily and with no expectancy as to economical rewards, to the creation of new datasets and for the updating or correction of errors in existing ones (Krumm 2007, Maguire 2007, Silva and Davis Jr 2008).

The costs involved in creating new and detailed geographic information, typical of urban applications, are no longer entirely supported by governmental mapping agencies. In place of fully centralized and official urban GIS datasets that were relatively common ten to twenty years ago, there is currently the need to deal with a patchwork of data sources, in which official data producers participate only with themes that comprise their institutional responsibilities and require systematic updating (Craglia, Goodchild et al. 2008).

Citizens can be incorporated in this process, by supplying (1) indications about where the official data are disparate from the reality, (2) current information obtained on-site, using equipment such as smartphones, (3) dynamic, up-to-the-minute information on distributed urban events, and (4) their personal knowledge on details of their local reality, as urban dwellers or community members. This involvement of potentially large groups of citizens in data collection and production is identified by many names, as described in the next section, but we will use the term crowdsourcing from there on.

2.1. CONCEPTS AND TERMINOLOGY VARIATIONS

The literature that discusses crowdsourcing of geographic data and information brings various designations for the process. Neogeography (Turner 2006) reflects the set of techniques and tools that do not fit traditional geography, and are employed by regular citizens that are familiarized with their surroundings (Haklay, Singleton et al. 2008, Goodchild 2009). Public Participation Geographic Information Systems (PPGIS) seek incentivizing communities to participate politically, in governmental actions of their local interest (Drew 2003, Elwood 2006, Johnson and Sieber 2013, Lin 2013). Some initiatives use the expression Citizen Science to characterize data collected by non-professional people that focus on contributing to scientific projects (Haklay 2010, Haklay 2013).

The expression Volunteered Geographic Information (VGI) (Goodchild 2007a) is also used in thee context of allowing volunteers to contribute with initiatives of geographic data collection and map updating, in a perspective that employs "citizens as sensors" (Goodchild 2007b).

A useful taxonomy to differentiate and classify crowdsourcing techniques is proposed by Mateveli, Machado et al. (2015). This taxonomy considers two dimensions: the process used for data capture, and the level of participation of the user, or volunteer. Figure 1 shows the taxonomy schematically. Each quadrant consists in a context for the combination of the two dimensions. (Fig. 01).

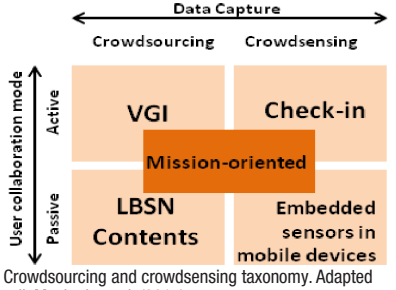

Figure 1 - Crowdsourcing and crowdsensing taxonomy. Adapted from Mateveli, Machado et al. (2015).

a) Active crowdsourcing - takes place when the user supplies information consciously, typically as a volunteer, and knows how the information is supposed to be used. Examples include the OpenStreetMap [1] initiative and the Flooding Points project (Hirata, Giannotti et al. 2015).

b) Passive crowdsourcing - occurs when information is obtained from material posted by the user for other ends, that is, useful information are extracted from content placed online by someone without the explicit involvement of such user. A common example is the extraction of information from location-based social networks, such as Twitter, based on the message text or on specific hashtags (Davis Jr, Pappa et al. 2011, Gomide, Veloso et al. 2011, Chatfield and Brajawidagda 2014).

c) Active crowdsensing - corresponds to the active collection and sharing of information that is captured by sensors embedded in mobile devices. For instance, a user checks in a point of her interest, whose position is determined by the smartphone's GPS (Ballatore, MacArdle et al. 2010, Mytilinis, Giannakopoulos et al. 2015). Another example involves the active collection of sound level data by a noise pollution denouncement application (Vellozo, Pinheiro et al. 2013).

d) Passive crowdsensing - happens when information is captured by sensors in mobile devices, but without direct interaction of involvement of the user. The collection of trajectory and speed data by urban traffic navigation tools such as Waze [5], and the capture of other environmental data using smartphone sensors (Chowdhury, Patwary et al. 2014) are examples of passive crowdsensing.

e) Mission-oriented - At the center of Figure 1 an element was used to incorporate mission-oriented crowdsourcing or sensing initiatives. Such applications start with a list of tasks that need to be executed, and then seek volunteers to complete them (Chen, Fu et al. 2014). Mission-oriented initiatives can take place within the scope of all four dimensions defined in the taxonomy, and can be aided by recommendation systems techniques that identify potential volunteers based on their online profiles.

Notice that some applications can be the source of more than one type of crowdsourced/crowdsensed data. Waze, for instance, involves both passive crowdsensing, in regard of vehicle speed and location, and active crowdsourcing, when its users inform others through the app about accidents, traffic jams or roadblocks.

2.2. VOLUNTEERED DATA

Within the scope of crowdsourcing, taken in a broader sense, as proposed in the taxonomy presented earlier, and incorporating the various definitions and names proposed in the past, the roles of contributors can vary widely. They can range from localized and precise map updates or the generation of novel data themes, to the denouncement of misconduct or environmental abuse. They can incorporate opinions, and foster discussions. Crowds can exert quality control, by expressing agreement or displeasure with individual contributions by peers.

Furthermore, contributions can be anonymous or publicly identified. Online tools can request simple contributions on single themes, or open up a variety of thematic possibilities for data collection at the same time. Contributions can be validated by peers, can be moderated by systems administrators, or be taken individually, with no restrictions. Comments and discussions can be allowed, or discouraged for the sake of simplicity and speed. Complementary multimedia information, such as images and video, can be collected simultaneously to geolocated contributions. Contributors can receive feedback, or praise, that can then be propagated to her friends and acquaintances using social media.

Anyway, the range of possibilities for volunteers is very broad, defying the developers of general-use crowdsourcing tools. Many frameworks or software platforms for the creation of geographic crowdsourcing or VGI tools have been proposed. Ushahidi [6] was proposed as a PPGIS initiative, but evolved towards becoming a framework for the development of data collection applications directed to crisis management scenarios. The framework allows people with some software development experience to customize its appearance and resources (Okolloh 2009).

Sheppard (2012) also presents a framework for mobile and Web applications, directed towards more experienced developers, and aiming at reducing the difficulty in the development of new applications with little loss of flexibility through the use of HTML 5.0 cross-platform features. The ClickOnMap project [7] (de Souza, Lisboa Filho et al. 2013) focuses on the creation of metadata and on the assessment of contributions by peers.

Based on a previous project by the same group, ClickOnMap also implements the description of contributions using wikis, which can be revised by other users, and a set of analysis tools, such as geographic visualizations and graphics, both with support to filtering contributions by type. ThemeRise [8], a framework for generating VGI applications, was implemented with current and flexible technologies, such as HapiJS, Sequelize, AngularJS, Leaflet and the C3 visualization library, seeking to ensure extensibility and evolution (Davis Jr, Vellozo et al. 2013, Pinheiro and Davis Jr. 2018).

The framework manages the structure and characteristics of data collection target themes individually. Features include requiring or dismissing user identification, allowing comments or wikis about each contribution, a scheme for the gamification of user contributions, with a reward scheme for volunteer actions, and many other configuration options that allow theme managers to implement multi-thematic VGI applications. As it currently stands, ThemeRise allows creating and publishing VGI resources in minutes, while preserving a sophisticated control over themes, contributions and user experience.

More challenges lie in the path of future development of crowdsourcing and crowdsensing tools. Gathering useful information from location-based social networks becomes more challenging as concerns for privacy rise, in the wake of unlawful user profile data sharing. Spatial analysis tools and techniques need to be customized and redefined, considering the known limitations of volunteered data as to coverage and reliability. Natural language processing and sentiment analysis, techniques usually employed in the study of social networks, can be adapted for geographic data collection. Data mining and machine learning algorithms, likewise, need to be adapted and evaluated in the context of geospatial crowdsourced data.

ISSN 1828-5961

One aspect of such challenges, however, comes to the foreground, especially when taking into consideration urban problems and urban computing techniques: the need to integrate official and crowdsourced data. This aspect is explored in the next section.

3. INTEGRATION OF VOLUNTEERED AND OFFICIAL DATA

Among their institutional responsibilities, several public organizations produce geo- graphic data of general interest. However, due to operational or technological difficulties, part of these data do not become accessible to the public, or is published in formats that preclude their dynamic integration to other data sources, thereby making it harder to accomplish more elaborate analyses.

In Brazil, even though the Law on Information Access [9] has been passed in 2011, most governmental data producers still lack a clear open data policy or practice. Some download sites are available, but most do not implement technological resources such as APIs or service-based spatial data infrastructures (SDI) to foster easy access to data that are relevant to the society at large. Maps are mostly available in PDF format, which is useless for further processing or uses other that direct viewing.

The Brazilian National Spatial Data Infrastructure contains mostly cartographic data, and is managed by the country's main mapping agency, the National Geography and Statistics Institute (IBGE), with a large participation by the Army's cartographic branch. The Law establishes mechanisms for requesting data and information, but the process is usually slow and ridden with bureaucratic hindrances.

Regardless of the degree of readiness by Brazilian institutions, the Law on Information Access is a fundamentally important resource for data access. It is based on a broader proposal, namely the Open Government Data initiative [10] , which dates back to 2005. This initiative has stipulated a number of principles for the sharing of public data, defined as all data that is not subject to valid privacy, security of privilege limitations, briefly summarized in Table 1

Other open government data principles were proposed, although not officially adopted. Some of these represent valid concerns on the availability of government data, such as principles on data being permanent, i.e., available at a stable Internet location indefinitely, or on data being trusted, i.e., digitally signed or including guarantees on publication data, authenticity and integrity. These principles are in line

Table 1 - Principles of Open Government Data

Principle	Description
Complete	All public data must be made available. Digital access is encouraged to the maximum. Bulk data should be made available, if existing APIs or other mechanisms allow for obtaining only small slices of the entire dataset at a time.
Primary	Data is available as collected at the source, with the highest possible level of granularity, not in aggregate or modified form. If the publisher chooses to transform data by aggregation or other technique to allow summarized reading by end users, it still must publish the data in its full resolution.
Timely	Data is made available as quickly as necessary, to preserve its value.
Accessible	Data is available to the widest possible range of users, for the widest range of purposes. Data should be available online, so that this access can be enabled. Publishers should consider how their choices in data preparation and publication can affect the access by the disabled, and how it may affect users of a variety of software and hardware platforms. Data must be published with current industry standard protocols and formats, and using alternative protocols and formats when the industry standard imposes burdens on the widespread use of the data. Data is not accessible if it can be retrieved only through navigating Web forms, or if there are any other technological restrictions to retrieve it.
Machine processable	Data is reasonably structured to allow automated processing. Free-form text is not a substitute for tabular and normalized records. Images of text do not substitute the text itself. Sufficient documentation of the data format and the meanings of normalized data items must be supplied.
Non-discriminatory	Data is available to anyone, with no requirement of identification or registration. Anonymous access must be allowed.
Non-proprietary	Data is available in a format over which no entity has exclusive control. Proprietary formats impose unnecessary restrictions over who can use the data, how it can be used and shared, and whether data will remain useful in the future. Their use is not acceptable. If non-proprietary data formats may not reach a wide audience, data should be made available in multiple formats.
License-free	Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

However, most governmental data producers in Brazil still do not entirely adhere to the Open Government Data principles, even with the Law on Information Access. Geographic data sharing resources often rely on proprietary file formats, such as Shapefile, maps are published in PDF format, statistical data are published using spreadsheets, and open file formats, such as GML and GeoJSON are practically never used. Even open formats, such as the Global Transit Feed Specification (GTFS) [11], are seldom used, and some published file sets, as the one for the city of Belo Horizonte [12], are incomplete at the time of this writing. The notion of publishing geospatial data under Web services is limited to the SDIs that are in operation, and public APIs for basic information elements such as urban addresses and transit routes are non existent.

In that regard, our group has been working on assessing the integration of official and volunteered data sources in many different subjects. Wenceslau, Davis Jr. et al. (2017) discuss the integration of the official catalog for businesses in the city of Belo Horizonte, supplied by the local government, and extracted form the city's tax databases, to the location and description of businesses from online sources such as Google Places, Foursquare and Yelp. Results show that in some categories up to 75% of businesses that are found in crowdsourced records match the official dataset, while in other categories the number of matches is near zero. The higher match in categories related to commerce and services indicates that the official data might be successfully replaced by crowdsourcing in many situations, including the calculation of the city's Urban Quality of Life Index (Smarzaro, de Lima et al. 2017, Smarzaro, Lima et al. 2017.

Santos, Davis Jr. et al. (2017) present a study that aims the integration of data on traffic accidents in the city of Belo Horizonte (an earlier version of the study can be found in (Santos, Davis Jr. et al. 2016). Sources are the official accident reports, supplied by the police and the transit authority, and user notifications of accidents in Waze (Figure 2). Results show that only about 9% of the accidents reported by Waze users could be matched to official reports. Within the same time period, of one year, deduplicated Waze reports outnumber official reports by 2.5 to 1, and the integrated dataset contains 3.6 times more accidents than the official reports only. Spatial analysis of each dataset indicates that official reports concentrate on higher-income areas, while Waze reports concentrate on the main thoroughfares. This allows the authors of the study to hypothesize that the official reports are mostly limited to situations in which involved drivers need the official document for insurance purposes, while Waze users mainly report accidents that cause a larger impact on traffic. Accident clusters, which might indicate locations for which authorities should improve local conditions as to signing and patrolling, are widely different between the two datasets.

These examples merely indicate the challenges and the potentially positive consequences of being able to find, obtain and integrate data from multiple and heterogeneous sources. The general idea is to combine the reliability and broad coverage that characterize official sources, which are based on definite and systematic administrative and organizational processes, to the dynamic, opinative and often imprecise characteristics of citizen-supplied information, which covers aspects that are left untouched by the official sources. Our preliminary studies demonstrate that the combination of such sources can be much broader than the official data, i.e., crowdsourced data can expand, enhance, complement or replace official data, depending on the theme and on the data collection methodology used (Borges, Jankowski et al. 2016).

Problems exist, however, with the timeliness of the approaches: official data can be harder to obtain, in a process that can take a significant time, especially until APIs and SDIs are not widely available (Borges, Jankowski et al. 2015). When this is the case, crowdsourced data can be used as a proxy, anticipating what the official data might show in the near future, when it is released (see (Gomide, Veloso et al. 2011), for a study on the use of georeferenced tweets – a type of passive crowdsourcing – about dengue fever as an early indicator for the outbreak of the disease, later confirmed by official records).

Consider an example, involving urban mobility.

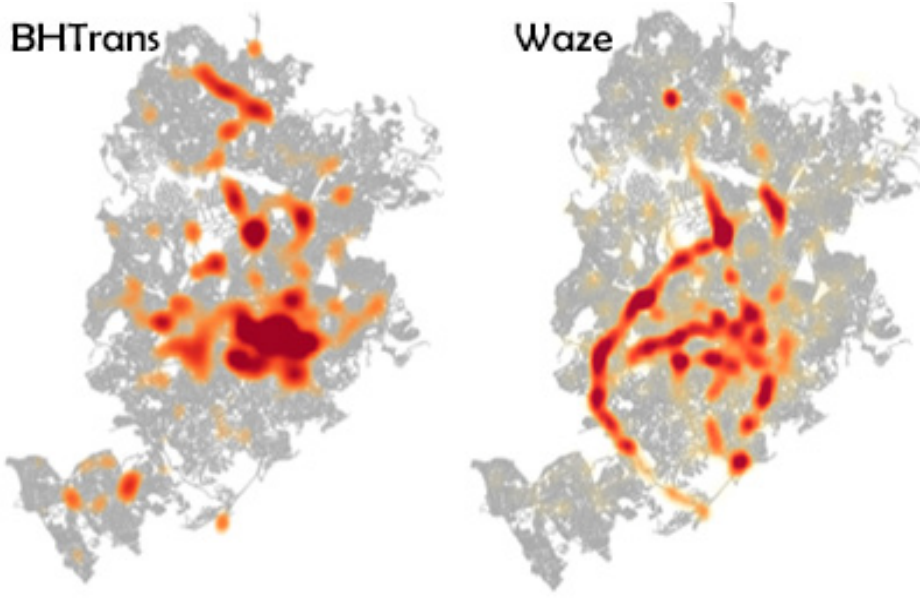

Figure 2 - Traffic accidents in Belo Horizonte, Brazil, from official (BHTrans) and crowdsourced (Waze) data. Source: Santos, Davis Jr. et al. (2017.)

ISSN 1828-5961

Table 2 lists a number of datasets that might be more adequately obtained from official sources, as compared to other data of interest for that problem that might only be obtained using crowdsourcing tools and techniques. A transit authority that needs to improve on their management of mobility assets would need to combine both columns, both by keeping their investment on systematic mapping and on geographic information associated to their work processes, and by creating crowdsourcing initiatives, to be disseminated throughout the users of the system and combined with official records. Furthermore, the public authority would need to provide feedback to citizens that contribute with valid data, both by demonstrating that crowdsourced data are actually used, and by promptly reacting to denouncements or indications of service quality issues. (Table 2)

4. DISCUSSION: CHALLENGES AND DIFFICULTIES

It may be tempting to place the entire burden of widely and systematically opening public data on the shoulders of public organizations. However, for governments, opening their data does not come without costs. Johnson, Sieber et al. (2017) identify both direct and indirect costs of open data provision, and propose a discussion on such costs. They argue that there is little evidence to support claims that opening data leads to citizen participation, even though transparency is beneficial in many established ways. The authors raise an important and valid point: publishing and keeping up-to-date large volumes of complex machine-readable data assumes the capacity of users handling such data, especially in a time where data literacy is widely recognized as a societal challenge. Opening data can reinforce the digital divide, since private organizations and more highly educated citizens are more likely to benefit from it than the average citizen. Standardization can have a positive effect, but there is a huge need to deal with heterogeneity of sources, in part caused by the variety of data collection methods, and create better integration techniques.

Furthermore, Johnson et al. point out that open data that fails to deal with privacy concerns might even lead to harmful consequences to citizens. This immediately raises a discussion on the balance of the need for governmental transparency versus the need to protect people from wrongful uses of data, and on the situations in which decision-makers might bend this scale towards their political interest. Janssen (2012) argues that open data must be viewed as an input for open government efforts in which actions are taken in the pursuit of transparency, openness, accountability and accessibility. Johnson, Sieber et al. (2017) conclude that "durable challenges to accessing data can have little to do with its openness, and are more focused on concerns of format, technical knowledge required, standardization across jurisdictions, as well as data completeness and quality concerns." In our point of view, such challenges must be actively pursued, and the design and implementation of easy to use apps that employ governmental open data may empower citizens with a tool with which to overcome individual limitations and enable their participation in the decision-making processes (Davis Jr., Fonseca et al. 2009).

Sieber and Haklay (2015) raise another important point on volunteered geographic information. They argue that crowdsourced data cannot be considered a data source as any other, since it is not free from societal implications. We agree with this point of view, and consider that both designers of crowdsourcing applications and users of the data generated by volunteers need to bear in mind the various limitations of such a data source. The literature on crowdsourcing and VGI strongly reinforces the notion that volunteered data can be biased, incomplete, imprecise and covering only a part of the space of interest. However, we must realize that, in many situations, and for many themes, the crowd may be the best – or the only – source of important data. As a result, any future research agenda on volunteered or crowdsourced geographic data must include the development of means to adequately deal with citizen-produced data, by designing around its limitations and by taking such limitations into account in its use.

We pointed out that one of the key potential advantages of crowdsourced data is its timeliness, in dynamic applications. However, we must recognize that dealing with real-time data streams can be even harder, especially when those data are to be used in analyses and accumulated to instruct the creation of public policies. There is an immediate contrast with big data obtained by extensive sensor networks, and the warning issued by Sieber and Haklay (2015) comes immediately to mind to remind us that citizens may serve as sensors (Goodchild 2007a), but their nature is fundamentally

Table 2 - Official versus crowdsourced datasets

Official Data	Crowdsourced Data
Street network	Dynamic bus speeds
Street addressing	Timetable verification
Vehicle circulation network	Vehicle quality and cleanliness
Bus stops, metro stations	Service quality assessment
Bus and metro itineraries	Ride comfort assessment
Bus and metro timetables	Unscheduled road blocks
Scheduled road blocks	Traffic jams
Scheduled public works and interventions	Safety incidents
Accidents	Accidents
Tariffs and fares	Points of interest for traveling
	Bicycle routes
	Street walkability and scenic routes
	Crime and security denouncements

different from that of widespread hardware devices.

While the outlook for achieving popular participation in crowdsourcing and crowdsensing initiatives is positive, the ready availability of public data from Brazilian governmental organizations is still limited to some success cases, and some ongoing projects. For the sake of fulfilling the requirements in the Law on Information Access, simple Web GIS sites or resources are not sufficient, since there are no metadata, and usually data can only be visualized and interacted with, but not downloaded in a non-proprietary and machine-readable format. Spatial Data Infrastructures are more adequate in that respect, but the excessively cartographic emphasis in some of them, such as the Brazilian National Data Infrastructure (INDE) [13] may inhibit some types of use. Other relevant and referential spatial data provision initiatives include the environmental SDI for the state of São Paulo [14], the SDI for the state of Bahia [15], and other resources that, while not organized as SDIs, provide access to broadly useful spatial data, as in the cases of the open data portals for the cities of São Paulo [16] and Belo Horizonte [17]. Internationally, the gold standard for SDI is the INSPIRE [18] project, from the European Commission, and open governmental data resources such as Data.gov from the USA and Data.gov.uk from the United Kingdom, among many other initiatives.

5. FINAL REMARKS

There is a growing access to multiple and heterogeneous sources of spatial data, produced by various agents, to fulfill various needs. We must research, develop and entrepreneur exploring ways to integrate data from such varied sources and bridge the virtual and real worlds, considering that humans are a valuable, albeit problem-ridden, source of valuable information. Dealing with such limitations is necessary, if we are to use the potential of today's bountiful sources of new digital data, along with our increasing capacity for processing and learning from it, in the search for solutions to actual problems of our society and our planet.

NOTES

[1]. http://www.openstreetmap.org

[2]. https://developers.google.com/places

[3]. https://foursquare.com

[4]. https://www.yelp.com

[5]. http://www.waze.com

[6]. http://ushahidi.com

[7]. http://www.dpi.ufv.br/projetos/clickonmap/

[8]. http://aqui.io/themerise

[9]. Lei 12.527, de 18 de novembro de 2011. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm (in Portuguese)

[10]. https://opengovdata.org

[11]. http://gtfs.org. The GTFS specification was initially proposed by Google and the transit authority for Portland, Oregon, as an open data standard. Its original name was Google Transit Feed Specification. See http://beyondtransparency.org/chapters/part-2/pioneering-open-data-standards-the-gtfs-story for a complete background on GTFS.

[12]. https://prefeitura.pbh.gov.br/bhtrans/informacoes/dados/dados-abertos

[13]. http://www.inde.gov.br

[14]. http://datageo.ambiente.sp.gov.br

[15]. http://geoportal.ide.ba.gov.br

[16]. http://dados.prefeitura.sp.gov.br

[17]. http://dados.pbh.gov.br

[18]. http://inspire-geoportal.ec.europa.eu

REFERENCES

Ballatore, A., MacArdle, G., Kelly, C., & Bertolotto, M. (2010). RecoMap: an interactive and adaptive map-based recommender. Paper presented at the *ACM Symposium on Applied Computing*, New York City (NY).

Borges, J., Jankowski, P., & Davis, C. A. (2015). Crowdsourcing for Geodesign: Opportunities and Challenges for Stakeholder Input in *Urban Planning Cartography-Maps Connecting the World* (pp. 361-373): Springer.

Borges, J. L. C., Jankowski, P., & Davis Jr., C. A. (2016). A study on the use of crowdsourced information for urban decision-making. *Revista Brasileira de Cartografia*, 68(4).

Chatfield, A., & Brajawidagda, U. (2014). Crowdsourcing hazardous weather reports from citizens via twittersphere under the short warning lead times of EF5 intensity tornado conditions. . Paper presented at the *47th Hawaii International Conference on System Science*, Waikoloa, Hawaii.

Chen, Z., Fu, R., Zhao, Z., Liu, Z., Xia, L., Chen, L., Zhang, C. J. (2014). gMission: a general spatial crowdsourcing platform. Paper presented at the *40th International Conference on Very Large Databases*, Hangzhou, China.

Chowdhury, M., Patwary, K. H., Imteaj, A., & Chowdhury, S. (2014). Designing a wireless sensor network using smartphone as data source. Paper presented at the *The 9th International Forum on Strategic Technologies (IFOST)*, Cox's Bazar, Bangladesh.

Craglia, M., Goodchild, M. F., Annoni, A., Câmara, G., Gould, M., Kuhn, W., Parsons, E. (2008). Next-Generation Digital Earth. *International Journal of Spatial Data Infrastructures Research*, 3, 146-167.

Davis Jr, C. A., Pappa, G. L., Oliveira, D. R. R., & Arcanjo, F. L. (2011). Inferring the location of Twitter messages based on user relationships. *Transactions in GIS*, 15(6), 735-751.

Davis Jr, C. A., Vellozo, H. S., & Pinheiro, M. B. (2013). A framework for Web and mobile volunteered geographic information applications. Paper presented at the *XIV Brazilian Symposium on Geoinformatics (GeoInfo 2013)*, Campos do Jordão (SP).

Davis Jr., C. A., Fonseca, F. T., & Câmara, G. (2009). Beyond SDI: Integrating Science and Communities to Create Environmental Policies for the Sustainability of the Amazon. *International Journal of Spatial Data Infrastructures Research*, 4, 156-174. doi:10.2902/1725-0463.2009.04.art9

Drew, C. H. (2003). Transparency - Considerations for PPGIS research and development. *URISA Journal*, 15(1), 73-78.

Elwood, S. (2006). The Devil is still in the Data: Persistent Spatial Data Handling Challenges in Grassroots GIS. In A. Riedl, W. Kainz, & G. A. Elmes (Eds.), *Progress in Spatial Data Handling* (pp. 1-16). Berlin Heidelberg: Springer.

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

Gomide, J., Veloso, A., Meira Jr., W., Benevenuto, F., Almeida, V. A. F., Ferraz, F., & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. Paper presented at the *Proceedings of the Third International Conference on Web Science*.

Goodchild, M. (2009). NeoGeography and the nature of geographic expertise. *Journal of location based services*, 3(2), 82-96.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221. doi:10.1007/s10708-007-9111-y

Goodchild, M. F. (2007). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2, 24-32.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1, 110-120.

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. (2013). *Citizen science and volunteered geographic information: Overview and typology of participation Crowdsourcing geographic knowledge* (pp. 105-122): Springer.

Haklay, M., Singleton, A., & Parker, C. (2008). Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass*, 2(6), 2011-2039.

Hirata, E., Giannotti, M. A., Larocca, A. P. C., & Quintanilha, J. A. (2015). Flooding and inundation collaborative mapping - use of the crowdmap / ushahidi platform in the city of São Paulo – Brazil. *Journal of Flood Risk Management*. doi:10.1111/jfr3.12181

Janssen, K. (2012). Open government data and the right to information: Opportunities and obstacles. *The Journal of Community Informatics*, 8(2).

Johnson, P. A., Sieber, R., Scassa, T.,

ISSN 1828-5961

Stephens, M., & Robinson, P. (2017). The Cost (s) of Geospatial Open Data. Transactions in GIS, 21(3), 434-445.

Johnson, P. A., & Sieber, R. E. (2013). *Situating the adoption of VGI by government Crowdsourcing geographic knowledge* (pp. 65-81): Springer.

Krumm, J. (2007). Exploiting users' map annotations. Paper presented at the *Workshop on Volunteered Geographic Information*, Santa Barbara, California, USA.

Kuhn, W. (2007). Volunteered Geographic Information and GIScience. Paper presented at the *Workshop on Volunteered Geographic Information*, Santa Barbara, California, USA.

Lin, W. (2013). *When Web 2.0 meets public participation GIS (PPGIS): VGI and spaces of participatory mapping in China Crowdsourcing geographic knowledge* (pp. 83-103): Springer.

Maguire, D. J. (2007). GeoWeb 2.0 and Volunteered GI. Paper presented at the *Workshop on Volunteered Geographic Information*, Santa Barbara, California, USA.

Mateveli, G. V., Machado, N. G., Moro, M. M., & Davis Jr, C. A. (2015). Taxonomia e Desafios de Recomendação para Coleta de Dados Geográficos por Cidadãos. Paper presented at the *XXX Simpósio Brasileiro de Bancos de Dados*, Petrópolis (RJ).

Mytilinis, I., Giannakopoulos, I., Konstantinou, I., Doka, K., Tsitsigkos, D., Terrovitis, M., Koziris, N. (2015). MoDisSENSE: A Distributed Spatio-Temporal and Textual Processing Platform for Social Networking Services. *Paper presented at the ACM SIGMOD,* Melbourne, Australia.

Okolloh, O. (2009). Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action*, 59(1), 65-70.

Pinheiro, M. B., & Davis Jr., C. A. (2018). ThemeRise: a theme-oriented framework for Volunteered Geographic Information applications. Open Geospatial Data, *Software and Standards*, 3(9), 1-12. doi:https://doi.org/10.1186/s40965-018-0049-4

Santos, S. R., Davis Jr., C. A., & Smarzaro, R. (2016). Integration of data sources on traffic accidents. Paper presented at the *XVII Brazilian Symposium on Geoinformatics (GeoInfo 2016)*, Campos do Jordão (SP), Brazil.

Santos, S. R., Davis Jr., C. A., & Smarzaro, R. (2017). Analyzing traffic accidents based on the integration of official an crowdsourced data. *Journal of Information and Data Management*, 8(4 (to appear)).

Sheppard, S. A. (2012). wq: A modular framework for collecting, storing, and utilizing experiential VGI. Paper presented at the *Proceedings of the 1st acm sigspatial international workshop on crowdsourced and volunteered geographic information.*

Sieber, R. E., & Haklay, M. (2015). The epistemology (s) of volunteered geographic information: a critique. *Geo: Geography and Environment*, 2(2), 122-136.

Silva, J. C. T., & Davis Jr, C. A. (2008). Um framework para coleta e filtragem de dados geográficos fornecidos voluntariamente. Paper presented at the *X Brazilian Symposium on GeoInformatics (GeoInfo 2008)*, Rio de Janeiro (RJ).

Smarzaro, R., de Lima, T. F. M., & Davis Jr, C. A. (2017). Quality of Urban Life Index From Location-Based Social Networks Data. *Volunteered Geographic Information and the Future of Geospatial Data*, 185.

Smarzaro, R., Lima, T. F. M., & Davis Jr., C. A. (2017). Could data from lcoation-based social networks be used to support urban planning? Paper presented at the *7th International Workshop on Location and the Web*, Perth, Australia.

Souza, W. D., Lisboa Filho, J., Vidal Filho, J. N., & Câmara, J. H. (2013). DM4VGI: A template with dynamic metadata for documenting and validating the quality of Volunteered Geographic Information. *Paper presented at the GeoInfo.*

Swan, M. (2015). *Blockchain: Blueprint for a new economy:* "O'Reilly Media, Inc.".

Turner, A. (2006). *Introduction to Neogeography: O'Reilly.*

Vellozo, H. S., Pinheiro, M. B., & Davis Jr, C. A. (2013). Strepitus: um aplicativo para coleta colaborativa de dados sobre ruído urbano. Paper presented at the *IV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA 2013)*, Maceió (AL).

Wenceslau, R., Davis Jr., C. A., & Smarzaro, R. (2017). Challenges for the integration of data on economic activities from official and alternative sources. Paper presented at the *XVIII Brazilian Symposium on Geoinformatics (GeoInfo 2017)*, Salvador (BA), Brazil.

ISSN 1828-5961

Desafios para Substituir ou Aprimorar Fontes Oficiais de Dados Geoespaciais através de Crowdsourcing

1. INTRODUÇÃO

Tradicionamente, dados que alimentam sistemas de informação geográficos (SIG) são produzidos por organizações públicas ou privadas cujas responsabilidades incluem cadastros tributários e multifinalitários, gerenciamento de infraestrutura, ou prestação de serviços públicos distribuídos, como no caso de companhias concessionárias de serviços públicos. Tais fontes tradicionais de dados, no entanto, são cada vez menos capazes de criar e manter conjuntos de dados detalhados sobre vários aspectos da vida urbana. Os cidadãos atuais, munidos de smartphones conectados à Internet e equipados com uma variedade de sensores, serviços e aplicações, demandam mais informação geográfica em suas atividades cotidianas.

Em consequência disso, uma ampla gama de novos provedores provê informação que não está conectada nem é derivada das fontes tradicionais. Muitos dos objetos que compõem o OpenStreetMap [1] (OSM), e o conteúdo de serviços baseados em localização tais como Google Places [2], Foursquare [3], Yelp [4] e outros vão além da cartografia básica e das fontes oficiais, complementando-as com dados sobre mobilidade, atividades comerciais, pontos de interesse e muitas outras classes de informação.

Grande parte do conteúdo encontrado nessas fontes inovadoras, no entanto, é provido por voluntários, através de plataformas de colaboração voluntária (ou crowdsourcing) tais como o OSM, ou pelos próprios proprietários de empresas, incentivados pelo seu interesse em figurar em aplicativos de uso amplo. Não existe garantia, para essas fontes alternativas, de cobertura ampla ou completa, ou de qualidade uniforme, ou de geração de conteúdo sem viés (Michael F. Goodchild, 2007). Se existirem mais voluntaries que conhecem uma dada região e que estão aptos a contribuir, a densidade e a qualidade dos dados naquela região tendem a ser mais altos (Flanagin & Metzger, 2008; M. F. Goodchild & Li, 2012; Mordechai Haklay, 2010). Empresas que não buscam atrair pedestres ou internautas, tais como turistas, não são atraídos por essas plataformas. Exemplos incluem escritórios de direito e empresas prestadoras de serviços para outras empresas (business-to-business) (Wenceslau, Davis Jr., & Smarzaro, 2017).

Em razão disso, observamos que, atualmente, dados geográficos relevantes e úteis não são sempre cartográficos por natureza, nem sempre são obtidos de fontes oficiais, e, em consequência, nem sempre são completos ou homogêneos. Mesmo assim, espera-se que esses dados possam complementar, melhorar ou mesmo substituir parcialmente dados de fontes oficiais. Nas próximas seções serão apresentadas ideias e limitações na direção indicada por este argumento.

ISSN 1828-5961

2. COLETA COLABORATIVA DE DADOS

A quantidade e variedade de dados espaciais disponíveis online para o cidadão comum aumentaram rapidamente. Desde o lançamento do Google Earth, em 2004, e do Google Maps, em 2006, existe um interesse crescente a respeito de ferramentas e recursos que ofereçam serviços baseados em localização, tais como localização de endereços e roteamento de veículos. Tamanho interesse expandiu a gama de possibilidades em outras direções. Acompanhando as ideias que sustentaram o movimento denominado Web 2.0, indivíduos comuns, equipados com hardware e software de baixo custo e amplamente disponível, se tornaram atores importantes na criação, disseminação e manutenção de informação geográfica (M. F. Goodchild, 2007; Kuhn, 2007). Muitas iniciativas recentes mostram que as pessoas se dispõem a contribuir, voluntariamente e sem expectativa de recompensa econômica, com a criação de novos conjuntos de dados e com a atualização ou correção de erros em dados existentes (Krumm, 2007; Maguire, 2007; Silva & Davis Jr, 2008).

Os custos da criação de informação geográfica nova e detalhada, típica de aplicações urbanas, não são mais integralmente suportados por agências governamentais de mapeamento. Em lugar dos SIG oficiais e totalmente centralizados que eram relativamente comuns dez ou vinte anos atrás, existe atualmente a necessidade de lidar com uma "colcha de retalhos" formada por fontes de dados, na qual produtores oficiais de dados participam apenas com temas que fazem parte de suas responsabilidades institucionais e que exigem atualização sistemática (Craglia et al., 2008). Cidadãos comuns podem ser incorporados a esse processo, provendo (1) indicações sobre em que lugares os dados oficiais discordam da realidade, (2) informação atualizada obtida em campo, usando equipamentos como smartphones, (3) informação dinâmica, up-to-the-minute, sobre eventos urbanos distribuídos espacialmente, e (4) conhecimento pessoal sobre detalhes de sua realidade local, como moradores ou membros de comunidades locais. Esse envolvimento de grupos potencialmente grandes de cidadãos na coleta e produção de dados é denominada de várias formas, como descrito na próxima seção, mas daremos preferência ao termo crowdsourcing daqui em diante.

2.1. CONCEITOS E VARIAÇÕES TERMINOLÓGICAS

A literatura que discute crowdsourcing de dados e informação geográfica traz diversas designações par o processo. Neogeography (Turner, 2006) é um conceito que reflete o conjunto de técnicas e ferramentas que não se encaixam na geografia tradicional, e são empregadas por cidadãos comuns que estão familiarizados com sua vizinhança (M. Goodchild, 2009; M. Haklay, Singleton, & Parker, 2008). Public Participation Geographic Information Systems (PPGIS, ou SIG de participação pública) busca incentivar comunidades a participar politicamente em ações governamentais de seu interesse local (Drew, 2003; Elwood, 2006; Johnson & Sieber, 2013; Lin, 2013). Algumas iniciativas usam a expressão Citizen Science (ciência cidadã) para caracterizar dados coletados por pessoas sem formação específica que se dispõem a colaborar com projetos científicos (Mordechai Haklay, 2010; Muki Haklay, 2013). A expressão Volunteered Geographic Information (VGI, informação geográfica provida por voluntários) (Michael F. Goodchild, 2007) é também usada no sentido de permitir que voluntários contribuam com iniciativas de coleta de dados geográficos e atualização de mapas, em uma perspectiva que emprega os cidadãos como sensores humanos (M. F. Goodchild, 2007).

Uma taxonomia útil para diferenciar e classificar técnicas de crowdsourcing é proposta por Mateveli, Machado, Moro, & Davis Jr (2015). Essa taxonomia considera duas dimensões: o processo usado para a captura de dados, e o nível de participação do usuário, ou voluntário. A Figura 1 apresenta esquematicamente a taxonomia. Cada quadrante consiste em um contexto no qual as duas dimensões são combinadas.

Crowdsourcing ativo ocorre quando o usuário provê informação conscientemente, tipicamente como voluntário, e sabe como a informação deve ser usada. Exemplos incluem a iniciativa OpenStreetMap [5] e o projeto Pontos de Alagamento (Hirata, Giannotti, Larocca, & Quintanilha, 2015).

Figura 1 – Taxonomia de crowdsourcing e crowdsensing. Adaptado de (Mateveli et al., 2015)

Crowdsourcing ativo ocorre quando o usuário provê informação conscientemente, tipicamente como voluntário, e sabe como a informação deve ser usada. Exemplos incluem a iniciativa OpenStreetMap [5] e o projeto Pontos de Alagamento (Hirata, Giannotti, Larocca, & Quintanilha, 2015).

Crowdsourcing passivo ocorre quando a informação é obtida a partir de material postado pelo usuário para outras finalidades, ou seja, informação útil é extraída de conteúdo colocado online por alguém sem que essa pessoa esteja explicitamente envolvida com o tema de interesse. Um exemplo comum é a extração de informação em redes sociais baseadas em localização, tais como o Twitter, baseada no texto das mensagens ou em hashtags específicas (Chatfield & Brajawidagda, 2014; Davis Jr, Pappa, Oliveira, & Arcanjo, 2011; Gomide et al., 2011).

Crowdsensing ativo corresponde à coleta ativa e compartilhamento de informação que é capturada por sensores embarcados em dispositivos móveis. Por exemplo, uma usuária registra presença (checks in) em um ponto de seu interesse, cuja posição é determinada pelo receptor GPS integrado ao smartphone (Ballatore, MacArdle, Kelly, & Bertolotto, 2010; Mytilinis et al., 2015). Outro exemplo envolve a coleta ativa do nível de ruído em uma aplicação voltada para a denúncia de poluição sonora (Vellozo, Pinheiro, & Davis Jr, 2013).

Crowdsensing passivo ocorre quando a informação é capturada por sensores embarcados em dispositivos móveis, mas sem interação ou envolvimento direto do usuário. A coleta de dados de trajetórias e velocidades por aplicativos de navegação no tráfego urbano, tais como Waze [6], e a captura de outros dados ambientais usando sensores de smartphones (Chowdhury, Patwary, Imteaj, & Chowdhury, 2014) são exemplos de crowdsensing passivo.

No centro da Figura 1 um elemento foi incluído para incorporar a noção de crowdsourcing ou crowdsensing orientados por missão. Tais aplicações começam com uma lista de tarefas que precisam ser executadas, e então buscam encontrar voluntários que as realizem (Chen et al., 2014). Iniciativas orientadas por missão podem ocorrer no escopo de todas as quatro dimensões definidas na taxonomia,

ISSN 1828-5961

e podem ser apoiadas por técnicas de sistemas de recomendação que identificam potenciais voluntários com base em seus perfis e postagens nas redes sociais.

Observe-se que algumas aplicações podem utilizar mais de um tipo de dados de crowdsourcing/ crowdsensing. Waze, por exemplo, usa tanto o crowdsensing passivo, pelo qual coleta dados de localização e velocidade de veículos, e crowdsourcing ativo, pelo qual recebe contribuições sobre acidentes de trânsito, engarrafamentos e bloqueios em vias.

2.2. DADOS PROVIDOS POR VOLUNTÁRIOS

No âmbito do crowdsourcing, tomado em um sentido mais amplo, como proposto na taxonomia apresentada anteriormente, e incorporando as várias definições e nomes propostos no passado, os papéis dos contribuintes podem variar bastante. Vão de atualizações de mapas localizadas e precisas até a geração de novos temas de dados, ou até a denúncia de mau comportamento ou abuso ambiental. Podem incorporar opiniões e fomentar discussões. Usuários em quantidade suficiente (crowd) podem exercer controle de qualidade, expressando concordância ou desagrado com as contribuições individuais de seus pares.

Além disso, as contribuições podem ser anônimas ou publicamente identificadas. As ferramentas online podem solicitar contribuições simples sobre temas isolados ou abrir uma variedade de possibilidades para a coleta simultânea de dados de múltiplos temas. As contribuições podem ser validadas por pares, moderadas pelos administradores do sistema ou tomadas individualmente, sem restrições. Comentários e discussões podem ser permitidos, ou desencorajados por uma questão de simplicidade e rapidez. Informações multimídia complementares, como imagens e vídeos, podem ser coletadas simultaneamente para contribuições geolocalizadas. Os colaboradores podem receber feedback, ou elogios, que podem ser propagados para seus amigos e conhecidos usando as mídias sociais.

De qualquer forma, a gama de possibilidades para os voluntários é muito ampla, desafiando os desenvolvedores de ferramentas de crowdsourcing para uso geral. Muitas estruturas ou plataformas de software para a criação de ferramentas de crowdsourcing geográfico ou VGI foram propostas. O Ushahidi [7] foi proposto como uma iniciativa de PPGIS, mas evoluiu para se tornar uma estrutura para o desenvolvimento de aplicativos de coleta de dados voltados a cenários de gerenciamento de crise. Sua estrutura permite que pessoas com alguma experiência em desenvolvimento de software personalizem a aparência e recursos da aplicação (Okolloh, 2009). Sheppard (2012) também apresenta uma estrutura para aplicativos móveis e baseados na Web, direcionada a desenvolvedores mais experientes, e visando reduzir a dificuldade no desenvolvimento de novos aplicativos com pouca perda de flexibilidade por meio do uso de recursos multiplataforma do HTML 5.0. O projeto ClickOnMap [8] (Souza, Lisboa Filho, Vidal Filho, & Câmara, 2013) enfoca a criação de metadados e a avaliação de contribuições por pares. Baseado em um projeto anterior do mesmo grupo, ClickOnMap também implementa a descrição de contribuições usando wikis, que podem ser revisados por outros usuários, e um conjunto de ferramentas de análise, como visualizações geográficas e gráficos, ambos com suporte para filtrar contribuições por tipo.

ThemeRise [9], um framework para geração de aplicações VGI, foi implementado com tecnologias atuais e flexíveis, como HapiJS, Sequelize, AngularJS, Leaflet e a biblioteca de visualização C3, buscando garantir extensibilidade e evolução (Davis Jr, Vellozo, & Pinheiro, 2013; Pinheiro & Davis Jr., 2018). O framework gerencia individualmente a estrutura e as características dos temas de interesse para a coleta de dados. Os recursos incluem exigir ou dispensar a identificação do usuário, permitir comentários ou wikis sobre cada contribuição, utilizar um esquema para a gamificação das contribuições do usuário, com um sistema de recompensas para ações voluntárias, e muitas outras opções de configuração, que permitem aos gerentes de temas implementar aplicativos VGI multitemáticos. Atualmente, o ThemeRise permite criar e publicar aplicações VGI multitemáticas em minutos, preservando um controle sofisticado sobre temas, contribuições e experiência do usuário.

Mais desafios estão colocados no caminho do desenvolvimento futuro de ferramentas de crowdsourcing e crowdsensing. A coleta de informações úteis a partir de redes sociais baseadas em localização torna-se mais desafiadora à medida que aumenta a preocupação com a privacidade, na esteira de denúncias de compartilhamento ilegal de dados de perfis de usuários. Ferramentas e técnicas de análise espacial precisam ser customizadas e redefinidas, considerando as limitações conhecidas de dados voluntários quanto à cobertura e confiabilidade. Processamento de linguagem natural e análise de sentimentos, técnicas usualmente empregadas no estudo de redes sociais, podem ser adaptadas para a coleta de dados geográficos. Os algoritmos de mineração de dados e de aprendizado de máquina, da mesma forma, precisam ser adaptados e avaliados no contexto de dados geoespaciais de crowdsourcing.

Um aspecto de tais desafios, no entanto, ascende ao primeiro plano, especialmente quando se leva em consideração problemas urbanos e técnicas de computação urbana: a necessidade de integrar dados oficiais e de crowdsourcing. Esse aspecto é explorado na próxima seção.

3. INTEGRAÇÃO DE DADOS PROVIDOS POR VOLUNTÁRIOS E OFICIAIS

Em meio a suas responsabilidades institucionais, várias organizações públicas produzem dados geográficos de interesse geral. No entanto, devido a dificuldades operacionais ou institucionais, parte desses dados não se tornam acessíveis ao público, ou são publicados em formatos que impedem sua integração dinâmica a outras fontes de dados, portanto dificultando a realização de análises mais elaboradas. No Brasil, embora a Lei de Acesso à Informação [10] tenha sido aprovada em 2011, a maioria dos produtores de dados governamentais ainda carece de uma política ou prática de dados abertos claros. Alguns sites para download estão disponíveis, mas a maioria não implementa recursos tecnológicos como APIs ou infraestruturas de dados espaciais (IDE) baseadas em serviços para facilitar o acesso a dados que são relevantes para a sociedade em geral. Os mapas estão disponíveis geralmente em formato PDF, o que é inútil para processamento posterior ou para usos que não sejam limitados à visualização direta. A Infraestrutura

ISSN 1828-5961

Nacional de Dados Espaciais (INDE) do Brasil contém principalmente dados cartográficos e é gerenciada pela principal agência de mapeamento do país, o Instituto Nacional de Geografia e Estatística (IBGE), com uma grande participação do ramo cartográfico do Exército. A lei estabelece mecanismos para solicitar dados e informações, mas o processo é geralmente lento e sobrecarregado com obstáculos burocráticos.

Independentemente do grau de preparo das instituições brasileiras para prover dados amplamente, a Lei de Acesso à Informação é um recurso fundamental para a garantia de acesso. Baseia-se em uma proposta mais ampla, a iniciativa Open Government Data [11], que data de 2005. Esta iniciativa estipulou vários princípios para o compartilhamento de dados públicos, definidos como "todos os dados que não estão sujeitos limitações válidas de privacidade, segurança ou restrição de acesso". limitações de privilégio, sumarizadas na Tabela 1.

Outros princípios de dados governamentais abertos foram propostos, embora não tenham sido adotados oficialmente. Alguns desses princípios representam preocupações válidas sobre a disponibilidade de dados governamentais, como princípios sobre a permanência dos dados, ou seja, dados disponíveis em um local da Internet estável indefinidamente, ou sobre confiabilidade, ou seja, dados assinados digitalmente ou que incluam garantias do produtor sobre autenticidade e integridade. Esses princípios estão alinhados com as ideias por detrás do blockchain, uma tecnologia que visa criar armazenamento de dados estável e seguro através de cópias distribuídas e partilhadas, e que é frequentemente discutida em relação às criptomoedas (Swan, 2015).

No entanto, a maioria dos produtores governamentais de dados no Brasil ainda não aderiu inteiramente aos princípios do Open Government Data, mesmo com a Lei de Acesso à Informação. Os recursos de compartilhamento de dados geográficos geralmente dependem de formatos de arquivos proprietários, como Shapefile, mapas são publicados em formato PDF, dados estatísticos são publicados usando planilhas e formatos abertos, como GML e GeoJSON, praticamente nunca são usados. Mesmo formatos abertos, como o GTFS (Global Transit

Tabela 1 – Princípios de dados governamentais abertos

Princípio	Descrição
Completos	Todos os dados públicos têm que estar disponíveis. Acesso digital é encorajado ao máximo. Dados em volume devem também ser oferecidos, se APIs ou outros mecanismos permitirem selecionar apenas partes do conjunto de dados em um acesso.
Primários	Dados estão disponíveis conforme coletados na fonte, com o mais alto nível de granularidade possível, e não em forma agregada ou modificada. Se o produtor decidir transformar os dados por meio de agregação ou outra técnica para permitir a leitura sumarizada pelos usuários finais, ele ainda tem que publicar os dados em sua resolução máxima.
Oportunos	Dados são disponibilizados tão rapidamente quanto possível, para preservar seu valor.
Acessíveis	Dados são tornados disponíveis para a mais ampla gama possível de usuários, para mais ampla gama possível de propósitos. Os dados devem estar disponíveis online, de modo que o acesso possa ser garantido. Produtores deve levar em conta a maneira em que suas escolhas na preparação e publicação podem afetar o acesso pelas pessoas portadoras de deficiências, e como podem afetar usuários de uma variedade de plataformas de hardware e software. Os dados devem ser publicados usando padrões e formatos atuais da indústria, e usando protocolos e formatos alternativos quando os padrões da indústria impuserem restrições ao amplo uso dos dados. Dados não são acessíveis se só puderem ser recuperados após navegar por formulários na Web, ou se existirem quaisquer outras restrições tecnológicas para recuperá-los..
Processáveis por máquina	Dados estão razoavelmente estruturados para permitir processamento automático. Texto desestruturado não é substituto para registros tabulares e normalizados. Imagens de texto não substituem o próprio texto. Tem que ser fornecida documentação suficiente sobre o formato de dados e sobre o significado de itens de dados normalizados.
Não-discriminatórios	Dados estão disponíveis para qualquer pessoa, sem exigência de identificação ou registro. Acesso anônimo tem que ser permitido.
Não-proprietários	Dados estão disponíveis em um formato sobre o qual nenhuma entidade tenha controle exclusivo. Formatos proprietários impõem restrições desnecessárias quanto a quem pode usar os dados, como eles podem ser usados e compartilhados, e se permanecerão úteis no futuro. O uso desses formatos não é aceitável. Se formatos de dados não-proprietários não forem capazes de alcançar uma audiência mais ampla, os dados devem estar disponíveis em múltiplos formatos.s.
Livre de licença	Dados não são sujeitos a nenhum copyright, patente, marca registrada ou regulação de segredo comercial. Restrições razoáveis quanto a privacidade, segurança e acesso restrito podem ser permitidas.

ISSN 1828-5961

Feed Specification) [12], raramente são usados, e alguns conjuntos de arquivos publicados, como os da cidade de Belo Horizonte [13], estão incompletos no momento da redação deste texto. A noção de publicação de dados geoespaciais em serviços da Web é limitada às IDEs que estão em operação, e não existem APIs públicas para elementos de informação básicos, como endereços urbanos e rotas de trânsito.

Nesse sentido, nosso grupo de pesquisa tem trabalhado na avaliação da integração de fontes de dados oficiais e voluntárias em diversos assuntos. Wenceslau et al. (2017) discutem a integração do catálogo oficial de empresas (contribuintes) na cidade de Belo Horizonte, fornecido pelo governo local a partir dos dados tributários da cidade, com a localização e descrição de atividades econômicas catalogadas em fontes on-line como Google Places, Foursquare e Yelp. Os resultados mostram que, em algumas categorias, até 75% das empresas encontradas em registros de crowdsourcing correspondem ao que consta no conjunto de dados oficial, enquanto em outras categorias o número de correspondências é próximo de zero. A maior correspondência em categorias relacionadas a comércio e serviços indica que os dados oficiais podem ser substituídos com sucesso pelo crowdsourcing em muitas situações, incluindo o cálculo do Índice de Qualidade de Vida Urbana da cidade (Rodrigo Smarzaro, de Lima, & Davis Jr, 2017; R Smarzaro, Lima, & Davis Jr., 2017).

Santos, Davis Jr., & Smarzaro (2017) apresentam um estudo que visa a integração de dados sobre acidentes de trânsito na cidade de Belo Horizonte (uma versão anterior do estudo pode ser encontrada em (Santos, Davis Jr., & Smarzaro , 2016)). As fontes são os boletins de ocorrência oficiais sobre acidentes de trânsito, fornecidos pela polícia e pela autoridade de trânsito, e as notificações do usuário sobre acidentes no Waze. Os resultados mostram que apenas cerca de 9% dos acidentes relatados pelos usuários do Waze correspondem a acidentes reportados nos boletins de ocorrência oficiais. Dentro do mesmo período de um ano, os registros do Waze deduplicados superam os relatórios oficiais em 2,5 vezes, e o conjunto de dados integrado contém 3,6 vezes mais acidentes do que os registros oficiais. A análise espacial de cada conjunto de dados indica que os registros oficiais se concentram em áreas de renda mais alta, enquanto os acidentes reportados no Waze se concentram nas principais vias de trânsito. Isso permitiu que os autores do estudo especulassem que os registros oficiais são limitados principalmente a situações nas quais os motoristas envolvidos precisam de um documento oficial para fins de seguro, enquanto os usuários do Waze relatam principalmente acidentes que causam um impacto maior no tráfego. Aglomerações de acidentes, que podem indicar locais nos quais as autoridades precisariam melhorar as condições locais quanto à sinalização e fiscalização, são amplamente diferentes entre os dois conjuntos de dados.

Figura 2 – Acidentes de trânsito em Belo Horizonte, Brasil, segundo dados oficiais (BHTrans) e voluntários (Waze). Fonte: (Santos et al., 2017)

Estes exemplos indicam apenas os desafios e as consequências potencialmente positivas de ser capaz de encontrar, obter e integrar dados de fontes múltiplas e heterogêneas. A ideia geral é combinar a confiabilidade e a ampla cobertura que caracterizam as fontes oficiais, baseadas em processos administrativos e organizacionais definidos e sistemáticos, com as características dinâmicas, opinativas e muitas vezes imprecisas das informações fornecidas pelos cidadãos, que abrangem aspectos que não são cobertos pelas fontes oficiais. Nossos estudos preliminares demonstram que a combinação de tais fontes pode ser muito mais ampla do que os dados oficiais, ou seja, dados voluntários podem expandir, aprimorar, complementar ou substituir dados oficiais, dependendo do tema e da metodologia de coleta de dados usada (J. L. C. Borges, Jankowski, & Davis Jr., 2016).

Existem problemas, no entanto, com o timing das alternativas: dados oficiais podem ser mais difíceis de obter, em um processo que pode levar um tempo significativo, especialmente enquanto APIs e IDEs não estiverem amplamente disponíveis (J. Borges, Jankowski, & Davis, 2015). Quando este for o caso, os dados voluntários podem ser usados como proxy, antecipando o que os dados oficiais poderão vir a mostrar em um futuro próximo, quando forem liberados (vide Gomide et al. (2011), para um estudo sobre o uso de tweets georeferenciados – um tipo de crowdsourcing passivo – sobre dengue como um alerta antecipado para o surto da doença, a ser mais tarde confirmado por registros oficiais).

Considere-se um exemplo envolvendo mobilidade urbana. A Tabela 2 lista vários conjuntos de dados que podem ser obtidos de forma mais adequada a partir de fontes oficiais, em comparação com outros dados de interesse para esse problema que podem ser obtidos usando apenas ferramentas e técnicas de crowdsourcing. Uma autoridade de trânsito que queira melhorar seu gerenciamento de ativos de mobilidade precisaria combinar ambas as colunas, mantendo seus investimentos em mapeamento sistemático e em informações geográficas associadas a seus processos de trabalho, e criando iniciativas de crowdsourcing, a serem disseminadas pelos usuários do sistema e combinadas com registros oficiais. Além disso, a autoridade pública precisaria fornecer feedback aos cidadãos que contribuem com dados válidos, tanto demonstrando que os dados de crowdsourcing são realmente usados, como reagindo prontamente a denúncias ou indicações de problemas de qualidade de serviço.

4.DISCUSSÃO: DESAFIOS E DIFICULDADES

Pode ser tentador colocar todo o fardo da ampla e sistemática abertura de dados públicos sobre os ombros das organizações públicas. No entanto, para os governos, a abertura de dados não é livre de custos. Johnson, Sieber, Scassa, Stephens & Robinson (2017) identificam os custos diretos e indiretos do provimento de dados abertos e propõem uma discussão sobre esses custos. Johnson et al. argumentam que há poucas evidências para sustentar alegações de que a abertura de dados leva à participação do cidadão, embora a transparência seja benéfica em muitos aspectos conhecidos. Os autores levantam um argumento importante e válido: publicar e manter grandes volumes, atualizados, de dados complexos, legíveis por máquina pressupõe a capacidade dos usuários de lidar com esses dados, especialmente em um momento em que a alfabetização em dados é amplamente reconhecida como um desafio social. A abertura de dados pode reforçar a exclusão digital, uma vez que as organizações privadas e os cidadãos com maior nível educacional tendem a beneficiar-se

ISSN 1828-5961

dela mais do que o cidadão comum. A padronização pode ter um efeito positivo, mas há uma enorme necessidade de lidar com a heterogeneidade de fontes, em parte causada pela variedade de métodos de coleta de dados, e criar melhores técnicas de integração.

Além disso, Johnson et al. ressaltam que dados abertos que não lidam com questões de privacidade podem até prejudicar os cidadãos. Isso imediatamente levanta uma discussão sobre o compromisso entre a necessidade de transparência governamental e a necessidade de proteger as pessoas do uso ilícito de dados, e sobre situações nas quais os tomadores de decisão poderiam desequilibrar esses critérios em direção ao seu interesse político. Janssen (2012) argumenta que os dados abertos devem ser vistos como um insumo para os esforços de abertura governamental, nos quais ações são tomadas na busca da transparência, abertura, prestação de contas e acessibilidade. Johnson et al. (2017) concluem que desafios duradouros à garantia de acesso aos dados podem ter pouco a ver com a sua abertura e estão mais ligados a questões de formato, conhecimento técnico necessário, padronização entre jurisdições, bem como completeza dos dados e preocupações com qualidade. Em nosso ponto de vista, tais desafios devem ser ativamente perseguidos, e a criação de aplicativos fáceis de usar e que empregam dados abertos governamentais pode empoderar os cidadãos com ferramentas com as quais podem superar limitações individuais, viabilizando sua participação na tomada de decisão (Davis Jr., Fonseca, & Câmara, 2009).

Sieber & Haklay (2015) levantam outro ponto importante a respeito de informação geográfica voluntária. Eles argumentam que dados de crowdsourcing não podem ser considerados uma fonte de dados como qualquer outra, já que não estão livres de implicações sociais. Concordamos com esse ponto de vista e consideramos que tanto os projetistas de aplicativos de crowdsourcing quanto os usuários dos dados gerados pelos voluntários precisam ter em mente as várias limitações dessa fonte de dados. A literatura sobre crowdsourcing e VGI reforça enfaticamente a noção de que dados providos por voluntários podem ser enviesados, incompletos, imprecisos e abrangendo apenas uma parte do espaço de interesse. No entanto, devemos perceber que, em muitas situações, e para muitos temas, a multidão pode ser a melhor – ou a única – fonte de dados importantes. Como resultado, qualquer agenda de pesquisa futura sobre dados geográficos voluntários ou de crowdsourcing deve incluir o desenvolvimento de meios para lidar adequadamente com dados produzidos pelos cidadãos, projetando de modo a levar em conta suas limitações e considerando tais limitações em seu uso.

Tabela 2 – Conjuntos de dados oficiais versus voluntários

Dados Oficiais	Dados de crowdsourcing
Malha viária	Velocidade instantânea de ônibus
Endereçamento urbano	Verificação do quadro de horários
Malha de circulação de veículos	Qualidade dos veículos e limpeza
Pontos parada de ônibus, estações de metrô	Avaliação da qualidade dos serviços prestados
Itinerários de ônibus e metrô	Avaliação do conforto das viagens
Quadros de horário de ônibus e metrô	Bloqueios de vias não planejados
Bloqueios de ruas planejados	Engarrafamentos de trânsito
Intervenções e obras públicas planejadas	Incidentes de segurança no transporte
Acidentes de trânsito	Acidentes de trânsito
Tarifas	Pontos de interesse para viagens
	Rotas de bicicletas
	Ruas caminháveis e pontos de visada
	Denúncias de crimes e problemas de segurança

Argumentamos que uma das principais vantagens potenciais dos dados de crowdsourcing é a sua disponibilidade em curto prazo, principalmente para aplicações dinâmicas. No entanto, devemos reconhecer que lidar com fluxos de dados em tempo real pode ser ainda mais difícil, especialmente quando esses dados são usados em análises e acumulados para instruir a criação de políticas públicas. Há um contraste imediato com o big data obtido por extensas redes de sensores, e o aviso emitido por Sieber & Haklay (2015) vem imediatamente à mente para nos lembrar que os cidadãos podem servir como sensores (Michael F. Goodchild, 2007), mas sua natureza é fundamentalmente diferente do comportamento dos dispositivos de hardware.

Embora as perspectivas de participação popular em iniciativas de crowdsourcing e crowdsensing sejam positivas, a disponibilidade imediata de dados públicos de organizações governamentais brasileiras ainda é limitada a alguns casos de sucesso e a alguns projetos em andamento. Para atender aos requisitos da Lei de Acesso à Informação, sites ou recursos de Web GIS simples não são suficientes, pois não há metadados e, geralmente, só se consegue visualizar e interagir, mas não baixar os dados em formato não proprietário e legível por máquina. Infraestruturas de Dados Espaciais são mais adequadas nesse aspecto, mas a ênfase excessivamente cartográfica em alguns deles, como a Infraestrutura Nacional de Dados Espaciais (INDE) [14] brasileira pode inibir alguns tipos de uso.

Outras iniciativas brasileiras relevantes e referenciais de provimento de dados espaciais incluem a IDE ambiental do estado de São Paulo [15], a IDE do estado da Bahia [16], e outros recursos que, embora não estejam organizados na forma de uma IDE, dão acesso a amplos conjuntos de dados espaciais úteis, como nos casos dos portais de dados abertos das cidades de São Paulo [17] e Belo Horizonte [18]. Internacionalmente, o padrão-ouro para IDEs é o projeto INSPIRE [19], da Comissão Europeia, e provedores de dados abertos governamentais tais

ISSN 1828-5961

como o Data.gov dos Estados Unidos e o Data.gov.uk do Reino Unido, entre muitas outras iniciativas.

5. CONSIDERAÇÕES FINAIS

Há um acesso crescente a fontes múltiplas e heterogêneas de dados espaciais, produzidos por vários agentes, para atender a diversas necessidades. Precisamos pesquisar, desenvolver e empreender explorando maneiras de integrar dados de fontes muito diversificadas para fazer a ponte entre os mundos virtual e real, considerando que os seres humanos são uma importante, ainda que problemática, fonte de informações valiosas. Lidar com essas limitações é necessário, se quisermos usar o potencial das fontes atuais de dados digitais, juntamente com nossa crescente capacidade de processamento e aprendizado, na busca de soluções para problemas reais da nossa sociedade e do nosso planeta.

AGRADECIMENTOS

Este texto é derivado de uma palestra convidada e motivada pela professora Ana Clara Mourão Moura, coordenadora do simpósio GeoDesign South America 2017, a quem o autor agradece. O autor agradece as contribuições de Michele Brito, Hugo Vellozo, Natalia Machado, Guilherme Mateveli, Salatiel Santos, Rodrigo Wenceslau, Rodrigo Smarzaro e Junia Borges, estudantes e pesquisadores cujo trabalho individual e colaborativo dentro do Laboratório Interdisciplinar de Ciência da Computação (LabCS+x) e na UFMG contribuíram muito para consolidar as ideias e discussões apresentadas neste texto. O autor também deseja agradecer ao CNPq e à FAPEMIG, agências de fomento à pesquisa e desenvolvimento no Brasil e em Minas Gerais, pelo suporte a suas atividades de pesquisa.

NOTAS

[1]. http://www.openstreetmap.org

[2]. https://developers.google.com/places

[3]. https://foursquare.com

[4]. https://www.yelp.com

[5]. http://www.waze.com

[6]. http://ushahidi.com

[7]. http://www.dpi.ufv.br/projetos/clickonmap/

[8]. http://aqui.io/themerise

[9]. Lei 12.527, de 18 de novembro de 2011. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm (in Portuguese)

[10]. https://opengovdata.org

[11]. http://gtfs.org. A especificação GTFS foi inicialmente proposta pelo Google e pela autoridade de trânsito para Portland, Oregon, como um padrão de dados abertos. Seu nome original era Google Transit Feed Specification. Ver http://beyondtransparency.org/chapters/part-2/pioneering-open-data-standards-the-gtfs-story for a complete background on GTFS.

[12]. https://prefeitura.pbh.gov.br/bhtrans/informacoes/dados/dados-abertos

[13]. http://www.inde.gov.br

[14]. http://datageo.ambiente.sp.gov.br

[15]. http://geoportal.ide.ba.gov.br

[16]. http://dados.prefeitura.sp.gov.br

[17]. http://dados.pbh.gov.br

[18]. http://inspire-geoportal.ec.europa.eu