Crowdsourced information from Tweets during the WorldCup in Brazil: A theme search

J. Borges[1]*, P. Jankowski[2], C.A. Davis Junior[3]
[1]Department of Urbanism, Federal University of Minas Gerais, Belo Horizonte, Brazil
[2]Department of Geography, San Diego State University, San Diego, CA 92182-4493, United States of America
[3]Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil

*Corresponding author: E-mail: juniaborges@yahoo.com.br, Tel: +5531 75835000

Abstract
Social Media have become an ordinary place for citizens. The popularity of social media affords a great opportunity to investigate how citizens experience landscape, how they feel about it, and what is its impact over citizens. Geographic Information Systems (GIS) have become an important source to support decision making as the use of geospatial technologies enables the construction of complex databases, quick access to data, and development of predictive scenarios. Combination of both has been called crowdsourced mapping and represents a propitious opportunity for their integrated study. In this paper we aim to evaluate contributed versus volunteered crowdsourced geotagged information. As case study we analyze information about the FIFA WorldCup (Brazil, June and July 2014) and Instagram's information from Pampulha Region in Belo Horizonte, Brazil, to show the analytical potential of contributed information. We seek to get prepared and inspired to identify possibilities to study landscape and its essential urban values reported by users of the analyzed area according to the stakeholder point of view.
Keywords: Crowdsourced Mapping; Volunteered Geographic Information; Social Media Geographic Information; Public Participation.

1. SOCIAL MIDIA, GIS AND CROWDSOURCING

Social Media have become an ordinary place for citizens. The popularity of social media affords a great opportunity to investigate how citizens experience landscape, how they feel about it, and what is its impact over citizens. Geographic Information Systems (GIS) have become an important source to support decision making as the use of geospatial technologies enables the construction of complex databases, quick access to data, and development of predictive scenarios.

Crowdsourcing is a set of techniques that allow the creation of datasets by collecting and joining contributions from citizens with no previous training or special expertise. Usually, citizens contribute voluntarily, and the Web is used as a platform for receiving contributions[1].The term Geographic Information System (GIS) is applied to systems that perform the computational treatment of geographic data and that store the geometry and the attributes of data that are georeferenced, that is, situated on the earth surface and represented in a cartographic projection [2].Combination of both has been called crowdsourced mapping and represents a propitious opportunity for their integrated study.

As well as the definition of Crowdsourcing and GIS, an observation of what is Social Mediais important to enlarge the crowdsourcing mapping comprehension and envision its development as a whole. Social Media represents a revolutionary new trend of communication among and with citizens that should be carefully analyzed by those who have interest about the opinions and trending values of the society. The current trend toward Social Media can be seen as an evolution that goes back to the Internet's roots, since it re-transforms the World Wide Web to what it was

initially created for: a platform to facilitate information exchange between users [3]. Thus, “Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content”[3].

Understanding the trends of the crowd communication and using GIS through Spatial Analysis enables researchers to emphasize the measurement properties and relationships, taking into account the spatial location of the phenomenon under study in a direct way [4].

The development of new geographic information favors the development of Spatial Data Infrastructures (SDI). As an example, the demand for advancements in the use of geographic information in Europe, mainly through the implementation of Directive 2007/02/CE, established the INfrastructure for Spatial InfoRmation in Europe (INSPIRE). Regional SDIs already represent the de-jure technical platform for the development of regional and local planning processes. The innovation also is promoted if it is considered that the process of data collection, mainly after the Web 2.0, can also include the participation of users, in the role of creators of content/information [5]. Furthermore, the results from collecting data in social networks, or Social Media Geographic Information (SMGI) and their integration with information from established data production institutions, or Authoritative Geographic Information (AGI), can foster the development of "smart city" strategies, since the needs and opinions are informed by local communities in a bottom-up approach [5]. The same principle shall be applied to volunteered geographic information (VGI) projects. Next section establishes the distinction between SMGI and VGI.

2. VGI x SMGI

Volunteered Geographic Information (VGI) is a concept linked directly to requesting and collecting geographic information from the crowd with a clear objective for the user’s participation.

However, relevant geographic information can also be obtained using an API (application programming interface) to tap into social media sources without the user’s explicit knowledge. In the first case, users must be mobilized to contribute with their individual knowledge, and in the second indications of such knowledge are automatically gathered from the user’s postings on social media. APIs are collections of functions that are used to query and retrieve information from such datasets or services in an organized way, i.e., so that the data owner’s servers are not overcome with excessive requests and the users does not need to know details on how data are organized and stored at the source [1]. Examples of geographic APIs include Google Maps Geocoding API (which includes functions for obtaining geographic coordinates from an urban address) and Twitter’s API (which allows capturing streams of tweets based on some selection criteria) [1].

We are faced with two very different methods of gathering information."Crowdsourced data collected with user control is volunteered, whereareas crowdsourced data collected with no or limited user control is contributed" [6]. Contributed geographic information extracted through API was coined as Social Media Geographic Information by Campagna et al. [5, 7]. The main difference between the sources is on whether citizens know or not about the purpose of the use of published data. SMGI represents a deviation from common vision of VGI [4] and for its nature may be classified as “implicit VGI” in respect to “explicit VGI” [4].

The nature of the data collection process is critical because, in the case of information from the SMGI, there is no objective response and data capture comes from generic manifestations. Often, the information is not directly applicable and it is up to the researcher to engage in identifying focus responses and behaviors, therefore the importance of its metadata. These data can mainly answer questions related to the user’s position as recorded automatically and related contributions, such as comments on places, paths taken, most often visited places, etc. This is comparable to dynamic

mappings, in which the elements are well defined in terms of time and space, and which has as one of its basic principles to reveal mobility and concentration of groups of people around certain places or events.

On the other hand, in a VGI project, the questions can be more complex, aiming to obtain answers directly related to the objectives of the project. For instance: given two alternative urbanistic projects under analysis by the city council, which one better suits the population? Where are the urban issues that the population encounters daily? Which and where are the values of a community? The nature of a question from a VGI Project can be more elaborated and directed to a research objective.

Some application examples of SMGI would be the origin of calls made from mobile phones during or after a given event[1] [8, 9], and the number of check-ins to a location at a given time or period (period of the day, month, year, etc.). SMGI records the concentration of activities in space and in response produces a picture of the territory use. These principles were stated in the article "Discovering Landmark Preferences and Movement Patterns from Photo Postings" [10]. The document shows mapped paths over a period of time, indicating preferred paths in Seattle, among other findings as shown the following excerpt from work:

> "Spatiality of people's interests; locations of landmarks and events that are of interest to photographers; Temporality of people's interests; dates of photographing places and events and the seasonality of people's interests; Spatial extent of people's interests; boundaries of areas and events represented on photographs; Connectivity between photographed places represented by a network of moves connecting places of interest; Travel patterns of photographers and their temporal characteristics." [10]

Other very interesting related article to the treatment of the data from SMGI is: "Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics"[11]. The authors found through territorial distribution of information what themes were discussed in Seattle and their clusters. They also found distribution patterns and listed the topics, concluding that Twitter's posts are very related to time and space. In other words, when and where certain events occupy the mind of the people are strictly related, for instance the food theme was posted during lunch and dinner time, and the majority of people posted about transportation during the traffic period of the weekdays.

Differentiating the types of crowdsourced mapping is important because, in SMGI, although information has been voluntarily disclosed in social media, it was collected by a data mining process. Therefore, information is used without users having objective knowledge of how it has been used. On the other hand, VGI requires involvement and participation of citizens in a voluntary and knowledgeable way. Therefore, crowdsourcing today means a powerful way to research on behaviors, trends and values. In the first case, the behavior and individual choices are automatically gathered from user postings on social media, and in the second users must be mobilized to contribute their individual knowledge.

Some questions should be highlighted to justify the integration of diverse methodologies of crowdsoucing mapping: why would citizens, who have no obvious incentive, be willing to spend time creating content or filling in forms in a VGI website? What kind of person is willing to participate, and what makes them to be precise or imprecise about their posts? [12] Notice also that,

[1]The Sensible City Lab at MIT (Massachusetts Institute of Technology) produced a video explaining the results available online [8, 9]

in many cases, there is a difficulty to keep participants in a VGI project motivated [13, 14] while the participation from SMGI is self-promoted, as it is the result of another Web activity.

3. CASE STUDIES

3.1 SMGI general characteristics

The first analysis of this paper uses a Twitter dataset collected during the FIFA WorldCup (Brazil, June and July 2014) to show the analytical potential of contributed information. The potential of using SMGI is also demonstrated through two Instagram datasets (one from a long period of time, extracted based on a set of key words, and the other from a given day, collected considering a given area). We demonstrate Twitter's API potential, by establishing a parallel between contributed and volunteered information, its possibilities, its reach, impact over analysis and coverage of specific events or phenomena. The analytical potential of the collected information's is directly related to a semantic filter, which could be used to extract the information from the Web. Other interesting method to download the information that could be related to the geolocation of the posts as demonstrated through Instagram. It is important to notice that area focus extraction can also be performed from Twitter's API. The geographic related area data mining process for collecting posts could help to improve the search and indicate what kind of people use social media in the area, and what are the keywords and tags used to indicate interests and values. Experiments will be detailed in the next sections.

Using a semantic filter, the collection of data for a detailed study should take place to indicate important keywords that can have a direct impact on the research success. In a social media environment, the main way to attach a category or a subject to a message is by adding hashtags to it. If a post is related to food in a positive way it could simply use a good food related tag such as #goodfood. A semantic analysis should follow the potential words used by the group of people that are the focus of the research agenda. This could also help to indicate which is the profile of the people related to the certain topic or what is considered a good food. The hashtags are dynamically created by users and their creation and use could be short lived (hours in the Twitter case), and the searcher should be attentive to the posts and follow the indication of words used when relating to certain topics, since tweets can only be downloaded within a 7 day window.

As in Twitter's case, other constraints are usually imposed on the API. In the case of the Instagram, only 5000 requests per hour can be read, and also the results of the API online only displays 20 answers to the queries, but some programming could make a loop consultation to download the requested information. This has been explored by the LETICIA API for Instagram, developed by LABIC (*Laboratório de Imagem e Cybercultura of Federal University of Espirito Santo Brasil*). A very interesting information about the use of Instagram as a source of SMGI is that it has the biggest average of geotagged (linked to coordinates) posts, as in the posting procedure the geotag option is set by default. An analysis performed on Instagram messages showed an average of 56% of posts with a geotag, as opposed to Twitter's average of around 2%.

3.2 SMGI from Twitter

Twitter's API has as its main function the provision of access to messages. In principle, many Twitter messages collected through the API may contain volunteered information, in textual and unstructured form. As such, the API provides opportunities for data gathering considering uses other than the ones intended by the message issuers. In Twitter the agreement between the service provider and user establishes an exchange of the contents for the service of sending messages. From the tweets we are able to observe trends, flows, vocabularies, slang, meaningful events, information propagation, community formation and many other social networking properties. Other social media with API availability may allow for gathering completely diverse patterns and information depending on their characteristics and their intended purposes [1, 15, 16, 17].

In the case study a two day Twitter dataset was collected during Brazil's World Cup(14th and 15th of June, 2014) using a combination of multilingual semantic filters related to sports, countries, and players.

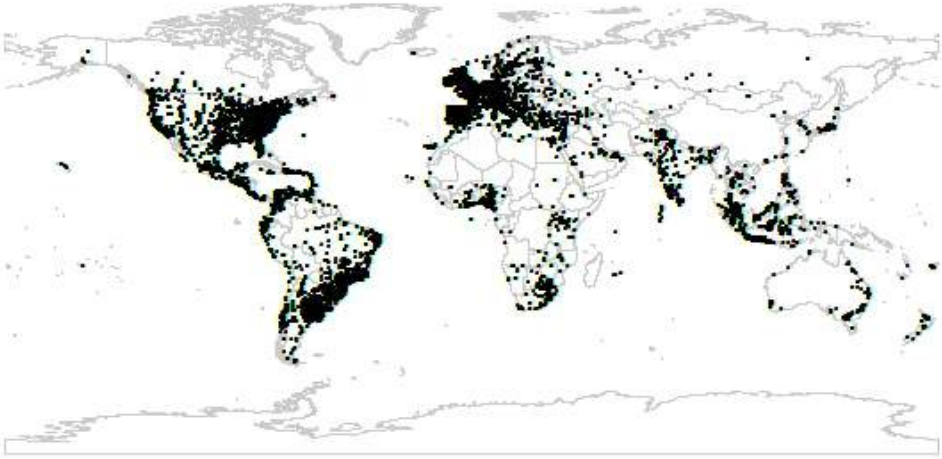

Figure 1: WorldCup Twitter posts from around the globe. National Institute of Science and Technology for the Web (InWeb), [18]

In Figure 1 can clearly visualize where the world cup was tweeted about the most. On the 14th of June there were four World Cup games: in Belo Horizonte Colombia versus Greece; in Recife Ivory Coast versus Japan; in Fortaleza Uruguay versus Costa Rica; and in Manaus England versus Italy. On the 15thof June there were three games: in Brasilia Switzerland versus Equator; in Porto Alegre France versus Netherland and in Rio de Janeiro Argentina versus Bosnia.

The geographical focus in this paper is on Pampulha (an administrative area in the City of Belo Horizonte, Minas Gerais, Brazil). Accordingly, tweets posted in Pampuhla during the two days of the World Cup were collected. Looking at Figure 2 with Pampulha and the georeferenced tweets one can see a relatively small number of postings originating from the study area as compared with a larger, surrounding area.

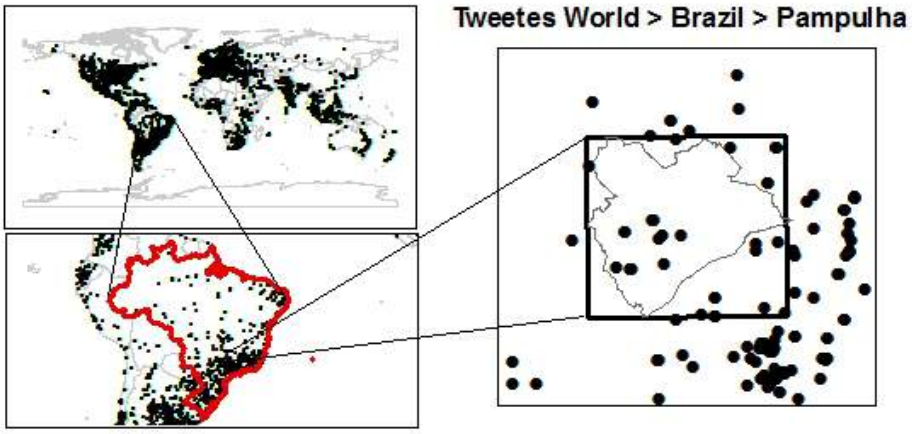

Figure 2: World Cup related tweets originating in Pampulha region of Belo Horizonte. National Institute of Science and Technology for the Web (InWeb), [18].

Using a word cloud tool to generate an overview of semantic content[2]it is possible to visualize the most frequently mentioned words in the comments posted from around the world.

[2] Using www.taxedo.com access in March 2015.

Figure 3: Word cloud performed from on Twitter messages from the Brazil's World Cup [19].

Not surprisingly Italy and England were the most frequently mentioned words as they were the two favorite teams playing during the two day tweet collection period.

3.3. SMGI from Instagram

As part of the research reported here, data analysis was performed on tweets with #pampulha (hashtag Pampulha). The analyzed dataset contained the data collected from the 14th of July 2014 to the 2nd of March 2015. There were 13.141 posts, of which 7.433 (56,5%) were georreferenced and fit for SMGI. The collection was performed using Leticia [20] developed by LABIC – *Laboratório de Cibercultura do Departamento de Comunicação da Universidade Federal do Espirito Santo*.

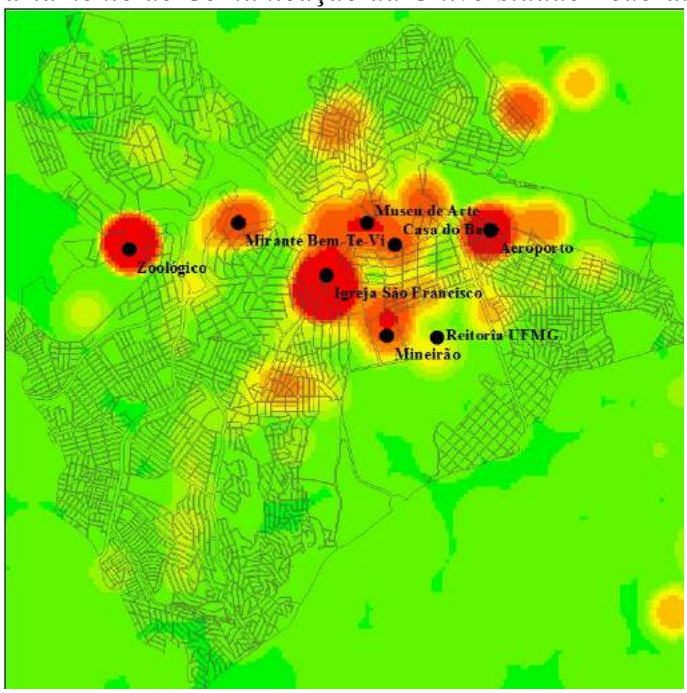

Figure 4: Spatial distribution of Instagram data of #pampulha. The courtesy of Andrei Bastos, *LABIC - Laboratório de Imagem e Cyber Cultura, Departamento de Comunicação, Universidade Federal do Espírito Santo*, Brazil. Warm colors represent higher density of postings.

In Figure 4 it is possible to observe the streets of Pampulha's area and that the postings of "#pampulha." Despite a relatively long data collection period (approximately seven months) the data is concentrated on Pampulha's landmarks; from west to east: the zoo, a sight view place called Bem-te-vi, The Church São Francisco de Assis, the Art Museum of Pampulha, Ball house (just beside the museum), the stadium Mineirão, the Federal University of Minas Gerais area, and the regional airport of Pampulha. All al these locations are Pampulha's landmarks.

Figure 5 depicts a second dataset, the courtesy of Pierangelo Massa, from the UrbanGIS Lab at the *Dipartimento di IngegneriaCivile, Ambientale e Architettura*, at *Università di Cagliari*, Italy. It represents only one day in March 2015 of Instagram posts from Pampulha area. The data set was collected using a rectangular area from the Pampulha region.

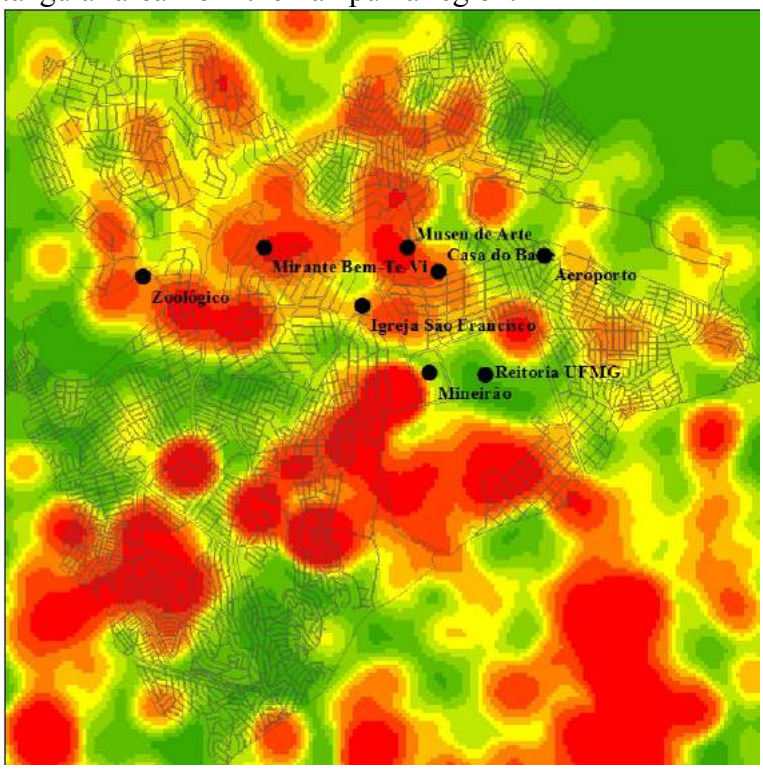

Figure 5: Collection of Instagram data postingsfor one day. The courtesy of Pierangelo Massa, UrbanGIS Lab, *Dipartimento di Ingegneria Civile, Ambientale e Architettura, Università di Cagliari,* Italy. Warm colours represent higher density of postings.

Without the use of keyword filters it became clear that the area of Pampulha, Belo Horizonte, Brazil, is very popular with Istagram postings, as the dataset represents 2703 posts collected during two ordinary days, making it a suitable area for future value and trend analysis from SMGI of Instagram posts.

Since the posts from #pampulha numbered over seven thousand and the harvesting of posts using a geographical filter (rectangular frame) netted nearly two thousand, given two very different time frames for both data collections (seven months and one day respectively) one can conclude that there are other (than World Cup and football) important words and topics being discussed in that area.

5. CONCLUSIONS

From the analysis performed it is possible to conclude that VGI and SMGI represent a great opportunity for analysis of specific events or understanding spatial phenomena. The potential of SMGI is amplified by the possibility of reaching a larger share of the population, whereas from a VGI project the type of analysis should attain a higher complexity and focus on the question asked

but with a more focused participation. The participation quality is a matter to be further addressed.

To perform a study about the landscape and its essential urban values reported by users requires specific preparation. There should be a preparation of the time frame of the collection of the dataset from Twitter, for instance, as well as the appropriate correlation analysis of the phenomena observed mainly through keywords and associated hashtags. However, it is necessary to conduct a comprehensive research on the semantic analysis of the keywords used, for example, as in the case of analysis of urban values at Pampulha, in which it is not reasonable to expect success while performing a search for technical keywords, as "#occupationrate", assuming that their use would be as popular as "#niceview" or "#sky". It should be interesting to make a long period and area-focused dataset collection, and perform a semantic analysis of the correlation and concentration of posting spots in the Pampulha region.
Transposed the barriers of understanding and expansion of user participation of Crowdsourcing Mapping, it is essential to think about the role of the urban planner for a full application of the techniques. In both collection methods of crowdsourced information (VGI and SMGI), gaining knowledge is clearly a goal. In order for the urban planner to leave the position of authorial designer to become a decoder of the collective will it is necessary to create an understanding of conditions of urban values that are representative as to what is valuable to society. Technicians can make use of local information because generally the spaces are perceived and understood more clearly by those who experience it, and can make use of appropriate tools to build this knowledge.

Normally the acknowledgement of the change of an era or a paradigm takes time to be noticed by people, also the change occurs when society is ready for it, meaning that it has been through hierarchical processes. Looking at history it seems that a specific fact leads to the paradigm change but it shall be only a boundary, as society as well as nature does not make jumps to skip steps necessary to promote change. As a new and faster way of communication in the World Wide Web, social media can catalyze social processes as it virtually brings people together and enables information to circulate faster. As it takes time to process and reflect on some information, this speed-up in circulation time poses a danger of people reacting to information without reflecting on it, which may eventually lead to loss of interest. Motivation should be brought through reflection and knowledge production, not through impulsiveness. This should be an interesting topic for further discussion about crowdsourcing.

Acknowledgements
Contribution to the Projects: "Parametric Modeling of Territorial Occupation: proposal of new resources of geo-technologies to represent and plan the urban territory", with the support of CNPq – National Council for the Scientific and Technological Development - Call MCTI/CNPq/MEC/CAPES Nº 43/2013, Process: 405664/2013-3, and "Geodesign and Parametric Modeling of Territorial Occupation: new resources of geo-technologies to landscape management of Pampulha Region, Belo Horizonte", with the support of CNPq – National Council for the Scientific and Technological Development – Call MCTI/CNPQ/MEC/CAPES Nº 22/2014, Process: 471089/2014-1. We thank FAPEMIG the financial support to the presentation.

The third author acknowledges the support from CNPq - National Council for the Scientific and Technological Development and FAPEMIG – *Fundação de Amparo a Pesquisa do Estado de Minas Gerais*, Brazilian agencies in charge of fostering research and development.

References

1. Bores, J., Jankowski, P., Davis Junior, C. A. *Crowdsourcing for Geodesign: Opportunities and Challenges for Stakeholder Input in Urban Planning*. Cartography - Maps Connecting the World: 27th International Cartographic Conference 2015 - ICC2015, Springer. Rio de Janeiro. No Prelo.

2. CÂMARA, G., MONTEIRO, A. M., FUCKS, S. D. & CARVALHO, M. S. 2004. *Spatial Analysis and GIS: A Primer* [Online]. http://www.dpi.inpe.br/gilberto/tutorials/spatial_analysis_primer.pdf. [Accessed in April 2015].
3. Kaplan, A.M., Haenlein, M. *Users of the word, unite! The challengers and opportunities of Social Media.* Business Horizons. Elsevier. Volume 53, Issue 1, January – February 2010, Pages 59 – 68. [Onlile http://www.sciencedirect.com/science/article/pii/S0007681309001232, Accessed in April 2015].
4. Campagna, M., Floris, R., Massa, P., Gircheva, A., Ivanov, K. The role of Social Media Geographic Information (SMGI) in Spatial Planning. In: Geertman S, Ferreira J, Goodspeed R, Stillwell J (eds) Planning Support Systems and Smart Cities. 2015. Springer.
5. Massa, P., Campagna M. *Social Media Geographic Information: Current developments and opportunities in urban and regional planning*. Proceedings of the 19th International Conference on Urban Planning and Regional Development in the Information Society GeoMultimedia 2014.
6. Harvey in Sui, Daniel. Elwood, Sarah. Goodchild, Michael. *Crowdsourcing Geographic Knowledge*. Dordrecht, Heidelberg, New York and London: Ed: Springer. 2013. 396 p.
7. Floris, R., Campagna, M. *Social Media Data in Tourism Planning: Analysing Tourists' Satisfaction in Space and Time.* Proceedings of the 19th International Conference on Urban Planning and Regional Development in the Information Society GeoMultimedia 2014.
8. http://senseable.mit.edu/realtimerome/ [Online: accessed January 20, 2015]
9. https://www.youtube.com/watch?v=RbhBz5UwRDQ access [Online: accessed January 20, 2015]
10. Jankowski, P. Andrienko, G. Andrienko, N. Kisilevich, S. *Discovering Landmark Preferences and Movement Patterns from Photo Postings.* In: Transactions in GIS, 2010, 14(6): 833–852. Blackwell Publishing Ltd.
11. Andrienko, G. Andrienko, N. Bosch, H. Ertl, T. Fuchs, G. Jankowski, P. Thom, D. *Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics*. In: Computing in Science & Engineering. Visualization Corner.Copublished by the IEEE CS and the AIP. 2013.
12. Goodchild, M. F. *Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0.* International Journal of Spatial Data Infrastructures Research 2: 24-32, 2007.
13. Borges, J.L.C., Zynger, C. *Crowdsourcing: a citizen participation challenge.* In: TeMA – Journal of Land Use, Mobility and Environment Special Issue Eighth International Conference INPUT Smart City Planning for Energy, Transportation and Sustainability of the Urban System, Naples, 46 June 2014
14. Campagna, M., Kudinov, A., Girsheva, A., Ivanov, K., Kopnov, M., Falqui, R. *Place, I Care! Crowdsourcing planning Information.* AESOP – Association of European Schools of Planning, conference. Dublin. 2013. 18p.
15. Frias-Martinez, V., Soto, V., Hohwald, H., Frias-Martinez,H. Characterizing Urban Landscapes using Geolocated Tweets. 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risc and Trust.
16. Muhammad, A., Longley, P., Featured graphic. Tweets by different ethnic groups in Greater London. Environment and Plannin A 2013, volume 45, pages 1524 – 1527.
17. Singleton, A. D., Longley, P. Geodemographics, visualisation, and social networks in applied geography. Applied Geography 29 (2009) 289–298.
18. http://www.inweb.org.br [Online: accessed November 2014]
19. http://www.taxedo.com [Online: accessed March 20, 2015]
20. https://github.com/andreibastos [Online: accessed March 20, 2015]