VGI contributors’ awareness of geographic information quality and its effect on data quality : a case study from Japan

山下, 潤
九州大学大学院比較社会文化研究院社会情報部門

瀬戸, 寿一
東京大学空間情報科学研究センター

西村, 雄一郎
奈良女子大学人文科学系研究院

岩崎, 亘典
農業環境技術研究所

https://hdl.handle.net/2324/2348515

出版情報：International Journal of Cartography. 5 (2/3), pp.214-224, 2019-05-17. Taylor and Francis
バージョン：
権利関係：

VGI contributors' awareness of geographic information quality and its effect on data quality: A case study from Japan

Jun YAMASHITA [a]*, Toshikazu SETO [b], Yuichiro NISHIMURA [c] and Nobusuke IWASAKI [d]

[a] *Faculty of Social and Cultural Studies, Kyushu University, Fukuoka, Japan. Email: yamasita@sca.kyuhsu-u.ac.jp*
[b] *Center for Spatial Information Science, The University of Tokyo, Tokyo, Japan. Email: tosseto@csis.u-tokyo.ac.jp*
[c] *Faculty of Letters, Nara Women's University, Nara, Japan. Email: nissy_yu@cc.nara-wu.ac.jp*
[d] *Institute for Agro-Environmental Sciences, NARO, Tsukuba, Japan. Email: niwasaki@affrc.go.jp*

* Corresponding author

Abstract: In many countries, geospatial data are typically provided by public institutions. Cities have been mapped using such public data. On the other hand, the demand for geospatial data has been diversifying, given the requirements for mapping cities. To respond to demands for new geospatial data, creation of citizen-generated open data and volunteered geographic information (VGI) have recently become popular. However, the quality of such open data and VGI are not always guaranteed. The number of studies on quality assessments of VGI have increased in recent years. The present study aimed to identify OpenStreetMap (OSM), one type of VGI, as well as contributors' awareness of data quality, and the relationships between their awareness and the positional accuracy of the OSM data contributed by them. The results showed that awareness or lack of the positional accuracy did not affect the quality of the OSM data created by the contributors. These findings suggest that the crowdsourcing approach might not guarantee the data quality of VGI.

Keywords: OpenStreetMap, volunteered geographic information, data quality assessment, positional accuracy

1. Introduction

In many countries, geospatial data on roads, railways, rivers, city blocks, and building outlines, which comprise the skeletons of cities, have been primarily provided by public institutions. Cities have been mapped using such public data. Urban planning is a potential example of the mapping of cities using public geospatial data. However, demands for geospatial data have been diversifying, with the mapping cities being one such requirement. Geospatial data on bike roads for bike users, rather than car users on ordinal roads, and data on sidewalks without steps intended especially for physically challenged persons, are other examples of diversified demands for geospatial data.

Citizen-generated open data (Meijer et al., 2018) and volunteered geographic information (VGI) represented by OpenStreetMap (OSM) were created in response to demands for new geospatial data that are not covered by public data. Route plans within buildings (Goetz et al., 2013), assessments of obstructions for the physically challenged (Prandi et al., 2014), and locational cognition of bike accidents on bike paths (Ferster et al., 2017) are examples of VGI utilization.

In addition to the rise of public interest in VGI in recent years, academic research on VGI has been gaining ground. One aspect of VGI research is the assessment of its data quality. As mentioned above, public institutions have mainly created and provided geographical data according to the data quality standards stipulated by laws. Such official procedures have secured the quality of geographic information. On the other hand, the quality of VGI is not always guaranteed, partly because the creators of VGI, OSM contributors (or OSM mappers in the case of the aforementioned OSM) typically comprise ordinary citizens who might not be well-acquainted with the data quality standards or laws regarding geospatial data (Senaratne et al., 2017).

Research on positional accuracy is particularly advanced among studies on the data quality assessment of VGI, especially map-based VGI (Al-Bakri et al., 2010, Ciepłuch et al., 2010, Vandecasteele et al., 2015, Mullen et al., 2015). As described in the next section, geospatial data quality is assessed in terms of five data quality elements, one of which is positional accuracy. Haklay's (2010) work is viewed as a pioneering study on the data quality assessment of map-based VGI. He examined the positional accuracy of OSM highway data as the test data using road data provided by the Ordnance Survey, UK, as the reference data. He concluded that approximately 80% of the OSM road data were accurate on average. In addition, Haklay et al. (2010) first mentioned Linus' Law and also examined the geospatial data quality of OSM road data using a crowdsourcing approach. Linus' Law is stated as follows: "Given enough eyeballs, all bugs are shallow" (Raymond, 1999, p. 30). The crowdsourcing approach to the quality assessment of geospatial data follows the same concept. Consequently, he showed that the number of OSM contributors was weakly positively related to the positional accuracy, but this positive relationship was not statistically significant. As per Haklay et al. (2010), map contributors' awareness of positional accuracy, rather than their number, may influence the quality of their contributed VGI. It might, therefore, be hypothesized that high awareness of geospatial data quality leads to high positional accuracy. To the best of our knowledge, however, few studies have focused on the awareness of OSM contributors with regard to VGI data quality or the relationship between such awareness and the positional accuracy of VGI created by them.

In view of the aforementioned research achievements, the purpose of the present study is to understand the awareness of VGI users and contributors regarding the quality of the geographical information. Then, the study assesses the relationship between the contributors' awareness of the VGI and its data quality via comparison of public data with contributor VGI.

The research methods used for the aforementioned purposes are described in the next section. We also refer to the quality assessment of geospatial data. The results of the study are presented in section 3. Finally, we draw conclusions about the relationships between the awareness of VGI contributors and VGI data quality, and discuss the limitations and scope for future research in the last section.

2. Methods

Although several international standards regarding data quality assessment of geographic information (ISO 19157:2013) exist, the Japan Profile for Geographic Information Standards Version 2.0 (JPGIS 2.0)[1] was utilized to analyze the awareness of VGI users and contributors residing in Japan with regard to geospatial data in this study. Taking international and domestic standards of geographic information into consideration, the Geographical Survey Institute in Japan presented JPGIS 2.0 in 2014. This standard comprises the core aspects of the above-mentioned standards, and is recommended for users and makers of geographic information in Japan. JPGIS 2.0 has based the quality of geographical information in "Annex 3: Quality" on the "quality principles" stated in the aforementioned international and domestic standards for geographic information. Annex 3 also indicates that detailed data evaluation and reporting methods are available in a technical note published in 2005 and partly revised in 2007 by the aforementioned institute.

The research data comprised OSM data among the map-based VGI. From September 1st to November 30th, 2017, a web survey was conducted to identify the awareness of OSM users and contributors with regard to the five data quality elements shown in Table 1. In this study, an OSM user is regarded as a person who has used OSM data at least once before the web survey was conducted, while an OSM contributor is a person who has made or edited OSM data at least once before the web survey was conducted. Regarding the awareness of these data quality elements, it was hypothesized that OSM mappers are more aware of data quality than OSM users. To verify this hypothesis, we examined differences in awareness of geospatial data quality between OSM users and contributors using the analysis of variance (ANOVA) on percentages of awareness for the five data quality elements listed in Table 1.

Completeness
Logical consistency
Positional accuracy
Temporal accuracy
Thematic accuracy

Table 1. Data quality elements in JPGIS 2.0.

Concerning the positional accuracy of the OSM data created by web survey respondents, we used a method similar to that of Haklay (2010). He employed public road data as a reference and used them to generate buffers. He regarded coverage rates of OSM data, which are test data, by these buffers as measures of the positional accuracy.

The test data in this study were the OSM road data created by the web survey respondents. However, we could not extract only the OSM data created by these respondents. Thus, the OSM road data were extracted from a part of the OSM main contribution area using the web analysis tool entitled "How did you contribute to OpenStreetMap? (HDYC-OSM)"[2]. The reference data were the road centerline data in the Base Map Information compiled by the Geographical Survey Institute in Japan[3].

Both road centerline and OSM road data in four vector-tiles (measuring approximately one square kilometer) at zoom level 16[4], which were included in the spatial range of the OSM contributors' main contribution area, were extracted for each OSM contributor. The extracted road centerline and OSM road data in the four vector-tiles were compared to assess the positional accuracy. Road centerline data were used as input data for the buffer analysis because 0.5 m-wide buffers were generated from the OSM road data in this study; the ratios of the road centerline data provided by the OSM road data were considered to be high before the buffer analysis was conducted. The minimum width of the road considered in this study was 3 m. The road was a single-lane road. Moreover, the maximum width of a typical car is 2.5 m in Japan. Therefore, if errors in the positional accuracy exceed 0.5 m, the car could collide with buildings or objects situated beyond the boundary of the 3 m-wide road[5]. Thus, the buffer was set to 0.5 m in this study. The

[1] http://www.gsi.go.jp/ENGLISH/page_e30210.html

[2] http://hdyc.neis-one.org/

[3] https://github.com/gsi-cyberjapan/vector-tile-experiment/blob/gh-pages/README_en.md

[4] In the Base Map Information, the whole of Earth is included within a $2^0 \times 2^0$ vector-tile, namely a single square tile, at zoom level 0. At zoom level n, it is covered by $2^n \times 2^n$ vector-tiles.

[5] The road centerline data are collected and maintained as per the requirements of the Survey Act (Act No. 188 of 1949). The data are collected using geodetic rather than GPS surveys. For 3 m-wide roads, this act requires the positional accuracy to be less than 0.25 m in the horizontal direction. In the questionnaire survey, OSM users and contributors were

	JPGIS 2.0	Completeness	Logical consistency	Positional accuracy	Temporal accuracy	Thematic accuracy
a) Observation						
Yes	17	22	19	27	20	16
No	49	44	47	39	46	50
b) Proportion (%)						
Yes	25.8	33.3	28.8	40.9	30.3	24.2
No	74.2	66.7	71.2	59.1	69.7	75.8

Table 3. OSM users' awareness on spatial data quality and data quality elements.

	JPGIS 2.0	Completeness	Logical consistency	Positional accuracy	Temporal accuracy	Thematic accuracy
a) Observation						
Yes	5	12	20	26	10	17
No	44	37	29	23	39	32
b) Proportion (%)						
Yes	10.2	24.5	40.8	53.1	20.4	34.7
No	89.8	75.5	59.2	46.9	79.6	65.3

Table 4. OSM contributors' awareness about spatial data quality and data quality elements.

proportion of the total road length centerline data included in the 0.5 m buffers from the OSM road data to the total length of the road centerline data in the aforementioned four vector-tiles was termed "coverage ratio" and used as the indicator of positional accuracy.

Before comparing the OSM road data with the road centerline data, we unified features included in both datasets. The OSM road data were limited to those having the eleven tags in the highway category (Table 2), while centerlines of less than 3 m were excluded from the road centerline data.

Motorway
motorway_link
Trunk
trunk_link
Primary
primary_link
Secondary
secondary_link
Tertiary
Unclassified
Road

Table 2. Utilized road tags in OpenStreetMap.

Finally, the ANOVA was conducted on coverage ratios with and without awareness about the data quality mentioned above, so as to examine the influence of the presence or absence of the awareness of positional accuracy.

3. Results

The web survey was completed by 84 OSM users, among which 49 respondents contributed to OSM creation. Of these 49 respondents, 27 supplied their OSM account names. On this basis, these 27 respondents were designated as research objectives for the analysis of positional accuracy.

The results revealed that the OSM users were not aware of the OSM data quality. Less than half of the OSM users indicated that they were aware about either the data quality requirements of JPGIS 2.0 or the five data quality elements (Table 3). It is worth noting that a relatively high proportion of the OSM users indicated that they were aware about positional accuracy among the five quality elements.

The results also revealed that the OSM contributors, like the users, were less aware of the quality of geographic information used while creating their OSM data. This finding is supported by the fact that a little more than 10% of OSM contributors were aware of JPGIS 2.0 (Table 4). With respect to the five data quality elements, the OSM contributors were not aware of the data quality related to these elements, except for positional accuracy; approximately half of them were aware of this aspect. Similar to the OSM users, the OSM contributors had relatively high awareness of the positional accuracy.

Source	*SS*	*df*	*MS*	*F*
Between	80.083	1	80.083	2.207
Within	362.833	10	36.283	

*** < 0.01, ** < 0.05, * < 0.1

Note. Between: Between the OSM users and contributors, Within: Within the OSM users or OSM contributors, *SS*: Sum of squares, *df*: Degrees of freedom, *MS*: Mean squares, *F*: F ratio.

Table 5. Results derived from the ANOVA.

queried about their awareness of the data quality, but this study did not examine their knowledge of the positional accuracy or survey method mandated by the Survey Act.

Number of Respond-ents	Aware-ness of position-al accuracy	Contribu-tion to road tag "high-way"	Per-centage	Cumula-tive percentage
1	Yes	35,167	35.3	35.3
2	Yes	18,799	18.9	54.2
3	Yes	14,180	14.2	68.5
4	Yes	13,321	13.4	81.9
5	Yes	6,824	6.9	88.7
6	Yes	4,312	4.3	93.0
7	Yes	4,288	4.3	97.4
8	Yes	1,456	1.5	98.8
9	Yes	755	0.8	99.6
10	Yes	392	0.4	100.0
11	Yes	33	0.0	100.0
12	Yes	0	0.0	100.0
13	Yes	0	0.0	100.0
14	No	688,211	76.7	76.7
15	No	112,253	12.5	89.3
16	No	24,887	2.8	92.0
17	No	24,365	2.7	94.7
18	No	23,093	2.6	97.3
19	No	9,734	1.1	98.4
20	No	4,309	0.5	98.9
21	No	4,018	0.4	99.3
22	No	1,822	0.2	99.5
23	No	1,757	0.2	99.7
24	No	1,531	0.2	99.9
25	No	820	0.1	100.0
26	No	0	0.0	100.0
27	No	0	0.0	100.0

Table 6. Respondents' contribution to the road tag "highway".

As determined by the ANOVA, there is no difference in awareness of the data quality between the OSM users and contributors. As shown by the ANOVA, using the observed numbers of OSM users and contributors who were aware of JPGIS 2.0 and the five data quality elements, the ratio of variance (*F*-statistic) is not statistically significant at the 10% level between the OSM users and contributors (Table 5). This result indicates that neither OSM users nor OSM contributors were aware of the data quality of VGI.

Number of Respondents	Total length of road centerlines (a)	Total length of road centerlines covered by OSM (b)	Coverage ratio (%) (b/a)
1	20,119.9	3,245.7	16.1
2	12,078.5	670.6	5.6
3	6,643.4	788.2	11.9
4	16,291.5	3,077.1	18.9
5	19,162.8	2,554.7	13.3
Average	14,859.2	2,067.3	13.9
SD	5,556.0	1,248.3	
14	14,762.0	1,892.4	12.8
15	23,065.1	8,003.8	34.7
Average	18,913.6	4,948.1	26.2
SD	5,871.2	4,321.4	

Table 7. Differences in positional accuracy between the aware and unaware contributors.

As mentioned above, 27 respondents had experience with OSM creation and supplied their OSM account names in the web survey. Among them, 13 respondents contributed OSM data and were aware of the positional accuracy, while the remainder were not. Table 6 shows these respondents' contributions to OSM highway data in descending order using the aforementioned HDYC-OSM. In this table, the 13 respondents with awareness of the VGI data quality and 14 respondents lacking awareness are placed in two different groups. Respondents were placed in descending order based on the degree of their contributions, and those who contributed just less than the first 90% of the cumulative contributions for each group were selected as research subjects for detailed analyses of the positional accuracy. These selected respondents comprise the top five, namely Nos. 1 to 5, in Table 6, in the group with awareness of the positional accuracy, and the top two, namely Nos. 14 and 15, in the group without awareness. For the seven respondents, we examined the relationships between awareness/unawareness of the positional accuracy and the positional accuracy of the OSM data they created.

The results of this analysis based on coverage ratios as an indicator of the positional accuracy showed that the average coverage ratio of respondents without awareness of the positional accuracy was higher than that of the aware respondents. Namely, the former comprise 23.1%, and the latter, 13.2% (Table 7 and Figure 1). To ascertain whether the difference in the average coverage ratios between the aware and unaware respondents was statistically significant, we conducted an ANOVA using coverage ratios of the seven respondents.

The results of this ANOVA revealed no difference in the coverage ratios between the groups with and without awareness of the positional accuracy. This is because the ratio of variance (*F*-statistic) of the coverage ratios is not statistically significant at the 10% level (Table 8). This finding suggests that differences in awareness of VGI data quality had little influence on the data quality.

Source	*SS*	*df*	*MS*	*F*
Between	0.016	1	0.016	2.359
Within	0.034	5	0.007	

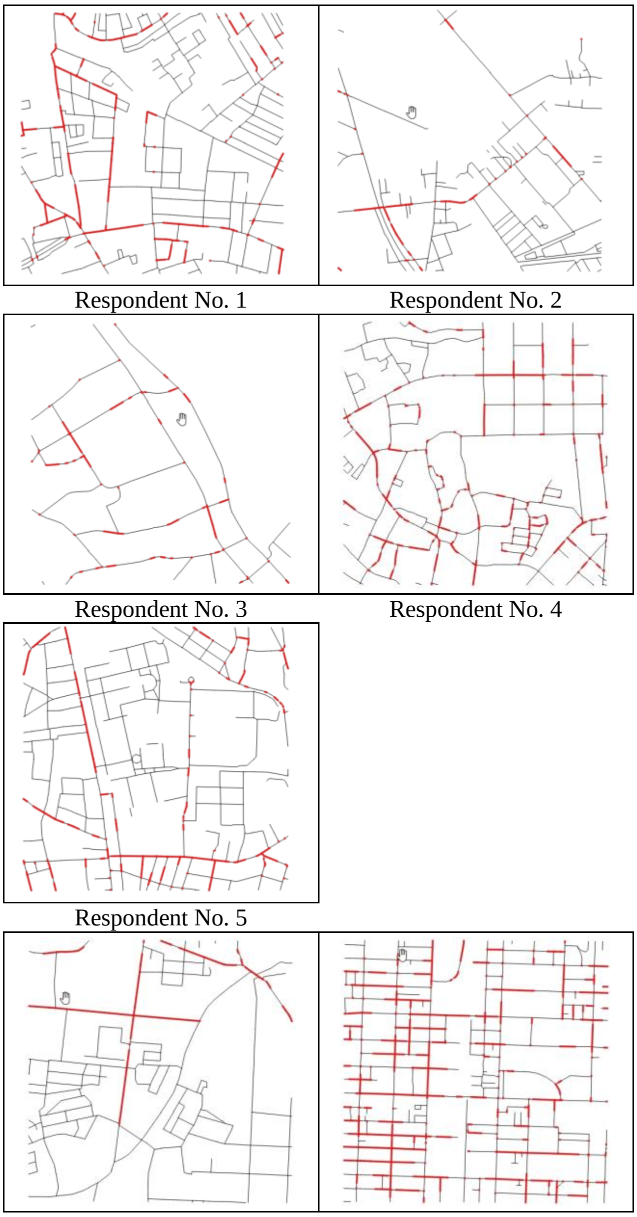
*** < 0.01, ** < 0.05, * < 0.1

Note. Between: Between two groups, Within: Within a group, *SS*: Sum of squares, *df*: Degrees of freedom, *MS*: Mean squares, *F*: F ratio.

Table 8. Results of the ANOVA.

4. Conclusions

In view of the current knowledge on VGI data quality, this study aimed to identify VGI users' and contributors' awareness of data quality and examine the relationships between awareness and data quality in comparison



Respondent No. 1 Respondent No. 2

Respondent No. 3 Respondent No. 4

Respondent No. 5

Respondent No. 14 Respondent No. 15

Figure 1. Coverage of road centerline data by OSM road data
Notes: Red lines represent road centerlines covered by OSM roads, while black lines indicate uncovered road centerlines.

between public data and OSM data. It was found that the OSM contributors' awareness of the geographic information data quality was no higher than that of the OSM users. In addition, it was revealed that the awareness of positional accuracy was relatively high among the five data quality elements. An analysis using the coverage ratio of publicly compiled road centerline data with OSM road data as the evaluation index of positional accuracy indicated that OSM contributors' awareness of data quality had little influence on the quality of the OSM road data created by them. However, the OSM road data employed in this study were not directly extracted by the web survey respondents, but were indirectly selected from the OSM road data to which these respondents chiefly contributed. If such OSM data are available, therefore, it might be necessary to identify the relationships between the awareness of data quality and the data quality of purely extracted OSM data to which only the respondent contributed.

The aforementioned findings contribute to current knowledge on positional accuracy. As mentioned in the introductory section, Haklay et al. (2010) assumed that map contributors' awareness of positional accuracy may influence the positional accuracy of their contributed VGI. However, the findings of our study do not support their assumption. We show that contributors' awareness of positional accuracy is less likely to affect the positional accuracy of their contributed VGI. Thus, the present study indicates the need for careful deliberation when considering this aspect in practical applications.

The aforementioned results also suggested that the crowdsourcing approach might not guarantee the data quality of VGI. Goodchild et al. (2012) proposed three approaches regarding the data quality of VGI, crowdsourcing being one of them. The crowdsourcing approach assumes that errors caused by a VGI contributor are verified and corrected by a group of other VGI contributors. However, the finding that the OSM data based on crowdsourcing did not fulfil the data quality in this study raises the question of whether this approach guarantees VGI data quality. Furthermore, according to this finding, it can be inferred that the data quality should be secured not by contributors but also by users. In such cases of VGI data quality assurance, the validity of the social approach (another approach presented by Goodchild et al. (2012)) should be considered. The social approach is defined as the way in which highly reputable and trustworthy persons, in terms of contribution to VGI, act as gatekeepers to maintain the data quality of VGI that other contributors create. In the context of the social approach, certain trustworthy individuals play leading roles regarding the data quality assessment of VGI. From this viewpoint, further studies will be necessary to identify how trustworthiness (Flanagin et al., 2008, Fogliaroni et al., 2018) and reputation (Resnick et al., 2000) affect VGI data quality. Lastly, although data quality depends on not only the awareness/knowledge of the OSM contributor but also the procedures described, tools used, specifications, and above all his/her engagement with regard to implementing these aspects, in the present study, the authors have focused solely on addressing contributors' awareness/knowledge about data quality. Thus, the aforementioned uncovered determinants of data quality will be considered in future studies.

5. Acknowledgements

The authors would like to express their gratitude to three reviewers for their valuable comments on early drafts of this article. This article was partly supported by the JSPS KAKENHI Grant No. 17K17658.

6. References

Al-Bakri, M., and Fairbairn, D. (2010). Assessing the accuracy of crowdsourced data and its integration with official spatial datasets. Ninth international symposium on spatial accuracy assessment in natural resources and environmental sciences. Leicester, UK.

Ciepłuch, B., Jacob, R. Mooney, P., and Winstanley, A. (2010). Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. Ninth international symposium on spatial accuracy assessment in natural resources and environmental sciences. Leicester, UK.

Ferster, C.J., Nelson, T., Winters, M., and Laberee, K. (2017). Geographic age and gender representation in volunteered cycling safety data: A case study of BikeMaps.org. *Applied Geography*, 88, 144-150.

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

Fogliaroni, P., D'Antonio, F., and Clementini, E. (2018). Data trustworthiness and user reputation as indicators of VGI quality. *Geo-spatial Information Science*, 21, 213-233.

Goetz, M., and Zipf, A. (2013). Indoor route planning with volunteered geographic information on a (mobile) Web-based platform. *Lecture Notes in Geoinformation and Cartography*, 211-231.

Goodchild, M.F., and 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*, 37, 682-703.

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

Meijer, A., and Potjer, S. (2018). Citizen-generated open data: An explorative analysis of 25 cases. *Government Information Quarterly*, 35, 613-621.

Mullen, W.F., Jackson, S.P., Croitoru, A., Crooks, A., Stefanidis, A., and Agouris, P. (2015). Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features. *GeoJournal*, 80, 587-605.

Prandi, F., Andreolli, M., Eccher, M., Di Staso, U., and De Amicis, R. (2014). Barriers survey: A tool to support data collection for inclusive mobility. *Lecture Notes in Computer Science, 8515 LNCS, PART 3*, 772-779.

Raymond, E.S. (1999). *The Cathedral and the Bazaar*. O'Reilly Media, Sebastopol, USA.

Resnick, P., Zeckhauser, R., Friedman, E., and Kuwabara, K. (2000). Reputation systems. *Communications of the ACM*, 43(12), 45-48.

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

Vandecasteele, A., and Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system: The case of OpenStreetMap. *Lecture Notes in Geoinformation and Cartography*, 59-80.