Assessment Formation of Open Data Sources During Their Aggregation For Analyzing Road Accidents

Sergey Savosin, Nikolay Teslya, Sergey Mikhailov
SPC RAS
St. Petersburg, Russia
{savosin, teslya, sergei.mikhailov}@iias.spb.su

***Abstract*—Analyzing data in a smart city often requires aggregating a large number of heterogeneous sources into a single system. For the selection of sources, it is necessary to first assess each source. This will provide informed data choices and improve the quality of decision making. The paper proposes a formalization of criteria for evaluating data and a method for evaluating a source based on the use of the proposed criteria. The performance of the method was evaluated by analyzing data on road accidents in St. Petersburg, Russia. Based on the analysis results, three data sources were selected, for which the analysis was carried out and the results were visualized.**

I. Introduction

Many tasks related to the development of a smart city require analysis of spatial and temporal data to predict how the decisions made will affect the current situation. [1], [2]. An example of such an analysis would be the analysis of the traffic of the road network [3] for the planning of a new or reconstruction of the existing infrastructure; recommendation of attractions for tourist routes [4], the use of spatial data for the analysis of objects on record, including outside the visibility of surveillance cameras [5], [6]; researching the needs of residents based on their feedback and posts on social networks [7]. For all the tasks mentioned, it is required to analyze data concerning space and time from a variety of sources to track the dynamics of the situation in city districts.

When analyzing data from different sources, their aggregation may be required to expand the context, as well as to increase the accuracy and relevance of data through cross-validation. Automatic integration requires the development of methods for assessing the quality of knowledge, which will allow ranking sources of knowledge and choosing for integration only those that are of sufficient quality according to the criteria of the problem being solved. Thus, the development of criteria and methods for assessing the quality of sources of spatio-temporal knowledge is an urgent task to improve the quality of integration of knowledge from various sources of knowledge.

An ontology-based approach can be used to aggregate spatio-temporal data and knowledge. In particular, for the representation of geospatial data, there is a standard for describing geospatial ontology - GeoSPARQL from the open geospatial consortium (The Open Geospatial Consortium, OGC) [8]. The Open Geospatial Consortium is an international consortium of more than 500 enterprises, government agencies, research organizations, and universities dedicated to making geospatial (location-related) information and services easily found, accessible, interoperable, and reusable. The OGC GeoSPARQL standard supports the presentation and querying of geospatial data on the Semantic Web. GeoSPARQL defines a vocabulary for representing geospatial data in RDF and defines an extension to the SPARQL query language for processing geospatial data.

To represent the concepts of time in the ontology, a standard for the OWL-DL language has been developed since 2016. To date, the standard is still being discussed and supplemented with new concepts and relationships that make it possible to reflect the features of entities related to time and use it in knowledge representation systems [9], [10]. Attempts have also been made to combine spatial and temporal ontologies in a common task to perform inference in a content management system [11].

Representation of data from sources in the form of ontology allows using existing methods of ontology merging and alignment for aggregating data from various sources [12]. As a result of such alignment, a common ontology or a top-level ontology is formed, with the help of which it is possible to solve the problems of searching for space-time slices by objects of interest from several sources at once. Since the standards for constructing the corresponding ontologies based on OWL are used to represent the concepts of space and time, both automatic methods [13]–[15] using space-time metrics of similarity [16], and experts can be used to merge private ontologies [17], [18].

In this work, the main attention is paid to the development of a method for analyzing the quality of data when they are aggregated from heterogeneous sources. Aggregation is considered based on the example of data analysis on road accidents in St. Petersburg, Russia. The main sources of data in the above study are open statistics from the traffic police website, road data from the OpenStreetMap portal, historical weather data from the Weather Underground portal, and OpenWeatherMap.

The article is structured as follows. Section 2 provides an overview of works related to the analysis of the quality of data from open data sources. Section 3 provides a method for forming an assessment of the source of knowledge, taking into account the priority of the criteria and normalizing the assessments into a single assessment. Section 4 contains an example of using the method to estimate sources when

analyzing road accidents data. Section 5 describes an example of data integration and visualization for road accidents analysis in St. Petersburg, Russia.

II. Related Work

Data quality is a major issue when using heterogeneous data sources. This is especially true for spatial and temporal data, which is due to the complexity of their processing and presentation. Such sources can contain geodata (various maps), the history of changes to various physical objects or characters, or a sequence of events. There are two main ways of forming spatio-temporal data sources: professional processing for solving a specific problem or volunteering [19].

Professional cartographers most often form very accurate knowledge, but they are limited to a narrow subject area. Also, such data sources are updated much less frequently. The ability to edit such sources is extremely limited and requires high qualifications and experience [20]. Examples of this type of source are knowledge provided through state portals, for example, the Russian Open Data Portal [21].

Unlike professionals, volunteers can provide much more information in a shorter time frame. At the same time, the filling of the data source will be more voluminous due to the broader expertise due to the larger number of participants. Filling of sources is carried out using publicly available tools (cheap sensors from mobile devices, free software), which have lower accuracy, however, thanks to averaging data over several volunteers, the quality of information becomes acceptable for use in various applications [22]. An example is the OpenStreetMap portal, which provides an open world map with additional information about geographic features. The detail of the map and the amount of available information about objects strongly depend on the population density, subject area, and external factors [23].

A review of studies devoted to assessing the quality of knowledge from open sources, using methods of systems analysis, showed that the concepts of data quality and knowledge quality are used interchangeably and most often imply multiple measures of compliance with the criteria for using data in a specific task [24]. In this regard, three main areas of research in this area can be distinguished: i) the definition of criteria for the quality of data and knowledge, ii) the definition of a combination of criteria that are relevant for a specific task, and iii) the formation of a general assessment for quality of data and knowledge [25], [26]. A detailed analysis of the measures applied is presented in the work [24]. It should be noted here that most of the presented criteria are subjective and require the involvement of experts to form an assessment. Only some of the criteria, for example, accuracy/inaccuracy, completeness/incompleteness, inconsistency, timeliness, can be objectively calculated algorithmically, which makes them especially valuable in assessing the quality of knowledge sources [26].

To assess open sources of knowledge, additional quality assessment measures can be used that clarify existing metrics, for example, time accuracy of an event/object, thematic accuracy, usability, logical consistency [27]. This is due to the lack of a single standard for the formation of open sources of knowledge, as a result of which the knowledge in them is rather heterogeneous and cannot be assessed as in the sources created by the professional community. Quality indicators may also include indicators of data and knowledge (for example, point density), demographic indicators (population density according to the geographic area for which knowledge is available), socio-economic indicators, characteristics of volunteers providing information [27].

When assessing knowledge from open sources [28] cross-validation can also be used or expert communities can be involved [23], [29], [30]. In the first case, automatic tools for assessing the quality of knowledge are being developed, the task of which is to find violations of consistency (consistency) in the response of the data source to the request, in the second, the assessment and comparison of the quality of data from the source is a separate complex scientific problem [30].

Separately, we can highlight the use of an expert opinion to form an assessment of the data source. Methods using experts are based on the formation of a questionnaire that eliminates ambiguity (for example, using a matrix of experts' answers [31]), combining assessments from several experts and forming a ranked list of sources. Also, probabilistic assessment methods are used, which are based on the distribution of the probabilities of obtaining high-quality data depending on their semantic consistency [32].

There are also works based on the use of fuzzy logic to assess the quality of data. For example, paper [33] presents a system that evaluates three parameters (type of source, quality of data extraction, age of source) for data sources in electronic patient health records (EHR) based on fuzzy rules.

III. Knowledge source assessment method

A. Criteria for evaluation

As a result of the systematization of existing studies on assessing the quality of data and knowledge, a list of the most frequently used criteria was identified [23]–[27]. Each of the criteria was additionally ranked according to the frequency of use and the complexity of the calculation. The estimation of the frequency of use is normalized by the number of mentions in the reviewed works (in normalization, the value "1" corresponds to the mention to each work, when the rating tends to 0, the purity of the mention decreases), and the complexity of the calculation was estimated from the possibility of automatic assessment of the quality criterion: the involvement of experts is required (0), partial automation is possible (0.5), fully automatic (1). The use priority of the score was calculated as the median value of the frequency of use and the complexity of the calculation (All estimates are rounded to the nearest hundredths).

- Accuracy / Uncertainty - how much data and knowledge correspond to real values . As a rule, this parameter can be estimated either by comparison with other sources or by expert judgment. It is one of the most important and frequently used criteria due to its objectivity.

- Relevance - how well the knowledge corresponds to the user's task. This criterion is irrelevant for a simple analysis of data sources, but has a high priority in the case of selecting data sources for a specific task;
- Time relevance - How old the knowledge is from the time it was introduced to the knowledge source until it was used. Also, due to objectivity, the priority of this criterion is high.
- Representativeness - how understandable the knowledge is to a non-specialist. The assessment of this criterion is objective, but can only be carried out with the involvement of a large number of volunteers. It is the participation of volunteers that is important here since for a specialist this criterion will rather express an assessment of the correctness of the data structure.
- Accessibility - whether there are restrictions on access to knowledge, whether the data source is open.
- Completeness - the degree of knowledge sufficiency to solve the problem. For this criterion, knowledge is considered relevant, and it is assessed to what extent it is sufficient to solve the problem.
- Correctness (no errors) - the presence or absence of inconsistencies in the data from the source. This parameter can be estimated both within one data source and with the involvement of other sources;
- Consistency - how well the knowledge formats match when re-accessing the source;
- Timeliness - The rate at which data ages over time.
- Misuse - the presence of data that should not be used (as a rule, service information: indexes, identifiers, metadata).

Table 1 shows the evaluation of the criteria mentioned above regarding to frequency, difficulty, and priority in reviewed scientific works. Cases with two values, such as "*0.33 / 0.44*", should be interpreted as values for "*source analysis / for a specific task*". For each of the criteria, units of measurement are determined, which makes it possible to assess knowledge in accordance with the criterion. Units are described in the next section.

TABLE I. VALUES OF THE CHARACTERISTICS

Criteria	Frequency	Difficulty	Priority
Accuracy / Uncertainty [23]–[27]	1	1	1
Relevance [24]–[27]	0.23 / 0.76	0.5	0.36 / 0.63
Time relevance [23]–[25], [27]	0.76	1	0.88
Representativeness [24]–[27]	0.84	0	0.42
Accessibility [23]–[25], [27]	0.46	0	0.23
Completeness [25], [26]	0.15 / 0.38	0.5	0.33 / 0.44
Correctness (no errors) [23]–[27]	1	1	1
Consistency [24], [25], [27]	0.23	1	0.62
Timeliness [23], [25]	0.23	1	0.62
Misuse [26]	0.15	1	0.57

B. Method description

A review of existing studies also found that evaluating a sufficiently large number of criteria requires work with experts. To evaluate sources with experts, it is proposed to use the questionnaire, followed by the calculation of the integral value for the data quality criterion. To formalize possible answers, it was decided to use the apparatus of fuzzy sets to formalize expert assessments. The rationale for this choice is the possibility of obtaining a numerical estimate corresponding to a peak from the sets of possible answers, and then using this estimate to form an overall estimate of the data or knowledge source. In other cases, the criteria are assessed by numerical characteristics from the range [0,1].

To calculate the estimates, linguistic variables were formed, which were evenly divided into disjoint sets from the range [0,1]. The following are the values of the variables for those criteria where expert judgment is required:

- Relevance. Expert judgment is required. For the assessment, 5 linguistic variables are introduced ("irrelevant", "rather irrelevant", "impossible to assess", "rather relevant", "relevant"), which are uniformly mapped to the range [0,1] using fuzzy sets.
- Representativeness. Also expert judgment is required. 5 linguistic variables ("unrepresentative", "rather unrepresentative", "impossible to evaluate", "rather representative", "representative"), which are uniformly mapped to the range [0,1] using fuzzy sets.
- Availability. Expert judgment is required. Three values are used, 0, 0.5, 1, which correspond to closed sources, sources with licensed use restrictions, and open sources.
- Completeness. For the assessment, 5 linguistic variables are introduced ("incomplete", "rather incomplete", "impossible to evaluate", "rather complete", "complete"), which are uniformly mapped to the range [0,1] using fuzzy sets.
- Timeliness. Expert judgment is required. Three variables are introduced: "low", "medium", "high".
- Misuse. Expert judgment is required. 5 linguistic variables are introduced ("wrong", "rather wrong", "impossible to evaluate", "rather correct", "correct"), which are uniformly mapped to the range [0,1] using fuzzy sets.

For the rest of the criteria, within the framework of the proposed method, a calculation method based on an automatic analysis of the source characteristics is determined:

- Accuracy / Uncertainty - can be considered as the ratio of the number of objects in the data source to the number of actually existing objects. The range of values is [0,1].
- Relevance. The older the information, the lower the relevance. It can be estimated as the inverse relationship to the difference between the date of inclusion in the source and the date of use. Range is [0.1]
- Correctness. It can be automatically estimated as the ratio of correct cases of data extraction and comparison from various sources to the total number of requests.
- Consistency. It can also be automatically estimated as the ratio of correct cases of data retrieval from one source to the total number of requests.

Thus, the entire method can be represented as the following set of actions:

1) Selection of criteria according to the priorities for the source.
2) Determination of the possibilities of automatic calculation of criteria. The ability is determined by the availability of data required for automatic calculation, such as the date of the last update, the ability to estimate the real number of objects, the presence of several versions of the dataset.
3) If a more complete analysis is required, then experts should be involved to carry out the assessment according to the relevant criteria.
4) All estimates are normalized for conversion to the range [0,1].
5) Normalized estimates are multiplied by the value of the priority of each criterion and summed up to obtain an integral estimate of the source.
6) Sources are ranked according to the given ratings and the sources with the highest ratings are selected.

IV. Example of analysis of data sources

The main data source is the traffic police statistics portal. On it every month, data on accident cards are downloaded from the internal traffic police systems filled in by employees at the scene of incidents. To assess the quality of data in this source, the following criteria will be considered: completeness, accuracy, timeliness, correctness, relevance, misuse. Criteria such as availability and representativeness are not considered in this case, because the data sources used were selected in advance in such a way that they always have free access, and the data form in them has complete decoding and can be automatically brought to any understandable representation.

One of the most important criteria for data quality is its relevance. In this case, the relevance of the data is associated with the city's road infrastructure, which is constantly changing, and therefore the data on dangerous road sections can be relevant for many years, if no adjustments are made to the traffic regime in some sections, and may become outdated after the road works. The relevance of such data can be assessed automatically using the analysis of emergency-dangerous locations. The method of calculating the relevance turns out to be quite simple in this case, namely, it is necessary to select emergency-dangerous places and, by them, find out the number of accidents occurring by months. If the average number of accidents in such a place has not changed for many years, then the data is still relevant, and if the number of accidents has dropped sharply, then adjustments have been made to the road network and the old data have become less relevant. For greater accuracy of cutting off data that are no longer relevant, an interval of at least 2-3 last months should be taken. Thus, the criterion of the timeliness of the data is also variable for different places in the city and it is impossible to calculate its total value, which would fit all the data. Since all of interval consist data, the relevance if the source is estimated as 1 as well as timeliness is also estimated as 1.

An equally important criterion for data quality is its correctness. During manual analysis of the extracted data, a frequent error in the data related to the address and coordinates was revealed: the coordinates do not always correspond to the address and the address does not always contain the correct value. These inaccuracies are associated with the human factor and malfunctions in the equipment of the accident employees, which determine the coordinates of the scene of the accident. In total, erroneous coordinates have 14.9% of all data. Therefore, the correctness can be estimated as 0.851.

However, it was found that most often one field from a pair of coordinates - the address is correct and on its basis, it is possible not only to check the correctness of the data but also to restore it in automatic mode. To do this, you need to use another data source. In this case, it was decided to choose the Geocoding API, which allows to get coordinates by address and vice versa. The reliability of this service in terms of providing correct geographic data is an order of magnitude higher than on the traffic police portal and using it you can improve the correctness of the data and, as a result, their quality.

The next data quality criterion to be considered is completeness. Although the accident cards contain quite a lot of different information about both the location of the accident and the participants, there are several gaps in them. One of them is the weather at the scene of the accident. In the accident cards, such information is presented rather scarcely, however, adding an additional source of information can increase the completeness of data on road accidents. In this case, the OpenWeatherMap service was chosen as such a source. Using it, full weather data can be acquired for the accident point by coordinates and time of the incident. Estimation of the completeness based on lacks in weather type (8.85%), incomplete information about road conditions (0.5%), and road types (0.01%) is 0.9109.

The criteria for data accuracy can also be checked automatically in this case, but the data are always provided by services in a standardized form and correspond to their fields, and the amount of service information in them is minimal and is also used within the system, so evaluation of accuracy is 1. Misuse of the data is also not possible due to datasource specific, so the evaluation is “correct” and the numeric value is 1. The result evaluation of the used data sources are presented in the Table II.

TABLE II. Estimation of used data sources

Criteria	Accidents Data	OSM	OpenWeatherMap
Completeness	0.9109	1	1
Accuracy	1	1	1
Timeliness	1	1	1
Correctness	0.851	1	1
Relevance	1	1	1
Misuse	1	1	1
Overall	4,4518	4,64	4,64

V. Integration and visualization of accident data

In order to aggregate different data sources, it is necessary to have an idea of the internal structure and what information is contained in each element of sources. A solution is to present the source metadata as an RDF graph. The road accident has an event nature and represents a point in space and time. In this way, the initial graph has two vertices: a location and a timestamp. The graph is extended when new data sources have new information about the road accident domain or remain unchanged in case there is no to add. The merging process is

based on searching the closest hyponym vertices. In this work, ontology is manually created for road accident cards (Fig. 1). In the future, the process of creating and expanding an ontology will be automated. And the expert's work will be to check the system operation.

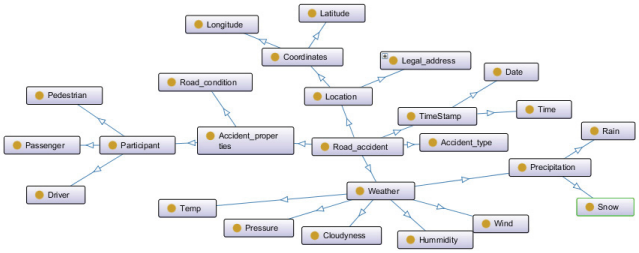

Fig. 1. Ontology for road accidents description

Data scrapper and data provider for microservices in the system are combined together into one web service. This service is written in Java using the Spring framework and the following libraries: Hibernate, Gson, Lombok. Data is stored in the PostgreSQL database with the PostGIS plugin. All interactions with the service are performed via REST API. As a starting dataset, accident cards were received for the last 5 years for Saint-Petersburg and Leningrad region. The number of received cards is approximately 55 000.

In a software implementation, Plotly for points visualization and PostgreSQL for data storage are used. The task was divided into the development of four microservices: a microservice for obtaining weather information, a microservice for obtaining traffic incidents cards, a clustering microservice, and a visualization microservice. The clustering algorithm is written in Python using the K-means algorithm. The algorithm minimizes the total squared deviation of cluster points from the centers of these clusters. The algorithm stops when there is no change in the intracluster distance on some iteration.

Fig. 2 shows the visualization of clustering results depending on the number of vehicles. The visualization shows the concentration of road accidents with the number of vehicles 1, 2, and 3, which may indicate safety problems in these areas. By using data from OSM and comparing it with data from the crash card, it is possible to identify the types of streets on which a large number of accidents occur. Integration with the weather service allows you to additionally categorize the accident areas according to weather conditions. For example, you can filter out accidents involving heavy rain or snow, bright sun, or fog.

Fig. 2. Traffic incidents map based on involved vehicles

Fig. 3 shows another example of data visualization. In this example, the main criterion is the type of incident. The correspondence between labels and types of incidents is presented in Table III.

Fig. 3. Traffic incidents clustering map based on involved vehicles

TABLE III. LABELS DESCRIPTION FOR FIGURE 3

Label	Description
1	Collision
2	Hit a pedestrian
3	Hitting an obstacle
4	Passenger fall
5	Hitting a cyclist
6	Hit a stationary vehicle
7	Exit the road
8	Rollover
9	Another type of accident
10	Hit on an animal
11	Collision with a person who is not a road user performing the service
12	Collision with a person who is not a road user performing the work
13	Throwing an Object
14	Hit a sudden obstacle
15	Collision with a person who is not a road user who carries out any other activity
16	Drop of cargo
17	Hitting a horse-drawn vehicle

The simple display of points on the map makes it very difficult to find places of concentration. In this regard, the DBSCAN clustering algorithm was used, which performs clustering based on the spatial proximity of points. Fig. 4 shows an example of filtering clusters depending on the number of accidents in it. As a result of clustering with distance parameters of 2 meters and the number of points in the cluster at least 20, more than 50 clusters were obtained, in which the dependence of the number of accidents on road conditions is clearly traced. Almost all the clusters obtained are concentrated at intersections and these roads are among the most heavily loaded roads in St. Petersburg.

VI. CONCLUSION

The criteria proposed in the work allow for a preliminary analysis of data sources for their inclusion in the analysis system. The source assessment method provides an objective assessment of each source and the formation of an integrated assessment, based on which a decision on the use of the data source can be made.

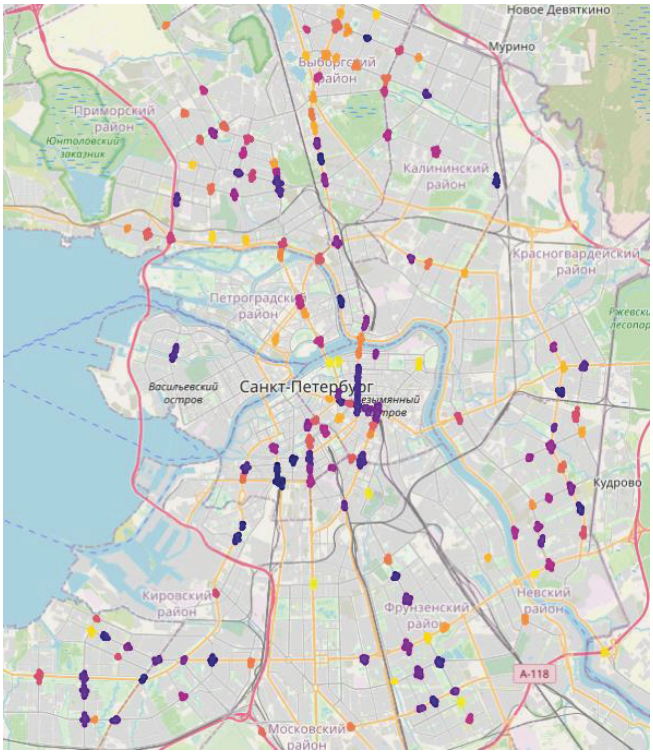

Fig. 4. Traffic incidents clustering map with distance less than 2 m and accidents count in a cluster more than 20

The method was tested on a system for analyzing data on the causes of road accidents. When using the method, data sources were assessed, among which were the accident statistics website, the OpenStreetMap map service, and weather services for obtaining historical weather data. As a result, services were selected that provide the required level of data quality and were aggregated in order to visualize the clusters of accidents and their connection with the causes.

Further work will focus on connecting more data sources and conducting a comprehensive analysis to identify dangerous road sections and developing a driver warning module when approaching an emergency dangerous section, in order to increase the attentiveness and accuracy of driving.

ACKNOWLEDGMENT

The reported study was funded by RFBR, according to the research project No. 20-07-00560 and research project 20-07-00904 (section 5).

REFERENCES

[1] M. Khan, M. Babar, S. H. Ahmed, S. C. Shah, and K. Han, "Smart city designing and planning based on big data analytics," *Sustainable Cities and Society*, vol. 35, pp. 271–279, Nov. 2017.

[2] M. Babar and F. Arif, "Smart urban planning using Big Data analytics to contend with the interoperability in Internet of Things," *Future Generation Computer Systems*, vol. 77, pp. 65–76, Dec. 2017.

[3] J. France-Mensah and W. J. O'Brien, "A shared ontology for integrated highway planning," *Advanced Engineering Informatics*, vol. 41, no. May, p. 100929, 2019.

[4] W. Zhang, T. Gu, W. Sun, Y. Phatpicha, L. Chang, and C. Bin, "Travel Attractions Recommendation with Travel Spatial-Temporal Knowledge Graphs," in *Computer Journal*, vol. 60, no. 3, Springer Singapore, 2018, pp. 213–226.

[5] W. Jin, Z. Zhao, Y. Li, J. I. E. Li, J. U. N. Xiao, and Y. Zhuang, "Video Question Answering via Knowledge-based Progressive Spatial-Temporal Attention Network," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2s, pp. 1–22, Jul. 2019.

[6] J. I. Olszewska, "Detecting Hidden Objects Using Efficient Spatio-Temporal Knowledge Representation," in *Communications in Computer and Information Science*, vol. 10162, J. van den Herik and J. Filipe, Eds. Cham: Springer International Publishing, 2017, pp. 302–313.

[7] N. Dilawar *et al.*, "Understanding Citizen Issues through Reviews: A Step towards Data Informed Planning in Smart Cities," *Applied Sciences*, vol. 8, no. 9, p. 1589, 2018.

[8] X. Lopez, "GeoSPARQL - A geographic query language for RDF data A proposal for an OGC Draft Candidate Standard," p. 13, 2012.

[9] S. J. D. Cox and C. Little, "Time Ontology in OWL." 2017.

[10] E. Baratis, E. G. M. Petrakis, S. Batsakis, N. Maris, and N. Papadakis, "TOQL: Temporal ontology querying language," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5644 LNCS, pp. 338–354.

[11] G.-A. Nys, M. Van Ruymbeke, and R. Billen, "Spatio-Temporal Reasoning in CIDOC CRM: An Hybrid Ontology with GeoSPARQL and OWL-Time," in *2nd Workshop On Computing Techniques For Spatio-Temporal Data in Archaeology And Cultural Heritage*, 2018.

[12] J. Euzenat and P. Shvaiko, *Ontology Matching*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[13] A. N. Tigrine, Z. Bellahsene, and K. Todorov, "Selecting Optimal Background Knowledge Sources for the Ontology Matching Task," in *Knowledge Engineering and Knowledge Management. EKAW 2016. Lecture Notes in Computer Science*, vol. 10024, 2016, pp. 651–665.

[14] T. Viana, C. Delgado, J. C. P. Da Silva, and P. Lima, "Ontology alignment with weightless neural networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10614 LNCS, pp. 376–384.

[15] N. Teslya and S. Savosin, "Matching Ontologies with Word2Vec-Based Neural Network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11619 LNCS, pp. 745–756.

[16] L. Wang, F. Zhang, Z. Du, Y. Chen, C. Zhang, and R. Liu, "A Hybrid Semantic Similarity Measurement for Geospatial Entities," *Microprocessors and Microsystems*, vol. 80, Feb. 2021.

[17] A. Smirnov, N. Shilov, N. Teslya, and A. Kashevnik, "Crowdsourcing-Based Multi-Layer Automated Ontology Matching An approach and Case Study," in *The Fourth International Conference on Intelligent Systems and Applications*, 2015, pp. 74–79.

[18] A. Smirnov, N. Teslya, S. Savosin, and N. Shilov, "Ontology matching for socio-cyberphysical systems: An approach based on background knowledge," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10531 LNCS, pp. 29–39.

[19] A. Mashhadi, G. Quattrone, and L. Capra, "The Impact of Society on Volunteered Geographic Information: The Case of OpenStreetMap," in *OpenStreetMap in GIScience: Experiences, Research, Applications*, J. Jokar Arsanjani, A. Zipf, P. Mooney, and M. Helbich, Eds. Cham: Springer International Publishing, 2015, pp. 125–141.

[20] A. Calafiore, "Designing an ontology of social place," in *CEUR Workshop Proceedings*, 2016, vol. 1769, pp. 23–27.

[21] "data.gov.ru | OPEN DATA RUSSIA." [Online]. Available: https://data.gov.ru/?language=en. [Accessed: 21-Sep-2021].

[22] M. Haklay, "How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets," *Environment and Planning B: Planning and Design*, vol. 37, no. 4, pp. 682–703, 2010.

[23] J. Cidália Costa Fonte, Vyron Antoniou, Lucy Bastin Jacinto Estima, Jamal Jokar Arsanjani and L. S. and R. V. Carlos Laso Bayas, "Assessing VGI data quality," *Mapping and the Citizen Sensor*, pp. 137–163, 2017.

[24] O. Azeroual, G. Saake, and J. Wastl, "Data measurement in research information systems: metrics for the evaluation of data quality," *Scientometrics*, vol. 115, no. 3, pp. 1271–1290, 2018.

[25] R. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[26] Q. Liu, G. Feng, X. Zhao, and W. Wang, "Minimizing the data quality problem of information systems: A process-based method," *Decision*

Support Systems, vol. 137, p. 113381, Oct. 2020.
[27] V. Antoniou and A. Skopeliti, "Measures and indicators of vgi quality: An overview," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 2, no. 3W5, pp. 345–351, 2015.
[28] H. Dorn, T. Törnros, and A. Zipf, "Quality evaluation of VGI using authoritative data-a comparison with land use data in southern Germany," *ISPRS International Journal of Geo-Information*, vol. 4, no. 3, pp. 1657–1671, 2015.
[29] D. Zielstra and A. Zipf, "A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany," *13th AGILE International Conference on Geographic Information Science*, vol. 1, pp. 1–15, 2010.
[30] M. Forghani and M. Delavar, "A Quality Study of the OpenStreetMap Dataset for Tehran," *ISPRS International Journal of Geo-Information*, vol. 3, no. 2, pp. 750–763, 2014.
[31] X. Chen and J. Lee, "The identification and selection of good quality data using pedigree matrix," in *Smart Innovation, Systems and Technologies*, 2021, vol. 200, pp. 13–25.
[32] B. Heinrich, M. Klier, A. Schiller, and G. Wagner, "Assessing data quality – A probability-based metric for semantic consistency," *Decision Support Systems*, vol. 110, pp. 95–106, Jun. 2018.
[33] C. Molina and B. Prados-Suarez, "Measuring the quality of data in electronic health records aggregators," in *IEEE International Conference on Fuzzy Systems*, 2020, vol. 2020-July.