Systematization of approaches to assessing the quality of spatio-temporal knowledge sources

S V Savosin, S A Mikhailov and N N Teslya

1 St. Petersburg Federal Research Center of the Russian Academy of Sciences, 39, 14th line, St. Petersburg, 199178, Russia

E-mail: savosin@iias.spb.su

Abstract. There are two main ways of creating and filling spatio-temporal knowledge bases: by professional cartographers and by volunteers using the crowdsourcing concept. The knowledge provided by professionals is highly accurate, but narrow in scope and domain-specific, as well as rarely updated. Editing of this kind of sources is impossible without permission, which makes them closed, despite the possibility of free use of the knowledge itself. On the opposite, ordinary users can provide a large amount of information in various fields of knowledge with minimal delay. However, the quality of such information and its accuracy will differ depending on the subject area and external factors, such as the accuracy of user devices, sensors, the distribution of users in the area, etc. The integration of knowledge from various sources will increase the overall accuracy and relevance of the data provided, which can be used to increase quality of decision making. Automatic integration requires the development of methods for estimating quality of knowledge sources, to rank them and choose only those that can provide sufficient quality in the criteria of the problem being solved. This paper presents a systematization of existing approaches to assessing sources of knowledge and, in particular, spatio-temporal knowledge.

1. Introduction

The formation of sources containing spatial and temporal data requires significant effort due to the complexity of the concepts of space and time. These sources can represent geodata (various maps), a history of various objects changes (objects history) or a sequence of events (events history). To create the structure of such data a long work of specialists is required, and before the digitalization process start, most of the work was performed only by professionals - cartographers, surveyors, archaeologists, historians, etc. The development and implementation of digital technologies allowed attracting a significant number of enthusiasts who are ready to contribute to the creation of sources of spatio-temporal data [1]. As a result, two main ways of forming spatio-temporal knowledge bases can be distinguished: by professionals and by volunteers.

The knowledge provided by professionals is highly accurate, but narrow in scope and focus in terms of the domain, and is rarely updated. Professionals can use very precise devices and all their knowledge and background to provide high-quality data. At the same time, editing these sources is impossible without permission, which makes them closed, despite the possibility of free use of the knowledge itself. Examples of this type of sources are knowledge provided through state portals, for example, the Russian Open Data Portal.
Unlike professionals, volunteers can provide a large amount of information in various fields of knowledge with minimal delay due to the development of mobile networks and sensors embedded in personal mobile devices. This data is not so accurate due to the hardware limitations of used devices, low background of participants[2]. However, open sources of knowledge are available for everyone and can be formed, edited, and used by any volunteer so the average result will be good enough for usage even in business purposes. An example is the OpenStreetMap portal, which provides an open world map with additional information about geographic features. The quality of such information and its accuracy will differ greatly depending on the considered subject area and external factors, such as the accuracy of the sensors of user devices, the density of distribution of users on the ground, and others [3].

The integration of knowledge from various sources (proprietary and open) could increase the overall accuracy and relevance of the data provided. The resulting integration can be used as a context in information system to make decision to support actions of a smart city residents. However, 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 in 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.

2. Overview of quality metrics for spatio-temporal data
Currently, studies are being conducted on both the development of methods and criteria for evaluating data and knowledge, and their practical verification on existing databases and knowledge. Quite often, the concepts of data quality and knowledge quality are used interchangeably and imply a multi-criteria measure of meeting the criteria for using data in a particular task. [4]. In this regard, three main areas of research in this area can be distinguished: the definition of criteria for the quality of data and knowledge, the definition of a combination of criteria that are relevant for a specific task, and the formation of an overall assessment of the quality of data and knowledge [5,6]:

• Accuracy/Inaccuracy - How much the data and knowledge correspond to the true values;
• Relevance - how much knowledge corresponds to the user's task;
• Representativeness - how understandable the knowledge is;
• Accessibility - are there any restrictions on access to knowledge.

The following criteria are often added to them [5,6]:

• Completeness/Incompleteness - how fully the knowledge corresponds to the task;
• Correctness (absence of mistakes) - knowledge is correct;
• Consistency - how much the formats of knowledge correspond when re-accessing the source;
• Timeliness - how old the data is over time.
• Mismembership – existence of data that should not exit in the source

For each of the criteria, units of measurement are determined to assess knowledge in accordance with the criterion. A detailed analysis of the measures used is presented in the work [4]. It should also be noted here that most of the presented criteria are subjective and require the involvement of experts to form an assessment. Some of them, for example, accuracy / inaccuracy, completeness /incompleteness, mismembership, timeliness can be formed in objective manner using algorithms, which makes them especially valuable when assessing knowledge sources [6].

Despite the universality of the considered criteria, for open sources of knowledge, additional quality assessment measures can be applied, clarifying the existing metrics, such as time accuracy, thematic accuracy, usability, logical consistency, [7]. The need to use additional measures and indicators arises 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 proprietary sources. Among the quality indicators, indicators of data and knowledge (for example, the density of points), demographic indicators (population density according to the geographic area for
which knowledge is available), socio-economic indicators, characteristics of volunteers who provide information are considered [7].

Proprietary sources can also be used to assess open source knowledge [8], or expert knowledge [3,9,10]. In the first case, automatic tools for assessing the quality of knowledge are being developed, in the second, assessing and comparing quality is a separate scientific task, as, for example, in the work [10].

Separately, we can highlight the use of expert opinion to form an assessment of the data source. The methods using experts are based on the formation of a questionnaire that allows you to eliminate ambiguity (for example, using pedigree matrix [11]), combining assessments from several experts and forming a ranked list of sources. Probabilistic estimation methods are also used, which are based on the probability distribution of obtaining quality data depending on semantic consistency. [12].

Fuzzy logic can also be used to assess data quality. Thus, the paper [13] presents a system that evaluates three parameters (type of source, quality of data extraction, age of source) for data sources on electronic health records (EHR) based on fuzzy rules.

3. Assessment of the quality of the result of knowledge integration
The consolidation of data sources into a single system allows not only to create a single point of access to them, but also to improve the quality of the data provided. This is achieved due to the possible overlap of the data provided by the sources and the formation of averaged data when combined.

For the merging of spatio-temporal data sources, a system is proposed, which is based on data representation using ontologies. Spatial and temporal concepts in the system are represented in the form of relationships between ontology concepts and can be used in queries to select data from sources corresponding to a certain area of space and / or time interval. Due to the use of logical and semantic constructions of time and space, such relations as “near”, “close”, “distant”, “now”, “before”, “after”, “during” and others can be modeled.

To connect data sources, it is proposed to create mediators (see figure 1.). Their main role is to provide an ontological description of the data provided by the source and translation of requests to the data source and the response from it from and to the ontology. The ontologies presented by mediators can be partially or completely integrated and combined into a common ontology of the problem domain. During the integration process, modules describing related concepts are allocated in ontologies. These modules are then linked through equivalent ontology classes. Due to this, the consistency of the common ontology is achieved.

![Figure 1. Integrating data from various data sources through ontology.](image-url)
Requests to an information system using a common ontology and mediators for accessing spatio-temporal data sources are made using the SPARQL language. For each source, the query is mapped according to the source type. For example, to query a spatial database running PostgreSQL with the PostGIS extension, the query from SPARQL is converted to SQL using specialized functions for finding geodata.

When the query results are returned, they are converted from a table view (or another one used in the given data source) to an ontology using a mediator. In the case of overlapping or conflicting data comes from a query from different sources, the source estimation described in the previous section should be applied. Sources are ranked according to the results of the assessment. The higher the position of the source in the ranked list, the higher the credibility of its data. Therefore, depending on the rules for combining data, they will either be accepted as the only possible ones, or they will have a higher coefficient in the convolution function.

The assessment of the completeness of the result of spatio-temporal knowledge integration was not carried out in any of the studied works. However, for the integration of knowledge, metrics have been developed, borrowed from the methods of mathematical statistics used to assess the classification and the process of extracting knowledge. [14]. These metrics are based on the analysis of sets and the assessment of the elements included in these sets. Among these metrics stand out: Precision, recall, F-measure [14–16]. To assess the quality of the resulting integration, as well as individual data sources, a set of metrics data can also be used, as suggested in the work [17]. The main quality characteristic in this metric is how good quality recommendations can be formed using data from certain sources.

4. Conclusion
In general, the analysis of the current state of research shows that the metrics for assessing data and knowledge are well developed, however, they require verification and adaptation for use in assessing spatio-temporal knowledge, due to their specificity arising from the use of complex concepts and relations of space and time. Much attention in research is paid to assessing the quality of spatial knowledge from open sources (Volunteered Geographic Information, VGI), such as OpenStreetMap, Wikidata, and others. At the same time, due attention has not been paid to the assessment of the quality of integration of spatio-temporal knowledge, although there are works on the integration of spatio-temporal knowledge [18].

5. Acknowledgments
The reported study was funded by RFBR, according to the research project No. 20-07-00560 and research project 20-07-00904 (section 3).

References
[1] Goodchild M F 2007 Citizens as sensors: The world of volunteered geography GeoJournal 69 211–21
[2] Haklay M 2010 How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets Environ. Plan. B Plan. Des. 37 682–703
[3] Cidália Costa Fonte, Vyron Antoniou, Lucy Bastin Jacinto Estima, Jamal Jokar Arsanjani J and Carlos Laso Bayas L S and R V 2017 Assessing VGI data quality Mapp. Citiz. Sens. 137–63
[4] Azeroual O, Saake G and Wastl J 2018 Data measurement in research information systems: metrics for the evaluation of data quality Scientometrics 115 1271–90
[5] Wang R and Strong D M 1996 Beyond Accuracy: What Data Quality Means to Data Consumers J. Manag. Inf. Syst. 12 5–33
[6] Liu Q, Feng G, Zhao X and Wang W 2020 Minimizing the data quality problem of information systems: A process-based method Decis. Support Syst. 137 113381
[7] Antoniou V and Skopeliti A 2015 Measures and indicators of vgi quality: An overview ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2 345–51
[8] Dorn H, Törnros T and Zipf A 2015 Quality evaluation of VGI using authoritative data-a comparison with land use data in southern Germany ISPRS Int. J. Geo-Information 4 1657–71
[9] Zielstra D and Zipf A 2010 A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany 13th Agil. Int. Conf. Geogr. Inf. Sci. 1 1–15
[10] Forghani M and Delavar M 2014 A Quality Study of the OpenStreetMap Dataset for Tehran ISPRS Int. J. Geo-Information 3 750–63
[11] Chen X and Lee J 2021 The identification and selection of good quality data using pedigree matrix Smart Innovation, Systems and Technologies vol 200 (Springer Science and Business Media Deutschland GmbH) pp 13–25
[12] Heinrich B, Klier M, Schiller A and Wagner G 2018 Assessing data quality – A probability-based metric for semantic consistency Decis. Support Syst. 110 95–106
[13] Molina C and Prados-Suarez B 2020 Measuring the quality of data in electronic health records aggregators IEEE International Conference on Fuzzy Systems vol 2020-July
[14] Euzenat J and Shvaiko P 2013 Ontology Matching (Berlin, Heidelberg: Springer Berlin Heidelberg)
[15] Zhang Y, Panangadan A and Prasanna V K 2014 UFOM: Unified fuzzy ontology matching Proc. 2014 IEEE 15th Int. Conf. Inf. Reuse Integr. IEEE IRI 2014 787–94
[16] Euzenat J, Roșou M E and Trojahn C 2013 Ontology matching benchmarks: Generation, stability, and discriminability J. Web Semant. 21 30–48
[17] Heinrich B, Hopf M, Lohninger D, Schiller A and Szubartowicz M 2019 Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems Electron. Mark. 1–21
[18] Subramanian A, RR P K, Vikkurthi M and Buttigieg P L 2019 Semantic Harmonisation of Numeric Data from Open Government Data Proceedings of the ACM India Joint International Conference on Data Science and Management of Data - CoDS-COMAD ’19 (New York, New York, USA: ACM Press) pp 238–44