https://doi.org/10.4467/21995923GP.24.007.20474 GEOINFORMATICA POLONICA
23: 2024

Natalia Dziuba[1]
https://orcid.org/0009-0001-0070-1988
Stanisław Szombara[2]
https://orcid.org/0000-0002-0205-7823

GDDKIA XML SERVICE, ROAD TRAFFIC INCIDENTS DATA: AVAILABILITY AND QUALITY RESEARCH

[1]AGH University of Krakow, Kraków, Poland
ndziuba@agh.edu.pl
[2]AGH University of Krakow, Kraków, Poland
szombara@agh.edu.pl

Abstract

With the rapid development of technology and the increasing amount of data being processed, smart cities use real-time information processing to optimise various processes, including navigation and traffic management. In this context, spatial data on traffic incidents plays a key role, enabling drivers to bypass congestion and navigate safely and efficiently through the city. This study focuses on assessing the quality and availability of spatial data on road obstructions provided by the General Directorate for National Roads and Motorways (pl.: Generalna Dyrekcja Dróg Krajowych i Autostrad, GDDKiA) in Poland. In order to analyse this data, a web application was developed to visualise the data on a map to identify problems related to the accuracy of locations and the lack of key information. During the study, it was found that many of the points that were supposed to represent the locations of road incidents were at a significant distance from the actual road axes, indicating errors in geolocalisation. To improve accuracy, additional algorithms were applied that automatically corrected the position of the points to match the nearest roads. Additionally, in cases where geographic coordinates were missing, a geocoding method based on OpenStreetMap data was used. The analysis showed that the data from the General Directorate for National Roads and Motorways is partly incomplete and problematic, which limits its usefulness in further spatial analyses. In addition, the irregular updating of the data at 10-minute intervals and the low number of reported incidents further affect their quality and usefulness. Furthermore, comparison of the visualisation results with the GDDKiA road information map revealed significant discrepancies in the location of obstructions, suggesting that data sources or processing methods may differ. The results suggest the need to improve the accuracy and completeness of the spatial data provided so that it can effectively support cartographic analyses and visualisations.

Keywords: GIS, real-time data, traffic incidents, geoinformation

SERWIS INTERNETOWY XML GDDKIA, DANE O UTRUDNIENIACH NA DRODZE: BADANIE DOSTĘPNOŚCI I JAKOŚCI

Abstrakt

W obliczu dynamicznego rozwoju technologii oraz coraz większej ilości przetwarzanych danych, inteligentne miasta (*smart cities*) wykorzystują przetwarzanie informacji w czasie rzeczywistym do optymalizacji różnych procesów, w tym nawigacji i zarządzania ruchem drogowym. W tym kontekście dane przestrzenne dotyczące utrudnień drogowych odgrywają kluczową rolę, umożliwiając kierowcom omijanie zatorów oraz bezpieczne i efektywne poruszanie się po mieście. Niniejsze badanie

koncentruje się na ocenie jakości i dostępności danych przestrzennych dotyczących utrudnień drogowych, udostępnianych przez Generalną Dyrekcję Dróg Krajowych i Autostrad (GDDKiA) w Polsce. W celu analizy tych danych opracowano aplikację webową, która wizualizuje dane na mapie, co pozwala na identyfikację problemów związanych z dokładnością lokalizacji i brakiem kluczowych informacji. Podczas badania stwierdzono, że wiele punktów, które miały reprezentować miejsca utrudnień drogowych, znajdowało się w znacznej odległości od rzeczywistych osi dróg, co wskazuje na błędy w geolokalizacji. Aby poprawić dokładność, zastosowano dodatkowe algorytmy, które automatycznie korygowały położenie punktów, dopasowując je do najbliższych dróg. W przypadkach, gdy brakowało współrzędnych geograficznych, zastosowano metodę geokodowania opartą na danych z OpenStreetMap, umożliwiając przypisanie utrudnień do najbliższych słupków kilometrowych. Analiza wykazała, że dane z GDDKiA są częściowo niekompletne i problematyczne, co ogranicza ich użyteczność w dalszych analizach przestrzennych. Ponadto nieregularna aktualizacja danych w interwałach 10-minutowych oraz niska liczba raportowanych zdarzeń dodatkowo wpływają na ich jakość i przydatność. Ponadto porównanie wyników wizualizacji z mapą informacji drogowej GDDKiA ujawniło istotne rozbieżności w lokalizacji utrudnień, co sugeruje, że źródła danych lub metody ich przetwarzania mogą się różnić. Wyniki sugerują potrzebę poprawy dokładności i kompletności udostępnianych danych przestrzennych, aby mogły skutecznie wspierać analizy i wizualizacje kartograficzne.

Słowa kluczowe: GIS, dane w czasie rzeczywistym, utrudnienia drogowe, geoinformacja

1. INTRODUCTION

With the rapid technological development, the amount of processed data in every area of life is increasing. Real-time data is very popular, often used, among others, in increasingly popular smart cities [1]. This is data acquired, then processed and analyzed immediately after its receipt in order to obtain an immediate result [2]. This means that it is updated on an ongoing basis, and its acquisition and processing takes place at the moment of its collection in the source database. This data is often used in navigation and routing [3].

Data of this type is divided into two types – static data and dynamic data. The first type is information that is rarely updated. It is easy to process, but at the same time it is less efficient and does not allow for performing complex actions. Dynamic data, on the other hand, is subject to continuous updating and changes at any time. Although access to specific information of this data structure is difficult, it allows for interactive action and advanced analyses [4]. Unlike static data, its essence is the continuous delivery of messages.

There are two ways to download data in real time, i.e. by creating an application that automatically updates streaming data or by creating a button to refresh the application [5]. The first of the methods mentioned does not require the user to interfere with the data update, but the application downloading the data must be constantly in "standby mode" to be able to detect changes in the database it uses. The second method is more convenient if there is no need for continuous data update, but only for the user to refresh it.

The extremely dynamic development of cities and communication systems has a huge impact on the number of unforeseen events on the roads. The constantly increasing number of people in the world results in an increase in the number of cars per capita, and consequently – the demand for publicly available navigation applications, which also feature the provision of information about events on the roads on an ongoing basis. The above-mentioned applications processing data in real time can be considered a type of cartographic visualization. The quality of the visualizations performed depends primarily on the way the data is presented on the map. However, they are also influenced by a number of other factors – one of them is the specificity of the data set itself.

The data on road obstructions that were analyzed in this paper had already been used to create a road information map commissioned by GDDKiA and are made available in the form of a publicly available web application that illustrates the situations on Polish roads. The application allows the user to view the obstructions of interest and display information about them [6]. Road obstructions are the subject of interest for many researchers. Wieczynski et al. [7] described in their work a system for supporting road transport in Europe. One of the services they used is the data described in this article, provided by GDDKiA. TrafficView, a web application visualizing events on roads in the eastern part of the United States, created and made available by the University of Maryland, also worked on a similar principle. Data for the website was provided by states and local road transport departments in real time. This allowed

for dynamic updating of the road situation by automatically refreshing the page at a one-minute interval [8].

This study focused on the analysis of the quality and availability of spatial data on road obstructions in Poland, which are made available by the General Directorate for National Roads and Motorways (pl: Generalna Dyrekcja Dróg Krajowych i Autostrad, GDDKiA)[1]. The analysis conducted allowed for the assessment of the usefulness of the data set and the possibility of using this data in cartographic studies

2. SERVICE OF INCIDENTS ON NATIONAL ROADS GDDKiA

The General Directorate for National Roads and Motorways (GDDKiA) provides data on current events on national roads, expressways and motorways throughout Poland [9]. The data was entered into the government's public data opening program on September 20, 2016 [10]. According to the data.gov.pl website [11], the data are available under a CC0 1.0 license.

The research for this article was conducted between December 2022 and February 2023. The data is made public by the institution in the form of an Application Programming Interface (API), in the XML format (Extensible Markup Language, Fig. 1), which is probably the most popular universal language for representing data recommended by the W3C (World Wide Web Consortium) [12].

The service provides information on incidents on roads such as: type of incident, number and kilometre of the road on which it occurred, length of the section on which the incidents occur and a number of other information describing the event. Objects also have assigned geographic coordinates in the form of geo_lat and geo_long attributes compliant with W3C standards [9]. Thanks to the coordinates, the data on events could be classified as spatial data and could be used to visualize them in the form of a map application, but individual objects in the website did not have coordinates.

The XML file is generated asynchronously at 10-minute intervals. However, during the period of work on the author's engineering project, which is the basis of this study, in December 2022, the file update time did not fit into equal time intervals. These were very irregular time intervals ranging from several to a dozen or so minutes. The file generation time is displayed in the header of the XML file. GDDKiA does not provide an XSD (XML Schema Definition) schema for the tested data file (the template published on the website [14] is not a template for the tested XML file). This prevents validation, which is necessary to check the correctness of the XML file. Only the service specification is available, which describes the individual data elements [9].

The data provided by the GDDKiA in the form of an XML service is useful in terms of using it to automatically feed transportation systems such as EPLOS (European Logistics Services Portal) [15].

3. METHODOLOGY

In order to examine the availability and quality of spatial data provided by the General Directorate for National Roads and Motorways, a simple web application was created that allows for the visualization of incidents on roads in Poland. The basis for this research is the author's engineering thesis, which describes in more detail the entire process of creating the application. It was developed in the Python programming language using the Leaflet library. Creating a map showing road traffic incidents in real time allowed for the visualization of the location of each point.

The source data used by the program contained major irregularities. The geographic coordinates of the visualized points were not located on the axes, or even on the edges, of national roads and motorways. The markers were placed on the map at a distance of several to several hundred meters from the roads, which is shown in the figure below (Fig. 2). Considering the fact that these were road disruptions, i.e. they should occur on the roads, an attempt was made to "drag" the markers to the axes of national roads or motorways during the creation of the web application. The road layer came from the General Geographic Objects Database (pl.: Baza Danych Obiektów Ogólnogeograficznych, BDOO) [16]. A function was written that moved the point symbolizing the obstruction so that it was on the road axis. The searched location was calculated as a function of the projection of the point onto a straight line, which was the nearest section of the road.

[1] This article is based on the engineering thesis of the author of the article Natalia Dziuba entitled "Application of geocoding services for real-time visualization of spatial data" written under the supervision of Stanisław Szombara defended 18.01.2023 at the Faculty of Geo-Data Science, Geodesy, and Environmental Engineering, AGH University of Krakow

Fig. 1. Fragment of an XML file containing data on road events (source: [13])
Ryc. 1. Fragment pliku XML zawierającego dane o utrudnieniach drogowych (źródło: [13])

The set of source data on incidents on roads in Poland is not fully complete, because individual points do not have latitude or longitude, and therefore it is not possible to visualize them on a map. The solution to this problem was to use geocoding coordinates by querying databases containing data on milestones and assigning the coordinates of the nearest milestone to road incidents.

The initial solution to the problem was to use the Universal Geocoding Service (pl.: Uniwersalna Usługa Geokodowania, UUG), which is provided by the Head Office of Geodesy and Cartography (pl.: Główny Urząd Geodezji i Kartografii, GUGiK) in the form of an application programming interface (API). The GetRoadMarker query allows you to obtain the coordinates of

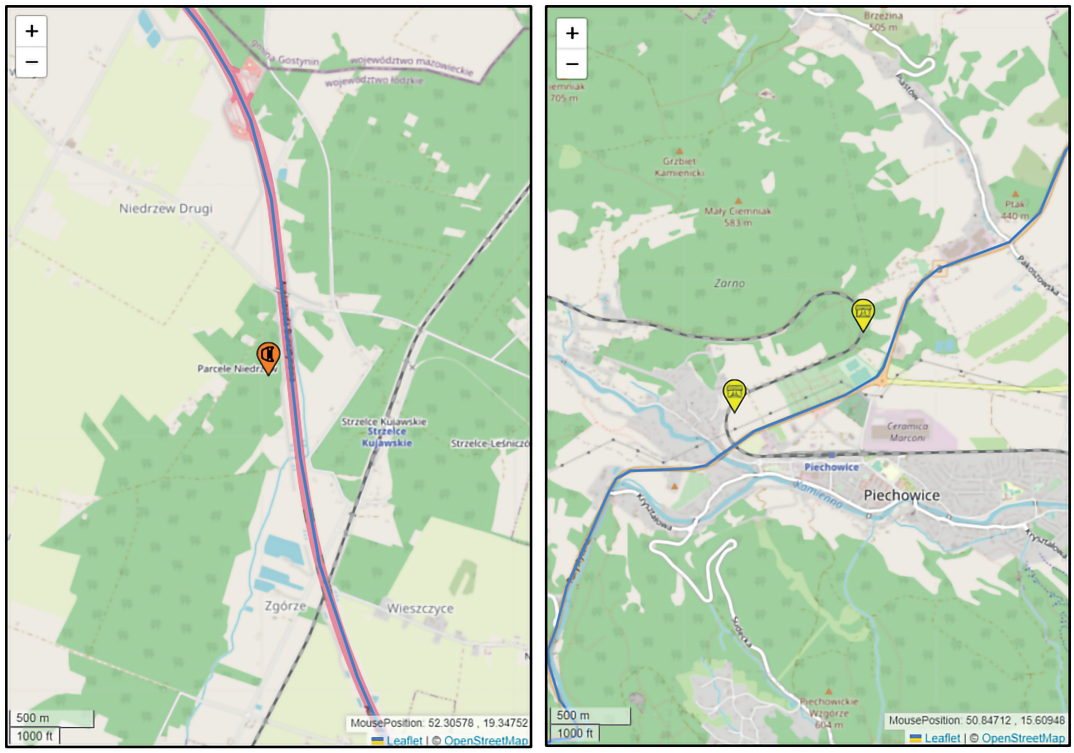

Fig. 2. Screen views of the showing an accident and roadworks outside the road area (source: own study)
Ryc. 2. Widoki ekranu aplikacji przedstawiające wypadek oraz roboty drogowe poza obszarem dróg (źródło: opracowanie własne)

the kilometer marker you are looking for after specifying the location parameter, which takes the value of the road number and the kilometer of the road separated by a space [17]. The road kilometer data provided by GUGiK through the Universal Geocoding Service was based on data from the OpenStreetMap [18], but the services are not fully compatible, because some kilometer markers appearing in the OpenStreetMap service do not appear in UUG. Therefore, the possibility of geocoding was used by directly querying the OpenStreetMap database (Fig. 3).

The web application created in this way was used as a tool for analyzing the availability and quality of the data set being studied. The focus was primarily on problems occurring during the application development process. It should be noted that the problems associated with the possibility of visualizing data affect their quality and availability. Additionally, raw source data ("not connected" to roads) were compared with data after "connecting" in the described manner and with events visualized on the GDDKiA road information map. The location of selected points on the map was compared.

The change in the daily availability of spatial data during one week (09–15.02.2023) in relation to individual days was also examined. The number of events that

Fig. 3. Diagram of the geocoding process (source: own study)
Ryc. 3. Schemat procesu geokodowania (źródło: opracowanie własne)

had coordinates, correctly geocoded, and those that, despite an attempt at geocoding, could not be placed on the map was analyzed. This allowed for the depiction of the approximate number of events made available daily on the service and the percentage of objects without coordinates.

4. RESULTS

Visualization of data provided via the XML service of the General Directorate for National Roads and Motorways allowed for a visual assessment of the usefulness of the data. Obstacles originating directly from the source file were located at distances of up to several hundred meters from the road axes (Fig. 4a). After carrying out a process that corrected the location of points so that they were located on the road axes, these obstructions were located in a way that was more close to the correct one (Fig. 4b). However, it is not possible to fully determine which location of obstructions is correct, because on the GDDKiA road information map (Fig. 4c) the same points were located on other road sections (slightly distant from those in the author's visualization). An analogous situation occurs for all road obstructions originating from the analyzed data file.

Visualization of individual points from the XML file was not possible due to the lack of coordinates, which prevented their direct placement on the map. In the case of the author's application, this problem was partially solved by geocoding and the events were marked on the map area in places corresponding to the milestones. However, if the milestone did not appear in the OSM database, the event remained without coordinates. The issue of these points (without coordinates) is different in the case of the map prepared at the request of GDDKiA – in this case, all points that did not have coordinates in the source file were marked on the map in an unidentified way. Probably based on the road number and kilometer, but not the milestone itself.

In the period under review, the average number of events with coordinates that allowed for direct placement of points on the map (excluding the fact of shifting coordinates) was 189, or 95.5% of all road incidents occurring in this period (Fig. 5a). 3% were objects for which it was possible to correctly geocode using the OpenStreetMap service. The remaining part were points that could not be correctly geocoded, which resulted from the lack of individual road milestones in the OSM database. Such points could not be plotted on the map in the author's application. The daily numbers of points were averaged and presented on the cumula-

Fig. 4. Comparison of the location of selected events: a) without improving the location of points, b) after improving the location of points, c) on the GDDKiA road information map (source: own study and [6])
Ryc. 4. Porównanie umiejscowienia wybranych utrudnień: a) bez poprawy położenia punktów, b) po poprawie położenia punktów, c) na mapie informacji drogowej GDDKiA (źródło: opracowanie własne i [6])

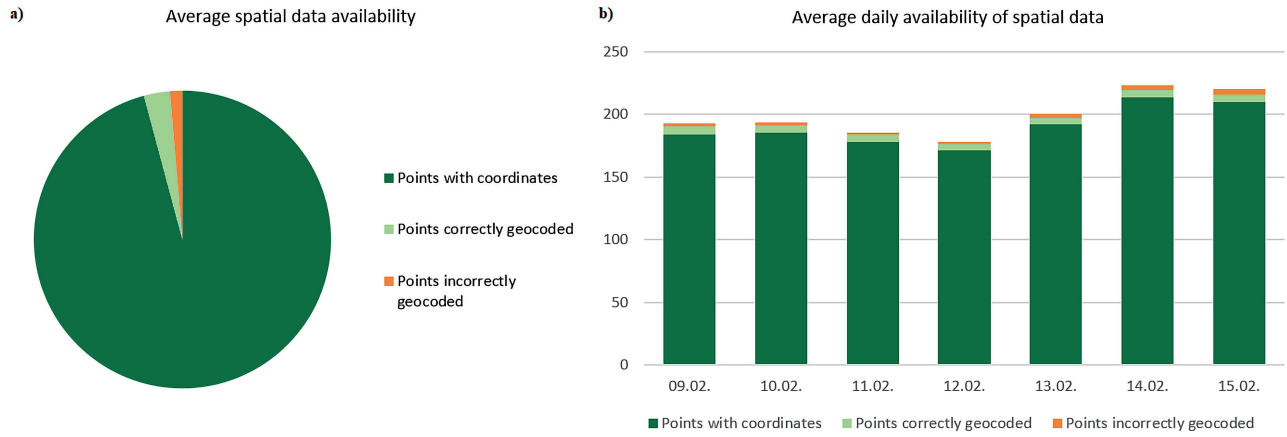

Fig. 5. Average availability of data on road incidents interpreted as the number of incidents with coordinates and subject to geocoding: a) for the entire week of 09–15.02.2023, b) on individual days (source: own study)
Ryc. 5. Średnia dostępność danych o utrudnieniach drogowych interpretowana jako liczba utrudnień ze współrzędnymi i podlegających geokodowaniu: a) dla okresu całego tygodnia 09–15.02.2023, b) w poszczególnych dniach (źródło: opracowanie własne)

tive graph below (Fig. 5b). On no day of the period under review were there any significant deviations in the number of data provided and their availability. The average number of all events throughout the period fluctuated around 200, and objects without coordinates constituted a small fraction of this number, which does not change the fact that they had an impact on the quality of the visualization.

5. SUMMARY

The initial assumption of the developed application which was the basis for this research, was only a simple visualization of data in real time. However, during the work it turned out that there were problems with the data that had to be solved, the source of which could not be clearly identified. The coordinates of the points that were to be placed on the map were not correct, i.e. the markers that were created in the original version of the program were not located on the axes (and generally outside the area) of the roads. Due to the fact that they were road incients, theoretically they should have been located on the roadway or on the axis of the communication route. What is more, these were not small errors, and the pickets were located in extreme cases about several hundred meters from the road.

As a result of the occurring irregularities, it was necessary to carry out a number of additional actions in order to fix the errors from the source data. Functions were implemented in the code that allowed for "dragging" points to roads. Several iteration methods of dragging points to roads were tested (including with the division of the country into provinces and without division). Unfortunately, none of the methods allowed for a complete solution to the problem. There were still points on the map that were not perfect (according to the source data, they were located outside the borders of Poland).

Some incidents did not have coordinates that would allow for their direct location. Geocoding was used to solve this problem. Initially, the Universal Geocoding Service (UUG) provided by GUGiK was to be used for this purpose, but in this case, the national database also failed. It was not possible to implement the geocoding function in the entire program code, due to the lack of a very large number of milestones in the GUGiK database. The problem was solved by performing a direct geocoding query to the OpenStreetMap database. An interesting fact here is that UUG data was based on data from OpenStreetMap. Geocoding from the OSM database was performed using the Nominatim tool.

The problem of geocoding road-related events has been noted in the literature for many years [19]. As early as 2014, Cichocinski included the multiplicity of ways to record geographic information among the main problems. The paper dealt with all kinds of road events, but the problems demonstrated in this paper (different ways of specifying the location in an XML service) seem to coincide with Cichocinski's work despite the passage of years. The actuality of the background data,

i.e. OSM, also continues to impinge on the quality of the final result.

The author's web application described in the work is very similar to the application provided by GDDKiA, because both applications are based on the same XML file. After analyzing this file, it can be assumed that there were discrepancies between the original data and the data provided by GDDKiA, the causes of which could not be identified. Despite the obstacles, it is possible to use the provided data after appropriate correction. The application of the General Directorate for National Roads and Motorways is a closed program that does not allow for the secondary use of data, so there is no option of their reuse or integration with other databases.

According to the definition of The Directive 2007/2/EC of the European Parliament and of the Council adopted on 14 March 2007 aims at establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), spatial data are data that can be identified with a given location or geographical area directly or indirectly [20]. Therefore, it can be considered true that if an object has specific coordinates or at least address data that allow it to be placed in space, it belongs to a spatial data set. On the other hand, visualization of spatial data consists in their transmission in such a way that they are fully legible and understandable to the recipient. Only then can they serve as a source for further spatial analyses [21]. In the case of the analyzed data, shifted coordinates, as well as their individual gaps, significantly hinder the performance of any analyses. Poorly located points subject to analysis can lead to erroneous results.

It is also worth noting the surprisingly low number of events on Polish national roads and motorways. This number oscillates around an average of 200 to 300 events. In this case, it was also impossible to clearly identify the problem. This significantly affects the usefulness of the analyzed data, because incomplete information about road incidents is of little use to road users. In addition, the service update in 10-minute time windows means that the data can only be used in a way that is close to reality.

The XML file is provided by the General Directorate for National Roads and Motorways and is publicly available, which may suggest that it is easily accessible and possible to use in further visualizations or applications via API. However, considering the issue related to the problem of object coordinates in the provided data set, as well as their local lack, one can discuss the availability and quality of this data. For this reason, their usefulness for further spatial analysis or cartographic visualizations may be limited, because it is impossible to directly use the source data in a correct way. In order to solve the problem related to the availability and quality of the analyzed data set, its source should first be clearly identified. This would allow for proposing a specific solution. Assuming that the coordinate errors result from the inaccuracy of the recorded events, the accuracy of their recording should be improved.

FOUNDING

This paper was made at AGH-UST within statutory research 16.16.150.545

REFERENCES

1. Pourghebleh, B., Hekmati, N., Davoudnia, Z., Sadeghi, M. A roadmap towards energy-efficient data fusion methods in the Internet of Things. Concurr Comput (https://doi.org/10.1002/CPE.6959), 2022; volume 34.
2. Tejada, Z. Real-time processing https://learn.microsoft.com/en-us/azure/architecture/data-guide/big-data/real-time-processing#architecture (accessed November 6, 2022).
3. Wade, T., Sommer, S. A to Z GIS: An Illustrated Dictionary of Geographic Information Systems. vol. Esri Press. 2006.
4. Introduction to Real-time Data. Memgraph 2021. https://memgraph.com/blog/real-time-data (accessed November 6, 2022).
5. Yang, S. Real-time dashboard in Python. Streaming and Refreshing . Towards Data Science 2021.
6. Mapa informacji drogowej https://drogi.gddkia.gov.pl/mapy/mapa-informacji-drogowej (accessed November 7, 2022).
7. Wieczyński, A, Perski, A, Baczyńska, M. Innowacyjny projekt satelitarny dla podniesienia bezpieczeństwa i komfortu podróży na europejskich korytarzach transportowych. Pomiary Automatyka Robotyka. 2012; volume 2/2012, pp. 101––105.
8. TrafficView Help https://trafficview.org/introduction/ (accessed January 4, 2023).
9. Specyfikacja plików XML zawierających dane udostępniane przez GDDKiA. 2014.
10. Uchwała nr 107/2016 Rady Ministrów z dnia 20 września 2016 r. w sprawie ustanowienia „Programu otwierania danych publicznych." Warszawa: 2016.
11. Utrudnienia na drogach – Otwarte Dane 2022. https://dane.gov.pl/pl/dataset/2752/resource/38739/table (accessed September 8, 2024).
12. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E. Extensible Markup Language (XML) 1.0 (Fifth Edition) 2008.
13. Utrudnienia na drogach krajowych https://www.archiwum.gddkia.gov.pl/dane/zima_html/utrdane.xml (accessed December 23, 2022).

14. Generalna Dyrekcja Dróg Krajowych i Autostrad. Dane xml i xsd https://www.gov.pl/web/gddkia/dane-xml (accessed August 4, 2024).
15. Wasiak, M., Szczepański, E., Jachimowski, R., Gołębiowski, P., Leleń, P. Structure of databases of the EPLOS system on road transportation and railway infrastructure. WUT Journal of Transportation Engineering (https://doi.org/10.5604/01.3001.0013.6582), 2019; volume 125, pp. 192–204.
16. Baza Danych Obiektów Ogólnogeograficznych https://www.geoportal.gov.pl/dane/baza-danych-ogolnogeograficznych-bdo (accessed December 25, 2022).
17. Uniwersalna Usługa Geokodowania https://services.gugik.gov.pl/uug/ (accessed December 26, 2022).
18. Uniwersalna Usługa Geokodowania – Aktualności https://www.geoportal.gov.pl/o-geoportalu/aktualnosci/-/asset_publisher/HCHq0YGNRszn/content/25-10-2019-uniwersalna-usluga-geokodowania?inheritRedirect=true (accessed December 26, 2022).
19. Cichociński, P. Problematyka geokodowania zdarzeń drogowych (Problems of road accidents geocoding). Rocznik Geomatyki, 2014; volume 2(64), pp. 205–216.
20. Dyrektywa 2007/2/WE Parlamentu Europejskiego i Rady z dnia 14 marca 2007 r. ustanawiająca infrastrukturę informacji przestrzennej we Wspólnocie Europejskiej (INSPIRE). 2007.
21. Suchecka, J. Statystyka przestrzenna. Metody analizy struktur przestrzennych. Wydawnictwo C.H.Beck, 2014; volume 1.