Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 72 (2023) 2237–2244

Transportation Research Procedia

www.elsevier.com/locate/procedia

Transport Research Arena (TRA) Conference

Making it easy for transport stakeholders to share mobility data

Thierry Chevallier[a1], Johannes Lauer[b], Chiara Renso[c], Alberto Blanco-Justicia[d], Didier De Ryck[e], Alexandros Papacharalampous[f]

[a]AKKA Research, Blagnac, France
[2] HERE Technologies, Schwalbach am Taunus, Germany, 3 Consiglio Nazionale delle Ricerche, Pisa, Italy, 4 Universitat Rovira i Virgili, Tarragona, Spain, 5 KISIO Digital, Paris, France, 6 AETHON Engineering, Athens, Greece

Abstract

With the emergence of new mobility services, an increasing amount of data is being produced. However, while it is recognized that data sharing can open up new opportunities and lead to more efficient processes and new products, there is still a lot of reluctance to share data. The EU-funded MobiDataLab project works to remove these limitations and to foster the sharing of data amongst transport authorities, operators and other mobility stakeholders. According to the FAIR principles (findable, accessible, interoperable and reusable), the MobiDataLab project provides a "transport cloud", that is an infrastructure to build new solutions with mobility data and services. With a close contribution between a reference group, the project team and contributors of virtual and living labs, the project will identify current challenges and work with the relevant interest groups on solutions.

© 2023 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the Transport Research Arena (TRA) Conference

Keywords: mobility data sharing; transport cloud; data privacy; data enrichment; journey planning; virtual and living labs.

1. Introduction - Challenges in mobility data sharing

The increase in actors producing data in the mobility sector, and the resulting growing number of data and diversity of data formats, is making it difficult to share and reuse data. Everyone agrees that data sharing can unlock new insights and lead to more efficient processes and new products, but there is still a lot of reluctance and issues to solve in order to achieve this. The MobiDataLab project, funded by the EU Horizon 2020 program, aims to foster data sharing in the transport sector. It provides mobility organizing authorities with recommendations on how to improve the value of their data, contribute to the development of open tools in the cloud, and organize hackathons to find innovative solutions to concrete mobility problems. For doing so, MobiDataLab builds an EU-wide cloud-based prototype platform for sharing transport data comprising different data sets provided by project partners, through open data portals and from other sources (public transport and road transport data in selected areas in addition to weather,

[1]* Corresponding author. Tel.: +33 (0) 6 6501 0274;
E-mail address: thierry.chevallier@akka.eu

2352-1465 © 2023 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the Transport Research Arena (TRA) Conference
10.1016/j.trpro.2023.11.711

pollution and other open data). In addition, the project creates a community to provide the necessary open innovation ecosystem for an effective improvement in the culture of data sharing.

2. The MobiDataLab Transport Cloud

2.1. Architecture

The MobiDataLab Transport Cloud platform aims to facilitate access to mobility data in an open, interoperable, transnational, and privacy-preserving way. The platform also aims to adopt the FAIR principles Wilkinson et al (2016) to the data access, i.e., mobility data available in a vast and diverse ecosystem that possibly encompasses many different sources should be findable, accessible, interoperable, and reusable. The vision behind these goals stems from the needs and interests of the stakeholders behind the MobiDataLab project. Indeed, these are public and private institutions that have an interest to either act as mobility data providers (for instance, provide real-time public transportation data, road-network data, vehicle data), or are interested in consuming mobility data through the access and services provided by the Transport Cloud platform. Several EU projects and initiatives face the data sharing problem and several approaches and solutions have been proposed. One prominent example is GAIA-X (GAIA-X, no date) which is still ongoing. After evaluation (Report on enabling technologies for Transport Cloud | MobiDataLab consortium, 2021), none of these generic solutions proved mature enough to be adopted for MobiDataLab without compatibility-related risks. A pragmatic approach was therefore adopted for the design of the platform, striving to meet a comprehensive list of previously identified functional and non-functional requirements.

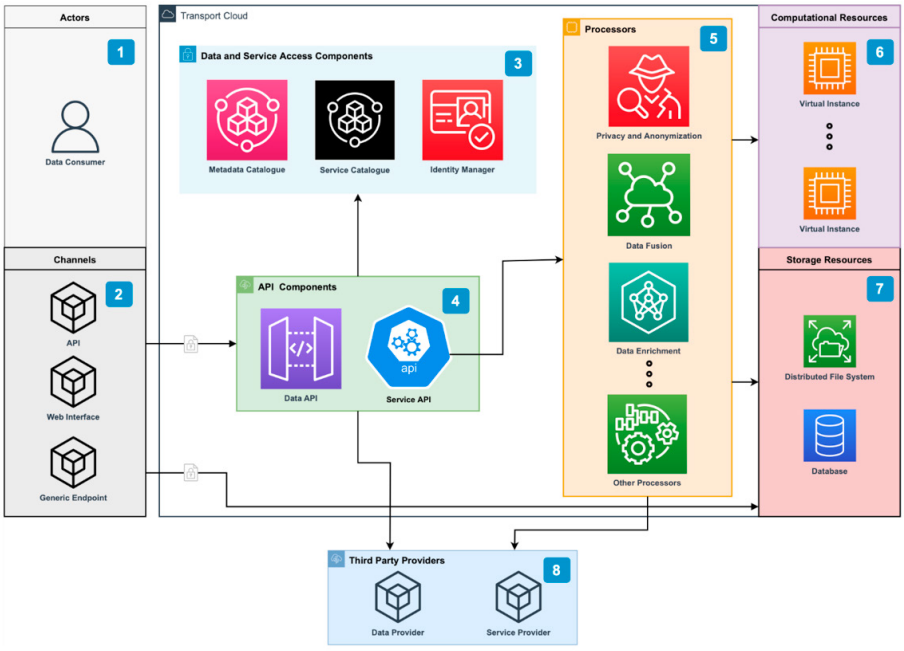


Fig. 1. MobiDataLab Transport Cloud architectural design.

Non-functional requirements include cloud federation, governance, etc. while functional requirements are use-cases based (see below). The elicited architecture covers them by using established standards and components, ensuring the maintainability and extensibility of the resulting platform, which is a collection of components

performing key operations (see Fig. 1). The aim is to promote mobility data sharing between data consumers and data providers:

- Data consumers can interact with the platform through: (1) API endpoints, (2) web interface endpoints and (3) generic endpoints – for instance, a SPARQL endpoint. They authenticate themselves via the identity manager and submit their requests via the API which in turn process the requests by querying the metadata and service catalogues (described below) to find out the appropriate data sources and services.
- Data providers represent the information entry point of the platform. Information retrieved from third-party providers can either be imported or accessed on the fly using the data and service APIs component.
- Processors encompass the Data Privacy & Anonymization, Data/Service Harmonization & Standardization, Data Processing, and Data Fusion and Enrichment. In generic terms, we define a processor as a component that models some function that inputs some data and produces an output according to a well-defined logic.

2.2. Metadata catalogue

An integrated and largely seamless data flow is a mandatory requirement not just for a better user experience. It is also contributing if not mandatory for machine-based processes and automation. Standardized data sets with self-speaking metadata and catalogues that can harvest, structure and list available data sets are supporting the accessibility and findability of data. It is comparable to a library, where one is using a data base to look for the books one wants to read. In MobiDataLab data sets from different domains are being used, where each domain has its own perspective, purposes and standards. Standards are not necessarily compatible to each other, which limits cross-domain interoperability. Due to different purpose and use cases for data, standards are mainly focusing their origin purpose (e.g. GTFS (General Transit Feed Specification, no date) for transportation data and OGC WFS (Web Feature Service | OGC, no date) for spatial vector data) and not primarily with the requirement of interoperability. Accessing the data and consuming the data is only one part of the chain. Finding required data is another important part to realize a seamless user experience. For this purpose, the data is getting "reduced" to its metadata, which enables a more efficient search on the key aspects of each data set via catalogues. The concept of data catalogues is commonly used especially in the open data ecosystem where several catalogue software systems are established. Within the MobiDataLab, a set of commonly used catalogues are being reviewed on their functionality (e.g. API, user interface), interoperability capabilities (e.g. supported meta data standards) and will be integrated into the Transport Cloud. Table 1 lists a set of commonly used meta data catalogues.

Table 1: Selection of data catalogues

Name	Link	Purpose
CKAN	https://ckan.org/	Metadata management system for data hubs led by the Open Knowledge Foundation (*Home | Open Knowledge Foundation*, no date). CKAN is an open-source solution with an active community. Many transport authorities use it as their open data portal.
GeoNetwork	https://geonetwork-opensource.org/	Reference implementation for geospatial data, harvesting options, network-based system, only stores metadata
GeoNode	https://geonode.org/	Focus on geospatial data, combination of geospatial software components, option to store data
OpenDataSoft	https://www.opendatasoft.com/	Used for many open data formats, option to store data and metadata.
Socrata	https://dev.socrata.com/	Popular open data solution in Northern America
STAC	https://stacspec.org/	Focus on geospatial raster data

The heterogeneous set of metadata catalogues are originating from different use cases and each of them providing specific capabilities for data search and managing. Supporting a broader set of standards for metadata is therefore key factor to enable interoperability between the catalogues and data users. This will help to follow that paradigm of FAIR (cf. (Wilkinson et al., 2016)).

2.3. Service catalogue

From the various data standards via a narrower set of metadata standards and tools, the interfaces for services are more heterogeneous and underlies less standardization. While interoperability is mainly driven by political initiatives s.a. INSPIRE (Eu, 2007) or (Geospatial Data Strategy - United States Department of State, no date), services and service interfaces are part of product portfolios and are underlying business decisions from companies. However, further initiatives about service standardizations are driven by the OGC (OpenGeospatial Consortium). The OpenLS specifications on location based services (Mabrouk et al., 2005) are being reviewed and initiatives, such as the OGC routing pilot project (Routing Pilot | OGC, no date) comparing and contributing to the alignment and interoperability of routing services, as part of the location based services portfolio. Besides geospatial related services, further services from the mobility sector need to be taken into account. This is where the MobiDataLab project wants to contribute further to review existing service interfaces and trigger discussions within the involvement of service providers and users to increase the interoperability and the usage of service functionalities.

2.4. Privacy and anonymisation

It is important for the proposed Transport Cloud to find the optimal balance between the encouragement of data sharing and complete regulatory compliance. In this respect, anonymization mechanisms will be developed and deployed, so that personal mobility data can still be collected, analyzed, shared, and released with privacy guarantees, focusing on performance, resistance to privacy attacks, and transparency of the protection techniques.

Mobility data in its simpler form are data about individuals that include their locations at specific times. Sources of real-time raw individual location data include, but are not limited to, cell towers, Wi-Fi access points, RFID tag readers, location-based services, or credit card payments. Historical location data, in form of data sets in which each of the records corresponds to an individual and includes her location data for some time periods are referred to as trajectory microdata sets. Such trajectory microdata sets are often of interest to transport authorities, operators, and other stakeholders to evaluate and improve their services, the state of the traffic, etc., and thus are often publicly released or shared. Sharing of mobility data is occasionally shared as aggregates (e.g., heat maps) instead of at an individual level. Whichever the form of these mobility data, they all share some statistical characteristics that make their sharing a potential privacy risk. Mobility data are highly unique and regular. Unicity refers to the data of different individuals being easily differentiable, particularly at some specific locations. The starting and ending locations of users' trajectories are often their home and work locations which, again are highly unique and can lead to reidentification. Studies show that user full trajectories can be uniquely recovered with the knowledge of only two locations, and knowledge of 4 locations can lead to full reidentification of 95% of users in a 1.5 million user data set, as shown by Monjoye et al. The regularity of trajectories means that for single individuals, their data follows periodic patterns. Namely, individuals tend to follow the same trajectories during workdays—home to work and back to home. The statistical particularities of personal mobility data, along with their special semantics (locations can both be quasi-identifiers and confidential information) make them vulnerable to a series of attacks that include, depending on how the data are shared or released, record linkage, re-identification, attribute inference, probabilistic attacks, and membership inference.

The Transport Cloud will include anonymization mechanisms for mobility data. Anonymization mechanisms can be classified as utility-first methods, in which data is modified iteratively until a privacy requirement (e.g. risk of reidentification) is fulfilled, and privacy-first methods, where a privacy model (e.g. k-anonymity, differential privacy) is chosen and parameters for masking methods are derived from the model. Namely, the Transport Cloud provides the SwapMob anonymization mechanism (Salas et al., 2018), which works on real-time location data with a trusted third party acting as a proxy (the service provider can act as the trusted party before sharing the data to other entities) and also on trajectory microdata sets. Additionally, we provide the SwapLocations mechanism which enforces k-anonymity on data sets of historical data (Domingo-Ferrer and Trujillo-Rasua, 2012). Finally, the Transport Cloud

will include mechanisms for synthetic trajectory data generation, leveraging on differential privacy to generate trajectories not linkable to the original sets.

2.5. Data enrichment

The MobiDataLab Transport Cloud includes a toolbox for data set enrichment and quality raising, containing so-called processors. Processors add further value to the platform, as they give the ability to create novel data and services.

- Processors for semantic enrichment exploits the capacity of linking semantically related data of the Resource Description Framework formalism (RDF) and, in general, of Linked Open Data (LOD) to create new knowledge that can be queried through the Transport Cloud. Here the explicit link from mobility data to contextual and related semantic information is performed based on common vocabularies. The task of semantically enrich mobility data consists in four phases: (1) preprocess mobility data to clean and remove noise; (2) segment mobility data into meaning segments to enrich (e.g. stops and moves) and (3) select one or more enriching data sources (4) create a link between segments and entities in the enriching data sources (e.g. based on the spatial or temporal distance or by semantic similarity).
- Processors for geographical enrichment (based on a common geometry) include geocoders to convert a human readable address into geographical coordinates, geo converters to simplify, convert or normalize geographical data (e.g. projections), spatial functions to run computations (e.g. distance computation) on them. In addition, MobiDataLab geographical processors will include connectors to open geodata popular APIs (e.g. OpenStreetMap Overpass API) allowing to generate new data out of original transport data sets.

3. Use-Cases

The MobiDataLab project is investigating data, services, standards, and interoperability by exploring typical mobility use cases. Within a “Version 1” of the project use-cases, the project team designed a wide set of example use-cases, that demonstrates the availability of data and services and help to trigger discussions, further investigations, and initiatives on standardization. The creation of the use-cases followed a data-driven approach and close collaboration with several stakeholders. Specifically, MobiDataLab executed a survey for determining the topics of relevance to stakeholders for data sharing. A total of ten topics were evaluated and the participants were requested to rank them based on how interesting they are to be solved in a data sharing environment. The survey carried out was a 3-round Delphi survey (Hasson et al., 2000) with 50, 29 and 20 per round. The interest of the participants was ranked on a Likert scale and the results can be found in the table 2 below. The results showed that topics 5,6 and 10 were voted as the most interesting (according to a weighted average). A further analysis on the actors involved in each topic (actor here means a data provider or consumer) showed that the highest voted topics contained the most actors compared to the rest. The following subchapters are describing a subset of the version one use cases. The second part of the project further involves a reference group of relevant mobility stakeholders (i.e. local/regional authorities, transport authorities, transport associations/clusters, transport operators and others), established by MobiDataLab. This reference group is designed to help co-create, explore, experiment, and evaluate the Transport Cloud. With their practical insights, their data provisioning and the real-world challenges, the “Version 2” set of use-cases will be defined. This process involved an extensive stakeholder engagement process initiated before the start of the project.

Table 2: Results of the Delphi survey

#	Topics for data sharing	Weighted Average
1	Transport Planning Activities for multimodal systems	4
2	Daily commuting congestion and low emission zones management	4
3	Real-time environmental data monitoring for Green Cities and Green Logistics	3.96
4	Machine Learning and Artificial Intelligence for better operations	3.78
5	Better journey planning through 3rd-party data integration	4.04
6	Decision support through data sharing	4.26
7	Transport planning activities to improve area accessibility	3.72
8	Facilitating connections for critical infrastructure and emergency vehicles	3.58
9	Georeferenced and geo-represented (better maps) data to support planning and operational activities	3.68
10	Real-time data sharing across modes for better operations	4.26

Through co-creation activities such as the Delphi and a set of regular workshops and meetings, the project partners and stakeholders devised the use-cases which provided requirements for the project and the basis for the challenges that will be presented to living labs' participants. Following challenge creation from use-cases, data users, hackers and coders invited in the Living Labs will explore data and services around the Transport Cloud to build creative mobility solutions by using shared mobility data.

3.1. Example Use case: Journey planners

When dealing with mobility data, one of the first use cases that usually comes to mind is journey planning (or trip planning), which we could define as the capability to determine one or several ways to travel between two or more locations, by combining transport data and geospatial data. There is a large panel of sub use cases associated to journey planning. It is expected that a journey planner should propose the following capabilities:

- multi-modality: ability to combine different transport modes, such as public transports (train, bus, tram, subway), personal transports (car, bike), walking, free floating or ride sharing
- multi-criteria: ability to propose different journeys depending on several criteria (time to destination, number of changes, accessibility, preferred transport mode, cost …)
- real-time: ability to take into account real time information (car traffic, network disruptions, equipment availability …) instead of theoretical information to propose optimal journeys

The Transport Cloud will integrate Navitia, an open-source passenger information platform that includes journey planning capabilities, exposes Navitia API and propose a way to integrate data sets. This way, instead of having to deal with raw data sets, users of the platform will be able to leverage built-in journey planning services for their own services and use cases. The Transport Cloud also makes use of journey planning historical data to build supplementary services, e.g. predictive models or recommendations based on past journey planning requests.

3.2. Example Use case: Urban mobility and tourism

In this case study we want to exploit the services offered by the Transport Cloud to enrich mobility data with semantic information provided by Linked Open Data (by the semantic enrichment processor) and use the resulting enriched data sets to perform analyses that can be then exploited by a Tourist Service. We assume that mobility data that need to be enriched can fall within two categories, i.e., (1) tracks generated by the use of position-enabled devices which collect the movements of tourists visiting a city, or (2) segments (i.e., sequence of geo-located points satisfying some criteria) suggested by some journey planner service. The goal to be achieved in this use case is to allow a tourist service, via the services offered by Transport Cloud, to offer personalized services that enrich and improve the tourists'

visiting experience by leveraging the mobility data that is being tracked and collected. In the context of MobiDataLab, this knowledge will be available and linked through the Transport Cloud, while the actual semantic enrichment will be executed by the data processors. Inspired by (Ruback et al., 2016) this use case study wants to describe a semantic enrichment process performed by the Transport Cloud, where data made available by tourism operators, transport data providers, and Linked Open Data providers, is exploited for this purpose. Also, Linked Open Data is employed to enrich mobility data thanks to the use of data processors (more specifically, semantic and geographic enrichment).

3.3. Example Use case: Inclusive and sustainable Mobility

The first use case identified for the MobiDataLab geospatial enrichment processor is the inclusiveness of transport, and in particular its accessibility to people with reduced mobility. Here the project proposes to assess, in partnership with one or more local authorities in the MobiDataLab reference group, if local public transport is usable by people with reduced mobility. In order to map accessibility data, we build on the OpenStreetMap ecosystem, which is a pillar of the data sharing culture in the geospatial world and beyond. Train stations and bus stops in the local area of partner municipalities will be assessed for their accessibility to people with disabilities, hence a contribution of MobiDataLab to the OpenStreetMap project.

The geospatial enrichment module can also be used in the environmental context and the EU Green Deal in particular. This use case aims at exposing a combination of data provided by the public (transport) authorities of the Reference Group with local environmental data (air quality, atmospheric conditions, weather, etc.) following the Geographical Information Systems formats and exchange standards (OGC, INSPIRE (Eu, 2007), GeoJSON (rfc7946, no date), open APIs, etc.). These data can be static (e.g. low emission zones), real-time (e.g. road traffic and and/or historical. The result of this analysis could be a geographical representation of the environmental impact of road traffic in a given territorial context.

4. Virtual and Living Labs

The Transport Cloud is a holistic tool for storing, transforming and exchanging data. Nonetheless, making data available is a requirement not only for the MobiDataLab project, but for the scientific, technical, regional and EU community. Besides data, other aspects relating to generating insights and services from said data are requirements stemming from stakeholders that dictate the “what” and “why” of an innovation. Those requirements should come directly from the people, the local/regional government, and other stakeholders. To that end, a Virtual Lab is created in the project to support and facilitate the collaboration of those organisations and individuals. The Virtual Lab is a digital version of a Living Lab and contains various functionalities that support discussion and promotion of solutions throughout their inception life cycle, from challenge to idea to a prototype. In simple terms, a stakeholder will be provided with tools to post challenges, innovators can propose and build solutions for the challenges, and users/innovators/stakeholders can provide suggestions, opinions and requirements through functionalities such as forums, comments or polls.

Facilitating the above process will require a tool that resembles a social network but that has a specific context and functionalities related to challenge solving. While social networks restrict access to certain elements, such as posts or pages, the Virtual Lab must empower participation. The methodology with which the Virtual Lab inspires participation is the execution of virtual and physical events. Three such events will be executed, a datathon, a hackathon and codagon during the MobiDataLab project life cycle. These three events are designed to allow the full spectrum of experimentation with data. The datathon will require participants to utilise data offered by the consortium focusing on exploration and experimentation, expecting insights and conclusions to be made, rather than developing services and applications. SMEs will be allowed to bring their existing application into the competition or create a new one that will support conclusion extraction and solutions provision. The hackathon requires participants to develop new data

analytics tools and services. The SMEs will be also required to demonstrate the application's business model and the added value it brings for the stakeholders of the challenge. The codagon will allow the scope of both the datathon and hackathon with the addition of live co-creation activities. Other teams, people, organisations and project partners will vote on the polls submitted by a coding team that assists brainstorming in real-time. This process is lengthier and will allow competitors to bring new requirements into their solutions that were not thought before. The events can be viewed as a continuation, a stepwise process to building innovations. Teams participating in any event will be requested to join the next one as a priority. They can also be viewed separately, meaning that a team can join an event even if they have not participated in the previous. Results from any event will be available for teams of the next event thus, enhancing continuation and building on existing data and insights.

5. Future work

The datathon, hackathon and codagon are tackled as instances of Living Labs. The outcomes developed by the participants during the events will demonstrate the impact that data availability can induce with respect to innovation and added value for data providers. Such a diverse impact can be captured by many stakeholders and by repetitive measurements. The three events will be executed sequentially in a matter of little more than a year. During and after each event, surveys will be issued and executed focusing on data collection first, from innovators and data users and second, from data providers (quantifying the usefulness of the outcome for their organization). The data will be used to improve the operation of the Virtual Lab and to provide a list of recommendations that promote data sharing in the context of collaboration among actors for the coming years, in terms of new business models and cooperation frameworks. This long-lasting study will provide definitive conclusions on the impact of data sharing in the transportation domain.

References

Wilkinson et al (2016) M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.7

EU (2007) 'Directive 2007/2/EC of the European Parliament and of the council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)', Official Journal of the European Union, 50(January 2006), pp. 1–14.

GAIA-X: A Federated Secure data infrastructure (no date). Available at https://gaia-x.eu/ (Accessed: 8 July 2022)

General Transit Feed Specification (no date). Available at: https://gtfs.org/ (Accessed: 5 May 2022).

Geospatial Data Strategy - United States Department of State (no date). Available at: https://www.state.gov/geospatial-data-strategy/ (Accessed: 8 May 2022).

Home | Open Knowledge Foundation (no date). Available at: https://okfn.org/ (Accessed: 6 May 2022).

Mabrouk, M. et al. (2005) 'OpenGIS Location Services (OpenLS): Core Services, OpenGIS® Im- plementation Specification'. OpenGeospatial Consortium.

Report on enabling technologies for the Transport Cloud | MobiDataLab consortium. Available at https://mobidatalab.eu/wp-content/uploads/2022/01/MobiDataLab-D2.6-ReportEnablingTechnoTranspCloud-v1.0DRAFT.pdf (Accessed: 8 July 2022)

Routing Pilot | OGC (no date). Available at: https://www.ogc.org/projects/initiatives/routingpilot (Accessed: 8 May 2022).

Web Feature Service | OGC (no date). Available at: https://www.ogc.org/standards/wfs (Accessed: 20 June 2021).

Salas, Julián, David Megías, and Vicenç Torra. "SwapMob: Swapping trajectories for mobility anonymization." In International Conference on Privacy in Statistical Databases, pp. 331-346. Springer, Cham, 2018.

Domingo-Ferrer, Josep, and Rolando Trujillo-Rasua. "Microaggregation-and permutation-based anonymization of movement data." Information Sciences 208 (2012): 55-80.

Hasson, F., Keeney, S. and McKenna, H., 2000. Research guidelines for the Delphi survey technique. Journal of advanced nursing, 32(4), pp.1008-1015.

Ruback, L. et al. (2016) 'Enriching mobility data with linked open data', in ACM International Conference Proceeding Series. Association for Computing Machinery, pp. 173–182. doi: 10.1145/2938503.2938550.