Deliverable

D1.3 Cities Guide to Legal Compliance for Data-Driven Decision Making It. 2

Project Acronym:	DUET	
Project title:	Digital Urban European Twins	
Grant Agreement No.	870697	
Website:	www.digitalurbantwins.eu	
Version:	1.0	
Date:	30 July 2021	
Responsible Partner:	GSL	
Contributing Partners:	AIV, IMEC, ISP	
Reviewers:	Lieven Raes (AIV) Nils Walravens (IMEC) Andrew Stott	
Dissemination Level:	Public	x
	Confidential – only consortium members and European Commission	

This project has received financial support from the European Union's Horizon 2020 Programme under grant agreement no. 870697

Revision History

Revision	Date	Author	Organization	Description
0.1	24.6.2021	Martina Piantoni, Tomas Pavelka	GSL	Initial structure
0.2	7.7.2021	Martina Piantoni, Tomas Pavelka	GSL	First draft
0.3	14.7.2021	Andrew Stott, Lieven Raes, Lea Hemetsberger, Jiri Bouchal	AIV, OASC, ISP	Comments and edits
0.4	19.7.2021	Martina Piantoni, Tomas Pavelka	GSL	Pre-final draft
0.5	29.7.2021	Andrew Stott	GSL	Comments and edits
1.0	30.7.2021	Tomas Pavelka	GSL	Final version

Table of Contents

Executive Summary

This deliverable seeks to address a request made by DUET partner organisations to provide a guide tailored to a critical stage in any data-based decision-making process: the **selection of datasets and models** for the decision-making use, and the reporting activities related to that selection.

This deliverable complements a more comprehensive guidance on legal necessities provided by deliverable D1.2 and works on the background of applicable laws as set out in deliverable D1.1 (Legal Landscape and Requirements Plan) submitted in Year 1 of the project.

As is usual in this stream of deliverables, before providing a structured guide, we first aim to provide some theoretical and conceptual background, putting forward the important building blocks to explain where the guidance is coming from and what it is seeking to achieve. Accordingly, **Section 2** first sets out some particular legal requirements and recommended practice related to the types of issues in dataset selection and use cases:

There is an important dichotomy between personal data and non-personal data, which can set anyone wishing to use this or that database type on a different path as regards the legal necessities and recommended practice. Sometimes databases can be mixed (including both personal and non-personal data), and if it is not feasible to separate the two then the whole mixed dataset will need to be considered as personal data. This is discussed by **Subsection 2.1**.

Subsection 2.2 asks whether a particular dataset is originally collected by a DUET partner organisation, in which case that organisation is fully responsible for its lawful collection and processing, or conversely, whether the dataset or model is sourced from a third party, in which case we recommend running a basic data audit and ask for guarantees. Specific caveats apply with regard to sourcing of third-party anonymized or pseudonymised datasets.

Subsection 2.3 provides an introduction to issues typically faced by those who negotiate access to commercial data sources or source data and models subject to an open licence. Open licences may make data more widely available but some other licence types are less open in that they may contain important restrictions which limit the options of integrating, reusing, or re-publishing the sourced data, for example a restriction that prohibits creation of derivative works.

Risks of some database types / models identified or pre-selected for DUET user epics are discussed in **Subsection 2.4**. This includes ANPR data, crowd-sourced traffic counts, FCD data, wifi/cellular network-sourced data, and other types. The bottom line is that while mitigation measures can be identified and applied, some data types are inherently privacy-sensitive and should be selected / used with caution, subject to an appropriate legal risk assessment, organisation-level (and DUET-level) approval, and potentially also only after consultation with competent Data Protection Authorities in respective jurisdictions.

Subsection 2.5 provides a high-level discussion of certain risks involved in simulation models, such as unrealistic model calculations or the drift caused by combining different datasets. General risk mitigators are recommended.

Finally, **Subsection 2.6** provides a step-by-step guidance in form of a checklist to meet critical legal necessities related to datasets selection and use processes. The guidance builds on the blocks identified in previous subsections and aims to provide a more easy-to-follow recommendation for each step and the issues identified. We would like this initial guidance to undergo further testing by the Pilot partner organisations as to its effect and user-friendliness.

A stand-alone **Section 3** is provided in response to the consortium request to come up with some recommendations on documenting and communicating DUET decisions related to database and model selection. These recommendations, which are fully subject to the emerging DUET data management plan and take into account DUET's commitments to the applicable Horizont 2020 and FAIR data principles rules on data management, suggest the creation of a centralized logging system. This would help to document relevant actions taken with regard to selection, sourcing and use of databases and models, and to enhance transparency and accountability by providing a reliable track record of such actions. Such an approach may also constitute a transferable recommended practice.

We conclude by suggesting that all these guidance efforts made to-date should be subject to further testing by the Pilots and other DUET partners and refined in the final upcoming deliverable in this working stream (D1.4) towards the end of the DUET project. The ambition is also to make these guidelines transferable for the purposes of the emerging DUET "book" of good practice.

1. Introduction and legal notice

This deliverable is the second out of three deliverables in this working stream aiming to build up an easy to understand Cities Guide to legal compliance for data-driven decision making.

The first iteration (deliverable D1.2) aimed at providing a more comprehensive guide going beyond the traditional topics of privacy protection (personal data collection and processing), and sought to touch on a variety of issues including the basis for the decision making, types of decision making and legal liability, data quality issues, Intellectual Property matters, dissemination, transparency and trust building, etc. In addition, the deliverable D1.5 (Ethical Principles for using Data-Driven Decision in the Cloud) supplemented these considerations with further information over and beyond purely legal requirements, such as the ethical aspects (and emerging laws) of decision-making processes involving automated means and AI. The connecting theme between all these angles is the aspect of transparency towards the data subjects (typically citizens) and winning their trust in smart city projects.

The present deliverable is more focused. It follows an explicit request made by DUET partner organisations to provide a guide tailored to a critical stage in any data-based decision-making process: the **selection of datasets and models** for the decision-making use, and the reporting activities related to that selection.

We understand the need for such specific guidance in the light of the fact that the majority of datasets or models used by the DUET project will be sourced “off the shelf” from third parties (such as other public authorities and law enforcement, or commercial data providers or vendors), as opposed to being collected by DUET partner organisations themselves. This imposes fairly specific requirements on the personnel responsible for selecting datasets and models for the use in DUET in order to ensure the overall compliance of the DUET project with the applicable legal requirements, mainly privacy and intellectual property laws. At the same time, good guidance may help to streamline these data selection processes and make them more cost-efficient, because third-party provided data may not always necessitate a full-scale legal compliance check in order to be used.

This deliverable therefore aims to supplement the “cities guide” with a guideline for cities’ approach on:

- selecting datasets and data models (meeting legal requirements); and
- documenting and communicating the use/selection of data in Digital Twin use cases.

This document and the emerging guideline is complementary to, and does not replace, existing or future applicable legislation in further detail described in deliverable D1.1 (Legal Landscape and Requirements Plan). Readers should make use of references back to deliverable D1.1 in order to get a fuller picture of the applicable law in the areas covered in the emerging guide. This document should be read in conjunction with the definitions and broader guide to legal requirements as set out in deliverable D1.2 (the first iteration in this series of deliverables on “easy to use guides”).

As regards the applicability of the guidance to DUET’s internal processes, this deliverable and the guidance provided is fully subject to the DUET Data Management Plan (WP8). As regards data collection, sourcing, use and any other data management activities, “*every* [DUET] *partner is responsible for the behaviour of all team members, which may also include subcontracted organisations.*” (deliverable D8.3, p. 18). Section 3 of this

deliverable seeks to propose a central logging system for documenting the selection and use of datasets / models by DUET partners to complement the existing data management plan (or be integrated with it). We anticipate further discussion among the consortium members as regards the usefulness and the ways of implementing such a central logging system.

Even though it is an ambition of this document to provide a useful guidance to any interested smart cities and other stakeholders out there, it is important to note that this document or deliverables D1.1 and D1.2 do not, and are not intended to, constitute legal advice to DUET partner organisations or any third parties. Instead, all information, content, and materials in these documents are for informational purposes only within the scope and objectives defined for the respective DUET project deliverables. Given that these documents got finalized at a certain point in time, information in these documents may not constitute the most up-to-date legal or other information at the cut off date. Readers of these documents and their organisations should contact their in-house team members (including their **Data Protection Officers** (DPOs)) or an attorney qualified in the concerned jurisdictions to obtain advice with respect to any particular legal matter.

2. Selecting datasets and models - legal requirements and recommended practice

The DUET Dataset Inventory (an internal document[1]) sets out various types of datasets and models pre-selected to enable and enhance the DUET system. This list is non-exhaustive and subject to further change:

- Geospatial datasets and models (digital building, surface and terrain models, location of road signs, location of hospitals and doctors, bike-sharing stations, etc.)
- Loop-based traffic data; multi-modal traffic models
- Floating car data, real-time position of public transport data
- Dynamic speed limit and lane indicator signs (RSS) traffic management data
- Trajectory control zones data
- Air quality sensor data, and air quality models
- Noise level sensor data (official source and citizen science source), and noise models
- Inventory of empty and neglected business premises
- NACE registers (such as company registers)
- Road accident statistics (anonymised data)
- Digital height model

Some further indicated dataset types are unclear if they may be available and/or whether they will be useful for any of the identified DUET user epics:

- ANPR data (anonymised / pseudonymised)
- Geospatial referenced register of installed cameras on public domain
- Crime and infringements geospatial register
- Parking data

We note at the outset that the *prima facie* most difficult data types from the legal perspective (pseudonymised ANPR data, public domain cameras and crime and infringements data) are less likely to be used in DUET.

[1] https://docs.google.com/spreadsheets/d/1xrlheOOE76aDtS1GWJR2j0FBrKJ_TI6pXCeTCMvBJPo/edit#gid=240358067

Nonetheless the text below provides some basic guidance on these, and these data types have also been addressed by the comprehensive guidance provided in deliverable D1.2.

There are three more general aspects of data selection that are discussed in more detail below in subsections 2.1 (personal vs. non-personal data), 2.2 (original/collected data vs. third-party data) and 2.3 (licence requirements). Subsections 2.5 and 2.6 discuss in more detail risks in some selected data types and in simulation models.

2.1 Personal data vs. non-personal data

Personal data is information relating to an identified or identifiable natural person. Where a database contains at least one data point which is personal data and which is inextricably linked with the other data points in that dataset (separating the two would either be impossible or considered by the controller to be economically inefficient or not technically feasible), the whole mixed dataset will need to be considered as personal data (see also Section 3.2.1 of D1.2).

Any sourcing and use (and any further re-use or re-distribution) of personal data must comply with the GDPR requirements. This means, mainly, that:

- Data must be sourced/collected and processed lawfully (legal basis must be identified)
- The purpose of any further use (processing) of the data must not be incompatible with the purpose for which data was originally collected. There are certain purposes which may be deemed compatible, such as archiving purposes in the public interest, scientific research or statistical purposes.
- The data minimisation principle should be observed at all stages, privacy-by-design and by-default approaches should be adopted and implemented.

For more details on this matter, consult Section 5 of D1.2 and Sections 2.3 and 2.4 of D1.1.

On the other hand, use of non-personal data and databases/models not containing personal data points do not have to conform to the GDPR, even though the deliverable D1.2. recommended some overarching principles to be followed in these cases as well to maximise risk mitigation. That deliverable also explained that, while fully anonymised data can be treated as non-personal data, pseudonymised data, on the other hand, is considered personal data and the GDPR fully applies.

In addition, the **European Data Protection Board (EDPB)** has recently stressed (again) that "*anonymisation of personal data can be difficult to achieve (and upheld) due also to ongoing advancements in available technological means, and progress made in the field of re-identification*" and that "t*hose parties which consider that they are using anonymous information [...] should be in a position to satisfy themselves – and when questioned also the competent [authorities] - on an ongoing basis that this continues to be the case, and that they have not inadvertently become data controllers of personal data for the purposes of the [GDPR].*"[2]

The cities are therefore advised to implement an approach that allows an **ongoing monitoring that the data they work with are still anonymous information**, and that they have not inadvertently become personal data

[2] EDPB Document in response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research, adopted 2 February 2021, available at https://edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf

controllers as a result of, for instance, combination of different datasets or the release of other datasets that would enable the data subjects to be identified.

2.2 Original/collected data vs. third-party data (sourced data)

The organisation that collects data from data subjects (or from the world of things) is responsible for clearing all legal requirements for the lawful collection and processing of such data (be it personal or non-personal data).

On the other hand, an organisation that sources data from third parties is not strictly responsible for the lawfulness of collection of the data[3], but assumes the responsibility to source the right dataset for the right purposes, and shares the responsibility for acquiring and using a dataset that is free of legal and factual quality defects. In Section 5 subsections E and F of deliverable D1.2, we have suggested that cities may make a careful assumption that a third-party data, when provided for further re-use, is free from legal defects, because such third party provider/vendor is obligated by law to collect, process, and share data lawfully and for legitimate purposes only.

However, each sourcing of a dataset/model should be, to the extent practicable, subjected to a **(basic) data audit.**[4]

Self-check questions for sourcing of datasets:

1. Does the data come from a reliable source? (reputable vendor or a public authority). Does the data come from a provider established in the EU?
2. Doesn't the data suffer from defects in important properties? (accurateness, timeliness, consistency, etc.)
3. Isn't the data manifestly unfit for the required purpose?
4. Is the data shared under reasonable licensing conditions?
5. Are there apparent issues with anonymization, aggregation or pseudonymization of the data?

Pro-active steps towards the third party provider before accessing, using, or feeding the data into the system:

1. Ask the data provider to declare that the information provided is in line with the applicable privacy legislation. Such a declaration of conformity may be provided in the data purchase specifications or licence conditions.
2. If you are sourcing *personal data*, ask also specifically for the legal basis and the permitted purposes for which the data may be further used (re-used).
3. If you are sourcing data that used to be personal data but now is provided as *anonymized or pseudonymized*, ask what measures / data sanitization techniques have been used for anonymisation, pseudonymisation or aggregation.
4. As a good practice we would recommend that each acquisition of a dataset or model that works with personal data (including pseudonymised data and mixed datasets) is subject to **approval by the organisation's Data Protection Officer (DPO)**.

[3] But note that if your organisation merely outsources the data collection to a third party, but determines the purposes for which and the means by which data is processed, your organisation may still be the "data controller" within the meaning of the GDPR.

[4] See also Section F of D1.2.

Note also the proviso in the DUET Data Management Plan: *In case of reuse of existing data, i.e. owned by someone else (a third party or another DUET partner), the* ***individual or joint responsibility is to check the nature of data […] and undertake the consequent actions [per the Data Management Plan]*** (D8.3).

In a similar vein, cities can assume that the data may be used for the purposes, which are explicitly allowed by the licencing conditions given by the licensor (the third party providing the data). However, even such purposes explicitly allowed by the licencing conditions may be prohibited by applicable law (cities should check with their legal department/DPOs) and some other purposes may be illegitimate on other grounds such as ethical considerations (see also deliverable D1.5).

One particular pitfall exists in the question whether data can be considered anonymised and therefore not personal data: if the originator party (data controller) provides anonymised or aggregated data to a third party recipient (such as DUET), but keeps the original (raw) personal data or the identification keys, the **anonymised or aggregated data may still need to be treated as personal data**. **WP29 opinion 5/2014 on Anonymisation Techniques**[5] stressed that:

"*it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data. Only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous.*

For example: if an organisation collects data on individual travel movements, the individual travel patterns at event level would still qualify as personal data for any party, as long as the data controller (or any other party) still has access to the original raw data, even if direct identifiers have been removed from the set provided to third parties. But if the data controller would delete the raw data, and only provide aggregate statistics to third parties on a high level, such as 'on Mondays on trajectory X there are 160% more passengers than on Tuesdays', that would qualify as anonymous data."

Therefore, cities would be advised specifically to check with the third party data providers whether they have fully anonymised the original dataset as well so that the anonymized / aggregated dataset provided to cities can be treated as non-personal data. Where such guarantees cannot be provided, cities would be advised to treat the dataset as personal data, following the precautionary principle.

2.3 Licence requirements

Closed or proprietary (commercial) licences: Copyright, neighbouring rights, database rights and other IP rights over particular content (data, database, simulation model/software) mean that the right holder has the exclusive right to commercialize the content and decide with whom the content can be shared, how it is used, etc. If you seek access to such content, the only way is to negotiate and conclude a licence agreement, or place a purchase order, with the right holder that will define the terms and conditions of your use of the database/model. Individual licence agreement negotiations, where they are necessary, are typically done with the help of cities' legal departments or outside legal counsel, as they can involve quite complex issues.

[5] ARTICLE 29 DATA PROTECTION WORKING PARTY Opinion 05/2014 on Anonymisation Techniques, April 201 4, p. 9., available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

The majority of databases and models indicated for DUET user epics appear to be provided on an open-licence basis. The right holder (third-party provider) under such licence grants you a worldwide, royalty-free, non-exclusive licence to "use" the dataset or model for the duration of any applicable copyright or other IP rights (e.g., database rights), without the need to individually and ad-hoc negotiate these licence conditions or place any purchase orders. In addition, an open licence will contain few if any obligations or limitations imposed on what you can do with the licensed content.

An "Open" licence will include the following **"baseline" rights** (see the "Open Definition" https://opendefinition.org/):

- extraction and re-use (or re-utilisation) of the whole or a part of the data in the set or model
- creation of derivative datasets
- creation of collective datasets (i.e., combining datasets)
- creation of temporary or permanent reproductions by any means and any form of the dataset
- communicating to public (within the meaning of IP laws) of the dataset

Some Open Licences may nevertheless place obligations on the data user:

- Attribution (e.g. CC BY ) - you must give the licensor (or the original creator/author) the credit (attribution) in the manner specified by the licence. For example, all Creative Commons licences require attribution except for the CC0-public domain variant
- Share-alike (e.g., SA ) - you may share or distribute derivative works only under a licence identical with (or not more restrictive than) the original open licence

It is important to understand that the widely used "Creative Commons" licences include both open and non-open types. The mission of Creative Commons is "Provide Creative Commons licenses and public domain tools that give every person and organization in the world a free, simple, and standardized way to grant copyright permissions for creative and academic works; ensure proper attribution; and allow others to copy, distribute, and make use of those works." However only the CC0, CC-BY and CC-BY-SA licences are regarded as open.

Typical **licence restrictions** to look out for in other Creative Commons licences (and in licences from other sources):

- Restricted data use or sharing purposes, (e.g, Creative Commons 4.0 ("**CC**") CC-NC  – only noncommercial uses of the work are permitted)
- Restriction on derivative works (ND) (e.g., CC-ND  – no derivatives or adaptations of the work are permitted)
- Restrictions on combination with other datasets

Licence conflict: note that two or more separate licences (for different datasets) may impact or even prohibit the possibility of their combination with other datasets, or their joint communication vis-a-vis third parties or the general public. This can be the case even with Open licences: for instance if dataset A is licensed under Creative-Commons-Attribution and dataset B is licensed under Creative-Commons-Attribution-ShareAlike then a combined dataset would have to have the ShareAlike condition applied to any re-use of that data.

Complex questions arising in the context of combination of datasets with different licence conditions should be resolved with the help of your legal department.

Bear in mind that other interests than intellectual property rights may get affected by publication of data, whether the data is accurate or inaccurate. For example, economic interests of third parties may get affected if the publication of data showing high pollution levels happen to depress real estate prices. However, it may be difficult to consider such potentially remote risks at the stage of data selection by city officials. As a rule of thumb, if the data you publish is accurate and up to date, it is less likely that your organisation would be held liable for any damages caused by the publication of such data (for example, damages that a real estate owner might claim if the price of its property deteriorates allegedly as a result of your organisation publishing pollution levels data). Conversely, publication of inaccurate or incorrect data would increase the likelihood that some liability for damages may arise (see also deliverable D1.1). Also other considerations than legal liability may impact your selection and use of databases or models, such as ethical considerations (see also deliverable D1.5).

2.4 Risks of selected database / model types

2.4.1 Automated Number-Plate Recognition (ANPR) data

ANPR technology is typically used to store the images captured by the cameras as well as the text of the car's number plate. These images may (or even typically do) capture also the person of the driver and potentially other passengers, or even third persons in other cars or passers-by.

ANPR technology is a powerful tool for law enforcement but raises specific concerns as regards privacy, government-citizen tracking, misidentification, or high error rates.[6]

A PoliVisu project deliverable D4.7 identified the following basic levers of access to ANPR data:

- Access to camera pictures itself and the vehicle linked crimes and infringements. Link with national- and international crime registers. This access is typically restricted to law enforcement authorities;
- Access to individual vehicle information as the owner and the owners address information. This information is typically used for access to parking zones (usually parking garages) and on-street parking control (parking overdue). Authorised officials can also use this data.
- Access to anonymised or pseudonymised information about traffic counts, traffic speeds over fixed stretches of roads, traffic movements (origin/destination) and linked number plate data as detailed vehicle type data, emission data, weight and height. This data could be used, if properly anonymised, by a wide group of users.

[6] See also PoliVisu deliverable D4.7, page 15.

We understand that, as a principle, only anonymized ANPR data will be used for DUET purposes. As an exception, pseudonymized ANPR data may potentially be used for specific user epics and models. .

There is one general caveat applicable to using pseudonymized ANPR data. While such data can be useful as they can give insights into the origin/destination of the traffic, this may raise particular privacy concerns and higher administrative burden on collection and use of such data, as **pseudonymized data** (as opposed to fully anonymized data) **is treated as personal data**. In other words, the GDPR fully applies to each stage of such data collection, sourcing and use.

Considering this, and in line with the precautionary principle, we suggest that **only fully anonymised data be allowed to be used for DUET purposes (and enter the system)**. If, exceptionally, pseudonymised data must be used in order to meet the specific needs of a DUET user epic, these may be selected and ingested only with the prior explicit approval by the DUET partner organisations' DPO and the DUET ethics manager (**Mr. Geert Mareels, AIV**)

The use(fulness) of ANPR data in DUET is still rather unclear, as is availability of such data from the originators (law enforcement authorities). If DUET decides to source and use pseudonymised ANPR data for piloting purposes, we suggest initiating consultations with Data Protection Authorities in each of the countries concerned to mitigate any residual risks.

2.4.2 Crowd sourced traffic counts (e.g. Telraam data)

Telraam project is a hardware-software solution for traffic measurements recording cars, heavy vehicles, public transport, cyclists and pedestrians. The Telraam device is a combination of a Raspberry Pi microcomputer, sensors and a camera. The device is mounted on the inside of an upper-floor window with a view over the street. The device sends the traffic count data straight to the central database.[7]

The data can be essentially qualified as citizen science data: Telraam develops high-tech and reliable measuring equipment, which it makes available to interested citizens (who bear the costs of acquiring the device and its installation). However, Telraam helps the citizens to set the camera up and correctly position it to aim at the street area. The Telraam data has information about the camera device itself and its location and the camera measurements (traffic counts).

A robust combination of privacy-by-design measures helps to mitigate privacy risks. Telraam processes the camera images immediately. The camera images are never stored. Telraam is designed in such a way that there is no possibility of viewing the camera images itself (not by the Telraam owner, nor by third parties). The camera images are only visible during the installation of the Telraam, only for the user (in order to be able to aim the camera properly), and for a maximum period of 10 minutes. After the data is processed and transmitted wirelessly to a central database, they produce results that are available as open data to the general public.[8]

In summary, Telraam submits that there are two tiers ensuring privacy protection:

[7] https://telraam.net/en/what-is-telraam
[8] https://telraam.zendesk.com/hc/en-us/articles/360025746472-What-about-privacy-

- The images themselves are processed locally and immediately. In other words, it is impossible to consult the images directly from the camera (the camera would simply not work).
- The data processed is generic and is not personal data: no number plates, no faces, no characteristics of persons.[9]

Some risk remains with regard to the potential hackability of the solution, but is limited by the fact that only minimal data (traffic counts) are sent to the central database, and the fact that API are used to transmit data from a local measuring device to the central server.[10]

This data type is currently linked to the city flows Digital Case implemented in Antwerp and linked to the DUET Digital Twin during the alpha version. In DUET we consider that the Telraam privacy-by-design robustness makes it an excellent data source for solutions such as DUET. In case of Telraam, the Belgian Data Protection Authority (GBA) had issued an advance comfort letter stating that it had no objections to this method of processing camera images (immediately and locally, as opposed to storing images & forwarding them to a central database for central processing as is the case with classic camera systems).

2.4.3 Floating car data (FCD)

This data is based on the collection of localization data, speed, the direction of travel and time information from mobile phones in vehicles that are being driven. In addition, floating car data can also be captured by navigation systems such as PNDs (Personal Navigation Devices) and Telematics systems. Floating car data is used by applications like Google Maps and Waze to predict travel times. Based on these data, traffic congestion can be identified, travel times can be calculated, and traffic reports can be rapidly generated. In contrast to traffic cameras, number plate recognition systems, and induction loops embedded in the roadway, no additional hardware on the road network is necessary.[11]

Floating car data typically has a **location and a timestamp**. If a **unique identifier** of the terminal equipment (mobile phone, personal navigation system, etc.) is collected as well, the collection and processing of such data amounts to personal data processing (or terminal equipment data processing) and **must be carried out in full accordance with the GDPR and the ePrivacy legislation** (see deliverables D1.1 and D1.2). In order to sanitize the data for use and publication in DUET, the data should not include the unique identifiers as well as any links to specific individuals or their devices (mobile phones, cars, etc).

We understood from questionnaire responses by DUET partner organisations that only aggregated and fully anonymized FCD data are intended to be used: in case of **Pilsen**, this data will be collected and provided for DUET purposes by a third party (National Directorate of Roads of the Czech Republic) on a license agreement basis, and will be provided anonymised.

However, as it may be difficult to guarantee that the data is fully anonymized or otherwise collected in line with privacy legislation[12], DUET partners as well as **cities should ask the third-party data providers for a guarantee / declaration of compliance with privacy legislation**.

[9] https://telraam.net/en/what-is-telraam
[10] See also PoliVisu deliverable D4.7, page 30.
[11] See also PoliVisu deliverable D4.7, page 31
[12] See also PoliVisu deliverable D4.7, page 33.

2.4.4. Wifi / cellular / app collected terminal equipment data

Data, even if pure statistical counts, obtained by collection of unique identifiers or other information from users' terminal equipment (mobile phones, tables, cars) **are by default privacy sensitive**.[13]

In addition to GDPR, the ePrivacy legislation will typically apply to the cases of collection and processing of such electronic communication data. While the legal requirements for using these sources for "statistical counting" may get somewhat relaxed in the future, it is not yet so under the currently applicable EU ePrivacy Directive (see for more detail deliverable D1.1. and Section 2.2.2 of deliverable D1.2).

All this type of data using terminal equipment information (e.g., mobile phone data, raw mobile phone data, or wifi sniffing data) can provide information about mobility behaviour, number of devices / persons passing by ("statistical counting"), but even on other spatial-temporal behaviour of users and their social interactions.

The data may come from individuals using their devices (e.g., an individual user placing or accepting a call, or simply by moving in the signal area with his/her device turn on ("sniffing")), but can also originate from communications between IoT devices (a machine-to-machine communication) or be collected and sent by apps installed on the users' devices. In each case, collection and processing of such data will involve complex privacy and ePrivacy legal issues and cities should obtain a **DPO approval for any such action.**

When sourcing such data from third parties (typically fully anonymised or aggregated), cities should request the **third-party data provider to declare that the information provided is in line with the applicable privacy legislation**.

Engaging in prior consultation with the national Data Protection Authority may be advisable, particularly in cases where the originator (the third-party that collected the raw data) still has access to non-anonymised raw data, even if the data sourced by cities are anonymised or aggregated.

2.4.5 Other data risks (e.g., road accident data or noise / pollution levels sensor data)

Some data may carry residual privacy infringement or other legal risks, typically in combination with other data or information. The road accident data and noise / pollution levels may be a good example.

The road accident statistics may contain data on exact location and relevant data about the conditions of a traffic accident. The PoliVisu deliverable D4.7 lists the following data items contained in the anonymised road accident data contained in the Belgian Federal Police VOAC data source:

- Time
- Location (street + house number; crossing between two streets, street + kilometre sign)
- Geocoding (calculation based on the location interpretation)
- Technical findings (weather conditions, lighting conditions, road condition, city limits)
- Obstacles
- Number of Involved road users (extra information for the different types of road users)
- Alcohol, drugs, speeding
- Dangerous products

[13] See also PoliVisu deliverable D4.7, page 33.

- Aggravating circumstances

Even if this data is anonymised (i.e., cannot as of themselves be linked to any specific individual person), the data can be linked to, for example, a newspaper article based on the time and space data points. Newspaper articles may often reveal the identity of the individuals involved, including the car crews or the police officers investigating the event. In combination with the data from the anonymous road accident database this may reveal a set of detailed information that can be linked to specific individuals.
This may be similarly the issue with noise and pollution data. Even if these databases contain only anonymised location, time and the measured levels, linking these with a newspaper article about a major pollution event, for example, may lead to linking that information with specific individuals.

These residual (and perhaps rather remote) risks may be difficult to mitigate against. The only truly effective measure would be not releasing this type of data to the public (or to interested stakeholders such as DUET), but that would be clearly disproportionate considering how useful this type of data is for policy making. Cities should, however, be made aware of these residual risks.

2.5 Risks of simulation models

Simulation models and other data processing software may involve risks. These may be inherent or natural or be caused by the nature of the data processed (for example, some elements are not taken into account in a typical noise sensor data, which means that not all of the influencing factors are covered by the model). These are the most common risks involved in simulation models and their use:

- Calculation of unrealistic results
- Not all of the influencing factors are covered by the model
- Drift caused by combining multiple models
- Lack of information about the correct interpretation of model outcomes

To the extent these risks can be addressed or minimized by the quality of the simulation model, these factors should play a role in your selection of a particular model for DUET purposes. In case these risks are inherent or simply no other simulation models are available for use, the risk should be mitigated by full transparency and good reporting both within your organisation, but also vis-a-vis third parties or the general public, where applicable.

2.6 Step-by-step guidance to selection of datasets / simulation models

The purpose of the following guidance on selection / use of datasets and simulation models is to provide a structured checklist that can be applied for each data selection action. Going through the checklist should help the responsible person to consider and mitigate the most significant legal risks related mainly to sourcing third-party datasets / models, but also to cases where the dataset is provided by a DUET partner organisation.

1. **Set the purpose for your action**. This is important to know what legal rules and limitations may be applicable to your action. For example:
    a. I am sourcing a dataset to create a simulation model

b. I am sourcing a simulation model to integrate it with the DUET system (combining with other datasets / models) in order to enable a DUET functionality
c. I am sourcing a dataset to create a derivative dataset (by changing the data or by combining multiple datasets)
d. Each such action should have an identified (meta) purpose, such as "*I want to create a simulation model in order to achieve objective X*". In DUET, these final objectives are often pre-defined by DUET user epics.

2. **Data availability** (licencing conditions, format compatibility, APIs). These conditions are important to ensure that you meet applicable IP laws requirements for sourcing and use of a dataset / model.
   a. If you need to negotiate individual access to a closed / proprietary database or simulation model, engage your legal department in the contractual negotiations.
   b. Check restrictions on use of open licenced databases / models, such as the attribution rule (CC BY ), or share-alike (CC SA ) condition. Some licences may prohibit you from creating derivative works (e.g., CC-ND ). Consult with your legal department if you are unfamiliar with the terms
   c. Check for any licence conflicts or restrictions on combining multiple datasets. Consult with your legal department if complex questions arise

3. **Dataset / simulation model quality.** These steps are important to mitigate risks stemming from lack of data's legal or factual quality.
   a. Self-check the following:
      i. Does the data come from a reliable source? (reputable vendor or a public authority). Does the data come from a provider established in the EU?
      ii. Doesn't the data suffer from defects in important properties? (accurateness, timeliness, consistency, etc.)
      iii. Isn't the data manifestly unfit for the required purpose?
      iv. Is the data shared under reasonable licensing conditions?
      v. Are there apparent issues with correct anonymization, aggregation or pseudonymization of the data? (e.g., use the Telraam example as a good practice to achieve a privacy-by-design solution that helps to minimize the risks to citizen's privacy).
   b. If sourcing third-party data, ask the provider for:
      i. declaration of conformity with privacy legislation
      ii. legal basis and permitted use purposes of any acquired personal data
      iii. if the data is anonymised/pseudonymised, ask what measures/techniques have been used to achieve this.
   c. Anonymised data preference (DUET-wide applicable guidance):
      i. Select and use only fully anonymised for DUET purposes. If, exceptionally, pseudonymised data must be used in order to meet the specific needs of a DUET user epic, these should be selected and ingested only with the prior explicit approval by the DUET partner organisations' DPO and the DUET ethics manager (**Mr. Geert Mareels, AIV**)

ii. Beware the issue of anonymisation of the original dataset. Check with the third party data provider whether they have fully anonymised the original dataset as well so that the anonymized / aggregated dataset provided to you can be treated as non-personal data. If they have not done this, the data is treated merely as pseudonymised, as opposed to anonymised. Where such guarantees cannot be provided by the third party provider, treat the dataset as personal data (principle of precaution).

d. Check for inherent risks of a simulation model (e.g., possibility of unrealistic results, gaps in coverage of influencing factors, drift caused by combining models, information on the correct interpretation of model outcomes).

4. **Transparency as a general risk mitigator.** A clear record of your action will help to mitigate and correctly attribute any residual risks in using a dataset / model.
   a. Note and report any issues with data quality, e.g. accuracy, non-up to date data, incorrect data
   b. Note and report any issues with the selected data model, e.g.
      i. Risks of calculation of unrealistic results
      ii. Risk that not all of the influencing factors are covered by the model
      iii. Risk of drift caused by combining multiple models
      iv. Lack of information about the correct interpretation of model outcomes.
   c. Report and disclose these risks vis-a-vis third parties or the general public, if they are the intended addressees of your actions (they are the addressees of your decisions based on the data / models, they are the next users of datasets/models created by your organisation, they are testers, etc.)
   d. If collecting or processing personal data, you may be required under the GDPR to issue a Data Protection Impact Assessment (DPIA) where your actions are "likely to result in a high risk" to privacy. In such cases, work closely with your DPO to take the necessary and timely action. The following list of typical "high risk" activities is not exhaustive and you should request further guidance from your DPO.
      i. use of new technologies (e..g, IoT applications)
      ii. systematic monitoring of a publicly accessible area on a large scale (including with CCTV cameras or sensors able to collect personal data)
      iii. Systematic and extensive evaluation of personal aspects based on automated processing, including profiling of individual persons,
      iv. processing on a large scale of special categories of data and data relating to criminal convictions and offences

5. **Meet intra-organisational requirements (partner-level approval process)**. A general good practice would be to check with your DPO each selection / use of personal data (including, for example, pseudonymized ANPR data and similar data with higher risk profile (e.g., wifi / cellular mobile "sniffing" data, app collected data)

6. **Meet DUET requirements (consortium-level approval process)** - follow the Data Management Plan

3. Documenting and communicating the use / selection of data in DUET use cases

This section contains a set of suggestions for improving the documentation and communication of use / selection of databases and models in the DUET data management lifecycle. These suggestions are made based on a specific consortium request for such guidance. They take full cognisance to the principle that the **data ownership goes hand in hand with the responsibility for data management**.[14]

All suggestions are subject to the existing version of the DUET Data Management Plan (deliverable D8.3) and are intended to inform the future versions of the data management plan. These suggestions take note of the fact that DUET adheres to the Open Research Data policy in Horizon 2020 and the FAIR Data Handling Principles.[15]

The suggestions for a centralized logging and communication system would apply on top of the data related actions provided by the data management plan to date. The proposed system would work with actions related to any of the data types identified by the DUET management plan, which include:

- Original data / existing data in DUET possession / existing sourced data
- Confidential data / anonymised and public data / non-anonymised (temporary) data
- Government data / business data / citizen data

In order to enhance the documentation and communication processes, we suggest to complement the DUET Management Plan with the following:

1. **Create a centralized logging and communication system**
    a. Define which partner should set up and oversee the system
    b. Define “responsible persons” and “users” at the level of individual person - user, at the level of DUET partner organisation, and at the consortium level

2. **Define steps and events that should be logged in,** for example:
    a. When (on what date) the data sourcing approval processes were met
    b. Effective database / model sourcing date (date of concluding the licence agreement, date of first access, date of downloading, date of last access)
    c. All data sanitization measures applied to the dataset (anonymisation, pseudonymisation, aggregation, statistical evaluation, metadata generation, etc.)
    d. All other defined “uses” of the database (what is a data use? See D1.2 Cities Guide to Legal Compliance for Data-Driven Decision Making It. 1 for definitions), e.g., integrating the data into the DUET system, updating the data, using the data to make a decision. If necessary to decrease the burden on users, define a level of “significance” for database use. Only a significant use would trigger then need to create a log.
    e. AI- or automated system-related decisions (including the use of data to “train” an AI), use of the “Digital Twin Data Broker” to process and aggregate data
    f. Data storage-related decisions (e.g., decision to upload to a Cloud-based service)

[14] Deliverable D8.3 (Data Management and Modeling Plan), page 10.
[15] Deliverable D8.3 (Data Management and Modeling Plan), page 10.

   g. Access to / use of a database or model by third parties (e.g., external DUET testers, DUET final users in the course of DUET data management lifecycle)

3. **Define and make mandatory reporting of other events and errors**, for example:
   a. Data breaches, security incidences, or similar adverse events
   b. Accidental data re-identification
   c. Reporting of data getting outdated, data getting beyond the permitted time for their storage
   d. Reporting of significant change of circumstance that may make a dataset / model obsolete or inaccurate

4. **Define when a sharing of data / model with other DUET partners should be recorded**, transfer of responsibilities under the logging system and the DUET data management plan.

5. **Reporting of gaps and data needs.** Define ways to log into the system a particular issue with the data or the lack thereof: missing dataset type, missing dataset set or subset, missing updates.

6. **Trust-building and transparency.** The logging system should allow seamless communication about dataset / model selection and use processes with
   a. DUET partner organisations
   b. DUET users and testers
   c. the general public (dissemination)
   d. Stakeholders and users should be allowed and motivated to conduct prior consultations with respective partners regarding the data needs and gaps for DUET user cases

7. **FAIR data principles: ensure full machine-actionability of the logs** (findable, accessible, interoperable, reusable)

4. Conclusions and future work

This deliverable sought to complement the emerging comprehensive guide to legal necessities (introduced in deliverable D1.2) by a more tailored guidance focused on a particular stage of a decision-making process: the selection and use of datasets and models. This stage is critical for the development of the DUET system and the entire project, but also highly relevant for a broad range of other smart city activities and thus potentially to many other interested stakeholders.

In addition to the guide, this deliverable also provided a basic set of suggestions for creation of a central logging system that would help to document decisions and choices related to data selection and use in DUET. Such a system could be integrated into the DUET data management plan or set up separately, or these suggestions could be taken up in order to enhance the existing and emerging data management plan and activities.

As a further major step in the WP1 package (Ethics, Privacy, Legal) towards the finalisation of this working stream (D1.2, D1.3, and future D1.4 Cities Guide to Legal Compliance for Data-Driven Decision Making), we plan to cooperate even closer with DUET pilot city partner organisations (AIV, SITMP, DAEM) and other partners involved in the decision making regarding data, databases and models integration in DUET, in order to test, improve and streamline the emerging guidelines. The result - we hope - will serve two main purposes: a) documenting the approach to legal necessities that has been adopted by DUET, and b) create an easy to share guide(s) that could be instructive for other smart city initiatives and the interested general public.

More specifically, we aim to:

- Test the guide in Section 2.6 of this deliverable with Pilots, and motivate the Pilots to provide additional feedback on the emerging comprehensive guide provided in deliverable D1.2.
- Consider merging these guides.
- Follow up with Pilots and other relevant DUET partners by means of questionnaires in order to map the development of their data needs and approaches (follow up on the work done in preparation done for deliverable D1.2).
- Discuss with consortium management the need for creation and integration of a logging system for documenting data-related decisions proposed in Section 3 of this deliverable.
- Consider recasting the guidelines in a more user-friendly design, using diagrams, decision-making trees, etc.
- Adapt and include smart versions of the guidelines provided in the D1.2 through D1.4 working stream into the emerging DUET “book” of good practice.