DELIVERABLE

D2.12 Data Management Plan (Iter. 3)

Project Acronym:	PoliVisu	
Project title:	Policy Development based on Advanced Geospatial Data Analytics and Visualisation	
Grant Agreement No.	769608	
Website:	www.polivisu.eu	
Contact:	info@polivisu.eu	
Version:	1.0	
Date:	14 November 2019	
Responsible Partner:	POLIMI	
Contributing Partners:	-	
Reviewers:	Tomáš Řezník (HSRS) Joran Van Daele (Gent) Matteo Satta (ISSY) Stanislav Štangl (SITMP) Berdien De Roo (GEOS) Andrew Stott	
Dissemination Level:	Public	X
	Confidential – only consortium members and European Commission Services	

This project has received financial support from the European Union Horizon 2020 Programme under grant agreement no. 769608.

Revision History

Revision	Date	Author	Organization	Description
0.1	16/10/2019	Francesco Molinari	POLIMI	Initial draft for partners inputs and comments
0.2	05/11/2019	Francesco Molinari	POLIMI	Final draft taking partners comments onboard
0.3	11/11/2019	Francesco Molinari	POLIMI	Final draft for partner approval
1.0	14/11/2019	Francesco Molinari	POLIMI	Final version submitted to the EC

Every effort has been made to ensure that all statements and information contained herein are accurate, however the PoliVisu Project Partners accept no liability for any error or omission in the same.

Table of Contents

Glossary of used terms **5**

Executive Summary **6**

1. Introduction **8**
1.1 The PoliVisu Data Management Lifecycle 8
1.2 Reference Framework and Perimeter of the DMP 10
1.3 Alignment to the Principles of FAIR Data Handling 12

2. Legal framework **14**
2.1 The PSI Directive 14
2.2 The EU Personal Data Protection Regulation (GDPR) 15
2.3 Open Access in Horizon 2020 15
2.4 Grant Agreement and Consortium Agreement provisions 16
2.4.1 Grant Agreement 16
2.4.2 Consortium Agreement 17
2.5 The PoliVisu licensing policy 19
2.6 Special provisions for big datasets 19
2.7 Special provisions protecting personal data 21
2.7.1 Overview of available approaches 21
2.7.2 Limitations and risks 22

3. PoliVisu Data Management Plan **23**
3.1 Data summary 23
3.2 Data collection 24
3.3 Data processing 25
3.4 Data storage 26
3.5 Data sharing 27

4. Conclusions and Future Work **29**

References **32**

Annex 1. Running list of data sources **33**

List of Tables

Table 1. Alignment between this DMP and the EC's requirements 12
Table 2. Building blocks of the PoliVisu licensing policy 19
Table 3. Summary of relevant data for the PoliVisu research agenda 23
Table 4. Summary of PoliVisu data collection procedures 24
Table 5. Summary of PoliVisu data processing procedures 25
Table 6. Summary of PoliVisu data storage procedures 26
Table 7. Summary of PoliVisu data sharing procedures 28

List of Figures

Figure 1 – The PoliVisu Data Management Life Cycle 9
Figure 2 – Open access to scientific publications and research data in the wider context of a project's dissemination and exploitation (source: European Commission, 2017) 10

Glossary of used terms

Acronym	Definition
ANPR	Automated Number Plate Recognition
CA	Consortium Agreement
CSV	Comma Separated Values
DMP	Data Management Plan
DoA	Description of the Action
DPO	Data Protection Officer
EC	European Commission
EEA	European Economic Area
EU	European Union
GA	Grant Agreement
GDPR	General Data Protection Regulation
H2020	Horizon 2020
HTTPS	Hyper Text Transfer Protocol Secure
ICT	Information and Communication Technology
IPR	Intellectual Property Rights
KOM	Kick Off Meeting
ORDP	Open Research Data Pilot
PDF	Portal Document Format
PSI	Public Sector Information
R&D	Research and Development
TOC	Table Of Contents
WP	Work Package

Executive Summary

PoliVisu aims to establish the use of big data and data visualisation as an integral part of policy making, particularly, but not limited to, the local government level and the mobility and transport policy domain. The project's relation with data is therefore essential and connatural to its experimental research objectives and activities.

Additionally, the consortium has adhered to the H2020 ORDP (Open Research Data Pilot) convention with the EC, which explicitly caters for the delivery of a DMP (Data Management Plan). This "describes the data management life cycle for all data sets that will be collected, processed or generated by the research project. It is a document outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved." (PoliVisu DoA, p.97)

This document is the third edition of the PoliVisu DMP scheduled at M24 according to the contractual work plan. Its basic structure and contents are confirmed with respect to the two previous deliverables (D2.10 & D2.11) but the analysis and descriptions of how its principles are implemented have become more concrete, due to the development of pilot activities in the three pilot cities (Gent, Issy-les-Moulineaux and Pilsen) as well as in three additional cases (Mechelen, Flanders and Leuven).

The reader interested in getting to know more about the overall data management strategy of the Polivisu consortium may use the information contained herein as well as in the following official project documents: D6.2 (Baseline Analysis), D1.1 (Ethical Requirement No. 4), D1.3 (Ethical Requirement No. 3), D2.2 (Project Management Plan), D2.3 (Quality and Risk Plan), D4.5 (Privacy Rules and Data Anonymisation), D6.1 (Pilot Scenarios), D7.1 (Evaluation Plan) and D8.1 (Impact Enhancement Road Map).

The remainder of this document is structured as follows:

- **Section 1** presents PoliVisu's data management lifecycle and frames the DMP within the EU H2020 Guidelines and FAIR data handling principles, thus setting the stage for the following parts.
- **Section 2** is a brief overview of the legal framework, including the EU regulation on personal data protection (GDPR), the H2020 provisions for open access to research data, the provisions of the PoliVisu Grant Agreement and Consortium Agreement, an outline of PoliVisu licensing policy to be further developed in the next and final edition of this DMP, some special provisions for big data management as well as an overview of available data protection approaches.
- The core of this document is **Section 3**, in which the data usage scenarios are presented and the key issues to be examined in relation to each scenario are discussed. These issues include decisions on e.g. data anonymization, privacy and security protection measures, licensing etc.
- **Section 4** concludes the document by anticipating the expected contents of the next and last edition of this DMP, which will mostly reflect the current structure, only with the variations dictated by the emergence of any new dataset in the project pilots and/or by the need to finalise and formalise the licensing rules of the consortium in the post-project phase.
- A specific **Annex** provides the running list of data sources used in all PoliVisu pilot sites.

In the fourth and last edition of the DMP, the post-project scenario, with a special eye on licensing and open standards, will be defined and communicated as the official standpoint of the PoliVisu consortium.

Background

According to the PoliVisu DoA (2017), data management planning, monitoring and reporting is part of WP2 - the Project and Quality Management work package - and foresees the delivery of four consecutive editions of the DMP at months 6, 12, 24 and 36.

This first edition, delivered in May 2018 as Deliverable 2.10, was not a mere collection of principles, as it set the stage for the ongoing and next activities handling with data, before and even after the PoiVisu project is completed. As per the DoA provisions: "*DMP describes the data management lifecycle for all data sets that will be collected, processed or generated by the research project. It is a document outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved".*

An open question, apparently unfulfilled by D2.10, was related to the identification of which data sets would be relevant for PoliVisu (action) research. An Annex to the Deliverable, entitled "Running list of data sources", was left blank with the promise of being taken care of at a later project stage.

In practice, what happened in the consortium's life was the release - in month 7 - of Deliverable 6.2 "Baseline Analysis", authored by the GENT team together with other colleagues from AIV, GEOS, ISSY and PILSEN. With that deliverable, which was coming up as a specification of D6.1, where the different needs and scenarios of the three pilot cities (namely Gent, Pilsen and Issy-Les-Moulineaux) had been described in detail, an extensive overview was provided of the already available datasets in each city, as well as the missing ones at the date of release of that Deliverable.

Therefore, the second edition of the PoliVisu DMP took stock of the baseline analysis carried out in D6.2 and used the information contained therein to fill in the Annex entitled "Running list of data sources". This was the major change with respect to the first edition.

In this third edition of the DMP the PoliVisu consortium consolidates the evidence already provided before, with a description of used datasets more adherent to the work actually being carried forward, including the more recent involvement of three pilots external to the consortium, namely in Mechelen, Leuven and the Flanders.

Globally speaking, the same structure and contents of D2.10 and D2.11 have been maintained in D2.12, in perfect coherence with the conception of the PoliVisu DMP as a "living Deliverable", being updated at periodic dates following the progress in the underlying project related activities.

In conclusion, and apart from minor corrections of (mainly) typos all along the preexisting text, only Sections 3 and 4 and Annex 1 of this edition include considerable changes in their original contents. Of course, and until further advice, all Sections of the DMP remain valid and enforceable - for what pertains to or influences the daily work of the PoliVisu partners.

1. Introduction

Visualisation and management of (big) data in a user friendly way for public administration bodies is one of the primary goals of the PoliVisu project. The intention is to support integration of (big) data into policy and decision making processes. The project's relation with data is therefore essential and connatural to its experimental research objectives and activities. Additionally, the consortium adhered to the H2020 ORDP (Open Research Data Pilot) convention with the EC, which explicitly caters for the delivery of a DMP (Data Management Plan).

According to the PoliVisu DoA (2017), data management planning, monitoring and reporting is part of WP2 - the Project and Quality Management work package - and foresees the delivery of four consecutive editions of the DMP at months 6, 12, 24 and 36. The first edition, however, was not a mere collection of principles, as it set the stage for the ongoing and next activities handling with data, before and even after the project is completed.

In the following text, we reproduce the contents of Section 1 of D2.11 almost unchanged, apart from some corrections of typos or equivalent modifications aimed at improving readability and understandability. In case the reader is already familiar with D2.11 contents, the remainder of this Section may be skipped at all.

1.1 The PoliVisu Data Management Lifecycle

As per the DoA provisions, the PoliVisu DMP "*describes the data management lifecycle*[1] *for all data sets that will be collected, processed or generated by the research project. It is a document outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved".*

This paragraph summarizes the management procedures that will be followed when dealing with the data of relevance for the PoliVisu project, and which will be further described in Section 3 of this document. To get a prompt overview of the relevant datasets for the three pilot cities, the interested reader is referred to Annex 1 to this document, entitled "Running list of data sources".

After an internal discussion, also during the May 2018 held in Issy Les Moulineaux, the partners agreed to envisage **three main data usage scenarios**:

a) Original data produced by the PoliVisu consortium and/or individual members of it (e.g. during a dissemination action or a pilot activity);
b) Existing data already in possession of the PoliVisu consortium and/or individual members of it prior to the project's initiation;
c) Existing data sourced/procured by the PoliVisu consortium and/or individual members of it during the project's timeline.

For each of the above scenarios, the key issues to be examined are displayed by the following logic tree:

[1] The definition of lifecycle used in this document mirrors that of the European legislation on e.g. geospatial data, see section 9.7 at page 58 in https://inspire.ec.europa.eu/file/1419/download?token=SJENygxP

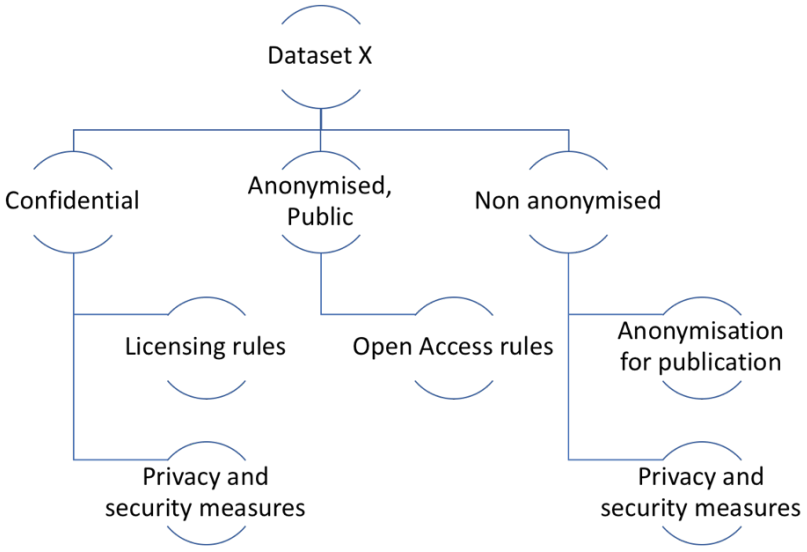

Figure 1 – The PoliVisu Data Management Life Cycle

For each dataset (or even data point) handled in the project, the first level of control/decision making must deal with its **nature**, notably whether it has been (or will be) deemed Confidential, or Anonymised and Public (it cannot be that the two latter things diverge, apart from very special occasions, which are coped with in the third logical category displayed in the picture).

Depending on the assessment of nature, the resulting, mandatory **action lines** can then be summarized as follows:

- For any acknowledged **Confidential**[2] dataset (or data point), the Consortium and/or each Partner in charge of its handling shall control (if existing) or define (if not) the **Licensing rules** and the **Privacy and security measures** (to be) adopted in the process.
- For any acknowledged **Anonymised and Public** dataset (or data point), the only relevant discipline to be clarified is the set of **Open Access rules** that apply to the case. This set is little controversial for PoliVisu, as the ODRP convention has been adopted, as specified above. Note that the use of open data across the PoliVisu pilots, including e.g. Open Transport Maps or Open Land Use Maps, falls in this category.
- Any dataset (or data point) that does not belong to any of the former two categories is subject to an additional level of action by the Consortium and/or Partner in charge, leading to its classification as either Confidential or Anonymised and Public. In that regard, the two, mutually exclusive action items belonging to this level are:
    - the **anonymisation for publication** action, leading to the migration to the second category of data, or
    - the adoption of appropriate **privacy and security measures** (very likely the same applied to the category of Confidential data) in case anonymisation is not carried out for whatever legitimate reason. Note that in this latter case, i.e. without anonymisation, **no licensing rules are applicable** (i.e. the PoliVisu consortium rejects the commercialisation of the personal profiles of human beings as a non-ethical practice).

[2] Confidential in a meaning as defined by the GDPR (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN)

1.2 Reference Framework and Perimeter of the DMP

The following picture – borrowed from the official EU H2020 information portal[3] - clearly identifies the positioning of the DMP in the context of projects that – like PoliVisu – have voluntarily adhered to the Pilot on Open Research Data in Horizon 2020[4].

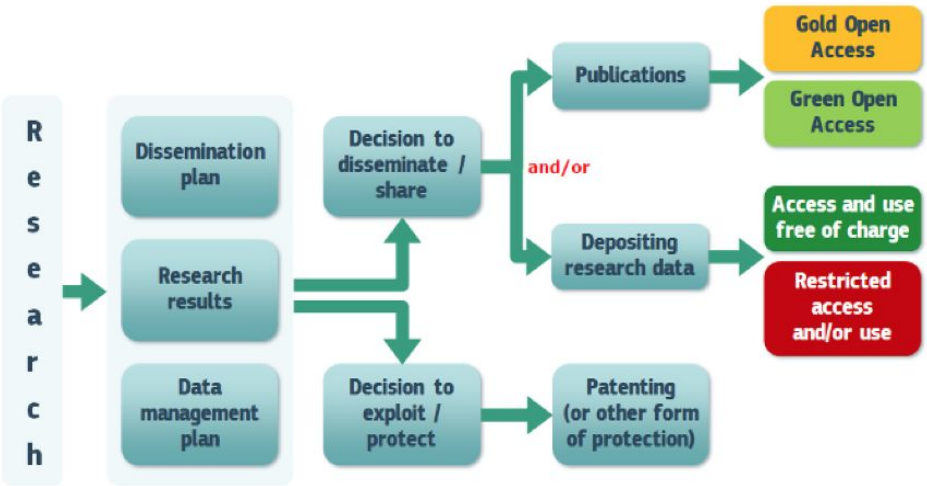

Figure 2: Open access to scientific publications and research data in the wider context of a project's dissemination and exploitation (source: European Commission, 2017)

As can be seen, a DMP holds the same status and relevance as the project's Dissemination Plan[5]. More specifically, in the former document, one should retrieve the full list of research data and publications that the project will deliver, use or reuse, as well as the indication of whether some data will be directly exploited by the Consortium, having been patented or protected in any other possible form. In the latter document, one should retrieve the Consortium's detailed provisions for all data and publications that can be shared with interested third parties, with or without the payment of a fee[6].

In particular, the following definitions – all taken from the aforementioned EU H2020 portal – shall apply to our discourse:

- **Access**: "*the right to read, download and print – but also the right to copy, distribute, search, link, crawl and mine*";
- **Research Data**: "*[any] information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form*";
- **Scientific Publications**: "*journal article[s], … monographs, books, conference proceedings, [and] grey literature (informally published written material not controlled by scientific publishers)*", such as reports, white papers, policy/position papers, etc.;

[3] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

[4] Extended to all H2020 thematic areas since the start of the 2017 Work Programme, the focus of the Pilot is on "*encouraging good data management as an essential element of research best practice*". See European Commission (2017).

[5] In the PoliVisu project, this deliverable is named Impact Enhancement Roadmap (D8.1) and has been issued at month 2 of the official work plan. This DMP is drafted at month 6.

[6] This means that, at least in principle, some research data might also remain undisclosed, without undermining the Consortium's participation in the ORDP.

- **Open Access Mandate**: "*comprises 2 steps: depositing publications in repositories [and] providing open access to them*". Very importantly, these steps "*may or may not occur simultaneously*", depending on conditions that will be explained below:
    - **"Green" Open Access (aka Self-Archiving)**: it is granted when the final, peer-reviewed manuscript is deposited by its authors in a repository of their choice. Then open access must be ensured within at most 6 months (12 months for publications in the social sciences and humanities). Thus, open access may actually follow with some delay (due to the so-called "embargo period");
    - **"Gold" Open Access (aka Open Access Publishing)**: it is granted when the final, peer-reviewed manuscript is immediately available on the repository where it has been deposited by its authors (without any delay or "embargo period"). Researchers can also decide to publish their work in open access journals, or in hybrid journals that both sell subscriptions and offer the option of making individual articles openly accessible. In the latter case, the so-called "article processing charges" are eligible for reimbursement during the whole duration of the project (but not after the end of it).

In the PoliVisu **DoA** (2017), the following provisions for Open Access were defined, which have become part of the Grant Agreement (GA) itself: *"PoliVisu will follow the Open Access mandate for its publications and will participate in the Open Research Data pilot, so publications must be published in Open Access (free online access). Following the list of deliverables, the consortium will determine the appropriate digital objects that will apply to the Data Management Plan. Each digital object, including associated metadata, will be deposited in the institutional repository of Universitat Politècnico Milano, whose objective is to offer Internet access for university's scientific, academic and corporate university in order to increase their visibility and make it accessible and preservable."* Evidently, these provisions belong to the **"Green" Open Access** case.

As far as patenting or other form of protection of research results is concerned (the bottom part of Figure 2), the ground for this has been paved by the PoliVisu Consortium Agreement (2017) - following the DoA, which recognises that *"formal management of knowledge and intellectual property rights (IPR) is fundamental for the effective cooperation within the project lifetime and the successful exploitation of the PoliVisu Framework and tools within and after the end of the project"*. Further steps towards a clarification of the licensing mechanisms will be taken in the context of the 3 foreseen editions of the Business and Exploitation Plan in the context of WP8 (deliverables D8.3 due at month 12, D8.6 due at month 24 and D8.10 due at month 34).
As a general principle, the GA article 26.1 is faithfully adopted in the PoliVisu Consortium Agreement (CA), according to which "*Results are owned by the Party that generates them*". This is what article 8.1 states. And in addition, article 8.2 specifies that "*in case of joint ownership, each of the joint owners shall be entitled to Exploit the joint Results as it sees fit, and to grant non-exclusive licences, without obtaining any consent from, paying compensation to, or otherwise accounting to any other joint owner, unless otherwise agreed between the joint owners*".

We take the above provisions also as a **guideline for the attribution of responsibilities of data management**, as far as PoliVisu research results are concerned. Very shortly, we posit that **ownership goes hand in hand with the responsibility for data management**. The latter involves the same project partner(s) who generate new data, individually or jointly. In case of reuse of existing data, i.e. owned by someone else (a third party or another PoliVisu partner), the individual or joint responsibility is to **check the nature of data** (as specified in Figure 1 above) and **undertake the consequent actions** as will be further described also in Section 3 below.

1.3 Alignment to the Principles of FAIR Data Handling

Generally speaking, a good DMP under H2020 should comply with the FAIR Data Handling Principles. FAIR stands for Findable, Accessible, Interoperable and Re-usable, as referred to a project's research outputs – notably those made available in digital form.

The FAIR principles, however, do not belong to H2020 or the EC but have emerged in January 2014, as the result of an informal working group convened by the Netherlands eScience Center and the Dutch Techcentre for the Life Sciences at the Lorentz Center in Leiden, The Netherlands[7].

Very pragmatically, the European Commission (2016) considers the FAIR principles fulfilled if a DMP includes the following information:

A. *"The handling of research data during and after the end of the project"*
B. *"What data will be collected, processed and/or generated"*
C. *"Which methodology and standards will be applied"*
D. *"Whether data will be shared/made open access", and*
E. *"How data will be curated and preserved (including after the end of the project)".*

In the case of PoliVisu, the above information is provided in Section 3 of this document, which consists of five paragraphs, respectively:

3.1 Data summary (*typologies and contents of data collected and produced*)
3.2 Data collection (*which procedures for collecting which data*)
3.3 Data processing (*which procedures for processing which data*)
3.4 Data storage (*data preservation and archiving during and after the project*)
3.5 Data sharing (*including provisions for open access*)

The following table matches the aforementioned EC requirements with the contents dealt with in Section 3 paragraphs.

Table 1. Alignment between this DMP and the EC's requirements

This document's Section 3 TOC EC requirements	3.1 Data Summary	3.2 Data Collection	3.3 Data Processing	3.4 Data Storage	3.5 Data Sharing
A. "The handling of research data during and after the end of the project"					
B. "What data will be collected, processed and/or generated"					
C. "Which methodology and standards will be applied"					
D. "Whether data will be shared/made open access"					
E. "How data will be curated and preserved (including after the end of the project)"					

[7] For more information, one can visit: https://www.force11.org/fairprinciples

This Introduction has presented PoliVisu's data management lifecycle and frames the DMP within the EU H2020 Guidelines and FAIR data handling principles. The remaining structure of this document comes as follows:

- **Section 2** is a brief overview of the legal framework, including the EU regulation on personal data protection (GDPR), the H2020 provisions for open access to research data, the specific provisions of the PoliVisu grant agreement and consortium agreement and some special provisions for big data.
- **Section 3** presents and discusses the data usage scenarios in the framework outlined in the above Table and examines the key issues in relation to each scenario. These issues include decisions on e.g. data anonymization, privacy and security protection measures, licensing etc.
- **Section 4** concludes the document by anticipating the expected contents of the next and last edition of this DMP.
- In **Annex 1** the interested reader can find a running list of utilized / relevant data sources, which will be further updated (if and where necessary) over the course of the project.

2. Legal framework

This Section briefly overviews the key normative references making up the DMP external context. The next paragraphs respectively deal with:

1. The PSI Directive and its recent modifications and revisions proposals (dated April 2018);
2. The General Data Protection Regulation, coming into force in May 2018;
3. The terms of the H2020 Open Research Data Pilot (ORDP) the PoliVisu consortium has adhered to;
4. The resulting, relevant provisions of both the Grant and the Consortium Agreements;
5. The special provisions for big data management mentioned in the DoA and thus become binding for all partners;
6. A general outline of PoliVisu's licensing policy.

In the following text, we reproduce the contents of Section 2 of D2.11 almost unchanged, apart from some changes in subsection 2.7 aligned to the subsequent release of Deliverable 6.2.

2.1 The PSI Directive

The Directive 2003/98/EC on the re-use of Public Sector Information (PSI) entered into force on 31 December 2003. It was revised by the Directive 2013/37/EU, which entered into force on 17 July 2013. The consolidated text resulting from the merge of these two legislative documents is familiarly known as the PSI Directive, and can be consulted on the Eur-Lex website[8].

On 25 April 2018, the EC adopted a proposal for a revision of the PSI Directive, which was presented as part of a package of measures aiming to facilitate the creation of a common data space in the EU. This review also fulfils the revision obligation set out in Article 13 of the PSI Directive. The proposal has received a positive opinion from the Regulatory Scrutiny Board and is now being discussed with the European Parliament and the Council. It comes as the result of an extensive public consultation process, an evaluation of the current legislative text and an impact assessment study done by an independent contractor[9].

The current PSI Directive and its expected evolution is noteworthy and useful to define the context of the PoliVisu project in general and of this DMP in particular. Thanks to the PSI Directive and its modifications and implementations[10], the goal of making government data and Information reusable has become shared at broad European level. In addition, the awareness has been remarkably growing that as a general principle, the datasets where PSI is stored must be set free by default. However, fifteen years after the publication of the original PSI Directive, there are still barriers to overcome (better described in the aforementioned impact assessment study) that prevent the full reuse of government data and information, including data generated by the public utilities and transport sectors as well as the results from public funded R&D projects, two key areas of attention for PoliVisu and this DMP.

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02003L0098-20130717

[9] Available online at: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51491

[10] For instance, the INSPIRE Directive (2007/2/EC) builds mechanisms for data and corresponding Web services discoverability on top of the PSI Directive. See: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32007L0002&from=en

2.2 The EU Personal Data Protection Regulation (GDPR)

Regulation (EU) 2016/679 sets out the new General Data Protection Regulation (GDPR) framework in the EU, notably concerning the processing of personal data belonging to EU citizens by individuals, companies or public sector/non government organisations, irrespective of their localization. It is therefore a primary matter of concern for the PoliVisu consortium.

The GDPR was adopted on 27 April 2016, and became enforceable on 25 May 2018, after a two-year transition period. By then, it will replace the current Data Protection Directive (95/46/EC) and its national implementations. Being a regulation, not a directive, GDPR does not require Member States to pass any enabling legislation and is directly binding and applicable.

The GDPR provisions do not apply to the processing of personal data of deceased persons or of legal entities. They do not apply either to data processed by an individual for purely personal reasons or activities carried out at home, provided there is no connection to a professional or commercial activity. When an individual uses personal data outside the personal sphere, for socio-cultural or financial activities, for example, then the data protection law has to be respected.

On the other hand, the legislative definition of personal data is quite broad, as it includes any information relating to an identifiable individual, whether it relates to his or her private, professional or public life. It can be anything from a name, a home address, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer's IP address.

While the specific requirements of GDPR for privacy and security are separately dealt with in other PoliVisu Deliverables - such as D1.1 on POPD Requirement No. 4 [June 2018] and D1.2 on POPD Requirement No. 6 [March 2018], as well as D4.5 [October 2019] & D4.6 [due in April 2020], both on Privacy Rules and Data Anonymization - it is worth noting here that the PoliVisu consortium has formed a working group composed of the partner organisations Data Protection Officers (DPOs). The DPO function and role has been introduced by the GDPR and better defined by a set of EC guidelines, given on 13 December 2016 and revised on 5 April 2017[11].

The GDPR text is available on the Eur-Lex website[12].

2.3 Open Access in Horizon 2020

As partly anticipated in Section 1, the EC has launched in H2020 a flexible pilot for open access to research data (ORDP), aiming to improve and maximise access to and reuse of research data generated by funded R&D projects, while at the same time taking into account the need to balance openness with privacy and security concerns, protection of scientific information, commercialisation and IPR. This latter need is crystallised into an opt-out rule, according to which it is possible at any stage - before or after the GA signature - to withdraw from the pilot, but legitimate reasons must be given, such as IPR/privacy/data protection or national security concerns.

With the Work Programme 2017 the ORDP has been extended to cover all H2020 thematic areas by default. This has particularly generated the obligation for all consortia to deliver a Data Management Plan (DMP), in which they specify what data the project will generate, if it will not be freely disclosed for e.g. exploitation related purposes or how it will be made accessible for verification and reuse, and how it will be curated and preserved.

[11] See: http://ec.europa.eu/newsroom/document.cfm?doc_id=44100

[12] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679

The ORDP applies primarily to the data needed to validate the results presented in scientific publications. Other data can however be provided by the beneficiaries of H2020 projects on a voluntary basis.
The costs associated with the Gold Open Access rule, as well as the creation of the DMP, can be claimed as eligible in any H2020 grant.
As already mentioned, the PoliVisu consortium has adhered to the **Green Open Access** rule.

2.4 Grant Agreement and Consortium Agreement provisions

The key GA and CA provisions worth mentioning in relation to our discourse on data management have already been introduced to a great extent in the previous Sections. Now we simply reproduce the corresponding articles.

2.4.1 Grant Agreement

24.1 Agreement on background
The beneficiaries must identify and agree (in writing) on the background for the action ('agreement on background').
'Background' means any data, know-how or information — whatever its form or nature (tangible or intangible), including any rights such as intellectual property rights — that:
(a) is held by the beneficiaries before they acceded to the Agreement, and
(b) is needed to implement the action or exploit the results.

26.1 Ownership by the beneficiary that generates the results
Results are owned by the beneficiary that generates them.
'Results' means any (tangible or intangible) output of the action such as data, knowledge or information — whatever its form or nature, whether it can be protected or not — that is generated in
the action, as well as any rights attached to it, including intellectual property rights.

26.2 Joint ownership by several beneficiaries
Two or more beneficiaries own results jointly if:
(a) they have jointly generated them and
(b) it is not possible to:
(i) establish the respective contribution of each beneficiary, or
(ii) separate them for the purpose of applying for, obtaining or maintaining their protection.

29.1 Obligation to disseminate results
Unless it goes against their legitimate interests, each beneficiary must — as soon as possible — 'disseminate' its results by disclosing them to the public by appropriate means (other than those resulting from protecting or exploiting the results), including in scientific publications (in any medium).

29.2 Open access to scientific publications
Each beneficiary must ensure open access (free of charge online access for any user) to all peer-reviewed scientific publications relating to its results.

29.3 Open access to research data

Regarding the digital research data generated in the action ('data'), the beneficiaries must:

(a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:

(i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;

(ii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan');

(b) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

(...)

As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex 1, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.

39.2 Processing of personal data by the beneficiaries

The beneficiaries must process personal data under the Agreement in compliance with applicable EU and national law on data protection (including authorisations or notification requirements).

The beneficiaries may grant their personnel access only to data that is strictly necessary for implementing, managing and monitoring the Agreement.

2.4.2 Consortium Agreement

Attachment 1: Background included

According to the Grant Agreement (Article 24) Background is defined as "data, know-how or information (...) that is needed to implement the action or exploit the results". Because of this need, Access Rights have to be granted in principle, but Parties must identify and agree amongst them on the Background for the project. This is the purpose of this attachment[13].

(...)

As to EDIP SRO, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)

Algorithms for the analysis of data characterizing the traffic flow from automatic traffic detectors.

Mathematical model of traffic network of roads in the Czech Republic, including car traffic matrix.

(...)

As to HELP SERVICE REMOTE SENSING SRO, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)

Metadata Catalogue Micka.

Senslog Web Server.

HSLayers NG.

Mobile HSLayers NG Cordova.

[13] Note: the partners not listed in this summary of Attachment 1 have declared not to have any background knowledge to share. While this can be true for software, models and algorithms, a comparable assessment is missing for the datasets that esp. the Cities intend to put at consortium's disposal for the execution of project pilots. Reasonably, that information was not available by the time the CA was prepared and signed. More details will be made available within the next editions of this DMP.

VGI Apps.
(...)
As to GEOSPARC NV, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
geomajas (http://www.geomajas.org).
INSPIRE>>GIS view & analysis component.
(...)
As to INNOCONNECT SRO, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
WebGLayer library (available at http://webglayer.org/).
(...)
As to CITY ZEN DATA, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
Warp10 platform (www.warp10.io).
(...)
As to ATHENS TECHNOLOGY CENTER SA, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
TruthNest, which will be integrated as a service within PoliVisu through an API to be provided by ATC
(...)
As to SPRAVA INFORMACNICH TECHNOLOGII MESTA PLZNE, PRISPEVKOVA ORGANIZACE, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
Mathematical model of traffic network of roads in the city of Pilsen, including a car traffic matrix (so-called CUBE software: http://www.citilabs.com/software/cube/).
(...)
As to MACQ SA, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
M3 Demo version in Macq's cloud for development, not allowed to put online or in production.
Excluded: background and especially data which is not owned by Macq or which it is not allowed to share.
(...)
As to PLAN4ALL ZS, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
Smart Points of Interest (http://sdi4apps.eu/spoi/).
Open Transport Map (http://opentransportmap.info/).
Open Land Use Map (http://sdi4apps.eu/open_land_use/).
(...)
As to STAD Gent, it is agreed between the Parties that, to the best of their knowledge, The following background is hereby identified and agreed upon for the Project: (...)
Any software developed for the publication, analysis, harmonisation and/or storage of data by the City, its ICT partner Digipolis, or any subcontractor thereof.
(...)

2.5 The PoliVisu licensing policy

There is at the moment no single licensing policy within the PoliVisu consortium, either for the software (so-called Playbox) or their individual components, some of which belong to the Background as mentioned in the previous subparagraph. However, a few building blocks can already be identified, based on the discussion done in this document, the GA provisions quoted above as well as others not quoted yet, and the individual partners declarations in the CA. These provisions have been implicitly accepted by the PoliVisu consortium members upon their signature of the aforementioned documents and are therefore totally enforceable. They are summarized in the table below, which also mentions in a footnote the examples currently dealt with in the PoliVisu pilot cities.

Table 2. Building blocks of the PoliVisu licensing policy

Typology of data	Licensees	During the project period	After the project period	Legal references
Pre-existing (e.g. part of the Background knowledge of PoliVisu, as listed in the CA Attachment 1)	Other members of the PoliVisu consortium	Royalty free usage No right to sublicense	Under fair and reasonable conditions	GA Art. 25.2 GA Art. 25.3
	Any interested third party	As per the Background commercial licence	As per the Background commercial licence	CA Attachment 1
Sourced from third parties for the execution of project activities (e.g. portions of large datasets)[14]	Other members of the PoliVisu consortium	Royalty free usage No right to sublicense	Within the scope of the third party's license	General rules on IPR and license details
	Any interested third party	No right to sublicense	No right to sublicense	General rules on IPR and license details
Freely available in the state of the art (e.g. Open Data)	Other members of the PoliVisu consortium	Royalty free usage	Royalty free usage	Within the scope of the data owner's license
	Any interested third party	Royalty free usage	Royalty free usage	Within the scope of the data owner's license
Newly produced[15] during the project (i.e. part of the Foreground knowledge of PoliVisu)	Other members of the PoliVisu consortium	Royalty free usage No right to sublicense	Under fair and reasonable conditions	GA Art. 26.2
	Any interested third party	Open access at flexible conditions	Open access at flexible conditions	GA Art. 29.3

The use of different licences has not materially inhibited work on the Polivisu project or the combination of different datasets into consolidation visualisation. Nevertheless in the final phase of the project this experience will be reviewed and then discussed at the broad consortium level. The results of this discussion and recommendations for future projects will be added to the next and final edition of this Deliverable.

[14] ANPR dataset (Flanders), Floating car data (Issy); Cellular data (Gent); Citizen science data (Mechelen); Accident data (Flanders, Pilsen); Traffic modelling data (Flanders, Pilsen); Traffic counting data (Pilsen).

[15] Special provisions must apply to the case, disciplined by the GA Art. 26.3, of a third party producing Foreground knowledge during the project activities.

2.6 Special provisions for big datasets

The PoliVisu DoA describes how big data from different sources – notably available at city level, in relation to the nature of the identified project pilots, dealing with mobility and traffic flows – can distinctively contribute to the three processes of policy experimentation belonging to its Framework: design, implementation and (real time) evaluation of policy solutions[16].

Big data, as defined in ISO/IEC CD 2046, is data stored in "extensive datasets − primarily in the characteristics of volume, variety, velocity, and/or variability − that require a scalable architecture for efficient storage, manipulation, and analysis". This may include 'smart data', i.e. coming from sensors, social media, and other human related sources. This obviously raises questions about data security and privacy, which are explicitly and extensively dealt with in a dedicated WP (1) and will ultimately become part of a policy oriented manual, issued in two consecutive editions as Deliverables D7.4 (due October 2019) and D7.6 (due June 2020).

In another WP (4), the PoliVisu DoA extensively deals with the smart data infrastructure for cities that is now going to be developed within the project. This is based on the Warp 10 big data architecture and will set up various data processing and analytical steps. The general principle and modus operandi is that any (big) data can be used in any application, can be analysed and correlated with other sources of data and can be used to provide detection of patterns to understand the effective functioning of infrastructures, transport systems, services or process within a city. The processed and analysed big data will be published as map services. Free and open source geospatial tools and services will be used to generate OGC standards (especially WMS-T and WFS), TMS and vector tile based open formats for integration in GIS applications.

The existing OTN traffic modelling tool will be automated and ported to a big data processing cloud to yield near-real-time traffic calculations. The process will be calibrated to make the traffic model algorithms more accurate (in space and time) using real time and historical traffic sensor data. System interfaces and GUI will be developed to interact with the traffic modelling software.

Existing crowdsourcing tools (such as Waze and plzni.to) will be adopted and complemented with standard interfaces, protocols and data models to turn user generated data into actionable evidence for policy making. New modules will be designed for the SensLog open source library to support its integration with big data technologies.

Data analytics functions and algorithms will be implemented to support policy making processes. Social Media analytics will be based on TruthNest or TrulyMedia as an alternative option. This tool will be extended with a monitoring mechanism for Twitter contents that gathers any information on mobility trends automatically and in real-time and sends alerts to users on possible events.

Open source geospatial software (such as WebGLayer) will be used to realise the big data visualisation. The tool will be extended with support for line and area features. Advanced visualisation components will be added in the form of multiple linked views, filters through interactive graphs, parallel coordinates relationship analysis, map-screen extent filters, and area selection. Focus will be set on the visualisation and filtering of mobility related information and the comparison between different scenarios, time periods and locations, in particular on mobile and touch devices.

The appropriate metadata will be defined for supporting the different tools and processes in real life decision making conditions. This includes the structures, services, semantics and standards to support big data, sensor data, advanced analytics and linked data. Two open source metadata tools will be considered in the project: GeoNetwork and Micka. The consortium will contribute to the definition of integrated metadata standards in the OGC metadata workgroup.

Considering the above scenario, as well as the DoA statement that "PoliVisu will treat the data as

[16] For more details on the PoliVisu framework, the reader is referred to Deliverable D3.2 "The Polivisu Policy Making Model (DRAFT)".

confidential and will take every precaution to guarantee the privacy to participants, i.e., ensuring that personal data will be appropriately anonymised and be made inaccessible to third parties" (Part B, p. 102) the resulting, natural implication is that a number of anonymization, aggregation, and blurring techniques must be tested well in advance, and applied to sourced and produced datasets in dependence of the requirements of the various project pilots. The results of this effort will be released as two WP4 Deliverables, notably a White Paper on data anonymisation issued in two consecutive editions, D4.5 at month 24 and D4.6 at month 30.

However, due to the key role played by anonymization in the context of the PoliVisu project and the need to balance privacy and security with the policy (end user) requirements of having usable datasets for e.g. traffic flows measurement, detection of trends, or sentiment analysis, in the first edition of the DMP it was highly recommended that the contents of this Section be updated and integrated when this edition of the DMP would be published, namey at month 12 of the work plan. This is the purpose of the following subsection.

2.7 Special provisions for protecting personal data

The main conclusion of D6.2 and the internal partner survey done in preparation for it, has been that the majority of relevant datasets for the three pilot cities are already available and those that aren't have the potential to become so. However, further analyses and data preparations are needed before the partners can combine available (and for the moment unavailable/unusable) datasets, because of the wide variety in data standards and the differentiating quality of the datasets. Among these additional steps towards achieving 'data readiness', anonymization is a key process component. According to Deliverable 6.2 conclusions, this aspect is particularly relevant in the case of Gent, and more specifically for the student housing datasets. In any case, the next statements are general in principle and need to be adopted and translated into practice by all the members of the PoliVisu consortium.

2.7.1 Overview of available approaches

Personal data is any information which is related to an identified or identifiable natural person. Wherever personal data is collected or acquired it is essential that measures are taken to protect it from unauthorised or unlawful use (including unlawful processing by entities authorised to have access to it for other purposes) or that the data is "de-personalised" so that none of the data can be related to an identified or identifiable natural person even with the use of other data. Indeed, many approaches are known from the state of the art, including (non exhaustively) the following:

- **Attribute suppression**: consisting in the removal of an entire part of a dataset (also referred to as "column" in spreadsheets);
- **Record suppression**: consisting in the removal of some entire records in a dataset (also referred to as "row" in spreadsheets);
- **Data perturbation**: consisting in a modification of some values from the original dataset to be unpredictably different from the original ones (this prevents these values being matched with the known values for identifiable individuals);
- **Data filtering**: the process of choosing a smaller part of a dataset and using that for viewing or analysis purposes, in more easily controllable conditions;
- **Data aggregation**[17]: achieved by converting a dataset from a list of records to summarised values for a sufficiently large group of individuals (although care needs to be taken that each group consists of

[17] There has been a substantial discussion of this technique during the preparation of D4.5, released in October 2018.

a sufficiently large number of people and that the aggregated data does not allow identification of sensitive information about the group, such as the location of a women's refuge);

- **Character masking**: achieved by using a constant symbol (e.g. "*" or "x") in the place of a data point - typically, a name/surname string or other identification elements of a record;
- **Anonymisation**[18]: the process of removing information that can be used to identify individuals from data sets (although just removing identification information itself may leave the possibility of using other information, in combination with other datasets, to identify the individuals to which the records relate);
- **Pseudonymisation**[18]: achieved by using a pseudonym in the place of a name/surname string, with or without the possibility of coming back to the original identity of the data subject(s) - depending on whether the data owner decides to keep track of the matching and store that information in a secure place. Pseudonymised datasets are usually still personal data, either because the pseudonymisation can be reversed or that other knowledge of the identification data of a data subject can be used to regenerate the pseudonymisation and so find person data relating to that individual;
- **Generalisation**: achieved by deliberately reducing the precision of available data through e.g. converting a person's age into an age range, or a precise location into a less precise location. This technique is also referred to as blurring;
- **Swapping**: consisting in the rearrangement of items in the dataset such that the individual attribute values are still represented therein, but generally, do not correspond to the original records. This technique is also referred to as shuffling and permutation (however if records can still be associated with individuals this may breach the data protection principle of accuracy of an individual's data);
- **Creation of a "synthetic" dataset**: achieved by using the original data to generate a wholly different and depersonalised collection, instead of modifying the original dataset.

2.7.2 Limitations and risks

Generally speaking, all the above techniques may be useful in achieving the goal of protecting the personal identity of the data subjects against some risks. However, the overall approach to any specific dataset needs to be carefully considered in two aspects:

- no single technique is foolproof in achieving protection of personal data by itself in all circumstances, particularly where other data could be used in combination to related data to identified individuals.
- each technique has an impact on the integrity and usability of the underlying dataset; although fully de-personalised data is the safest approach it may make some uses of the data (for instance linking records to be able to analyse whole episodes) more difficult.

A further consideration is that all the aforementioned approaches are **static** - in the sense that they refer to an ideal (perhaps too idealistic) situation whereby the dataset to be anonymised is fully available at the time, has a standard format and is clean and complete; thus, the process of anonymisation can be successfully performed rather easily with the support of one or more methods and tools and especially once for all. Unfortunately this is not the case in the normal practice, where datasets are incomplete, "dirty", subject to continuous revision and upgrade, including in few (but not infrequent) cases the addition of new records/attributes and/or the initiation of new forms of data collection such as audio/video, images, texts, geolocation, biometrics etc.

In such a **dynamic** - not to say all chaotic - situation that characterizes real life, even the implementation of a sound initial design for the protection of personal data could prove inappropriate. In this difficult context, the best (and perhaps the only) defence is procedural: which means that in the case of PoliVisu pilots, the

[18] There has been a substantial discussion of this technique during the preparation of D4.5, released in October 2018.

preparatory stage outlined in Deliverable 6.2 should be complemented by the roll-out of a visibly and transparently implemented process for the protection of personal data. This is expected to integrate the use of other techniques mentioned in D6.2 such as data cropping, data cleansing and ETL, social media scraping and time series creation.

In Deliverable 4.5 this conclusion has been further developed into some key principles for the protection of privacy including:

- the need to seek expert advice;
- the need to have a holistic privacy design encompassing processes, technologies and skills (rather rely on an individual technical means);
- the adoption of the precautionary principle - it is better to over-protect privacy than under-protect it;
- de-personalisation should take place as close to the origin of the data as possible, to minimise the amount of personal data held and to allow techniques and procedures to protect personal data to be focused on where the holding of personal data is unavoidable and because it is most needed;
- careful consideration of the privacy aspects of data acquired from third parties.

3. PoliVisu Data Management Plan

In this Section, the data usage scenarios presented in the Introduction are used as a basis for discussing the key issues to be examined in relation to each distinct paragraph of the PoliVisu DMP. As a reminder, the three scenarios, which jointly compose the PoliVisu's data management lifecycle, are:

- Original data produced by the PoliVisu consortium and/or individual members of it (e.g. during a dissemination action or a pilot activity);
- Existing data already in possession of the PoliVisu consortium and/or individual members of it prior to the project's initiation;
- Existing data sourced/procured by the PoliVisu consortium and/or individual members of it during the project's timeline.

On the other hand, the datasets handled within the three above scenarios can belong to either of these three categories:

- Confidential data (for business and/or privacy protection);
- Anonymised and Public data (as explained in the Introduction, these two aspects go hand in hand);
- Non anonymised data (the residual category).

In the following text, we reproduce the contents of Section 3 of D2.11 almost unchanged, apart from some modifications dictated by the need for better alignment with the recently released D4.5 (Privacy Rules and Data Anonymisation). More specifically, we qualify the data usage scenarios with reference to the six main project pilots, namely: ANPR dataset (Flanders), Floating car data (Issy); Cellular data (Gent); Citizen science data (Mechelen); Accident data (Flanders, Pilsen); Traffic modelling data (Flanders, Pilsen); Traffic counting data (Pilsen). For a more detailed description of these scenarios, the reader is referred to Annex 1 to this document as well as to Deliverable 4.5 itself.

It is interesting to note here that, although inspired by different aims - i.e. to give more insights into the practical decisions to be taken for the preservation of integrity and security of the personal data owned by the PoliVisu consortium and/or individual partners thereof - the research stream on privacy has led to similar conclusions than the one of data management, notably that careful planning is required from data owners, starting from the very first collection act, and always leading to minimise the amount of personal data that is needed for subsequent processing.

3.1 Data summary

The following table summarizes the typologies and contents of data collected and produced. For each distinct category, a consolidated list is provided in Annex 1 to this document.

Table 3. Summary of relevant data for the PoliVisu research agenda

Nature of datasets Data usage scenarios	Confidential	Anonymised and Public	Non anonymised
Original data produced by the PoliVisu consortium	Raw survey/interview/sensor data	Summaries of surveys/interviews Data in reports of pilot activities End user data on public display	Photos/videos shot during public events Audio recordings (e.g.

	Evidence from project pilots Personal data of end users New contacts established	Contact data within deliverables	Skype) Data in internal repositories
Existing data already in possession of the PoliVisu consortium and/or partners	Data embedded in some of the Background solutions (see par. 2.4.2 above) Contact databases	Data embedded in some of the Background solutions (see par. 2.4.2 above) Website logs and similar metrics	N/A
Existing data sourced/procured by the PoliVisu consortium and/or partners	Raw data in possession of the Cities or of any third party involved in the pilots	Free and open data (including from scientific and statistical publications)[19]	N/A

The main implications of the above table for the three usage scenarios are the following, in **decreasing order of urgency** for the related action lines as well as **increasing order of gravity** for the consequences of any inadvertent behaviour by the members of the consortium:

- The organisation of Living Lab experimentations (as foreseen by the project's work plan) implies that personal data handling of the end users acting as volunteers must be carefully considered, also for their ethical implications.
- For any photos/videos shot during public events, it is crucial to collect an **informed consent note**[20] from all the participants, with an explicit disclaimer in case of intended publication of those personal images on e.g. newspapers, internet sites, or social media groups. This will bring the data back into the Confidential category, where it is legitimate to store and/or process it for legitimate reasons.
- For any audio recordings stored, e.g. in the project's official repository (currently Google Drive) or in individual partners' repositories, care must be taken of the risk of involuntary disclosure and/or the consequences of misuse for any unauthorized purpose. Same goes for the personal data of each partner in the consortium.
- Informed consent forms must be signed (also electronically) by all participants in surveys, interviews and/or pilot activities. As an alternative option, the partner in charge will commit to anonymisation and other related measures as a way to protect the identity of the respondents/pilot users.
- Informed consent forms are also required when using available contacts (be they preexisting to the project or created through it) to disseminate information via e.g. newsletters or dedicated emails. In this respect, the GDPR provisions are particularly binding and must be carefully considered, at least in any doubtful case.
- As a general rule, access conferred to Background knowledge on a royalty free basis during a project execution does not involve the right to sublicense. Therefore, attention must be paid by each partner of PoliVisu to ensure the respect of licensing conditions at any time and by any member of the team.
- This also applies to any dataset sourced or procured from third parties during the PoliVisu project's lifetime.

[19] ANPR dataset (Flanders), Floating car data (Issy); Cellular data (Gent); Citizen science data (Mechelen); Accident data (Flanders, Pilsen); Traffic modelling data (Flanders, Pilsen); Traffic counting data (Pilsen).

[20] Informed means that consent cannot be given before reading and understanding the provisions of extant legislation on personal data protection and the way the partner asking for signature commits to enforcing / abiding by it.

3.2 Data collection

The following table summarizes the procedures for collecting project related data. Annex 1 to this document ("Running list of data sources") provides some concrete examples of data usage scenarios.

Table 4. Summary of PoliVisu data collection procedures

Nature of datasets Data usage scenarios	Confidential	Anonymised and Public	Non anonymised
Original data produced by the PoliVisu consortium	Surveys Interviews Pilot activities F2F / distant interaction	Newsletters Publications Personal Emails Open Access repositories	Events coverage - directly or via specialised agencies A/V conferencing systems Internal repositories
Existing data already in possession of the PoliVisu consortium and/or partners	Seamless access and use during project execution	Seamless access and use during project execution	N/A
Existing data sourced/procured by the PoliVisu consortium and/or partners	Licensed access and use during project execution	Free and open access and use during project execution	N/A

An implication of the above table that may not have been evident in the previous one, is that **every partner is responsible for the behaviour of all team members**, which may also include subcontracted organisations (e.g. specialised press agencies) or even volunteers. The latter circumstance does not exempt the delegate of a certain job in case of improper application of extant norms and rules.
All data will be collected in a digital form – therefore CSV, PDF, (Geo)JSON, XML, Shape, spreadsheets and textual documents will be the prevalent formats. In case of audio/video recordings and images, the most appropriate standards will be chosen and adopted (such as .gif, .jpg, .png, .mp3, .mp4, .mov and .flv). Ontologies will be created in Protégé file format (.pont and .pins) or .xml/.owl can also be used. Website pages can be created in .html and/or .xml formats.
Individually, each research output will be of manageable size to be easily transferred by email. However, it is important to note that email transfer can become a violation of confidentiality under certain circumstances.

3.3 Data processing

The following table summarizes the procedures for processing PoliVisu related data that can be envisaged at this project's stage. As one can see, most of them make reference to the contents of paragraph 2.6 above. In this sense, more can probably be added to the cells of the table. Annex 1 to this document ("Running list of data sources") provides some meaningful case descriptions.

Table 5. Summary of PoliVisu data processing procedures

Nature of datasets Data usage scenarios	Confidential	Anonymised and Public	Non anonymised

Original data produced by the PoliVisu consortium	Anonymisation Visualisation	Statistical evaluation Visualisation	Selection/destruction Blurring of identities
Existing data already in possession of the PoliVisu consortium and/or partners	Anonymisation Statistical evaluation Metadata generation	Visualisation Analytics Publication as map services	N/A
Existing data sourced/procured by the PoliVisu consortium and/or partners	Anonymisation Statistical evaluation Metadata generation	Visualisation Analytics Publication as map services	N/A

Apart from the specific software listed in paragraph 2.6 above, state of the art productivity tools will be used to process/visualize the data used or generated during the project. Typically, the partners are left free to adopt their preferred suite (such as Microsoft Office™ for PC or Mac, Apple's iWork™ and OpenOffice™ or equivalent). However, the following tools are the ones mainly used by the consortium[21]:

- Google's shared productivity tools (so-called G-Suite™) are used for the co-creation of outputs by multiple, not co-located authors.
- Jira™ is a proprietary issue tracking product developed by Atlassian that allows bug tracking and agile project management.
- Adobe Acrobat™ or equivalent software is used to visualise/create the PDF files.
- Protégé™ or equivalent software is used to generate the ontologies.
- Photoshop™ or equivalent software are used to manipulate images.
- State of the art browsers (such as Mozilla Firefox™, Google Chrome™, Apple Safari™ and Microsoft Internet Explorer™) are used to navigate and modify the Internet pages, including the management and maintenance of social media groups.
- Cisco Webex™ or Skype™ (depending on the number of participants) are the selected tools for audio/video conferencing, which may also serve to manage public webinars.
- Tools like Google Forms™, and optionally SurveyMonkey™ and LimeSurvey™, are used for the administration of online surveys with remotely located participants.
- Dedicated Vimeo™ or YouTube™ channels can help broadcast the video clips produced by the consortium to a wider international audience, in addition to the project website.
- Mailchimp™ or equivalent software is helpful to create, distribute and administer project newsletters and the underlying mailing lists.

3.4 Data storage

The following table summarizes the procedures for storing project related data, during and after the PoliVisu lifetime, and the most frequently used repositories. As for the previous paragraphs, we limit ourselves now to listing the headlines. Annex 1 to this document ("Running list of data sources") provides some contents for the pilot descriptions.

Table 6. Summary of PoliVisu data storage procedures

Nature of datasets	Confidential	Anonymised and Public	Non anonymised

[21] This list is non-exhaustive and can be updated at all times without prior advice.

Data usage scenarios			
Original data produced by the PoliVisu consortium	Individual partner repositories Common project repository	Project website Open access repository	Individual partner repositories Common project repository
Existing data already in possession of the PoliVisu consortium and/or partners	Specific software repositories	Playbox components Map services	N/A
Existing data sourced/procured by the PoliVisu consortium and/or partners	Individual partner repositories Third party repositories Cloud repositories	Playbox components Map services Cloud repositories	N/A

Google Drive™ is the selected tool as Polivisu's data and information repository. This include both the project deliverables (including relevant references utilised for their production or generated from them as project publications, e.g. journal articles, conference papers, e-books, manuals, guidelines, policy briefs etc.) and any other related information, including relevant datasets. This implies that the privacy and security measures of Google Drive™ must be GDPR compliant. The verification of such circumstance is the responsibility of the coordinator.

Additionally, the coordinator will make sure that the official project repository periodically generates back-up files of all data, in case anything may get lost, corrupted or become unusable at a later stage (including after the project's end). The same responsibility goes to each partner for the local repositories utilised by them (in some cases, these are handled by large organisations such as Universities or Municipalities; in others, by SME or even personal servers or laptops).

Collectively, we expect the whole set of outputs to reach the size of 500-600 Gb all along the project duration. This range will particularly depend on the number and size of the received datasets to be utilised for the execution of PoliVisu pilots.

Whatever the license that the consortium establishes for final datasets, their intermediate versions will be deemed as **business confidential**, and restricted to circulating only within the consortium.

Finally and as stipulated in the DoA, each digital object identified as R&D result, including their associated metadata, will be stored in a dedicated open access repository, namely OpenAIRE/Zenodo[22], under the care of POLIMI, to the purpose of both preserving that evidence and making it more visible and accessible to the scientific, academic and corporate world.

In addition to the above, other datasets may be stored on the following repositories:

- Cordis, through the EU Sygma portal
- The PoliVisu website (with links on/to the Social Media groups)
- Individual Partner websites and the social media groups they are part of
- The portals of the academic publishers where scientific publications will be accepted
- Other official sources such as EUDAT[23]
- Consortium's and Partners' press agencies and blogs
- PoliVisu official newsletters.

[22] https://www.zenodo.org/communities/ecfunded/?page=1&size=20
[23] https://eudat.eu/what-eudat

3.5 Data sharing

Last but not least, the following table summarizes the procedures for sharing PoliVisu related data in a useful and legitimate manner. When sharing, it is of utmost importance to keep in mind, not only the prescriptions and recommendations of extant rules and norms (including this DMP), as far as confidentiality and personal data protection are concerned, but also the risk of voluntary or involuntary transfer of data from the inside to the outside of the European Economic Area (EEA).

In fact, while the GDPR applies also to the management of EU citizens personal data (for business or research purposes) outside the EU, not all the countries worldwide are subject to bilateral agreements with the EU as far as personal data protection is concerned. For instance, the US based organisations are bound by the so-called EU-U.S. Privacy Shield Framework, which concerns the collection, use, and retention of personal information transferred from the EEA to the US. This makes the transfer of data from the partners to any US based organisation relatively exempt from legal risks. This may not be the same in other countries worldwide, however, and the risk in question is less hypothetical than one may think, if we consider the case of personal sharing of raw data with e.g. academic colleagues being abroad for the purpose of attending a conference. It is also for this reason that the sharing of non-anonymised data is discouraged for whatever reason, as shown in the table.

Table 7. Summary of PoliVisu data sharing procedures

Nature of datasets Data usage scenarios	Confidential	Anonymised and Public	Non anonymised
Original data produced by the PoliVisu consortium	Personal email communication Shared repositories	Project website Open access repository	N/A
Existing data already in possession of the PoliVisu consortium and/or partners	Personal email communication Shared access to software repositories	Shared access to Playbox components Map services	N/A
Existing data sourced/procured by the PoliVisu consortium and/or partners	Personal email communication Shared repositories	Shard access to Playbox components Map services	N/A

As for the previous tables, Annex 1 to this document ("Running list of data sources") provides some relevant case descriptions.

4. Conclusions and Future Work

This document is the third in a row of four planned deliverables concerning the PoliVisu Data Management Plan (DMP) in fulfillment of the requirements of WP2 of the project's work plan. The main reason for planning four versions of the DMP (at months 6, 12, 24 and 36) and particularly two of them during the first project year, is evidently related to the need to hold on until the development as well as piloting activities of PoliVisu gain further momentum, in order to:

- Secure the current, proposed structure of contents against any changes suggested by the gradual and incremental start up of the core project activities, and
- Colour the already existing contents with important add-ons based on the learning process that the Polivisu partners will activate throughout the project's lifetime, considering also that most of project work will be oriented to operationalizing the connection between data handling (including analytics and visualization) and the policy making cycle outlined in deliverable D3.2 (also resting under POLIMI responsibility, like the present one).

The progress achieved with this incremental development of versions can be summarized as follows.

In the first edition of the DMP (Deliverable 2.10) we basically envisaged three main data usage scenarios, which jointly compose PoliVisu's data management lifecycle:

- Original data produced by the PoliVisu consortium and/or individual members of it (e.g. during a dissemination action or a pilot activity)
- Existing data already in possession of the PoliVisu consortium and/or individual members of it prior to the project's initiation
- Existing data sourced/procured by the PoliVisu consortium and/or individual members of it during the project's timeline.

That edition of the DMP has, in our opinion, fulfilled the immediate goals of such a stepwise approach to data management, by:

- Presenting the legislative and regulatory framework, shaping the external context of this DMP in a relatively immutable manner, at least within the timeframe of the PoliVisu project;
- Identifying the fundamental principles of FAIR data handling according to the EC requirements and that the PoliVisu consortium and individual partners are bound to respect;
- Proposing a unitary description of the PoliVisu data management lifecycle, a precise requirement of the DoA and that has been the leitmotif and conceptual architrave of the whole document;
- Summarizing the key aspects of data collection, processing, storage and sharing (the typical contents of a DMP) within the proposed lifecycle elements and particularly highlighting - first and foremost, to the attention of the partners - some key aspects of data management that go beyond the operational link with open access policy (the likely reason why this deliverable has been assigned to POLIMI) and interfere with privacy and security policies (an ethical topic falling under the competence of WP1) as well as with the way background knowledge and tools will be developed, deployed and customised to serve the needs of the city pilots (a topic entirely covered by the WP4 team).

All the above contents originated in D2.10 have not been removed from the second edition of the DMP, D2.11, but only aligned with and supplemented by some extra items, notably reflecting progress in the consortium's and partners' reflections on some aspects left unattended and explicitly mentioned in the conclusions of the Deliverable 2.10, namely:

- Integration of the baseline analysis done in D6.2 until month 7;
- Compilation of Annex 1 "Running list of data sources";

- Provision of case descriptions for Section 3;
- A more careful consideration of data anonymisation related aspects in Section 2.

In so doing, some of the contents left unattended or only partly covered by the previous edition of this DMP have been covered, which reflected:

- A stronger integration of partners contributions in the texts composing this document, although with the twin filter of the responsible author (POLIMI) who provided most of the contents by now, with other partners acting as external reviewers, and the table provided as Appendix to Deliverable 6.2, which has been "scrapped" by the responsible author to create Annex I to this DMP and then cross checked and integrated by the individual partner contributions.
- A deeper connection with the topic of data handling in other deliverables that this one only, and particularly: D1.1 (Ethical Requirement No. 4), D1.3 (Ethical Requirement No. 3), D2.2 (Project Management Plan), D2.3 (Quality and Risk Plan), D4.5 (Privacy Rules and Data Anonymisation), D6.1 (Pilot Scenarios), D7.1 (Evaluation Plan) and D8.1 (Impact Enhancement Road Map). This in order not to miss any precious information while at the same time avoiding duplications and inconsistencies in the respective structure and contents.

In this third edition of the DMP and thanks to the concurrent research work on Deliverable 4.5, significant progress has been made in the definition of the project pilots and therefore the provision of a concrete answer to the question of "who is in charge of" collecting, processing and storing data for each. This aspect includes the identification, again in Annex 1, of the responsible persons for data management in each pilot and a deeper analysis - particularly evident in D4.5 - of the approaches, methods and tools to be used for the preservation of data integrity and the protection of personal privacy to the maximum possible extent allowed by existing techniques.

More generally, it is possible to conclude that the goal set in the previous editions of this DMP has been achieved - enabling all partners engaged in the pilots (including some from outside the PoliVisu consortium) to understand the legal and procedural implications of handling data of different nature, origin and "size".

And even more importantly, this generalised level of awareness is not the result or byproduct of the DMP in itself, which continues to be more a collection of principles than operational guidelines, but of a broader and diffused inspiration that pervades the whole consortium: designing the most appropriate way of dealing with data is essential for the successful achievement of any pilot and project of this kind, and must be done since the early stages of it, not when it gets too late.

In the fourth and last edition of the DMP, two more aspects will be covered in full:

a) The post-project scenario. A missing aspect in this document is the consortium's and individual partners' foresight of the management of pilot related datasets and more generally, of all the datasets created during the project's lifetime that - for legitimate reasons, first and foremost exploitation related - are not subject to immediate publicity and may nonetheless require considerable attention and care to be maintained and preserved.
b) In relation to the above, the definition of an official stance and (to some extent) forward looking plan at consortium level, dealing with the specific licenses associated with the various data categories making up the project results. In that context, the issues of open standards and open data licenses (other than those belonging to the open access scheme) will be more extensively dealt with.

References

European Commission, Directorate-General for Research & Innovation (2016). *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0, July 26th.

European Commission, Directorate-General for Research & Innovation (2017). *H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*, Version 3.2, March 21st.

PoliVisu Consortium Agreement (2017). *Research and Innovation action No. 769608. Project Title: Policy Development based on Advanced Geospatial Data Analytics and Visualisation*. August 29th.

PoliVisu DoA (2017). *Grant Agreement No. 769608. Associated with document Ref. Ares(2017)4445074. Project Title: Policy Development based on Advanced Geospatial Data Analytics and Visualisation*. September 12th.

PoliVisu Deliverable D2.10 *Data Management Plan (Iter. 1)*, April 2018 (contractual delivery date).

PoliVisu Deliverable D6.2 *Baseline Analysis*, May 2018 (contractual delivery date).

PoliVisu Deliverable D2.11 *Data Management Plan (Iter. 2)*, October 2018 (contractual delivery date).

PoliVisu Deliverable D4.5 *Privacy Rules and Data Anonymisation*, October 2019 (contractual delivery date).

Annex 1. Running list of data sources

The following table has been initiated by the PoliVisu partners during a brainstorming session held on 18 May 2018 during a project meeting held in Issy-Les-Moulineaux, France. Its current contents derive from those in the Appendix to Deliverable 6.2 (Baseline Analysis) as described in the text above. They have been updated for the progress of time until the release of this Deliverable. Account has also been taken of the three new cases mentioned in Deliverable 4.5 (Privacy rules and data anonymisation), namely the Mechelen traffic model including ANPR (Automated Number Plate Recognition), the map of traffic accidents in the Flanders and the citizen data sources from Leuven - all separately displayed from the partner cases. As the list of data sources is ongoing, a final update will appear in the next edition of this Deliverable, by project's end.

Partner cases (source: Deliverable 6.2 and modifications)

Nature of datasets	**Used in Gent pilot**	**Used in Issy pilot**	**Used in Pilsen pilot**	**Used in the smart data infrastructure**
Confidential	*New:* n/a *Existing:* 1) Results of screening houses in the centre to identify possible student residences. 2) Student housing inspection list. *Third party's:* 1) Student addresses declared to the Universities. 2) Data on energy usage by address. 3) Telecom data by address. 4) Data of the wifi sniffing equipment	*New:* n/a *Existing:* n/a *Third party's:* 1) Mediamobile data on traffic congestion in IDF (2017 - 2018) 2) TomTom data analysis provided by MyAnatol	New: n/a *Existing:* Data from Traffic Intensity Model (OTM). *Third party's:* 1) Traffic model of the city in CUBE software. 2) National Uniform Traffic Information System data.	*New:* Raw data from the PoliVisu survey on big data for policy making. *Existing:* Data on road quality. *Third party's:* WMS-t containing the current traffic model data.
Anonymised and Public	*New:* Data from periodic bike counts. *Existing:* 1) Internal road database.	*New:* Number of people entering train/metro/RER network (by station) *Existing:* 1) Localisation of parking places for 2	*New:* n/a *Existing:* 1) Location of large parking areas and	*New:* Results of the European survey on big data for policy making. *Existing:* n/a

	2) Bike storage places. *Third party's:* 1) Spatial representation of cadastral parcels. 2) Bus routes and stop locations.	wheel vehicles (bikes & motorbikes). 3) Localisation of bike sharing stations. *Third party's:* 1) Timetables of buses and trains in the Paris area. 2) Real time timetable of buses in the Paris region. 3) Real time timetable of trains in France. 4) Bus lines in the Grand Paris Seine Ouest area. 5) Localisation of stations (Metro, RER and train) in Grand Paris Seine Ouest area. 6) Accessibility of stations (Metro, RER and train) of Grand Paris Seine Ouest area. 7) Map of bus stops with access for people with handicaps. 8) Roads with particular limits (30 Km/h and pedestrian zones) with geolocalisation. 9) Public bike lanes in Grand Paris Seine Ouest. 10) Paths accessible to people with handicaps on the Grand Paris Seine Ouest territory. 11) Localisation of stations of the future automated metro line 15 south (passing by Issy) 12) Number of commuters between municipalities in the Paris Region for work reasons.	car park buildings 2) Locations and types of parking zones 3) Locations of paying machines *Third party's:* 1) Street network from city GIS. 2) Traffic loops positions from city GIS. 3) Traffic signs localisation.	*Third party's:* A composite map enabling to create detailed land-use maps of various regions based on pan-European datasets such as CORINE Land Cover, Urban Atlas enriched by available regional data.

		13) Roadworks on departmental roads. 14) Number of vehicles on roads managed by the department. 15) Off-street parking locations with number of potentially available parking spots. 16) Available parking spots (real time) on off-street parkings managed by Indigo.		
Non anonymised	New: 1) Locations of shops and other economic activities. 2) Locations of gyms, sports centres etc. *Existing:* 1) Data on inhabitants from the Civil Registry. 2) Lists of student addresses in the Civil Registry. 3) Locations of bike repair services. 4) Locations of study places in the city. 5) University and college locations. 6) Locations of the train stations. *Third party's:* 1) Data on Real estate properties from the Cadastre.	New: n/a *Existing:* n/a *Third party's:* n/a	*New:* n/a *Existing:* 1) Traffic sensors data from traffic control panel office. 2) Data from SUPERDIO including an update of building dates, locations, and new plans (prospects for the future). *Third party's:* 1) Municipal Police fines and accidents.	*New:* n/a *Existing:* n/a *Third party's:* n/a

The following case descriptions can be detected from D6.2 and have been updated during the preparation of this Deliverable:

Gent

A) Map the students housing locations.
B) Assess the share of student houses against total houses.
C) Define the main traffic routes between student residencies and their destinations.
D) Perform geo-referenced analyses in preparation for the infrastructure to be deployed.
E) Explore students mobility and its possible links with traffic and accident reports.

F) Explore the correlation between various locations (being economic, social, cultural, educational attractors or "hub") and the mobility patterns of the students.
G) Analyse the opportunities from the "Student economy" for new local initiatives.

Within the City of Gent, everyone is responsible for respecting the GDPR. The Data Protection Officer (DPO) for the city works in the data and information unit and oversees all GDPR-related activities.

Issy-les-Moulineaux
H) Generate a map of the past traffic.
I) Create a map showing the matching between the past traffic and roadworks.
J) Produce a real time visualisation of traffic in the Issy region.
K) Improve communication effectiveness using a dedicated mobility app proposing behaviour improvement.

The City of Issy-les-Moulineaux responsible person for GDPR is the head of the "Assemblies" (Assemblées in French) department, while for Issy Média, holding confidential data, is the innovation department.

Pilsen
L) Increase efficiency of traffic planning (restrictions, additional lanes, new roads, SUMP measures, roadworks coordination, etc.).
M) Visualize the current status of city traffic in real time (closures, roadworks and accidents) and show predicted roadworks including impact to traffic intensities.
N) Identify problem areas, show results of the analysis, produce charts, etc.

The City of Pilsen Data Protection Officer for GDPR is Security Director. Data Protection Officer in SITMP (Information Technology Administration of the City of Pilsen) is the Security Manager from the Back Office department. In PoliVisu project are used no personal data.

Additional cases from outside the partnership (source: Deliverable 4.5)

Nature of datasets	**Used in Mechelen pilot**	**Used in Flanders pilot**	**Used in Leuven pilot**
Confidential	*New:* n/a *Existing:* n/a *Third party's:* License plate data: Registration information, Generic vehicle characteristics, Technical motor characteristics, Vehicle mass	*New:* n/a *Existing:* n/a *Third party's:* n/a	New: n/a *Existing:* n/a *Third party's:* n/a
Anonymised and Public	*New:* n/a *Existing:* ANPR Camera data:	*New:* n/a *Existing:* n/a	*New:* n/a *Existing:* 1) Camera device data:

	Camera id, Back-office id, Time stamp, Location GPS coordinates, Plate number, Vehicle nationality, Speed, Vehicle type *Third party's:* n/a	*Third party's:* Anonymized road accident data: 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, Aggravating circumstances	Mac address, Use, Road segment ID, Direction, Status 2) Camera measurement data: Time (start and end), Segment id, Percentage up, Headcount of pedestrians, bikes, cars and lorries *Third party's:* n/a
Non anonymised	New: n/a *Existing:* n/a *Third party's:* n/a	New: n/a *Existing:* n/a *Third party's:* n/a	*New:* n/a *Existing:* n/a *Third party's:* n/a

The following case descriptions can be detected from D4.5:

Mechelen

- What-if analysis on planned road works;
- Implementation of school streets;
- Impact measurement of road works on the different transport modes;
- Impact of local distribution;
- Detection and tackling short cuts (more specific the installation of truck locks);
- Design and expansion of car-free zones;
- Traffic flows speed management;
- More sustainable modal split by promoting walking, cycling and public transport.

Flanders

- Develop a public, freely accessible accident datasets.

Leuven

In Kessel-Lo, the Telraam initiative has rolled out a local network of 100 counting points so far.