DELIVERABLE

D1.3 – IPR & Data management approach

Project Acronym	TT
Project Title	Transforming Transport
Grant Agreement number	731932
Call and topic identifier	ICT-15-2016-2017
Funding Scheme	Innovation Action (IA)
Project duration	30 Months [1 January 2017 – 30 June 2019]
Coordinator	Mr. Rodrigo Castiñeira (INDRA)
Website	www.transformingtransport.eu
Project Acronym	TT

Funded by the European Union's H2020 GA - 731932

Document fiche	
Authors:	Florian Flocke [Fraunhofer IML], Julian Eggemann [Fraunhofer IML], Luise Weickhmann [Fraunhofer IML], Gianluca Ripa [CEFRIEL], Roberto Spigolon [CEFRIEL], Oscar Corcho [UPM]
Internal reviewers:	Sami Koskinen [VTT], Leyre Merle [INDRA]
Work Package:	WP1
Task:	T1.3
Nature:	R
Dissemination:	PU

Document History			
Version	**Date**	**Contributor(s)**	**Description**
1.0	14/06/2017	Fraunhofer IML, CEFRIEL, UPM	Created and filled Deliverable Document for D1.3
2.0	28/06/2017	Fraunhofer IML	Finalization and integration of review comments

Keywords:	Data management, Data Governance, Guidelines, IPR, Data Protection, Ecosystem, Data asset cards, Role management
Abstract (few lines):	This deliverable reports on the work performed in WP1/T1.3 "Management of Knowledge, IPR, Data Governance and Data Protection" with respect to the creation of a concept for data management within Transforming Transport. It sums up the decisions that were made and gives an insight in the concept and the structure of data management. It contains the approach to Data Governance and includes all Governance Guidelines. A summary of the pilot data assets regarding role management and data governance is made.

ACKNOWLEDGEMENT

This document is a deliverable of TT project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 731932

Table of Contents

Table of Figures

Definitions, Acronyms and Abbreviations

Acronym	Title
DX.X	Deliverable
DMA	Data management approach
DMP	Data management plan
EEA	European Environment Agency
EUODP	European Union Open Data Portal
IPR	Intellectual property rights
JRC	Joint Research Center
TT	Transforming Transport
WP	Work package

Executive Summary

The Data Management Approach (DMA) is the central document for data management in Transforming Transport and summarizes information on the data used within the pilots regarding form, quality, governance and legal or financial issues. It specifies the decisions in data management and all roles and decision areas within Transforming Transport.
It also provides a list of several guidelines concerning Data Governance and Data quality management that help the pilot partners when they want to use data assets within their pilots.
First, a general overview about the data management strategy and the Transforming Transport ecosystem is given. The concept of Data asset id cards is displayed and the workflow and work distribution within the data management work packages in Transforming Transport is described. The concept of user and role management is summarized and a pilot definition of data management roles is added.
Then all guidelines developed by the data management team are presented and an overview about all governance assets in the pilot is given.
The document concludes with an overview of the D1.4 – Report. The D1.4-Report is planned for Month 18 and will follow up the development in the Data Management.
The list of provided guidelines within this document is as follows:

- Data Quality Management Guidelines (chapter 4)
  This guideline supports the management of the data asset in terms of quality, homogenization and integration
- General Governance Guideline (chapter 5)
  Data Governance Guideline specifies a set of recommendations on how to use the data accordingly.
- Specific Data Governance Guidelines (chapter 6).
  Next to the general guideline, the chapter 6 outlines various specific approaches on the Data Governance (e.g. Location Data). Some datatypes require specific recommendations.

This document will be updated based on the new data assets and new information in Transforming Transport in months 12 and 18. The guideline itself requires frequent updates. Whenever changes occur, the updated version of this document will be circulated to the responsible pilots.

1 General Data management strategy

1.1 Transforming Transport Ecosystem

To assure a general approach to Data Management in Transforming Transport, that enables interoperability and direct data sharing, the three WPs that are related to data management, 1.3, 2.3 and 3.3, created a holistic approach to data management and built up a **Digital Ecosystem** for Transforming Transport.

The key aspects of the Transforming Transport Ecosystem are Interoperability, Coopetition and the avoidance of a central infrastructure as shown in figure 1.

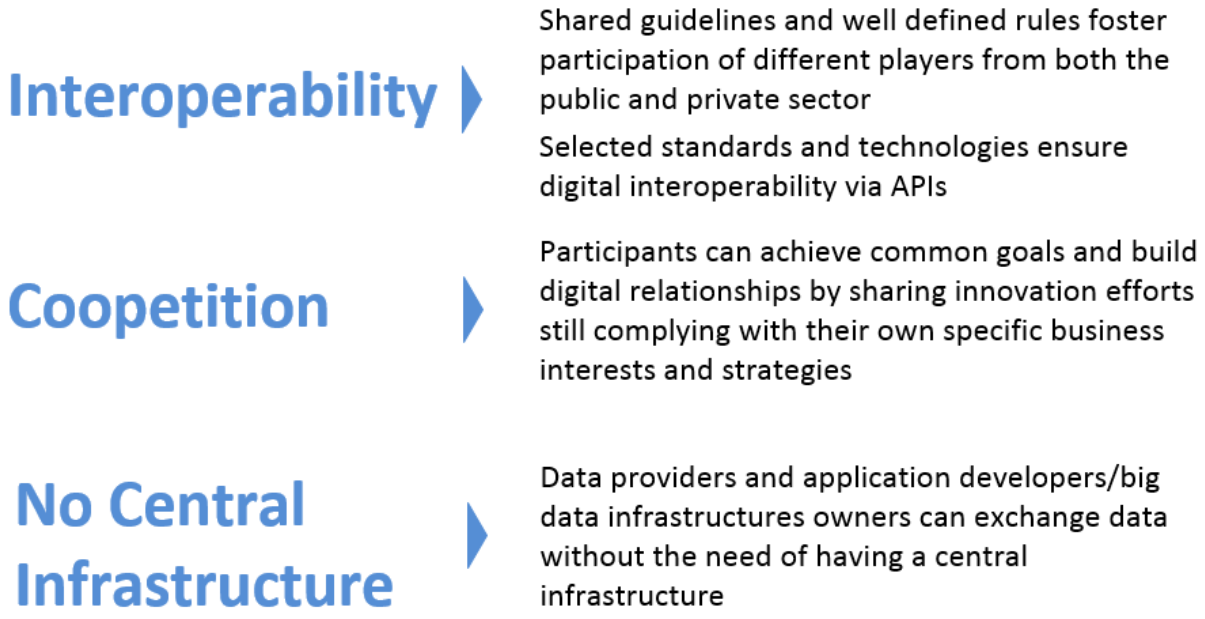


Figure 1: Key aspects of TT Ecosystem

Following the key aspects the digital ecosystem of Transforming Transport is a **set of guidelines and procedures that enable the direct data sharing between different parties in a coopetition way.**

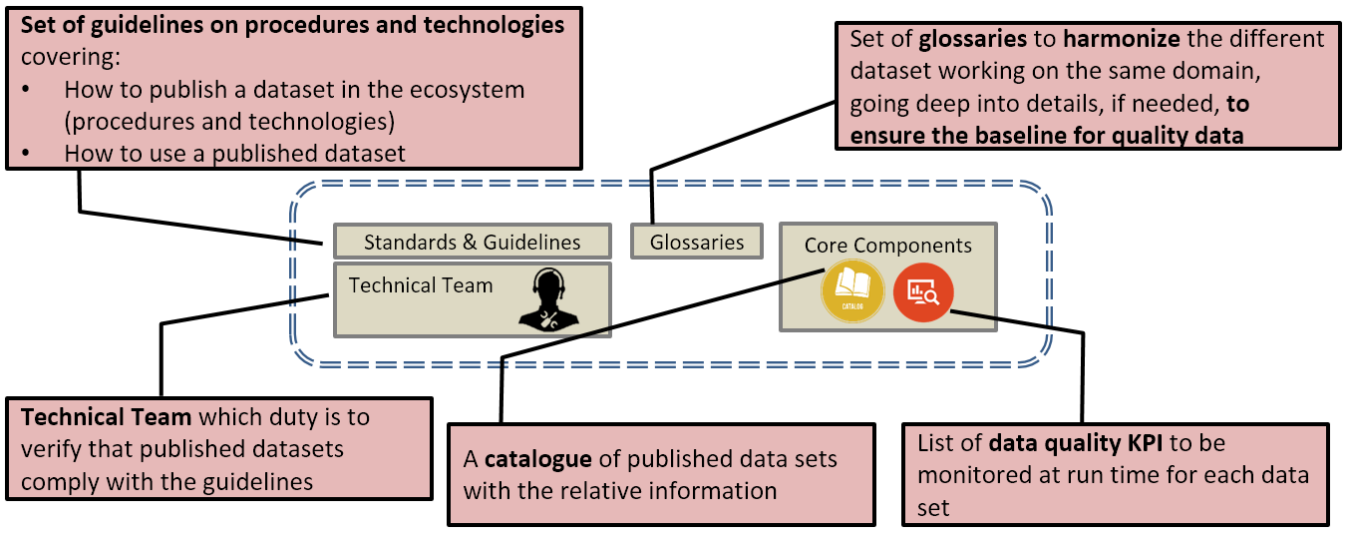


Figure 2: Core components of a digital ecosystem

Figure 2 shows the central components of a digital ecosystem and their connection:

- Set of guidelines and technologies
  - How to publish and use datasets in TT
  - How to solve problems in data management (e.g. governance, quality)
- Technical Team
  - Do the datasets comply with the guidelines and standards
- Glossaries
  - How can datasets be harmonized for the different pilots
  - How can a baseline data quality be assured
- Catalog
  - Which datasets were published and what is the metadata
- Data quality KPIs
  - Which quality KPIs describe and value a dataset

To display the advantages of an interoperable ecosystem without a central infrastructure figure 3 shows a set of pilots working on the same digital ecosystem baseline and giving the possibility for easy exchange between the pilots.

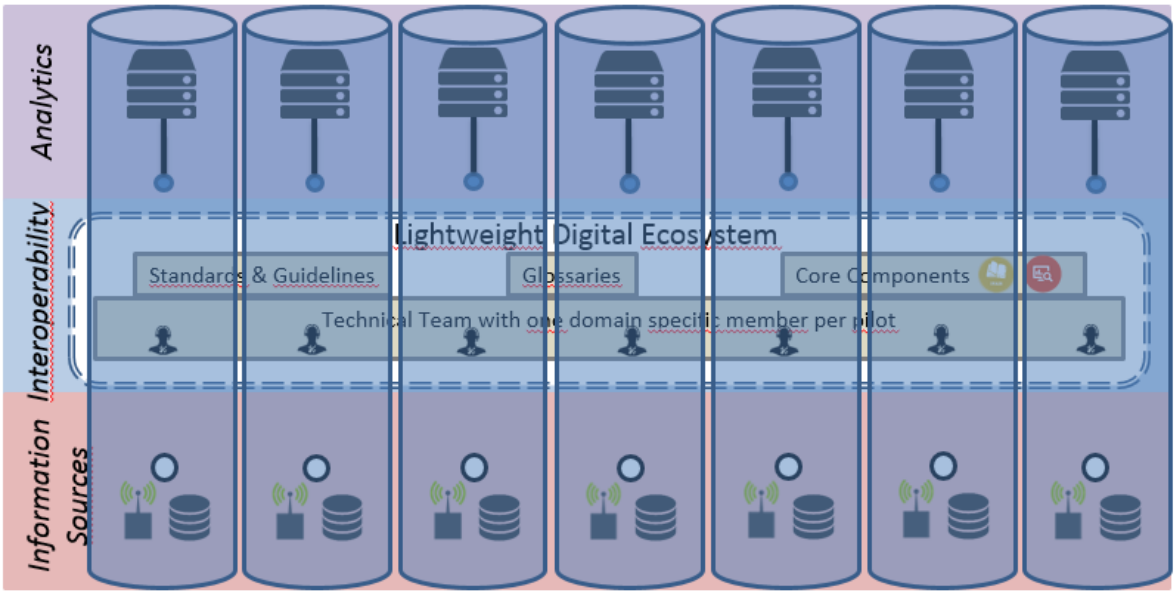


Figure 3: Combination of pilots and ecosystem

The detailed task distribution on work package level in Transforming Transport is shown in chapter 1.3.

1.2 TT Data Assets Ecosystem

In a Digital Ecosystem, participants can describe and publish their Data Assets, in terms of both functionalities and usage policies, in order to let other participants discover and leverage them, in agreement with the respective usage policies, for obtaining deeper insight in the driving-business KPI[1]. The increased ability to combine information by exploiting individual Data Assets can improve the services offered to the consumers by different players, through the principle of

[1] Michele Bonardi, Maurizio Brioschi, Alfonso Fuggetta, Emiliano Sergio Verga, and Maurilio Zuccalà. 2016. Fostering collaboration through API economy: the E015 digital ecosystem. In Proceedings of the 3rd International Workshop on Software Engineering Research and Industrial Practice (SER&IP '16). ACM, New York, NY, USA, 32-38. DOI: https://doi.org/10.1145/2897022.2897026

coopetition, i.e. cooperative competition[2]. Starting from the partner's experience (CEFRIEL) with Expo Milano 2015 in building the E015 ecosystem, they adopted a similar approach and customized it to the Transforming Transport domain, with the aim of obtaining a set of methodologies and guidelines for the stakeholders involved in the pilots.

The **TT Data Assets Ecosystem** will lead the Pilots to describe and publish their Data Assets, in terms of both functionalities and usage policies, in order let other participants discover and leverage them, in agreement with the respective usage policies, for obtaining deeper insight in the driving-business KPI.

1.3 Work package distribution

The overall data management in Transforming Transport is distributed in three work packages:

- WP1.3 Management of Knowledge, IPR, Data Governance and Data Protection
- WP2.3 Data Quality, Integration & Homogenization
- WP3.3 Open Data & Shared Data

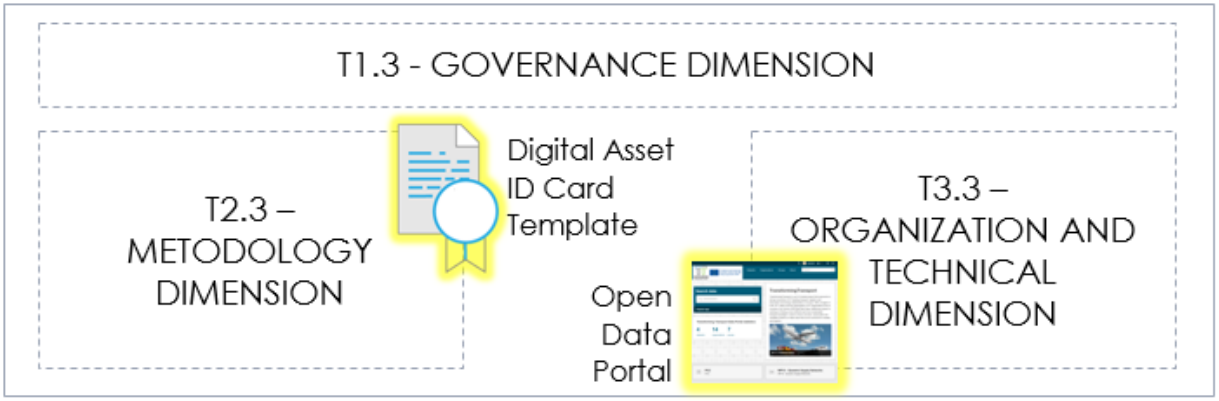


Figure 4: Task Distribution in Data management

[2] Zuccalà M., Celino I. (2015) Fostering Innovation Through Coopetition: The E015 Digital Ecosystem. In: Cimiano P., Frasincar F., Houben GJ., Schwabe D. (eds) Engineering the Web in the Big Data Era. ICWE 2015. Lecture Notes in Computer Science, vol 9114. Springer, Cham

Figure 4 shows the general distribution between the three tasks.

In **WP1.3 Management of Knowledge, IPR, Data Governance and Data Protection**, the task will ensure the management of IPRs and Data governance for all pilots. The task supports the pilots and their decision making when using data assets in Transforming Transport and has the following main goals:

> Summary of the Data Management Approach (DMA)Creation of Governance GuidelinesCreation of a Report on Data Management

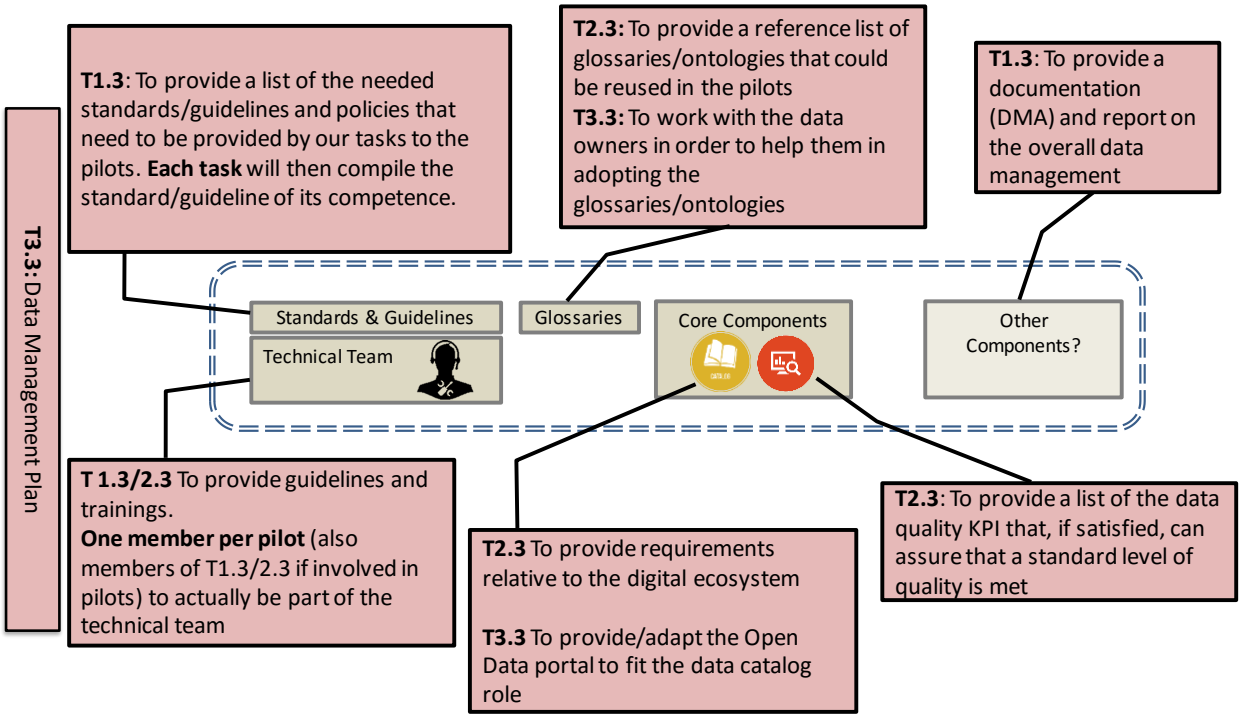


Figure 5: Distribution of work in the digital ecosystem

In **WP2.3 Data Quality, Integration & Homogenization** the task focuses on two aspects: data variety and data veracity. The goal of the task is to define:

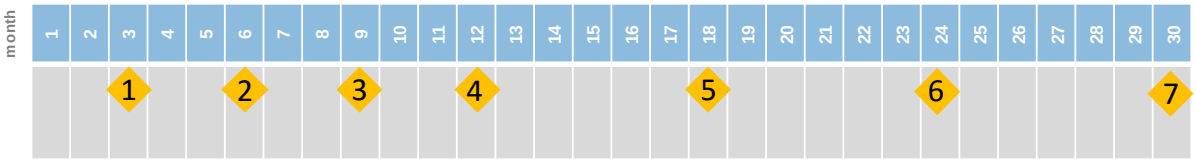
- Data quality procedures
- Data integration methodologies
- Data homogenization for interoperability

In **WP3.3 Open Data & Shared Data** the task focuses on gathering and managing all data assets in Transforming Transport. The main goals are:

- Creation of a Data Management Plan (DMP)
- Management of the TT open data portal (Reference D3.9 Data Management Portal)
- Guidance for the homogenization of data sources

When combined with the approach for a digital ecosystem, the different tasks are distributed among the components of the ecosystem as shown in figure 5.

The three mentioned work packages (WP1.3, WP 2.3 and WP3.3) summarize the data management activities in Transforming Transport. The achived work in these packages is summarized in a number of deliverables. Figure 6 shows the Timeline and the Milestones with the Deliverables by WP1.3 2.3 and 3.3.



#	Mo	WP 3.3 Open Data	WP 1.3 Data Governance	WP 2.3 Data quality, homogenization and integration
MS1	3	Data Management Plan [D3.2]		
MS2	6	Open Data Portal [D3.3]	Data Management Approach [D1.3]	
MS3	9	Updated DMP		Recommendations for data quality and integration [D2.3]
MS4	12	Updated DMP	Updated DMA	
MS5	18	Updated DMP	Updated DMA	
MS6	24	Updated DMP	Report on Data Management [D1.4]	
MS7	30			Lessons Learned for data quality and integration [D2.5]

Figure 6: Timeline and Deliverables for data management

1.4 Cooperation with other projects

The Transforming Transport Data management team cooperates with other European projects and initiatives to make sure that all crosscutting collaboration can add value to the Transforming Transport project.

For Data Governance, social and ethics-related topics a collaboration with the e-SIDES project (H2020 Coordination & Support Action) is initiated and further deepened in the future. e-SIDES is aiming at complementing the Research and Innovation Actions (RIAs) of the ICT-18 call on privacy-preserving big data technologies by exploring the societal and ethical implications of big data technologies and providing a broad basis and wider context to validate privacy-preserving technologies. The goal of the collaboration between the two projects is to exchange all results and best practices that were discovered in both project and strengthen both project results. The collaboration will be implemented through different approaches:

1. Organize a joint workshop by:
   - Co-locating the workshop in the next big Big Data Conference;
   - Ensuring high visibility in the wider Big Data community;
   - Bringing together e-SIDES key experts and Transforming Transport Pilots to discuss the key issues related to privacy-preserving big data technologies emerging in the pilots.
2. Contribute to e-SIDES Community Papers by:
   - Being involved in the community-wide discussion on the topic of privacy-preserving big data technologies;
   - Making sure that the project' views are known to the community;
   - Getting access to relevant, novel and continually updated content, validated by the community.
3. Open approach:
   - Fill-in the template developed by e-SIDES to collect input and open questions:
   - What you want from e-SIDES, What is your issue, Where would you map it in our assessment framework

The data management team will always look out for new and existing projects that can contribute or benefit to the Transforming Transport project results during the project lifetime.

2 Data Asset Information

The core element of summarizing data asset information in Transforming Transport is the **DATA ASSET ID CARD.** The id card represents the overall reference for the description of data assets within the Pilots. This description card provides many non-technical information related to the Data Assets, such as information about organizational aspects, information relating to the terms and conditions that regulate the use of the Data Asset, etc. Moreover, it addresses technical issues, providing all the necessary information to allow potential Data Assets users to proceed with integration activities within their Analytics Applications.

A higher level of completion of the card lead to a higher level of maturity of the Data Asset, growing from a simple set of raw data to a rich asset of information. This, in turn, can be discovered more easily and can be consumed in a more convenient way by many users as possible once it has been published on the Data Portal, see Figure 7.

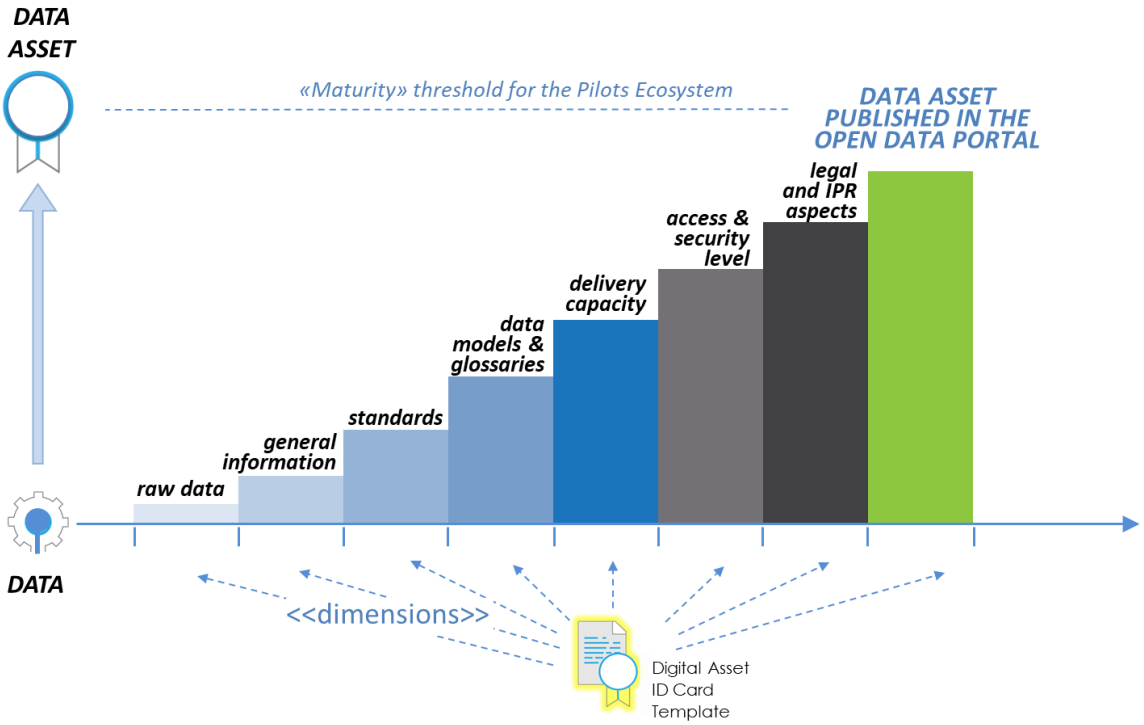


Figure 7: The cumulative completion of the ID Card dimensions carries the digital asset to the maturity level for being published on the Open Data Portal

Some dimensions are mainly related to organizational and/or business aspects, other dimensions are mainly related to technical aspects; for this reason, it is possible that many different roles can be involved to fill out the Data Asset ID card, as defined in the User and Role Management section.

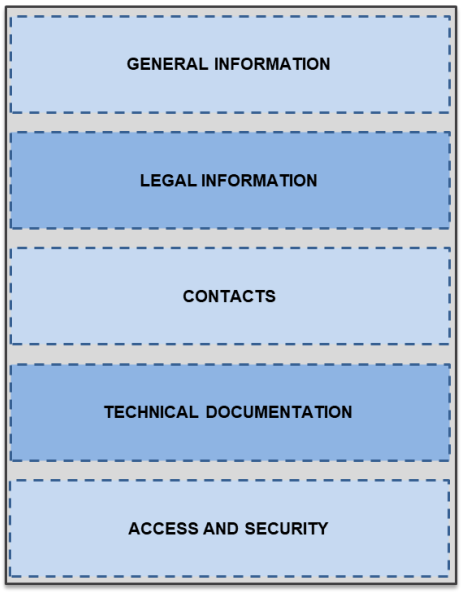


Figure 8: The dimensions over which the Data Asset metadata to be collected in the ID Card can be grouped in

Figure 8 illustrates the dimensions we proposed for the Data Asset ID Card:

- The *General Information* section contains information regarding the name, the content, the version number, the availability date and the type of data.
- The *Legal Information* section contains information regarding whether the data asset contains personal data, the license and the rights holder.
- The *Contacts* section contains information about the data owner's name and their contacts.
- The *Technical Documentation* contains information regarding the technology used, the source system's name, the data model glossaries, the data identifier standard used, the data volume and frequency, the data archiving and the geographical and time span.
- The *Access and Security* section contains information regarding:
    - whether the data asset is open to public access or restricted;
    - the access mechanism;

- the data asset URI (if available).

It is necessary that each dimension is well documented with a self-contained approach for a Data Asset to be published in the Open Data Portal. This information can be viewed as meta-data describing the data. This meta-data have been subject themselves to a data quality procedure depicted in Figure 9:

1. After the proposal of the ID Card template (Microsoft Excel file) on M1, Pilots began to fill in the files for each data asset and to upload them on the project's repository.
2. At the end of M4 we analyzed the information collected and gave feedback to the Pilots regarding the quality and completeness of the cards.
3. At the end of M5 Pilots uploaded an updated version of the files, where needed.
4. The Data Asset metadata collected in this way will be used to characterize the Data Asset to be published within M6.

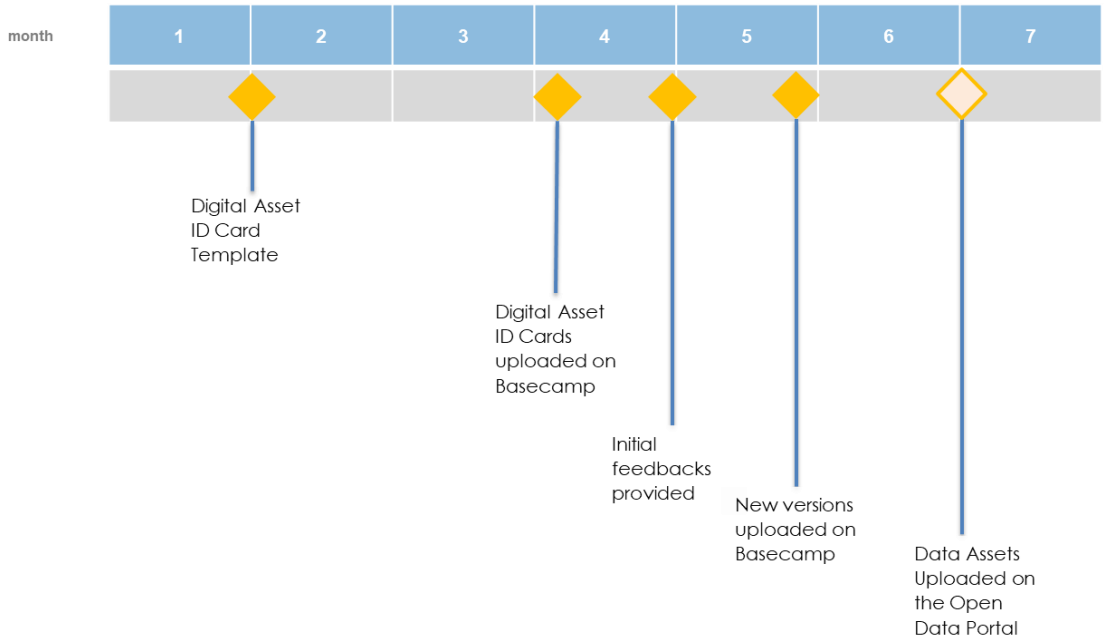


Figure 9: Timeline of the Data Asset ID Card process.

In parallel to the quality procedure on the data asset id cards, the data committee (Reference Chapter 3 User and Role Management) has conducted an analysis targeted to build a common taxonomy starting from the information collected about the type of data.

	WP4	WP5	WP6	WP7	WP8	WP9	WP10	TT
ID CARDS (on Basecamp)	26	4	65	5	12	13	11	136
DATA ASSETS (from DMP)	19	5	65	5	13	18	11	136

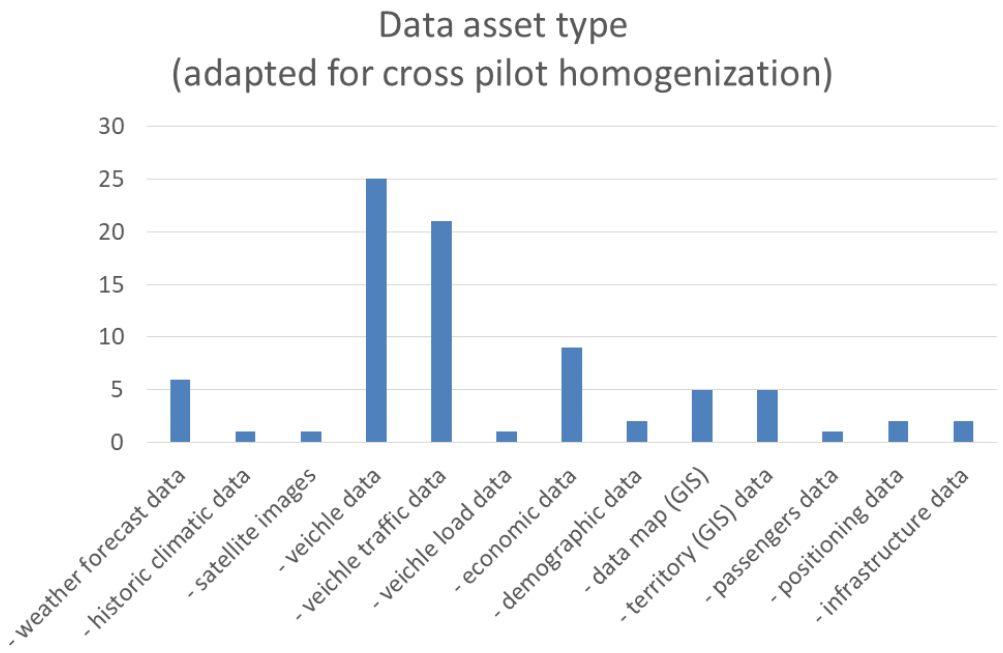


Figure 10: Data Type distribution of the Data Assets with respect to the cross context taxonomy proposed

Based on this information, the data committee defined an initial generic, i.e. domain independent, taxonomy to classify the type of data involved:

- weather forecast data,
- historic climatic data,
- satellite images,
- vehicle data,
- vehicle traffic data,
- vehicle load data,
- economic data,
- demographic data,
- data map (GIS),
- territory (GIS) data,
- passengers data,

- position data,
- infrastructure data

Figure 10 shows the distribution of the data assets with respect to this taxonomy, where information was available for about 60% of the sample.

This work is part of a wider activity aimed at obtaining a set of guidelines for the management of the data asset from the point of view of quality, homogenization and integration. The result of this work will be condensed and described in the deliverable "D2.3 - Guidelines for Data Quality, Homogenization and Integration" by the end of M09.

3 User and Role Management

Following Boris Otto, Data Governance must find answers to several questions including

- Which roles are involved in the decision making progress?
- How are the roles involved in the decision-making progress?[3]

According to Khatri and Brown, the area of decisions related to Data Governance range from general data management, data quality management, data access management/data lifecycle management to metadata management.[4]

Deriving from these decision areas, a lot of different roles are mentioned in literature, often related to the RACI notation if company specific roles should be defined (e.g. Weber et al 2009)[5].

Following RACI, roles should define whether they are

- **R**esponsible
- **A**ccountable
- **C**onsulted
- **I**nformed

[3] Otto, B (2011) "Data Governance," *Business & Information Systems Engineering*: Vol. 3: Iss. 4, 241-244
[4] Khatri V, Brown CV (2010) "Designing data governance", Communications of the ACM 53(1):148–152
[5] Weber K, Otto B, Österle H (2009) One size does not fit all – a contingency approach to data governance. ACM Journal of Data and Information Quality 1:1

for the respective decision area.

General role designs that are often mentioned are Steering or Data Committee, Head of Data Governance / Data Governance Lead, Data Owner / Business Process Owner, Data Steward and Technology Steward[6]. A committee defines the central authority and decision-making in data governance. The roles data owner and data steward represent the two different views on the data – from a business perspective (data owner) and from the data management perspective (data steward). Also the technology steward ensures data quality, availability of technology and the security and availability of related IT architectures.

For Transforming Transport pilots, these roles can be merged depending on the initial pilot data asset and the resulting requirements to create different role sets suitable for the pilot background.

Following the presented concepts the following roles are defined for Transforming Transport pilots:

- Data Committee
- Pilot Governance Lead
- Data Asset Owner
- Data Asset Steward
- Pilot Technology Steward
- (Pilot user)

The RACI roles of the decision areas will be defined as followed:

	Data Committee	**Pilot Governance Lead**	**Data Asset Owner**	**Data Asset Steward**	**Pilot Technology Steward**	**Pilot user**
General data management	A	R	I	I	I	I
Guideline management	A	R	I	I	I	I

[6] Khatri V, Brown CV (2010) "Designing data governance", Communications of the ACM 53(1):148–152

Data quality management	C	I	C	A	R	
Data lifecycle management	C	I	A	R	R	I
Metadata management	C	I	R	A	R	I
	Data Committee	**Pilot Governance Lead**	**Data Asset Owner**	**Data Asset Steward**	**Pilot Technology Steward**	**Pilot user**

A=Accountable, R=Responsible, C=Consulted, I=Informed

The described distribution results in a detailed description of the tasks of a role in TT:

Data Committee

- Creation of best practices and guidelines
- Helping resolve critical issues
- Participating in Data Committee Calls

Pilot Governance Lead

- Participating in Data Committee Calls
- Circulation of the data governance best practices, standards and guidelines
- Contact for pilot partners
- Exchange with data committee regarding problems in the pilot

Data Asset Owner

- Making sure that data access is possible
- Clarifying the legal requirements
- Responsible for license and costs control
- Coordinate lifecycle management (availability and archiving)
- Contact the pilot governance lead if any problems occur
- Ensure that business requirements are met

Data Asset Steward

- Responsible for the data quality and metadata management

- Responsible for the data asset representation in data asset id cards / open data portal
- Contact the pilot governance lead if any problems occur

Pilot Technology Steward

- Coordination of the needed technology platform
- Making sure that the used technology supports and enables data quality requirements
- Technical support for Data Security and Data Archiving
- Assisting in metadata management

The **Data Committee** contains the responsible persons from WP1.3 from Fraunhofer IML, CEFRIEL and UPM:

- Gianluca Ripa (CEFRIEL)
- Andrea Fiano (CEFRIEL)
- Oscar Corcho (UPM)
- Francisco Yedro Martinez (UPM)
- Victor Rodriguez-Doncel (UPM)
- Julian Eggemann (IML)
- Florian Flocke (IML)

The **Technical Team** as described in the digital ecosystem of TT contains the Data Committee and Pilot Governance Leads and coordinates within the Data Committee calls.

For each pilot, based on the pilot data, some roles may be merged to enable a smoother coordination. The following assignment summarizes the general roles and related persons for TT pilots:

	Pilot Governance Lead	**Pilot Technology Steward**
WP4.2 Ausol	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es
WP4.3 Norte Litoral	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es
WP5.2 Sustainable Cars	David Cobo dcobo@answare-tech.com	Not yet needed
WP5.3 Sustainable Trucks	Michael Schygulla Michael.schygulla@ptvgroup.com	Not yet needed

	(confirmation pending)	
WP6.2 Proactive Rail Assets	Phil Flynn phil.flynn@uk.thalesgroup.com	Nathan Marlor nathan.marlor@uk.thalesgroup.com
WP6.3 Predictive Network Maintenance	Ángel Pérez Bartolome aperezb@indra.es	Not yet needed
WP7.2 Valencia Sea Port	Paco Valverde fvalverde@iti.es	Rodrigo Garro rgarro@orbitaports.com
WP7.3 duisport inland port	(confirmation pending)	(confirmation pending)
WP8.2 Smart passenger flow	Juan Antonio Ubeda juan.antonio.ubeda@gmail.com	Not yet needed
WP8.3 Smart turnaround	Niels Stark niels.stark@jeppesen.com	Not yet needed
WP9.2 Tampere	Sami Koskinen sami.koskinen@vtt.fi	Not yet needed
WP9.3 Valladolid	Marta Galende margal@cartif.es	Marta Galende margal@cartif.es
WP10 Dynamic Supply Networks	Florian Flocke florian.flocke@iml.fraunhofer.de	Not yet needed
	Pilot Governance Lead	**Pilot Technology Steward**

The data committee together with the pilot governance leads form a clear communication structure which is shown in Figure 11.

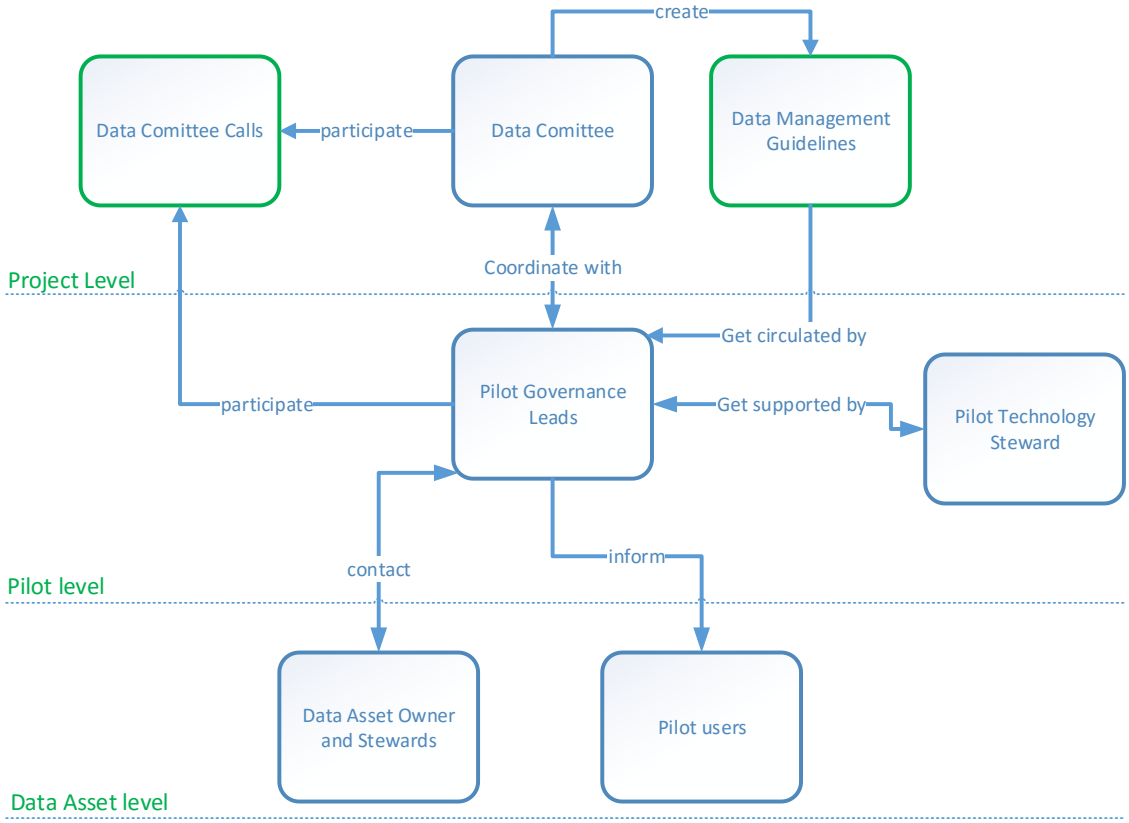


Figure 11: TT Role Management

The project level addresses pilot-overlapping coordination and guidelines. On pilot level the intermediate stage coordinates between project and data asset level. On data asset level the direct connection to data assets and metadata management is defined.

The defined roles are a suggestion for the pilots to clarify responsibilities and tasks within their pilots. However, if for some reason a differing role distribution is needed or advised by pilot members, the pilot governance lead is free to coordinate this with the data committee to help enabling a fast and coordinative approach on data management.

The following section provides a list of all data assets of the single pilots and makes suggestions for data asset level role management based on the initial data asset id cards filled by the pilot members.

Important Notice: Not all pilots finished the description of their initial data assets as id cards for the Data Management Plan deliverable. Since the analysis of this Deliverable is based on the Data Management Plan all late data assets will be added with a new version of this deliverable (and also the DMP).

Whenever new data asset or new responsibilities occur, this list will be updated and submitted in later versions of this Deliverable. For detailed Contact Information please refer to the related data asset id card.

The following paragraphs describe the Data Assets existing including the related responsibilities in each pilot use-case. The Data Asset is always connected to a Data Owner who is the representative of an organization. Together with the Data Asset Steward both roles are responsible for the correct usage of the data.

3.1.1.1 WP4.2 Ausol

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_WP4.2_Wheather Stations_AusolHighway_v1.0.xlsx	CINTRA	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.2_License Plate Recognition_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_Variable Message Sign - Highway_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_APPmobile_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_Traffic Flow - Outside Highway (surroundings)_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_Traffic Flow - Highway_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_Traffic events - Highway II_AusolHighway_v1.0.xlsx	CINTRA			

TT_WP4.2_Traffic events - Highway I_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_Maintenance Vehicle Routes - Highway_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_CCTV Cameras - Highway_AusolHighway_v1.0.xlsx	CINTRA			
TT_WP4.2_TrafficEvents_AusolHighway_v1.0.xlsx	Dirección General de Tráfico (DGT)	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.2_SpeedRadar_AusolHighway_v1.0.xlsx	Dirección General de Tráfico (DGT)			
TT_WP4.2_CCTVcameras_AusolHighway_v1.0.xlsx	Dirección General de Tráfico (DGT)			
TT_WP4.2_VariableMessageSign_AusolHighway_v1.0.xlsx	Dirección General de Tráfico (DGT)			
TT_WP4.2_SocioEconomic_AusolHighway_v1.0.xlsx	INE	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.2_YahooWeatherAPI_AusolHighway_v1.0.xlsx	Yahoo! Inc.	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.2_Twitter4j_AusolHighway_v1.0.XLSX	Yusuke Yamamoto	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**

3.1.1.2 WP4.3 Norte Litoral

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_WP4.3_Traffic events - Highway I_NorteLitoralHighway_v1.0.xlsx	CINTRA	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.3_CCTV Cameras - NorteLitoralHighway_v1.0.xlsx	CINTRA			
TT_WP4.3_TrafficFlow_NorteLitoralHighway_v1.0.xlsx	AENL	Hmendes@Nortelitoral.pt	Miguel Carpio mcarpio@cintra.es	
TT_WP4.3_METEO_NorteLitoralHighway_v1.0.xlsx	AENL			
TT_WP4.3_PMV_NorteLitoralHighway_v1.0.xlsx	AENL			
TT_WP4.3_SocioEconomic_NorteLitoral_v1.0.xlsx	INE	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.3_SpeedRadar_NorteLitoralHighway_v1.0.xlsx	Dirección General de Tráfico (DGT)	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.3_TrafficEvents_NorteLitoralHighway_v1.0.xlsx	Dirección General de Tráfico (DGT)			
TT_WP4.3_Twitter4j_NorteLitoral_v1.0.XLSX	Yusuke Yamamoto	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	
TT_WP4.3_YahooWeatherAPI_NorteLitoral_v1.0.xlsx	Yahoo! Inc.	Miguel Carpio mcarpio@cintra.es	David Collado Pradillo dcollado@indra.es	

3.1.1.3 WP5.2 Sustainable Cars

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_WP5_Referencial_SustainableConnectedVehicles_1.0.xlsx	SoFleet	Jérome Fenwick jfenwick@synox.io	David Cobo dcobo@answare-tech.com	
TT_WP5_Geo-tracking_SustainableConnectedVehicles_1.0.xlsx	SoFleet			
TT_WP5_CarSensors_SustainableConnectedVehicles_1.0.xlsx	SoFleet			

3.1.1.4 WP5.3 Sustainable Trucks

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_T2 3_Data AssetAPI_Id Card_PTV_5 3_Truck_Traffic.xlsx	Axxes	Michael Schygulla Michael.schygulla@ptvgroup.com (confirmation pending)	Michael Schygulla Michael.schygulla@ptvgroup.com (confirmation pending)	
TT_WP5_SatelliteImages_SustainableConnectedVehicles_1.0.xlsx	DLR	Joachim Rix joachim.rix@igd.fraunhofer.de	Michael Schygulla Michael.schygulla@ptvgroup.com (confirmation pending)	

3.1.1.5 WP6.2 Proactive Rail Assets

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
110103_01-TT_Thales_DataIDCard_Ellipse.xlsx	Network Rail	Phil Flynn phil.flynn@uk.thalesgroup.com	Nathan Marlor nathan.marlor@uk.thalesgroup.com	
110103_02-TT_Thales_DataIDCard_FMS.xlsx	Network Rail			
110103_03-TT_Thales_DataIDCard_ITPS.xlsx	Network Rail			
110103_04-TT_Thales_DataIDCard_TRUST.xlsx	Network Rail			
110103_05-TT_Thales_DataIDCard_IntelligentInfrastructure.xlsx	Network Rail			
110103_06-TT_Thales_DataIDCard_IntelligentInfrastructure.xlsx	Network Rail			
110103_07-TT_Thales_DataIDCard_379RollingStock.xlsx	Network Rail			
110103_08-TT_Thales_DataIDCard_SSI_Record.xlsx	Network Rail			
110103_09-TT_Thales_DataIDCard_MetDesk.xlsx	Network Rail			
110103_10-TT_Thales_DataIDCard_EarthWorks.xlsx	Network Rail			
110103_11-TT_Thales_DataIDCard_OLEEx.xlsx	Network Rail			
110103_12-TT_Thales_DataIDCard_LADS.xlsx	Network Rail			
110103_13-TT_Thales_DataIDCard_NRWS.xlsx	Network Rail			

110103_14-TT_Thales_DataIDCard_OLETempMonitoring.xlsx	Network Rail			
110103_15-TT_Thales_DataIDCard_OLEALP.xlsx	Network Rail			
110103_16-TT_Thales_DataIDCard_NetworkModel.xlsx	Network Rail			
Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**

3.1.1.6 WP6.3 Predictive Network Maintenance

Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**
TT_6.3_TramificationIdeadif_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_Tramification_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_TampingTasks_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_Sios_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	

TT_6.3_S&CInspections_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_Pidame_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_Icecof_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_GeometricalInspection_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_DynamicInspection_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_Davinci_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_Clima_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**

TT_6.3_CGRH24_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_AEMET_1_0.xlsx	To be determined	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_WeatherInfo#1_1.0.xlsx	ADIF	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_FerroDesignDigital_1.0.xlsx	Ferrovial	Ana Comas acomas@ferrovial.com	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_Ferrodrone_1.0.xlsx	Ferrovial Agroman	Antonio García antoniogarcia@ferrovial.com	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_TrainNoiseMalaga_1.0.xlsx	Ayto. Málaga	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_EURaildB_1.0.xlsx	EC	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**

TT_6.3_FerroDrainageOpPaper_1.0.xlsx	Ferrovial	Ignacio Jardí ignacio.jardi@ferrovial.com	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_FerroDrainageOpDigital_1.0.xlsx	Ferrovial			
TT_6.3_FerroDesignPaper_1.0.xlsx	Ferrovial			
TT_6.3_FerroTrackOpPaper_1.0.xlsx	Ferrovial			
TT_6.3_FerroTrackOpDigital_1.0.xlsx	Ferrovial			
TT_6.3_FerroTrackGeometry_1.0.xlsx	Ferrovial			
TT_6.3_FerroSlopesOpPaper_1.0.xlsx	Ferrovial			
TT_6.3_FerroSlopesOpDigital_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineTampingTemp_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineTamping_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineSpec_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineModeAut_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineMode_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineGPS_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineFuelAut_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineFuel_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineEngineData_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineEngineAut_1.0.xlsx	Ferrovial			
TT_6.3_FerroMachineEngine_1.0.xlsx	Ferrovial			

TT_6.3_FerroFenceOpPaper_1.0.xlsx	Ferrovial			
TT_6.3_FerroFenceOpDigital_1.0.xlsx	Ferrovial			
TT_6.3_VegetationIndex_1.0.xlsx	Junta de Andalucía	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_VecTerrFile_1.0.xlsx	Spanish Government	To be determined	Ángel Pérez Bartolome aperezb@indra.es	
TT_6.3_TopoTerrRast_1.0.xlsx	Spanish Government			
TT_6.3_SeismogenicZone_1.0.xlsx	Spanish Government			
TT_6.3_RecentSeismic_1.0.xlsx	Spanish Government			
TT_6.3_Ortoimages _1.0.xlsx	Spanish Government			
TT_6.3_NatTopoBase_1.0.xlsx	Spanish Government			
TT_6.3_HydrogeoMap_1.0.xlsx	Spanish Government			
TT_6.3_DigiTerrModel_1.0.xlsx	Spanish Government			
TT_6.3_WeatherInfo#2_1.0.xlsx	Spanish Government			
Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**

3.1.1.7 WP7.2 Valencia Sea Port

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_T2.3_Data Asset & API_Id Card v0.3 CATOS.XLSX	Noatum	Franciso Blanquer fblanquer@noatum.com	Paco Valverde fvalverde@iti.es	
TT_T2.3_Data Asset & API_Id Card ORBITA v0.2.xlsx	Noatum			
TT_T2.3_Data Asset _AIS FVAL v0.1.xlsx	Valenciaport	Miguel Llop mllop@fundacion.valenciaport.com	Paco Valverde fvalverde@iti.es	
TT_T2.3_Data Asset _PCS FVAL v0.1.xlsx	Valenciaport			
TT_T2.3_Data Asset _SCADA FVAL v0.1.xlsx	Valenciaport			

3.1.1.8 WP7.3 duisport inland port

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
(pending)	(pending)	(pending)	(pending)	

3.1.1.9 WP8.2 Smart passenger flow

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_T8.1_PASSENGER_v1.0.xlsx	AEGEAN	To be determined	Juan Antonio Ubeda juan.antonio.ubeda@gmail.com	
TT_T8.1_BAGGAGE_v1.0.xlsx	AEGEAN			

TT_T8.1_AIRPORT SLOT_v1.0.xlsx	Athens International Airport	To be determined	Juan Antonio Ubeda juan.antonio.ubeda@gmail.com	
TT_T8.1_MOBILE PHONE LOCATION_v1.0.xlsx	Athens International Airport			
TT_T8.1_BOARDING PASS READING_v1.0.xlsx	Athens International Airport			
TT_T8.1_FLIGHT PLAN_v1.0.xlsx	Eurocontrol	To be determined	Juan Antonio Ubeda juan.antonio.ubeda@gmail.com	

3.1.1.10 WP8.3 Smart turnaround

Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**
TT_T8 3_Data Asset API_Id Card Template_RTFP.xlsx	BR&T-E	David Scarlatti david.scarlatti@boeing.com	Niels Stark niels.stark@jeppesen.com	
TT_T8 3_Data Asset API_Id Card Template_RTS - Copy.xlsx	BR&T-E			
TT_T8 3_Data Asset API_Id Card Template_WFCST.xlsx	BR&T-E			
TT_T8.3_Data Asset & API_Id Card Template_FP.xlsx	Jeppesen	Niels Stark niels.stark@jeppesen.com	Niels Stark niels.stark@jeppesen.com	
TT_T8.3_Data Asset & API_Id Card Template_fuel.xlsx	Jeppesen			
TT_T8.3_Data Asset & API_Id Card Template_schedule.xlsx	Jeppesen			
TT_T8.3_Data Asset & API_Id Card Template_ground-tasks.xlsx	Jeppesen			

3.1.1.11 WP9.2 Tampere

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_T9.2_Data Asset & API_Id Card_Sensior.xlsx	To be determined	To be determined	To be determined	
TT_T9.2_Data Asset & API_Id Card_TreTrafficLights.xlsx	To be determined	To be determined	To be determined	
TT_T9.2_Data Asset & API_Id Card_TreTrafficCameras.xlsx	To be determined	To be determined	To be determined	
TT_T9.2_Data Asset & API_Id Card_SocialMedia.xlsx	To be determined	To be determined	To be determined	
TT_T9.2_Data Asset & API_Id Card_ProlongedDrivingTimes.xlsx	To be determined	To be determined	To be determined	
TT_T9.2_Data Asset & API_Id Card_Digitraffic.xlsx	To be determined	Tomi Lapinlampi tomi.lapinlampi@liikennevirasto.fi	To be determined	
TT_T9.2_Data Asset & API_Id Card_TreIncidentsAndRoadworks.xlsx	To be determined	? asiakaspalvelu2@infotripla.fi	To be determined	
TT_T9.2_Data Asset & API_Id Card_TreParking.xlsx	Finnpark	To be determined	To be determined	
TT_T9.2_Data Asset & API_Id Card_TrePublicTransport_Journeys.xlsx	ITS Factory	To be determined	To be determined	
TT_T9.2_Data Asset & API_Id Card_TrePublicTransport_SIRI.xlsx	ITS Factory	To be determined	To be determined	

3.1.1.12 WP9.3 Valladolid

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_WeatherData.xlsx	AEMET	To be determined	To be determined	
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_GPStraces.xlsx	Grupo Lince	Daniel Clavero daniel.clavero@grupolince.com	To be determined	
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_MagneticLoops.xlsx	Ayuntamiento de Valladolid	Roberto Riol Martinez rriol@ava.es	To be determined	
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_TrafficLights.xlsx	Ayuntamiento de Valladolid			
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_TrafficIncidences.xlsx	Ayuntamiento de Valladolid			
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_PneumaticLoops.xlsx	Ayuntamiento de Valladolid			
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_ORAdata.xlsx	Ayuntamiento de Valladolid			
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_FCDtomtom.xlsx	TomTom	To be determined	To be determined	

3.1.1.13 WP10 Dynamic Supply Networks

Data Asset	Organization	Data Asset Owner (Role)	Data Asset Steward (Role)	Comment
TT_10_InternetAccess_1.0.xlsx	European Commission, Eurostat	Florian Flocke florian.flocke@iml.fraunhofer.de	Florian Flocke	

TT_10_ECommercesales_Company_1.0.xlsx	European Commission, Eurostat		florian.flocke@iml.fraunhofer.de	
TT_10_ECommercepurchases_Individual_EU_1.0.xlsx	European Commission, Eurostat			
TT_10_ECommercepurchases_Individual_3M_1.0.xlsx	European Commission, Eurostat			
TT_10_ECommercepurchases_Individual_1.0.xlsx	European Commission, Eurostat			
TT_10_PrimaryIncome_1.0.xlsx	European Commission, Eurostat			
TT_10_PopulationDensity_1.0.xlsx	European Commission, Eurostat			
TT_10_Population_1.0.xlsx	European Commission, Eurostat			
TT_10_ECommercepurchases_Company_1.0.xlsx	European Commission, Eurostat			
TT_10_DisposableIncome_1.0.xlsx	European Commission, Eurostat			
TT_10_Logika_data_1.0.xls	Logika	George Karaganis gk@logika.gr	Florian Flocke florian.flocke@iml.fraunhofer.de	
Data Asset	**Organization**	**Data Asset Owner (Role)**	**Data Asset Steward (Role)**	**Comment**

4 Data Quality Management and Guidelines

The data quality management in Transforming Transport is part of a wider activity aimed at obtaining a set of guidelines for the management of the data asset from the point of view of quality, homogenization and integration. The result of this work will be condensed and described in the deliverable "D2.3 - Guidelines for Data Quality, Homogenization and Integration" by the end of M09.

However, a general overview on the guidelines for data quality management will be given within D1.3.

The table below shows some initial example of guidelines that will be provided in D2.3. A full Data Quality Managemeng and Guideline approach can be found in the Deliverable D2.3.The table names the guidelines and identifies possible benefits to increase the Data Quality. The costs of implementing the outlined benefits is analyzed in the 3rd column:

Guideline	**Benefits**	**Costs**
Shares domain specific data schemas and taxonomies • share, adopt and possibly extend a common data schema • avoid defining new data schemas if one already exists • if possible, prefer standards	• **Integration** of different sources adopting the same schema is straightforward • **Homogenization**: Data sources implementing the same glossary are homogeneous by construction • **Quality**: o shared domain specific data schemas drive the emergence of the most valuable data from wide data sources, o standards evolve over time exploiting the knowledge of a larger community, this simplifying the	• Requires big effort on existing data sources (especially on legacy ones) • On new data sources to be developed the cost of scouting existing data schemas pays back itself in the short term

	evolution of the data assets.	
Shares domain specific taxonomies • share, adopt and possibly extend common taxonomies • avoid defining new data taxonomies if one already exists • if possible, prefer standards	• Simplify integration and homogenization of heterogeneous data sources	• Existing data sources the effort required for adopting a common taxonomy is lower than adopting a common data schema
Guideline	**Benefits**	**Costs**

Table 1 - initial example of guidelines

4.1.1 Management and Integration of Geospatial Data

Geospatial data are a special form of data with an explicit information about spatial reference and can be available in different forms. The **Management and Integration of Geospatial Data** deals with the general guidelines on management and integration of geospatial data to manage and deliver them uniform across the entire project. The aim is to make geospatial data available and usable across different pilots.

Since geospatial data are a special form of data with an explicit information about spatial reference, there is a special focus on the requirements of metadata, since this is crucial for the implementation of standards.

Moreover, since geospatial data can be available in different forms, special attention must be paid on how to manage the data, which includes access, storage and backup. Then an overview of standards and their usage is given. These standards serve the provision of interoperable geospatial data and draw a distinction between standards for data and for services, e.g.:

- Standards for data
  - INSPIRE-based Guidelines on Metadata etc.
  - Metadata standard for Geographic information: ISO 19100 series
- Standards for services
  - OGC Services (WMS, WFS, WCS, …)
  - DIAS

4.1.2 Semantic Integration of Real-time Data

The **Semantic Integration of Real-time Data** is necessary when building scalable processing engines for real-time data. This guideline will first provide an overview of the approaches for semantic integration and data fusion for real-time data processing. Then a comparison of stream processing[7] engines will be provided along with a set of general and pilot specific guidelines.

4.1.3 Data Quality Indicators and Standards

Data quality guidelines are already existing in different areas and industry sectors. For transport data management and data quality requirements different data classes and quality indicators are to be considered. Data quality therefore is not a classic KPI which can be measured on the basis of raw data. The **Data Quality Indicators and Standards** guideline will lead to the assessment of the data quality with regard to the specific use case and user requirements. With this guideline a classification of relevant data (sources) within TT in relation to the different pilot sites and use cases will be elaborated. The quality indicators can then be set up and evaluated on the basis of results from the different pilot domains/ pilots.

4.1.4 Big Data Infrastructures and Data Modeling

Data Integration in Big Data environments major challenge is how to make available great quantities of data from various sources in a way that allows enough flexibility to be used for new purposes. While all the pilots have already defined their software architectures and IT infrastructures, the **Big Data Infrastructures and Data Modeling** guideline will focus on some of the major Big Data practices and technologies and, in general, other tools and technologies that are relevant for the Transforming Transport pilots. In particular, we will introduce the most common approaches for Big Data integration, including the constructions of Data Lakes, in which the raw data is collected for a later processing.

Attention will be put on infrastructure, persistence and other control tools for quality management, integration and homogenization of data in Big Data environments. For the given technologies, guidelines will be provided on when to use them and which are the main purposes

[7] Della Valle, Emanuele, Daniele Dell'Aglio, and Alessandro Margara. "Taming velocity and variety simultaneously in big data with stream reasoning: tutorial." Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems. ACM, 2016.

of them along with common practices, recommendations and common pitfalls. Furthermore, answers to a survey across all pilots will be analyzed in order to access the maturity of the technologies in use and to share the experiences across the different pilots.

5 General Governance Guideline

5.1 Introduction

In modern society more and more data is used for complex services. The processing of information is necessary to enable the Big Data Analytics in services like self-driving cars, individualized mass production, precise weather forecasts, individualized route guidance or individual medication.

This leads to additional responsibilities in modern business models. Companies and in particular business models increasingly depend upon confidential data such as intellectual property, market intelligence, and personal information (from customers). The challenge in this is to maintain the privacy and confidentiality of this data, as well as meeting the requirements of a growing list of related compliance obligations.

The following Data Governance Guideline is an approach that addresses many aspects of data management, including information privacy and security as well as compliance, for the Transforming Transport Project.

The guideline is based on the European Guidelines/Laws regarding data privacy, as well as a number of accepted guidelines as stated in the text.

5.2 Usage of this Guideline

Data Governance is the specification of decision rights and an accountability framework to encourage desirable behavior in the valuation, creation, storage, use, archiving and deletion of information. It includes the processes, roles, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals. The following guideline will support by outlining problematic approaches and name possible solutions.

The Importance of Data Governance is based on the fast growth of data. It is growing much faster than at any other stage in human history. Much of this information is based on new sources, such as social media feeds, blogs, web pages, web traffic logs which require new approaches in Data

Governance. The Guideline presents the challenges (blockades) and risks from Politics and Social Environment as well as from several Business Approaches.

Data Governance is a complex process and involved all Stakeholders. The definition of basic policies or guidelines has to be followed by the implementation as well as the daily usage within the enterprise architecture. Therefore, the project is ongoing and requires alignment during the Transforming Transport Projects-Lifetime. This Deliverable is a snapshot of the current developments. The Guideline for Data Governance, IPR and Data Protection will be updated throughout the project. This Guideline outlines a number of Goals for the Governance Process:

- A reduced operational friction within the Transforming Transport Project
- Outline the Needs and Requirements of data related stakeholders
- Improved quality and usability of the data
- Meet regulations and compliance & build awareness within the Project partners about common data issues
- Define a standard, repeatable process within TT
- ensure transparency of data related processes
- an advanced decision making on data handling.

This document is a general guideline for all data assets but for some reoccurring issues there also exist specific more detailed guidelines.

The Guideline is outlining common approaches on Data Governance, IPR and Data Protection. The Guideline does not constitute legal advice. The completeness can also not be guaranteed. In case of any questions or queries regarding the completeness of the information, the authors are always at your disposal.

5.3 General support / Best Practices

5.3.1 Clarify the requirements and goals

First of all, the data responsible (see chapter „Role management" below) needs to have a very clear idea of the (business) requirements that are related to the data asset so that an evaluation and comparison of the possible solutions is productive. Additionally, the goals that the pilot tries

to achieve by using the data should be detailed and it should be evaluated if the data asset adds enough value to the pilot to be considered.

5.3.2 Make sure to use the right data asset

Before starting to go into detail, the data responsible should check if the right data asset is used within the pilot. Depending on the requirements, goals and the proprietary status there can be alternative data source that need to be considered.

That may include:

- Usage of Open Data (e.g. EUODP, Eurostat, EEA, JRC, World Bank, regional Open Data Portals)
- Usage of Data sources without connection to the TT consortium

You can easily check the status of open data in the needed country on the internet (e.g. the Global Open Data Index https://index.okfn.org/ or at the EUODP).

5.3.3 Clarify the license for usage

Before using a data asset, the data responsible should try to identify the corresponding license and clarify if there are some clear issues for the usage within TT. More detailed interpretation of the license can be found as a solution below.

5.3.4 Validate the availability date

The data responsible should make sure to validate the estimated date of availability to make sure that the further going steps are based on secure information. If the availability of the data is unclear the data asset owner should be contacted. More detailed interpretation of the availability can be found as a solution below.

5.4 Problems & Solutions with data asset

The decision for a general guideline for all data asset types was done to avoid multiplications within the problems and solutions section. When a problem is only specific for a data asset type, then it will be mentioned within this guideline or addressed within a single specific guideline.

The following Data Asset Types are defined based on the initial data assets from Transforming Transport pilots:

- weather forecast data,
- historic climatic data,
- satellite images,
- vehicle data,
- vehicle traffic data,
- vehicle load data,
- economic data,
- demographic data,
- data map (GIS),
- territory (GIS) data,
- passengers data,
- position data,
- infrastructure data

5.4.1 General Recommendations

For better communication, organization and a successful cooperation within the project there are some common practiced principles[8] which are recommended to follow:

- Integrity:
  Every participant of Transforming Transport is recommended to maintain a fair and trustful dealing with each other.
- Transparency:
  The different steps of the development process need to be transparent and declared separately (as already practiced in TT they are likely to find on the first pages of the documentation dedicated to the projects). It should be comprehensible to other participants which decisions were made, who worked on the different steps and what improvements were done.
- Accountability:
  It is necessary for a project to name its persons and teams in charge.
  Therefore, the accountable persons and institutions for TT and its sub-projects and information about their responsibilities are listed in section “Role management”.
- Auditability:

[8] Data Governance Goals http://www.datagovernance.com/adg_data_governance_goals/

The data-related decisions, processes and controls need to be verifiable by an independent third party. To ensure the auditability of those data-related work steps it is necessary to maintain a meaningful documentation, which satisfies the compliance-based requirements.

- Checks-and-Balances:
  For the basic safety of the progress of TT the principle of checks and balances brought into action. The accountabilities for the reciprocal review between the participant teams are defined in section „Role management".
- Stewardship:
  Information about the responsibilities for the stewardship activities within the different projects and for the data administration are given in chapter „Role management".
  Standardization:
  For the consistency of the different projects it is recommended to develop a standardization of enterprise data, which is binding for all teams and projects within TT.
- Change Management:
  Due to the natural change of data a change management will be provided, which declares rules for the reference of data values and the structure and use of master data and metadata

5.4.2 Anonymization of personal data

5.4.2.1 General Restrictions / Challenges

"Under EU law, personal data can only be gathered legally under strict conditions and for a legitimate purpose. Furthermore, persons or organisations which collect and manage personal information must protect it from misuse and must respect certain rights of the data owners which are guaranteed by EU law."[9]

The EU law is often used as a guideline for the law of the participating states. For Example, The German Federal Data Protection Act recommends that companies store, process and use as little personal data as possible. In general, the anonymization or pseudonymisation is the standardized way to store personal data. Furthermore, the use of personal data is highly restricted and follows a number of certain rules:

- A law has to allow the collecting, processing and usage of data.

[9] http://ec.europa.eu/justice/data-protection/

In this case, the usage is controlled by the existing law. The system has to follow the restrictions by law strictly. This includes legal obligations such as tax or prosecution as well as responsibilities regarding legal contracts.

- If no law is controlling the data usage, the data consumer (e.g. companies) needs to request a consent from all involved partners. The consent is mandatory and illegal usage of data will be prosecuted by the local authorities.
- The consent has to include details of the required data. The data, the amount, as well as the way of the data usage have to be clearly indicated/marked. Furthermore, the collected data can only be used for the agreed purpose of use. Any other use not covered by the consent is forbidden.
  Here, the anonymization is a popular way to use personal data without requesting the usage rights from all individuals.
- In the case of anonymization, the data are processed in such a way that the identification of the individual person can be prevented. In general, data are considered anonymous if an irreversible identification of the person is not possible. Nor should it be possible to derive a connection to an individual by processing further data.The regulations may vary between different countries and need to be reconfirmed.

More details are stated in the EU Data Protection Directives[10] and from May 2018 will be followed up by the General Data Protection Regulations (GDPR)[11].

5.4.2.2 Specific Problems for Data Asset Types

Most data-asset-types can be addressed by general approaches. The following data types require specific approaches:

- Location data / floating car data
- Social media data
- Passenger data
- E-commerce customer data

[10] EU Data Protection Directives http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&from=en
[11] General Data Protection Regulations http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0089.01.ENG&toc=OJ:L:2016:119:TOC

To provide a precise location of vehicles it needs to be anonymized. Otherwise the position of the driver can be detected which violates the drivers privacy. The other cases are slightly stricter in terms of protecting the individual's privacy, because of the directly gathered user data. A detailed guideline for anonymization of locations is found in chapter 7.

5.4.2.3 Solutions

Any data processing company has legal responsibilities towards the individual person. Here some common statements[12] regarding the responsibilities of the company:

- Collecting, processing and using personal data requires a permissive law or an approval of the concerned person (like the German Federal Protection Law -BDSG§4) which must be in a written form
- In some exceptional cases the approval can be skipped. This has to be reconfirmed with national/European restrictions (e.g. German Federal Data Protection Act §4 Abs.2)
- The concerned person needs to see the form, content, scale of the measurement clearly
- The contract needs to point out the purpose of the data collection and processing and usage mustn't go further than allowed by the concerned person, otherwise its necessary to set up an supplement contract
- The responsible party must be directly named and pointed out (e.g. in the beginning of the contract)
- Consent can be cancelled at any time by the concerned person
- Company has to delete the data directly after request by the concerned person
- The locking of information (instead of deleting) is only accepted for certain reasons
- The concerned person has the right to correct false information
- Incorrect data needs to be corrected instantly
- Companies may only process, use and collect as less data as possible. Therefore, the data should be anonymized or pseudonymized. The process-effort needs to be in scale with the companies possibilities
- Unnecessarily stored data has to be deleted immediately.
- In case of revocation, expired right for holding data, illegal processing/using/collecting the according data needs to be deleted or blocked immediately.
- The concerned person has the right to request and be informed about the own stored data at any time and free. (in case of protection of the business plans its allowed to deny

[12] Handelsblatt Research Institute: "Datenschutz und Big Data: Ein Leitfaden für Unternehmen"

the request; in case of stored data for the purpose of (credit)transfer the information can be fee-based)

Anonymization Techniques

In general there are two kinds of techniques to anonymize a dataset: Randomization and Generalization. They have several specializations which should be named in the following[13]:

- Noise addition: adding imprecision to a certain degree to the original data
- Permutation: shifting the values of attributes in a table so that artificial linking occurs
- Differential privacy: determining the amount, form and amplitude of the added noise
- Aggregation: grouping an individual with others (see „Aggregation of data")

Pseudonymisation

In the new GDPR, another possible solution will be outlined: pseudonymisation. The pseudonymisation is defined as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information."[14] The idea is to enable more valuable data for a specific purpose without the risk of leading it back to the origin person. The identification of the person must be very hard. For example, the name and other personal identification data will be changed, so that the affected person or its reasons cannot easily be identified.

There are some recommended techniques[13] for the pseudonymisation of datasets:

- Encryption with secret key: encrypted data is stored in a table but can be resolved by the key-holder
- Hash-function: returns a fixed size output and cannot be resolved (specialized forms are: keyed-hash function with stored key or deterministic encryption with deletion of the key)
- Tokenization: replacing e.g. ID numbers with less useful values with an one-way encryption

Location Data

[13] "Opinion 05/2014 on Anonymisation Techniques" by the Data Protection Working Party of the EU: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[14] General Data Protection Regulations (GDPR): http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

This solution approach mainly covers the problem for vehicle-tracking. Some parts may be transferred to other location data problems. Location Data is of high interest in several Transforming Transport pilots.

Without anonymization or pseudonymisation of the gathered location data it is simply possible to track a vehicle's route and consequently the driver's route and driving habits. This is nowadays a huge lack of privacy and increases the risk of physical and privacy attacks.

There are developed approaches to close those data gaps[15]:

- Using "pseudonym pools" for vehicles to switch between different identities.
  Each vehicle gets an amount of pseudonyms by its ID by a central authority (CA). Those pseudonyms are created by classic approaches like cryptography and PKI. Concluding, the CA is able to retrace any pseudonym to its original ID.
- Using a big amount of pseudonyms for every vehicle to cover cases like not using a vehicle for a long time.
  This approach causes a big overhead, due to the frequent pseudonym-change. Furthermore, it requires a bigger storage in every vehicle for the pseudonyms and updating data takes more time.
- Holding a fix amount of pseudonym in the storage and matching every pseudonym to a time-slot.
  The pseudonym gets circularly active at its given time-slot and deactivates whet its time-slot is over. The time-slots get synchronized by GPS.
- Allowing vehicles to exchange pseudonyms, which are dedicated to the same time-slot, at any time.
  - Due to the exchanged pseudonym it get much more complicated to retrace a pseudonym to a special vehicle. Therefore, this method will gain more privacy.
  - If the exchanged pseudonym is the current active one the exchange partner needs to be in the vehicles environment. It is selected by e.g. their position, speed or heading direction.
  - Strict similarity bounds should provide that the exchange is not trackable (since the position of both vehicles is effectively identical)
  - The driver's privacy is improved by introducing "random silent periods" in which the vehicle will not send any location data for a period of time.

[15]D. Eckhoff, C. Sommer, T. Gansen, R. German, F. Dressler: "Strong and Affordable Location Privacy in VANETs: Identity Diffusion using Time-slots and Swapping"(2010): http://ieeexplore.ieee.org/document/5698239/?reload=true

- Since every used pseudonym is a possible lack of privacy, a used pseudonym is marked and exchanged when possible. Then there is no need to select an exchange partner by specific attributes.
- The vehicle is prohibited to exchange too often, because too frequent exchanges are contra productive.

In urban and high-frequency environments the vehicles enjoy a much higher level of privacy than in the countryside. That is because the vehicles are more likely to find exchange partners in a big amount of other vehicles than on less-used tracks. Nevertheless with the possibility of the vehicles to hold a certain amount of pseudonyms and exchange the expired ones at any time if there is a partner in the area, they also reach a high level of privacy there.
A detailed guideline for anonymization of locations is found in chapter 7.

Safeguards

Create valid Safeguards[16] as kind of restraints to protect the project.

- Safeguards to protect personal information should be established and in appropriate cases also listed in the contract
- Duration of the storage of the measured data should be appropriate and only as long as needed
- contact point for user's to report abuse should be provided
- Compliance „with applicable laws regarding the use and disclosure of location information, and in particular, laws regarding the protection of minors."
- Clarify the user with regard to the responsible use of the offered service, the freely given access to their location and „other risks associated with the disclosure of location information to unauthorized or unknown third parties."

5.4.3 Aggregation of data

5.4.3.1 General Problem

Data aggregation is defined as the process of generating non-redundant metadata out of collected information or a compilation of gathered data in groups, so that it is not possible to

[16] CTIA – "Best Practices and Guidelines for Location-Based Services"(2008)

retrace the information to the single values[17]. Unless for single values (e.g. maximum value) in the context of aggregated data.

This aggregated data can be used for statistical analysis. For example, collecting information like age, occupation or income from a specific user group. Users which fit in the thereby arising user profile get displayed a customized advertisement.

Certain data compilations like business data aggregations may require higher or additional security restrictions than individual data, due to the bigger damage that is caused if business data is not correctly aggregated. Some specific cases are aggregated data from databases, IT-systems, information systems and discrete data from business projects.

To manage aggregated data over a longer period it is necessary to identify the different compilations. Companies and agencies that work with data aggregation are recommended to specify security requirements e.g. for authentication, authorization and access rights and name user and owner of the aggregated dataset. Furthermore, it is good practice to define a consistent compilation description and certain amount of processed data. In addition, the location where and the media on which the data is stored, transported and, in case of further processing, processed should be determined. [18]

One popular upcoming problem with data aggregation is that specific user-profiles are build. This happens intentionally by the company who collect the data and unintentionally by transferring the data to the company who aggregates the data. In both cases the privacy of the user is not necessarily given. Therefore, the company or agency who store, transfer and proceed aggregated data need to build up a solid security and safety program to protect the users privacy.

Since personal data is as valuable as never before there are third parties who are interested in those data, too. This is the second great aspect which needs to be considered when it comes to data aggregation.

5.4.3.2 Specific Problems for Data Asset Types

Most data-asset-types can be addressed by general approaches. In Transforming Transport the following data assets require a specific approach to enable the data assets:

[17] http://www.itwissen.info/Aggregation-aggregation.html

[18] https://www.protectivesecurity.gov.au/informationsecurity/Documents/PSPF-InformationSecurityManagementGuidelines-ManagementOfAggregatedInformation.pdf

- Location data
- Social media data
- Passenger data
- E-commerce customer data
- Long latency because of a frequent transmission of large datasets
- Distortion of the actual situation by mismatched data due to inaccurate aggregation criteria
- Building user-profiles out of data transfer-patterns
- Secure information aggregation without privacy restrictions

For example, over many transfers of account data from an account provider to a database server the transfer process itself could develop the most efficient way to transmit the data. With that it creates a pattern with which it recognizes the sent data. Then it is possible to read out transmitted data by this pattern and build up detailed user-profiles with all the information which is given by the account.[19] Most of the time those discrete data additionally display indirect information, which extend the already detailed user-profile. Therefore, it is not only possible to read all stored data from a bank account but imply social-environment, information about the employment relationship and movements of an individual from that. To secure the individual privacy, those correlations must not be possible to detect.

5.4.3.3 Solutions

One of the main aspects why data aggregation is used is reducing datasets, so that they keep their core statements, but redundant and personal data is removed. Accordingly, data aggregation itself is already an option for reducing latency in data transmission. But to be even more efficient it is recommended to use hierarchical networks with short-range transmissions and data protocols which are geared to have low latency (e.g. HEED or CLUDDA approach). [20]

Furthermore, the low-latency approach in combination with aggregated data reduces the number of transmitted datasets. This saves energy and bandwidth, which is a benefit for vehicle tracking since long-term transmission of location data produces often large datasets and they should be transmitted to the server. In some Industries it is common to send the data at the end of the trip and only some location data for monitoring in between.

[19] Thomas P. Vartanian, Robert H. Ledig, “Scrape It, Scrub It, and Show It: The Battle Over Data Aggregation”(2002)
[20] Rajagopalan, Ramesh and Varshney, Pramod K., "Data aggregation techniques in sensor networks: A survey" (2006) http://surface.syr.edu/eecs/22

To solve a distortion it is necessary to understand where it can arise. Common reasons why distortion occurs are a wrong selection of the sample or the wrong weighting of some groups or incorrectly collected data. Since an incorrect collection of transfer data is in general caused by defective or misaligned sensors, the main reason for distortion is the wrong selection of data. Therefore, the data needs to be extracted homogenously and fair-weighted.

Due to different structures or formatting of the data, it has to be transferred in to a consistent scheme for further evaluation. This topic is covered in the next section "Homogenization of different structured data". In the aggregation step of the translation, the dataset is summarized and separated in more or less coarse-grained groups of data with the same characteristics. Also dimensional hierarchies of the groups are developed. It is possible to aggregate subtotals, which are stored in the database, too. Those can be used for pre-computations, if there is the need of a rough estimation. On this basis it is possible to perform short- or long-term statistical analysis and discover trends or correlations. Due to the granularity of the aggregated data the significance of the evaluation is limited.

Further techniques are K-anonymity and L-diversity. K-Anonymity is a method, where gathered values get suppressed by replacing them with a placeholder. For example, the gathered value is "age 29" and the placeholder is "X" or some kind of generalization like "age between 25 and 35". L-Diversity works additionally to K-anonymity by assigning every attribute with at least l different values. [21]

Please note that different evaluation purposes for one dataset can lead to parallel data hierarchies with different subtotals. Moreover, such data processing abandons the neutrality of the dataset and can distort the explicit and implicit information.[22]

When it comes to aggregating location data it is necessary under aspects of privacy protection, that the actual location is not stored. One way to protect the user's privacy is to encrypt the location information of the environment in a defined radius with a homomorphic encryption. [23]

[21] http://www.wsgrdataadvisor.com/2015/09/personal-data-anonymization-and-pseudonymization-in-the-eu/
[22] https://www.tecchannel.de/a/bi-datenmanagement-teil-1-datenaufbereitung-durch-den-etl-prozess,1746250,7
[23] F. Li, B. Luo, P. Liu, "Secure Information Aggregation for Smart Grids Using Homomorphic Encryption"(2010)

5.4.4 Homogenization of different structured data

5.4.4.1 General Problem

The integration of data, in the present case of logistics data, usually poses major challenges. The challenges of the logistics industry, such as inhomogeneous data sources and complex, sometimes enormous datasets, place special demands on the systems to be implemented in the research project Transforming Transport.

The format of data can be distinguished by so-called data structures. For this purpose, a distinction is made between four different data structures. Structured data is used to combine traditional data with a fixed structural structure, a uniform format or a defined data type. This group includes, for example, relational data banks, CSV files, or simple tabular documents. These data structures form simple and fixed data areas and provide the content using the defined structure.

While in the past the majority of data from this category arose, non-structured data are now responsible for the strong growth in data volumes. Today, more than 90% of the new data belong to the non-structured data. These are classified into the following three categories: Data with ambiguous structural structure belong to the group of semi-structured data sources. This data structure includes, for example, XML documents. The files or documents have a fixed structure, which can, however, be used variably.

The group of so-called quasi- or multi-structured data categorizes unstable or irregular textual data that can only be structured with great effort. These multi-structured data types include, for example, web-based (textual) data streams with possibly inconsistent structure and format.

All data without predefined structures are assigned to the group of unstructured data. This includes, for example, documents, PDFs and other non-pre-structured data. Newer document formats in Microsoft Office, such as .docx, can be classified into the semi-structured category because of the integrated XML structure.

The challenges are to first homogenize the data into a system-wide defined structure and second to integrate the data into the system.

For additional Information on homogenization please refer to the Homogenization Guidelines of WP2.3.

5.4.4.2 Specific Problems for Data Asset Types

Data is an extremely valuable business asset, but often the data is of different structure. Therefore, it is mostly difficult to integrate, access and interpret the data. The data isn't always in a standard format, so data integration aims to make data agnostic and usable quickly across the business. This enables the data to be accessed and handled by its constituents.

Specific Problems for Data Asset Types in Transforming Transport are:

- weather forecast data,
- historic climatic data,
- satellite images,
- vehicle data,
- vehicle traffic data,
- vehicle load data,
- economic data,
- demographic data,
- data map (GIS),
- territory (GIS) data,
- passengers data,
- position data,
- infrastructure data

The specific Problems will be outlined individually whenever they occur.

5.4.4.3 Solutions

The following paragraphs outline solutions for the Transforming Transport Project. In order to make that data usable more quickly, a number of data integration patterns can be used to standardize the integration process. This Guideline focuses on the following solutions:

1. Filtering
   correction of syntactic and content defects of the data
2. Standardization
   Operational adjustment of the filtered data
3. Aggregation
   Compression of the filtered and harmonized data

4. Enrichment
   Calculation, combining data with other data sources and storage of business metrics

Filtering

The filtering sub process describes the elimination of the extracted data from syntactic and semantic deficiencies. As a rule, syntactic deficiencies are formal errors such as false control characters and semantic flaw errors are content-related errors, for example, in a business-economic nature obviously false turnover figures. The literature also distinguishes deficiencies of the below listed categories. Deficiencies of the first category can be identified automatically and then automatically corrected during the extraction process. In case of failures of the 2nd category: The defect detection is automatic within the content, but the correction must be made manually after the extraction process. Defects of the third category will just be recognized. The detected defects then need to be corrected fully manual.[24]

1. Category: automatic Approach (e.g. missing values or automatic format change)
2. Category: automatic recognition with manual correction (e.g. identification of inherent formats or inconsistent value constellations)
3. Category: manual recognition with manual correction (unrecognized semantic errors in operational sources)

Standardization

In this second step it is all about aligning the data in terms of coding, synonyms and homonyms. Therefore, all schemes and identifier have to be standardized. Furthermore, it is necessary to resolve key disharmonies. Those appear when data from multiple databases is merged. The joined data need one common primary key. The problem is usually resolved by means of an allocation table, which, for example, generates a new, artificial primary key for each customer. [25] The primary keys of the operating systems are then carried as foreign keys, so that overlapping data evaluations are possible. Finally, content-sensitive data is unified. This is less a technical, than a business-organizational problem. In doing so, the operational data - for example, the scope of the different datasets, the currency or period count - have to be converted into uniform values.
For resolving key disharmonies and for unifying context data, transformation rules can be

[24] https://www.tecchannel.de/a/bi-datenmanagement-teil-1-datenaufbereitung-durch-den-etl-prozess,1746250,3
[25] https://www.tecchannel.de/a/bi-datenmanagement-teil-1-datenaufbereitung-durch-den-etl-prozess,1746250,8

implemented that do this harmonization. At the end of this process, the data are available on a level that would be directly usable for analysis purposes.

Aggregation

This topic is covered in the previous section "Aggregation of data".

Enrichment

Enrichment is the final step of the transformation process. In this context value has to be added to the integrated data referencing the Use-Case-Scenarios. For example, further computation on the basis of existing data in terms of some given aspect like taxes.

In general, it is recommended to store the information, which is important for several use-cases or questions in the scenario, to minimize the computation effort. As an example, the average weekly numbers of deliveries can be calculated and integrated at the distributors' level or on the product level. Therefore, the information can be useful for the management in the distribution center and for the product manager.[25]

Basic Examples for the use of data enrichment are[26]:

- Geographic: such as post code, county name, longitude and latitude, and political district
- Behavioral: including purchases, credit risk and preferred communication channels
- Demographic: such as income, marital status, education, age and number of children
- Psychographic: ranging from hobbies and interests to political affiliation
- Census: household and community data

Additionally, using the same coding (e.g. UTF-8) for all data in the system is required. Furthermore, the integration of meta-data that describes the data is helpful to use the data within the Big Data Infrastructures.

[26] https://www.captechconsulting.com/blogs/Guiding-Principles-for-Data-Enrichment

5.4.5 Unclear License

5.4.5.1 General Problem

Often, the license of a data asset is not given or the detailed license name or description is unclear. Sometimes the data owning party has created an own license type but also some of the most common licenses like CreativeCommons or GnuGPL are used.

Since the license is the key element of validating if a data source is valid and can be used, it is necessary to clarify all open questions as soon as possible to avoid later legal problems or the lack of usable data assets.

5.4.5.2 Solutions

For the clarification of licenses it is essential to make sure to know

- What is licensed? The data itself or a database structure?
- Can the data owner give additional license information?
- Can a different data asset be used?

A good overview of open data licenses and their comparison can be found at http://opendefinition.org/licenses/ and can be a good guidance when looking for the right data source.

In general the licenses are often individual (sometimes even from country to country) and need to be legally proven. This document can not give a legally binding guidance, so for most problems a bilateral communication needs to be started to solve and help with licensing issues. Also if available the project partner should include its legal department early or if not possible contact the data management team.

The general approach to license should be as follows:

- Try to find data assets based on open data licenses
- If not possible try to find data assets with well described and complete license descriptions
- Contact legal departments to solve any problems with the license
- If no clear license can be found, contact the data management team and data rightsholder to discuss a clear approach for the data asset

5.4.6 Availability unknown or pending

5.4.6.1 General Problem

If the availability of a dataset is set in the future or unclear this may cause problems in the TT pilots if the data will not become available. Often times the later available data can have differing quality (see chapter quality issue).

5.4.6.2 Solutions

The data owner should be contacted and any information about the data availability plan should be gathered:

- Planned Availability date
- Possible delay reasons
- Definition of the data that will be made availability
- Clear requirements on data quality and size of data asset

If any of the aspects remain unclear, the pilot governance lead should be contacted and depending on the information quality and the importance of the data asset for the pilot an alternative data asset might be better suitable.

Please always start the communication with the data owner to prevent unforeseen consequences for the pilot.

If possible agree with the data owner on a metadata description that later be relied on and any prework can be adjusted to.

5.4.7 Cost issue

5.4.7.1 General Problem

The usage of additional data sources is often limited and based on additional costs. The additional costs for the consumption of external data sources can be based on:

- Effort of the data collection: sensors (temperature, GPS, cameras...), server/database space/holding data over a longer period of time
- Installation and maintenance of the hardware on transportation
- licence for using data: traffic maps, weather forecast

- effort for evaluating the data/ development of evaluation-algorithms

5.4.7.2 Solutions

The additional costs of the data assets has to be discussed within the Pilots-Team. In some cases regarding high licence fees, a free open-source alternative might be available. The quality of the free data source has to be verified then.

The effort of extracting data from sensors or other devices is in several use-cases from high importance. In this case the costs are necessary and in general justifiable. With outlining the costs for the data extracting all Stakeholders are informed.

Also compare the data costs with the values you declared in the Transforming Transport proposal and whether the amounts are beyond the scope.

If the usage of a data source remains too expensive it can be suitable to contact the data owner and discuss their participation within Transforming Transport so that the data can be used without a fee.

5.4.8 Legal issue

5.4.8.1 General Problem

In some cases the processing of data might be unclear. In most cases the legal issues are situated in the following aspects:

- regard the licensing of data itself
- protection of privacy (EU law and laws of particular countries – at least in those, where the application is used)
- licences/ licence-fees
- data as assent of a business → paying taxes

5.4.8.2 Specific Problems for Data Asset Types

Most data-asset-types can be addressed by general approaches. The following data types require specific approaches:

- location data
- tracking vehicle routes and protecting the privacy of the driver
- processing accidentally gathered data, e.g. camera recordings of cars/pedestrians who are not the object being observed but moving in the recording area

5.4.8.3 Solutions

For the protection of the privacy of the individual it is necessary to anonymize and/or pseudonymize the collected and stored data, especially in terms of established law. More detailed information is in section “Anonymization of personal data”.

In some cases the processing of data might be unclear. This can first regard the licensing of data itself. In case the legal and rights of usage are not clearly outlined, the data source provider has to be contacted and the legal aspects need to be clarified.

Other legal issues are often based on protection of privacy and the restricted licenses. The data usage needs to be aligned with the allowances and stated guidelines by the European Commission or its member state in which the Pilot is focused on. If the Pilot is used in more than one country, the legal situation needs to be verified in all particular countries. The restricted licenses can only be discussed within the pilot-team and risks have to be outlined. In most cases alternative data providers with a different license can be found.

5.4.9 Quality issue

5.4.9.1 General Problem

The Data Quality is often unclear. Therefore, the Quality issue is known as a complex assessment to identify the value of the data. The Data Quality approach in Transforming Transport includes the following aspects:

- Accuracy
- Completeness
- Update status
- Relevance
- Consistency across data sources
- Reliability
- Appropriate presentation
- Accessibility

Within the project, a certain level of data quality is crucial to operational and transactional processes and to the reliability of the analytics results. Basically the data quality is affected by the way it is maintained – from integration, storing and managing – all steps have a high impact in the quality of the data assets. Therefore, Data quality assurance (DQA) is a common process of verifying the reliability and effectiveness of data. Once the quality of data is identified, the

involved roles need to discuss if the process requires further data sources to gain data quality level.

As mentioned before, this is an ongoing process which has to be followed up frequently.

5.4.9.2 Specific Problems for Data Asset Types

One data category that has to accomplish quality requirements is location data. The gathered data should be as precise as possible to read as much information as possible out of it. To process those huge datasets they need to be in a similar or identical format and should be complete. To benefit from the data and react to the detected situation is it necessary to have the latest data and a regular updates from the location-sending device.

Another data asset type are social media data. Those are more likely to contain spam messages and therefore need a special evaluation.

5.4.9.3 Solutions

In WP2.3 there will be detailed guidelines and procedures on how to establish and measure data quality. For any questions please refer to them or contact the data management team.

5.5 Role management for TT pilots

To avoid duplicate information in the D1.3 deliverable the role management part of the guideline was cut. Please refer to chapter 3 for a detailed overview.

6 Specific Governance Guidelines

Next to the Gerneral Governance Guideline, there is a need for customized Guidelines with special requests within the Transforming Transport Project. These Guidelines will be added in this chapter and handled once they occur. The following sub-chapter describes a specific guideline for the anonymization of location data in Work Package 5 named Sustainable Connected Vehicles. The WP is referenced as S521 in this Guideline.

6.1 Guideline S521 for the anonymization of location data

6.1.1 Usage of this Guideline

This document is a detailed guideline for a specific data asset where the general guidelines are not sufficient and an in depth examination needs to be done.

The document gives a summary about the details of the data asset and the particular problem. Then a general support guideline sums up the first steps that the data responsible should do in order to make sure that the right data is identified and all requirements are valid.

Then the guideline addresses the common problems with regard to Data Protection, IPR and Data Governance. Depending on the specific problem, only the relevant topics are addressed. If you need general guidance, please refer to the general guidelines.

For each topic, a detailed overview of possible solutions and the effect on the use case and data asset is given.

6.1.2 Data Asset

This paragraph describes the Data Assets existing including the related responsibilities in each pilot use-case. The Data Asset is always connected to a Data Owner who is the representative of an organization. Together with the Data Asset Steward both roles are responsible for the correct usage of the data.

WP	**Pilot**	**Data Asset ID card**	**Pilot Contact for Guideline**
WP5	Sustainable Connected Vehicles	TT_WP5_Geo-tracking_SustainableConnectedVehicles_1.0	Jerome Fenwick jfenwick@synox.io

6.1.2.1 Description of data asset "SoFleet Geo-tracking"

The data asset describes the GPS coordinates of vehicles during their trips. The data is proprietary and owned by SoFleet.

The GPS coordinates could be truncated on the three last digits to meet personal data standards. The obtained precision would then be approximately 500m.

The availability of the data asset is given on 30.07.2017.

The data type is “location data” which can also be referred to in the corresponding general guideline.

The geographical coverage of the data asset is France and Europe.

An access to the data asset is only possible with the approval of SoFleet.

6.1.3 General support

To avoid duplicate information, the general support section has been cut. Please refer to the general governance guideline in chapter 5 for all information.

6.1.4 Specific Problems & Solutions with data asset

6.1.4.1 Problem: Anonymization of personal data

Data provider SoFLEET from Pilot “Sustainable Connected Vehicles” informed Answare (Pilot lead) that they are not able to provide location of vehicles precisely since this information is personal data even if it was anonymized. The problem is that if you know precisely the location of a person you can know where he/she lives, works, the school of his/her children, etc.

SoFleet's customers are companies that own car fleets and, in turn, these companies rent the car to people and/or their workers use the vehicles both for personal and professional issues.

Answare needs the geo-tracking data to achieve the following objectives related with location and visualization:

- **CC_O_0.3.2:** Real-time geolocation of vehicles in a map, with the possibility of showing the current journey path already covered by a given vehicle
- **CC_O_0.3.3:** Visualization of previous journeys of each vehicle
- **CC_O_0.3.4:** For every journey (both real-time and historical), it will be possible to select the kind of event to be displayed. They will be displayed over the map according to their location and timestamp
- **CC_O_0.3.5:** For every historical journey, it will be possible to replay the path of the vehicle in the map according to geolocation, timestamp and other values such as speed. To allow a soft replay, intermediate values will be interpolated

The data should be displayed in a dashboard that can be used by fleet managers (and the Answare team). Only a representative subset of the vehicle data is needed.

In Addition, the following objectives related to traffic detection need to be achieved:

- **CC_O_3.1:** Applying algorithms to carry out descriptive analysis to identify situations where a car is in a traffic jam and evaluation of the traffic congestion in real-time
- **CC_O_3.2:** Extension of SoFLEET smartphone application to allow drivers to be notified about nearby traffic jams
- **CC_O_3.3:** Building up the service that will send detected traffic jams to the drivers' smartphone application
- **CC_O_3.4:** Real-time visualization of traffic jams in a map
- **CC_O_3.5:** Heatmap visualization of historical traffic jams by hour

The goal is to achieve as high data precision as possible.

Legal Linking

Corresponding EU-law: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

6.1.4.2 Solution: Reduce data location accuracy by shortening longitude/latitude

This is the most common solution. SoFleet can remove decimals from the fields to provide a precision of 500 meters and to remove the possibility of personal linking. With the given objectives, the location of the GPS tracking need to be mapped to Geodata and maps so the referencing based on shorter long/lat itself is not sufficient. There needs to be an algorithm that maps the shortened locations back to the traffic network and existing roads.

This might result in significantly different tours when the reachability of locations (e.g. on a motorway) is limited. Also reoccurring locations might be mapped to single vehicles and drivers primarily in rural areas.

Additionally the reduction of the accuracy disables a meaningful data usage.

(Possible violation of objectives **CC_O_3.4, CC_O_3.5, CC_O_0.3.2)**

Summary: **Not Feasible**

6.1.4.3 Solution: Reduce data location accuracy by geomasking

This is also a common solution. A common geomasking practice is to aggregate or relocate individual-level information to pre-existing administrative or political boundaries, such as census tracts or zip codes.

Possible approach: One approach to facilitate this is the usage of pseudonym pools, which allow vehicles to autonomously switch between different identities. This removes the possibility of displaying journey paths but anonymises the vehicle identities. It would be possible to create a traffic jam system where vehicles send traffic data using their pseudonym and request traffic information back from the server. The consolidated information can be shown on a map. However, a high amount of overall vehicles are needed, otherwise the lack of other local vehicles makes a pseudonym mapping possible again.

(Violation of objectives **CC_O_0.3.2, CC_O_0.3.3, CC_O_0.3.4, CC_O_0.3.5)**

(Solves the issues for **CC_O_3.1, CC_O_3.2, CC_O_3.3, CC_O_3.4, CC_O_3.5)**

Summary: **Feasible anonymization for the traffic detection system**

Possible approach: "Donut" Method for Geomasking

There are several scientific articles regarding donut geomasking. For example, Hampton, Fitch et al. say "In the donut method geomasking, each geocoded address is relocated in a random direction by at least a minimum distance, but less than a maximum distance (10). In addition, each point is moved a distance inversely proportional to the underlying population density, which provides privacy protection while minimizing the introduced spatial error (10, 12, 13). For example, persons in high-density urban areas do not need to be moved as far as persons in low-density rural areas to achieve the same magnitude of anonymization. The donut method is an adaptive geomask because the dimensions of the mask around each point vary to meet specified anonymity constraints based on the underlying population density (16, 17). Accounting for population density variation also optimizes the donut method by minimizing the distances required for privacy protection while maximizing analytical validity."[27]

[27] Hampton, K.H.; Fitch, M.K.; Allshouse, W.B.; Doherty, I.A.; Gesink, D.C.; Leone, P.A.; Serre, M.L.; Miller, W.C. Mapping health data: Improved privacy protection with donut method geomasking. Am. J. Epidemiol. 2010, page 1063

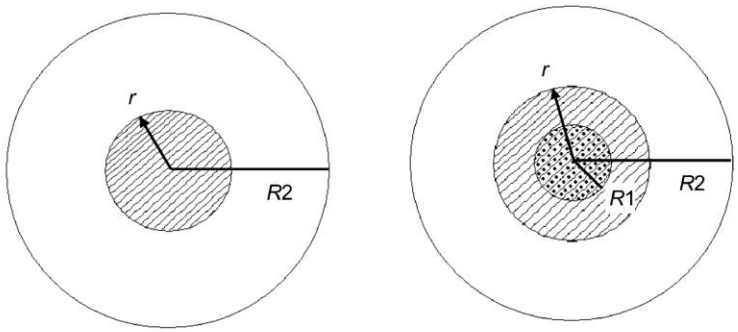


Figure 1. Comparison of random perturbation (left) and donut method (right) geomasks. For a given Max *k* geoprivacy level, the Euclidean distance *R*2 is calculated for each point from the underlying population density. The population within a circular region of radius *R*2 around a point is equal to Max *k*, with *R*2 being the maximum distance the point may be displaced from its original location. For the donut method (right), a Min *k* (dotted) is also given that defines the minimum displacement *R*1. The actual distance displaced, *r*, ranges in value from 0 to *R*2 for random perturbation (left) and from *R*1 to *R*2 for the donut method (right). The population within the circular region of radius *r* (striped) is the actual *k* achieved by the geomask.

Figure 12 Donut geomasking following Hampton, Fitch et al.[28]

For any more details please refer to the referenced article[29]

With donut geomasking (in comparison to shortening longitude and latitude) the problem of reoccurring locations can be solved because locations are randomly anonymized. To rudimental achieve the pilot objectives a detailed concept needs to be developed that addresses the following

- Identification of critical locations, that need to be anonymized (starting point, end point, points with long stay/without engine running)
- Randomized donut geomasking (based on donut ring locations that reference the needed geo data base and result in similar journey distances)
- Adaption of frequency of the GPS data input to the donut radius(to avoid journeys like displayed below, where the difference between the GPS tracks hints an anonymized position and extends the journey)

[28] Hampton, K.H.; Fitch, M.K.; Allshouse, W.B.; Doherty, I.A.; Gesink, D.C.; Leone, P.A.; Serre, M.L.; Miller, W.C. Mapping health data: Improved privacy protection with donut method geomasking. Am. J. Epidemiol. 2010, page 1063

[29] https://www.researchgate.net/publication/46149148_Mapping_Health_Data_Improved_Privacy_Protection_With_Donut_Method_Geomasking

This would result in anonymized journeys, which could be used for the work in a display tool. Also there would be no need to use pseudonyms since all critical locations are anonymized.

However the real time usage cannot be guaranteed because the evaluation if a location is critical can be difficult in real time. Also the approach weakens the results of the traffic detection system because the vehicle location is not always accurate.

Keep in mind: The implementation of this method is non-trivial and might outreach the scope of the pilot.

(Possible Violation of the real time component of objectives **CC_O_0.3.2, CC_O_0.3.4)**

Summary: **Feasible for historical data**

6.1.4.4 Solution: Address Real time issue by delaying the input data

The described real time issue with the donut geomasking can be addressed by delaying the anonymization and therefore usage of the data by a margin that clearly identifies a critical location. With this approach the data forwarding from SoFleet would be delayed and the monitoring of the current journeys would not be real-time. In the end, there cannot be an easy solution that enables the use of anonymized locations directly in real-time, so the pilot partners should decide on which aspect to focus on.

Summary**: Feasible, but TBD if it is useful in TT**

6.1.4.5 Solution: User-specific identification of critical locations

At least in theory it is possible for SoFleet to assign special critical locations for every user/vehicle with contract signing or something similar. When a vehicle is near that location, the location is anonymized with donut geomasking or just cut and then sent to the TT pilot. Of course this needs extreme effort.

Summary: **Feasible but only with extreme effort**

6.1.4.6 Solution: Cut critial locations

As a general alternative to the geomasking the critical locations could also just be removed. This needs the same algorithms to define critical locations, but then skips the location adjustment. Since the remaining journeys are often the same, a pattern recognition is easier compared to when we have randomized relocated positions. So this approach is easier to implement, but not

as efficient as donut geomasking. It depends on the frequency of GPS tracking and the amount of vehicle within the pilot if the anonymization would be sufficient.

The real-time issues also have to be addressed.

Summary: **Feasible for historical data**

6.1.4.7 Solution: User allows data usage for TT

It would be a possibility to ask for user approval regarding data usage. That could for example be done by a "traffic detection app" for TT or similar. This could include a direct location update initiated by the user itself.

There are several guidelines on how to provide location-enabled application (e.g. GSMA 2012 Guideline)

Summary: **Feasible but only with extreme effort**

7 Pilot Summary on IPR, Data Protection and Governance

In this chapter, the Data Management Plan gives a detailed overview about the single data assets in the Transforming Transport pilots and their information regarding IPR, Data Protection and Governance based on the data asset id cards filled by the pilot partners.

Not all pilots finished the description of their initial data assets as id cards for the Data Management Plan deliverable. Since the analysis of this Deliverable is based on the Data Management Plan all late data assets will be added with a new version of this deliverable (and also the DMP).

Whenever new data assets or new responsibilities occur, this list will be updated and submitted in later versions of this Deliverable. Also any new information from the pilot regarding missing content will be added when available.

All information that is missing or worked on is marked in grey.

7.1.1.1 WP4.2 Ausol

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_WP4.2_YahooWeatherAPI_AusolHighway_v1.0.xlsx	Yahoo! Inc.	no	https://policies.yahoo.com/us/en/yahoo/terms/product-atos/apiforydn/index.htm	01. Jan 17	OPEN DATA
TT_WP4.2_Twitter4_AusolHighway_v1.0.XLSX	Twitter4j	**YES**	http://twitter4j.org/en/index.html#license	01. Jan 17	WITHOUT APPROVAL
TT_WP4.2_Variable Message Sign - Highway_AusolHighway_v1.0.xlsx	CINTRA and Dirección General de Tráfico	NO	**TBC**	01. Jan 16	WITH APPROVAL
TT_WP4.2_APPmobile_AusolHighway_v1.0.xlsx	CINTRA	**YES**	**Private use (License under development)**	**30.09.2017**	OPEN DATA
TT_WP4.2_Traffic Flow - Highway_AusolHighway_v1.0.xlsx	CINTRA	NO	INTERNAL - Must be considered as confidential as per the terms of the Grant Agreement and the Consortium Agreement of the Transforming Transport Project.	01. Jan 16	WITH APPROVAL
TT_WP4.2_Traffic events - Highway II_AusolHighway_v1.0.xlsx	CINTRA	NO	INTERNAL - Must be considered as confidential as per the terms of the Grant Agreement and the Consortium	01. Jan 16	WITH APPROVAL

			Agreement of the Transforming Transport Project.		
TT_WP4.2_Traffic events - Highway I_AusolHighway_v1.0.xlsx	CINTRA	NO	INTERNAL - Must be considered as confidential as per the terms of the Grant Agreement and the Consortium Agreement of the Transforming Transport Project.	01. Jan 16	WITH APPROVAL
TT_WP4.2_Maintenance Vehicule Routes - Highway_AusolHighway_v1.0.xlsx	CINTRA	NO	**TBC**	01. Jan 15	WITH APPROVAL
TT_WP4.2_CCTV Cameras - Highway_AusolHighway_v1.0.xlsx	CINTRA	**YES**	INTERNAL	01. Jan 17	WITH APPROVAL
TT_WP4.2_SpeedRadar_AusolHighway_v1.0.xlsx	DGT	NO	Public Information http://infocar.dgt.es/datex2/dgt/	01. Jan 17	OPEN DATA
TT_WP4.2_CCTVcameras_AusolHighway_v1.0.xlsx	DGT	NO	Public Information http://infocar.dgt.es/datex2/dgt/	01. Jan 17	OPEN
TT_WP4.2_VariableMessageSign_AusolHighway_v1.0.xlsx	DGT	NO	Public Information http://infocar.dgt.es/datex2/dgt/	01. Jan 17	OPEN
TT_WP4.2_TrafficEvents_AusolHighway_v1.0.xlsx	DGT	NO	Public Information http://infocar.dgt.es/datex2/dgt/	01. Jan 17	OPEN
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

TT_WP4.2_Traffic Flow - Outside Highway (surroundings)_AusolHighway_v1.0.xlsx	Direccion General de Tráfico - ES	NO	INTERNAL - Must be considered as confidential as per the terms of the Grant Agreement and the Consortium Agreement of the Transforming Transport Project.	01. Jan 05	WITH APPROVAL
TT_WP4.2_Wheather Stations_AusolHighway_v1.0.xlsx	Direccion General de Tráfico - ES	NO	Public, no restriction	01. Jan 16	OPEN
TT_WP4.2_License Plate Recognition_AusolHighway_v1.0.xlsx	Direccion General de Tráfico - ES	NO	INTERNAL - Must be considered as confidential as per the terms of the Grant Agreement and the Consortium Agreement of the Transforming Transport Project.	01. Jan 15	WITH APPROVAL
TT_WP4.2_SocioEconomic_AusolHighway_v1.0.xlsx	INE	NO	Public use	01. Jan 10	OPEN
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

7.1.1.2 WP4.3 Norte Litoral

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_WP4.3_TrafficEvents_NorteLitoralHighway_v1.0.xlsx	-	NO	Public	01. Jan 17	OPEN DATA
TT_WP4.3_TrafficFlow_NorteLitoralHighway_v1.0.xlsx	AENL	NO	Public		WITH APPROVAL
TT_WP4.3_Twitter4_NorteLitoral_v1.0.XLSX	Twitter4j	**YES**	http://twitter4j.org/en/index.html#license	01. Jan 17	WITHOUT APPROVAL
TT_WP4.3_YahooWeatherAPI_NorteLitoral_v1.0.xlsx	Yahoo! Inc.	NO	https://policies.yahoo.com/us/en/yahoo/terms/product-atos/apiforydn/index.htm	01. Jan 17	OPEN DATA
TT_WP4.3_SpeedRadar_NorteLitoralHighway_v1.0.xlsx	-	NO	Public	01. Jan 17	OPEN DATA
TT_WP4.3_METEO_NorteLitoralHighway_v1.0.xlsx	AENL	NO	Public	01. Jan 17	WITH APPROVAL
TT_WP4.3_PMV_NorteLitoralHighway_v1.0.xlsx	AENL	NO	Public	01. Jan 17	WITH APPROVAL
TT_WP4.3_Traffic events - HighwayI_NorteLitoralHighway_v1.0.xlsx	CINTRA	NO	INTERNAL - Must be considered as confidential as per the terms of the Grant Agreement and the Consortium Agreement of the	01. Jan 16	WITH APPROVAL

			Transforming Transport Project.		
TT_WP4.3_CCTV Cameras - NorteLitoralHighway _v1.0.xlsx	CINTRA	**YES**	INTERNAL	01. Jan 17	WITH APPROVAL
TT_WP4.3_SocioEconomic_NorteLitoral_v1.0.xlsx	INE	NO	Public use	01. Jan 10	OPEN DATA
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

7.1.1.3 WP5.2 Sustainable Cars

Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**
TT_WP5_Referencial_SustainableConnectedVehicles_1.0.xlsx	SoFleet	NO	Proprietary	**30. Jul 17**	WITH APPROVAL
TT_WP5_Geo-tracking_SustainableConnectedVehicles_1.0.xlsx	SoFleet	NO	Proprietary	**30. Jul 17**	WITH APPROVAL
TT_WP5_CarSensors_SustainableConnectedVehicles_1.0.xlsx	SoFleet	NO	Proprietary	**30. Jul 17**	WITH APPROVAL

7.1.1.4 WP5.3 Sustainable Trucks

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_WP5_SatelliteImages_SustainableConnectedVehicles_1.0.xlsx	DLR	NO	https://lta.cr.usgs.gov/sites/default/files/Sentinel_Data_Terms_and_Conditions.pdf	03/04/2014 and 25/04/2016	OPEN
TT_T2 3_Data AssetAPI_Id Card_PTV_5 3_Truck_Traffic.xlsx	Axxes	NO	**TBC**	2015	WITH APPROVAL

7.1.1.5 WP6.2 Proactive Rail Assets

Data Asset	Rightsholder	Personal Data	License	Availability	Access
110103_16-TT_Thales_DataIDCard_NetworkModel.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_01-TT_Thales_DataIDCard_Ellipse.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_02-TT_Thales_DataIDCard_FMS.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_03-TT_Thales_DataIDCard_ITPS.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_04-TT_Thales_DataIDCard_TRUST.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL

110103_05-TT_Thales_DataIDCard_IntelligentInfrastructure.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_06-TT_Thales_DataIDCard_IntelligentInfrastructure.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_07-TT_Thales_DataIDCard_379RollingStock.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_08-TT_Thales_DataIDCard_SSI_Record.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_09-TT_Thales_DataIDCard_MetDesk.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_10-TT_Thales_DataIDCard_EarthWorks.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_11-TT_Thales_DataIDCard_OLEEx.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_12-TT_Thales_DataIDCard_LADS.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_13-TT_Thales_DataIDCard_NRWS.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_14-TT_Thales_DataIDCard_OLETempMonitoring.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL
110103_15-TT_Thales_DataIDCard_OLEALP.xlsx	Network Rail	NO	Network Rail Data License	**TBC**	WITH APPROVAL

7.1.1.6 WP6.3 Predictive Network Maintenance

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_6.3_GeometricalInspection_1_0.xlsx	ADIF	NO	**TBC**	**TBC**	WITH APPROVAL
TT_6.3_DynamicInspection_1_0.xlsx	ADIF	**TBC**	**TBC**	**TBC**	WITH APPROVAL
TT_6.3_Davinci_1_0.xlsx	ADIF	**TBC**	**TBC**	**TBC**	WITH APPROVAL
TT_6.3_Clima_1_0.xlsx	JUNTA DE ANDALUCIA	NO	**TBC**	**TBC**	WITHOUT APPROVAL
TT_6.3_CGRH24_1_0.xlsx	ADIF	**TBC**	**TBC**	**TBC**	**TBC**
TT_6.3_AEMET_1_0.xlsx	AEMET	NO	NO	1981	OPEN DATA
TT_6.3_WeatherInfo#2_1.0.xlsx	AEMET	NO	**TBC**	01. Jan 1920	OPEN DATA
TT_6.3_TramificationIdeadif_1_0.xlsx	ADIF	NO	**TBC**	January 2010	WITH APPROVAL
TT_6.3_Tramification_1_0.xlsx	ADIF	NO	**TBC**	JANUARY 2010	WITH APPROVAL
TT_6.3_TampingTasks_1_0.xlsx	ADIF	NO	TBC	TBC	WITH APPROVAL
TT_6.3_Sios_1_0.xlsx	ADIF/INECO	YES	**TBC**	**TBC**	WITH APPROVAL
TT_6.3_S&CInspections_1_0.xlsx	ADIF	**TBC**	**TBC**	**TBC**	WITH APPROVAL
TT_6.3_Pidame_1_0.xlsx	ADIF	YES	**TBC**	January 2010	WITH APPROVAL
TT_6.3_Icecof_1_0.xlsx	INECO	YES	TBC	TBC	WITH APPROVAL
TT_6.3_WeatherInfo#1_1.0.xlsx	ADIF	NO	NO	01. Jan 15	WITH APPROVAL

TT_6.3_Ferrodrone_1.0.xlsx	Ferrovial Agroman / ADIF	NO	NO	Depending on the flights planning and their number that must be approved by the competent authority	WITH APPROVAL
TT_6.3_FerroTrackGeometry_1.0.xlsx	Ferrovial Agroman / ADIF	NO	NO	01. Jan 08	WITH APPROVAL
TT_6.3_FerroSlopesOpPaper_1.0.xlsx	Ferrovial Agroman	NO	NO	01/01/2002-01/01/2008	WITH APPROVAL
TT_6.3_FerroSlopesOpDigital_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 08	WITH APPROVAL
TT_6.3_FerroMachineTampingTemp_1.0.xlsx	Ferrovial Agroman	NO	NO	06. Jan 17	WITH APPROVAL
TT_6.3_FerroMachineTamping_1.0.xlsx	Ferrovial Agroman	NO	NO	06. Jan 17	WITH APPROVAL
TT_6.3_FerroMachineSpec_1.0.xlsx	Ferrovial Agroman	NO	NO	04. Jan 17	WITH APPROVAL
TT_6.3_FerroMachineModeAut_1.0.xlsx	Ferrovial Agroman	NO	NO	06. Jan 17	WITH APPROVAL
TT_6.3_FerroMachineMode_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 15	WITH APPROVAL
TT_6.3_FerroMachineGPS_1.0.xlsx	Ferrovial Agroman	NO	NO	06. Jan 17	WITH APPROVAL
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

TT_6.3_FerroMachineFuelAut_1.0.xlsx	Ferrovial Agroman	NO	NO	06. Jan 17	WITH APPROVAL
TT_6.3_FerroMachineFuel_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 15	WITH APPROVAL
TT_6.3_FerroMachineEngineData_1.0.xlsx	Ferrovial Agroman	NO	NO	06. Jan 17	WITH APPROVAL
TT_6.3_FerroMachineEngineAut_1.0.xlsx	Ferrovial Agroman	NO	NO	06. Jan 17	WITH APPROVAL
TT_6.3_FerroMachineEngine_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 15	WITH APPROVAL
TT_6.3_FerroFenceOpPaper_1.0.xlsx	Ferrovial Agroman	NO	NO	01/01/2002-01/01/2008	WITH APPROVAL
TT_6.3_FerroFenceOpDigital_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 08	WITH APPROVAL
TT_6.3_FerroDrainageOpPaper_1.0.xlsx	Ferrovial Agroman	NO	NO	01/01/2002-01/01/2008	WITH APPROVAL
TT_6.3_FerroDrainageOpDigital_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 08	WITH APPROVAL
TT_6.3_FerroDesignPaper_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 02	WITH APPROVAL
TT_6.3_FerroDesignDigital_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 02	WITH APPROVAL
TT_6.3_FerroTrackOpPaper_1.0.xlsx	Ferrovial Agroman	NO	NO	01/01/2002-01/01/2008	WITH APPROVAL
TT_6.3_FerroTrackOpDigital_1.0.xlsx	Ferrovial Agroman	NO	NO	01. Jan 08	WITH APPROVAL
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

TT_6.3_VegetationIndex_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-nc-nd/4.0/	01. Jan 93	OPEN DATA
TT_6.3_TrainNoiseMalaga_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/3.0/es/	01/01/1991 - 01/01/1999	OPEN DATA
TT_6.3_SeismogenicZone_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/4.0/	10. Jan 10	OPEN DATA
TT_6.3_Ortoimages_1.0.xlsx	-	NO	NO	01/01/1990	OPEN DATA
TT_6.3_EURaildB_1.0.xlsx	-	NO	NO	01. Jan 92	OPEN DATA
TT_6.3_DigiTerrModel_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/4.0/	01. Jan 17	OPEN DATA
TT_6.3_VecTerrFile_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/4.0/	01/01/1990	OPEN DATA
TT_6.3_TopoTerrRast_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/4.0/	01/01/1990	OPEN DATA
TT_6.3_RecentSeismic_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/4.0/	01. Jan 17	OPEN DATA
TT_6.3_NatTopoBase_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/4.0/	**TBC**	OPEN DATA
TT_6.3_HydrogeoMap_1.0.xlsx	-	NO	https://creativecommons.org/licenses/by-sa/4.0/	01/01/1991 - 01/01/1999	OPEN DATA
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

7.1.1.7 WP7.2 Valencia Sea Port

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_T2.3_Data Asset & API_Id Card v0.3 CATOS.XLSX	NOATUM CONTAINER TERMINAL VALENCIA	NO	Internal CUSTOMER shall have the non-exclusive, non-transferable License only for use CATOS at the Site under the terms and conditions of License Agreement.	01. Jan 06	WITH APPROVAL
TT_T2.3_Data Asset & API_Id Card ORBITA v0.2.xlsx	NOATUM CONTAINER TERMINAL VALENCIA	NO	**TBC**	13. Feb 17	WITH APPROVAL
TT_T2.3_Data Asset _AIS FVAL v0.1.xlsx	Valenciaport	NO	CC BY-SA	**01. Jun 17**	WITH APPROVAL
TT_T2.3_Data Asset _PCS FVAL v0.1.xlsx	Valenciaport	NO	CC BY-NCSA	01. Mrz 17	WITH APPROVAL
TT_T2.3_Data Asset _SCADA FVAL v0.1.xlsx	Valenciaport	NO	CC BY-NCSA	01. Mrz 17	WITH APPROVAL

7.1.1.8 WP7.3 duisport inland port

Data Asset	Rightsholder	Personal Data	License	Availability	Access
(pending)	(pending)	(pending)	(pending)	(pending)	(pending)

7.1.1.9 WP8.2 Smart passenger flow

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_T8.1_PASSENGER_v1.0.xlsx	AEGEAN	NO	**TBC**	01. Apr 17	WITH APPROVAL
TT_T8.1_MOBILE PHONE LOCATION_v1.0.xlsx	ATHENS INTERNATIONAL AIRPORT	NO	**TBC**	**TBC**	WITH APPROVAL
TT_T8.1_FLIGHT PLAN_v1.0.xlsx	Eurocontrol	NO	Terms and Conditions of the NM Access Services	01. Apr 17	WITH APPROVAL
TT_T8.1_BOARDING PASS READING_v1.0.xlsx	ATHENS INTERNATIONAL AIRPORT	NO	**TBC**	**TBC**	WITH APPROVAL
TT_T8.1_BAGGAGE_v1.0.xlsx	AEGEAN	NO	**TBC**	01. Apr 17	WITH APPROVAL
TT_T8.1_AIRPORT SLOT_v1.0.xlsx	ATHENS INTERNATIONAL AIRPORT	NO	**TBC**	01. Apr 17	WITH APPROVAL

7.1.1.10 WP8.3 Smart turnaround

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_T8 3_Data Asset API_Id Card Template_RTS - Copy.xlsx	FlightRadar 24	NO	**Licence details for TT to be discussed with FR24**	2017	WITH APPROVAL
TT_T8.3_Data Asset & API_Id Card Template_ground-tasks.xlsx	**TBC**	**TBC**	**TBC**	**TBC**	**TBC**
TT_T8 3_Data Asset API_Id Card Template_RTFP.xlsx	Eurocontrol	NO	**Licence details for TT to be discussed with Eurocontrol**	2017	WITH APPROVAL
TT_T8.3_Data Asset & API_Id Card Template_schedule.xlsx	Jeppesen	NO	**TBC**	03. Jan 17	WITH APPROVAL
TT_T8.3_Data Asset & API_Id Card Template_FP.xlsx	Jeppesen	NO	**TBC**	03. Jan 17	WITH APPROVAL
TT_T8.3_Data Asset & API_Id Card Template_fuel.xlsx	Jeppesen	NO	**TBC**	03. Jan 17	WITH APPROVAL
TT_T8 3_Data Asset API_Id Card Template_WFCST.xlsx	NOAA	NO	The data may be used and redistributed for free but is not intended for legal use, since it may contain inaccuracies.	10Oct2006–Present (approx. two years to present online)	OPEN DATA

7.1.1.11 WP9.2 Tampere

Data Asset	Rightsholder	Personal Data	License	Availability	Access
TT_T9.2_Data Asset & API_Id Card_TrePublicTransport_Journeys.xlsx	City of Tampere/ ITS Factory	NO	http://www.tampere.fi/tampereen-kaupunki/tietoa-tampereesta/avoin-data/avoin-data-lisenssi.html	01. Jun 13	OPEN DATA
TT_T9.2_Data Asset & API_Id Card_TrePublicTransport_SIRI.xlsx	City of Tampere/ ITS Factory	NO	http://www.tampere.fi/tampereen-kaupunki/tietoa-tampereesta/avoin-data/avoin-data-lisenssi.html	01. Jun 13	OPEN DATA
TT_T9.2_Data Asset & API_Id Card_Sensior.xlsx	**TBC**	**TBC**	**TBC**	**TBC**	**TBC**
TT_T9.2_Data Asset & API_Id Card_TreTrafficLights.xlsx	**TBC**	**TBC**	**TBC**	**TBC**	**TBC**
TT_T9.2_Data Asset & API_Id Card_TreTrafficCameras.xlsx	**TBC**	**TBC**	**TBC**	**TBC**	**TBC**
TT_T9.2_Data Asset & API_Id Card_SocialMedia.xlsx	**TBC**	**TBC**	**TBC**	**TBC**	**TBC**
TT_T9.2_Data Asset & API_Id	**TBC**	**TBC**	**TBC**	**TBC**	**TBC**

Card_ProlongedDrivingTimes.xlsx					
TT_T9.2_Data Asset & API_Id Card_Digitraffic.xlsx	Finnish Transport Agency	NO	Creative Commons Attribution 4.0	**TBC**	OPEN DATA
TT_T9.2_Data Asset & API_Id Card_TreIncidentsAndRoadworks.xlsx	City of Tampere/ ITS Factory	NO	Tampere City Open Data License / ITS Factory	**TBC**	OPEN DATA
TT_T9.2_Data Asset & API_Id Card_TreParking.xlsx	Finnpark	NO	**TBC**	2/2015	OPEN DATA
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

7.1.1.12 WP9.3 Valladolid

Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_FCDtomtom.xlsx	TomTom	NO	**Due to strict privacy rules this will be handled completely inside TomTom and is not be accessible from the outside.**	**TBC**	WITH APPROVAL
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_TrafficLights.xlsx	City of Valladolid	NO	Project-wide access	**TBC**	WITH APPROVAL
TT_T2.3_Data Asset & API_Id Card T9.3-	City of Valladolid	NO	Project-wide access	**TBC**	WITH APPROVAL

Valladolid_TrafficIncidences.xlsx					
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_PneumaticLoops.xlsx	City of Valladolid	NO	Project-wide access	**TBC**	WITH APPROVAL
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_ORAdata.xlsx	City of Valladolid	NO	Project-wide access	**TBC**	WITH APPROVAL
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_MagneticLoops.xlsx	City of Valladolid	NO	Project-wide access	01. Nov 16	WITH APPROVAL
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_GPStraces.xlsx	LINCE	NO	Project-wide access	01. Mrz 17	WITH APPROVAL
TT_T2.3_Data Asset & API_Id Card T9.3-Valladolid_WeatherData.xlsx	AEMET	NO	Pilot-wide access	01. Jan 16	WITH APPROVAL
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

7.1.1.13 WP10 Dynamic Supply Networks

Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**
TT_10_InternetAccess_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2005	OPEN DATA

TT_10_ECommercesales_Company_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2005	OPEN DATA
TT_10_ECommercepurchases_Individual_EU_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2008	OPEN DATA
TT_10_ECommercepurchases_Individual_3M_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2005	OPEN DATA
TT_10_ECommercepurchases_Individual_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2005	OPEN DATA
TT_10_ECommercepurchases_Company_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2003	OPEN DATA
TT_10_DisposableIncome_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2003	OPEN DATA
TT_10_PrimaryIncome_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2003	OPEN DATA
TT_10_PopulationDensity_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2004	OPEN DATA
TT_10_Population_1.0.xlsx	Eurostat	NO	http://ec.europa.eu/eurostat/about/policies/copyright	2005	OPEN DATA
TT_10_Logika_data_1.0.xls	Logika	YES (personal data filtered out)	Internal CUSTOMER shall have the non-exclusive, non-transferable License only for the project	01. Okt 2010	WITH APPROVAL
Data Asset	**Rightsholder**	**Personal Data**	**License**	**Availability**	**Access**

As a first report on the data, we give pilot-wise summaries and created a risk analysis for the different pilots and their data assets.

	Risk	Personal Data	Access approval	Availability	Licensing
Pilot					
WP4.2 and 4.3		**High**	Medium	Low	Medium
WP5.2		Low	Medium	**High**	Medium
WP5.3		Low	Low	Low	**High**
WP6.2		Low	Medium	High	Medium
WP6.3		Medium	Medium	Medium	**High**
WP7.2		Low	Medium	Low	Low
WP7.3		Pending	Pending	Pending	Pending
WP8.2		Low	Medium	Medium	**High**
WP8.3		Low	Medium	Low	**High**
WP9.2		Low	Low	Medium	Medium
WP9.3		Low	Medium	**High**	**High**
WP10		Medium	Low	Low	Low

Special focus during the project lifetime of Transforming Transport needs to be put onto the following pilots:

WP4.2 and 4.3: The pilots contain the most data assets with personal data and need to put specific focus on anonymization. Since some data assets are used in both pilots a detailed coordination is needed.

WP6.2: Availability date not given (although possibly already available).

WP6.3 and 9.2: Some data asset id cards are not complete (although possibly without risk information).

WP7.3: Data assets missing (will be added once available and analyzed).

WP8.2 and 8.3: Licensing needs to be clarified (although possibly without risk information).

WP9.3: The licensing and usage of the TomTom data asset needs to be clarified.

As a general conclusion, most of the open topics for Transforming Transport IPR, Data Governance and Protection are focused around the licensing of the data assets and the in depth-clarification regarding the anonymization of personal data. Generally there is already a high number of open data assets used within the pilots.

A detailed report on the data management and used data asset will be done with Deliverable D1.4 in Month 18.

8 Summary

This deliverable summarizes the work performed in WP1/T1.3 "Management of Knowledge, IPR, Data Governance and Data Protection" with respect to the creation of a concept for data management within Transforming Transport. It sums up the decisions that were made and gives an insight in the concept and the structure of data management. It contains the approach to Data Governance and includes all Governance Guidelines.

This document defines recommendations for Transforming Transport on the Data Management in the following chapters:

- Data Strategy
- Data Asset Information
- Role Management
- Data Quality Management Guideline
- General Governance Guideline
- Specific Data Governance Guidelines

Nevertheless, due to the early stage of this project and the priority of managing data first, this document focuses on the Data Management and Usage. In later stage, the Data Committee will focus on the use of generated data within the Transforming Transport Project to deliver adequate recommendations for this purpose.

The concept of Reporting and Visualization is already described within D1.3 as following: The goal of the reporting is to give a complete overview analysis and summary on all data assets that were used within the Transforming Transport pilots and project as a whole. It should give a detailed information source for the European commission on the usage of data assets and their types in large scale Big Data projects. The analysis contains the following aspects for data assets:

- Technical specifications (Data format and interfaces, volume, …)
- Data type
- Data origin
- Data quality
- Access and Security
- IPR, Governance and Protection

Based on the analysis any occurring special characteristics will be displayed.

The technological baseline for the analysis are a data asset database and the TT Open Data Portal that enable a fast filtering and querying of the data assets. The analysis results will then be displayed in tabular form or with schematic charts.

This document will be updated based on the new data assets and new information in Transforming Transport in months 12 and 18. The guideline itself requires frequent updates. Whenever changes occur, the updated version of this document will be circulated to the responsible pilots.