DELIVERABLE

D3.2 – Data Management Plan

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

Funded by the European Union's H2020 GA - 731932

Document fiche	
Author:	Francisco Yedro [UPM] – editor; Oscar Corcho [UPM] – contributor; Víctor Rodríguez Doncel [UPM] – contributor;
Internal reviewers:	Michael Schygulla [PTV] Pablo Campillo [ANS]
Work Package:	WP3
Task:	T3.3
Nature:	Report
Dissemination:	PU

Document History			
Version	**Date**	**Contributor(s)**	**Description**
0.1	13/02/2017	UPM	Document Creation
0.2	03/03/2017	UPM	Inclusion of the first DMP
0.3	13/03/2017	UPM	Document sent for internal revision
0.4	14/03/2017	UPM	New sections added
0.5	15/03/2017	UPM	Final document for submission to internal peer review process
0.6	30/03/2017	UPM	Document reviewed, ready to submission

Keywords:	Pilot, Data Management Plan, Datasets, Backup, Storage, Standards and Metadata, Data Sharing, Archiving, Preservation, Ethical Aspects.
Abstract:	This deliverable provides the Data Management Plan of the project.

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 № 731932

Table of Contents

Definitions, Acronyms and Abbreviations

Acronym	Title
DGT	Dirección General de Tráfico
DMP	Data Management Plan
PCC	Project Coordination Committee
TT	TransformingTransport
WP	Work Package
WP3	WP Impact
WP4	WP Smart Highways
WP5	WP Sustainable Connected Vehicles
WP6	WP Dynamic Railway Infrastructures
WP7	WP Ports as Intelligent Logistics Hub
WP8	WP Smart Airport Turnaround
WP9	WP Integrated Urban Mobility
WP10	WP Dynamic Supply Networks

Table of Figures

Executive Summary

This document belongs to the framework of WP3 (Impact) of the TransformingTransport (TT) project and describes the Data Management Plan (DMP). It provides the first version of the deliverable.

The deliverable of the DMP gives an overview of how data are generated and collected, which standards and methodology for data generation and collection will be followed, what parts of datasets will be shared for verification or reuse and how data will be archived and preserved. The use of a DMP is required for all participating pilots.

The purpose of the DMP is to provide an analysis of the main elements of the data management policy that will be used with regard to the project research data and covers the complete research data life cycle. This deliverable is not a fixed document, it is a living document which will be updated over the lifetime of the project. This first version, delivered on March 2017, is a general overview of the DMP which will be refined in the next deliveries. Four more versions (M9, M12, M18 and M24) are expected to be delivered.

All pilots from the project have provided an initial version of their DMPs according to the template structure that has been elaborated. As a result, for the time being, 146 datasets have been identified overall for the complete set of TT pilots (10 out of 13 sent), 33 of which can be considered as open data and 113 of which will be maintained confidential in some form or another, as described in their corresponding sections (some of them present some controversy because the datasets are defined as Open Data but in the section “Access and Security” approval access by the owner is required, will try to solve in the following versions). These figures may change in the future, depending on the pilot’s progress.

1 Introduction

This document belongs to the framework of WP3 (Impact) of the TransformingTransport project and describes the Data Management Plan (DMP). It provides the first version of the deliverable.

A DMP is a formal short plan that outlines what data will be generated or collected, how data will be managed (access, storage, backup…), the standards in use, the workflow to make the data accessible for use, reuse and verification and which plans for data sharing and preservation exist ensuring that data are well-managed. In addition, the purpose of this DMP is to help pilots to manage the data and help other potential users to use the shared data.

This deliverable outlines the initial DMP, which is in line with the H2020 guidelines for data management plan creation and identifies the initial classes of datasets of the project.

As said before, this DMP is not a fixed document and it covers the whole research data life cycle.

Figure 1: Data life cycle [4]

At this early stage in the project, this document includes an overview of the datasets produced by the project, but not yet a complete description, something that will be provided in the next delivery of this document.

2 Guidelines for the Generation of DMPs for Pilots

2.1 DMP General Definition [1]

Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on:

- the handling of research data during and after the end of the project
- what data will be collected, processed and/or generated
- which methodology and standards will be applied
- whether data will be shared/made open access and
- how data will be curated and preserved (including after the end of the project)

2.2 Guidelines for the description of DMPs

This section describes the main guidelines that have been provided to pilot leaders for the provision of information on the datasets and data management strategies that they foresee for the successful delivery of the pilot implementations. Such guidelines have been used to collect the most relevant information at this stage for the generation of this DMP.

In this first version of the deliverable, two DMPs for each Work Package (WP4, WP5, WP6, WP7, WP8, and WP9) have been proposed, except for WP10, where there is only one DMP.

Each individual DMP is divided into the following sections:

- Context: contains a brief introduction of each pilot, as described in the Grant Agreement of the project (this will evolve in the following versions of the DMP, as pilot's progress).
- Data Description: summary of the dataset, including name, description, technology, volume, update frequency and so on. In this fist version of the deliverable, all the datasets will be included into the same data description. This information will be refined and improved in the following versions, dividing each dataset with the aim of offering the maximum amount of information for each of them.
- Standards and Metadata: standards and metadata used to describe the data.
- Data Sharing: conditions under which data are shared, including how users will be able to access these data.
- Archiving and Preservation: how data will be archived, and, if possible, why such option has been chosen. It will also be defined if the digital repository provides digital preservation.

- Ethical Aspects: whether there are any ethical or legal issues than can have an impact on data sharing such as whether consent for data preservation is gained or not, whether the data is protected by database law or copyright, personal data and so on.

All these sections are based on the information gathered in the template Data Asset & API_Id Card Template_draft produced as a result of task T2.3 (CEFRIEL), and which is described in the next paragraphs.

2.3 Template for Data Description

A template "Data Assets and API ID card" has been developed by CEFRIEL in task T2.3 for the description of all information relevant to each dataset to be used by any TT pilot. This template has been delivered to all the pilots, which have filled it in with all the metadata available.

This template consists of six sheets and, in each one, there are some attributes to be completed. The attributes have a description, notes for guidance and an example of allowed values for helping the pilots to fill in the information, also, on the right side of the table, there is a column called "Value" where de pilot can write down the information required. The different sheets listed below are being depicted in the following figures:

- General Information: This sheet contains the whole information related to the data description (composed of Name of the Data Asset / API, Internal name of the Data Asset / Data Asset / API, Name of the Data Asset / API Provider, Short Description, Extended Description, Version, Initial Availability Date and Data Type).

ATTRIBUTE	DESCRIPTION	NOTES FOR GUIDANCE	ALLOWED VALUES / EXAMPLES	MANDATORY / OPTIONAL / CARDINALITY	VISIBILITY (INTERNAL / PUBLIC)
Name of the Data Asset / API	Public Name of the Data Asset / Data Asset / API within the Ecosystem (or data catalogue)	The name must be univocal in the Ecosystem. The name should include only alphanumeric characters and the maximum length of the field is 50 characters, spaces included.	e.g., "Athene Airport Flight Plans"	MANDATORY	PUBLIC
Internal name of the Data Asset / Data Asset / API	Internal name of the Data Asset / Data Asset / API adopted within the Pilot/WP/Project	The name should include only alphanumeric characters and the maximum length of the field is 50 characters, spaces included.	e.g., "Athene NOP"	OPTIONAL	INTERNAL
Name of the Data Asset / API Provider	Complete name of the provider of the Data Asset / API		e.g., Athens International Airport	MANDATORY	PUBLIC
Short Description	Plain text and concise description of the content of the Data Assets / Data Asset / APIs from a business perspective	It is recommended to provide a business overview of theData Asset / API. This description will be visible into the Shared Data Portal of the Ecosystem. The length of the text cannot overcome the threshold of 500 characters (spaces included).	Plain text	MANDATORY	PUBLIC
Extended Description	Plain text description, intended to provide a complete overview of theData Asset / API	It is recommended to provide a functional overview of the Data Asset / API. This description will be visible into the Shared Data Portal of the Ecosystem.	Plain text	OPTIONAL	INTERNAL
Version	Identifier of the Version of the Data Asset / API into the Ecosystem	The value associated to the Version must be coherent with potential previous versions.	x.y, e.g., 1.0 x.y.z, e.g. 1.0.1	MANDATORY	PUBLIC
Initial Availability Date	Date from which the Data Asset / API Provider guarantees to make available this specific version of the Data Asset / API.		e.g., 30/09/2017	MANDATORY	INTERNAL
Data Type	Closed list of data types that classify the Data Asset / API	The taxonomy of data types will be defined by the pilots in collaboration with T1.3, T2.3 and T3.3. It may be used to support the description and the search of the Data Asset / API in the Shared Data Portal of the Ecosystem.	e.g., "Car traffic flows", "Flights traffic flows", Speed	OPTIONAL, MULTIPLE	INTERNAL

Figure 2: General Information Sheet

- Legal Information: This sheet contains the whole part concerning the legal aspects related to the data description (composed of Personal data, License, Rightsholder and Other Rights Information).

ATTRIBUTE	DESCRIPTION	NOTES FOR GUIDANCE	ALLOWED VALUES / EXAMPLES	MANDATORY / OPTIONAL / CARDINALITY	VISIBILITY (INTERNAL / PUBLIC)
Personal data	Whether the dataset contains personally identifiable information	Contains the dataset personal information? (e.g. full names, personal ID, credit card number)	yes / no	MANDATORY	INTERNAL
License	Public license if the "Access Level" is public or internal terms and conditions under which the dataset is provided to the consortium otherwise.	If public... which is the the well known-license (e.g. CreativeCommons, etc.)? else... which are the terms and conditions?	Either a URI pointing to a license: https://creativecommons.org/licenses/by/4.0/ Or plain text	MANDATORY	INTERNAL
Rightsholder	Legal rightsholder of the dataset.	If the rightsholder matches the data asset owner field, then this is not necessary	"ACME Inc."	OPTIONAL	INTERNAL
Other Rights Information	Plain text with additional information about the legal status of the dataset	Only if relevant: - Whether the dataset is protected not only by database law but also by copyright - If the dataset has been registered in a IP registry, which IP registry and which id - If with personal data, in which DPA (data protection agency) is registered - If with personal data, until which date consent has been given by individuals - If with personal data, for which purpose individuals have given consent - If with personal data, which level of privacy (low, medium, high) the data is - If with personal data, who is the controller and who is the processor -If getting access to the data requires a NDA.	Plain text. Usually empty.	OPTIONAL	INTERNAL

Figure 3: Legal Information Sheet

- Contacts: This sheet contains the whole information about the owner (or contact) of the dataset (composed of Data Asset/API Owner/Responsible and Data Asset/API Owner/Responsible contacts).

ATTRIBUTE	DESCRIPTION	NOTES FOR GUIDANCE	ALLOWED VALUES / EXAMPLES	MANDATORY / OPTIONAL / CARDINALITY	VISIBILITY (PRIVATE / INTERNAL / PUBLIC)
ata Asset/API Owner/Responsible	Name and Surname or Company of the Data Asset / API owner, i.e., the Business Contact for the Data Asset / API	This information won't be publicly available, but they will be used for internal communications.	e.g., Mario Rossi	MANDATORY	INTERNAL
ata Asset/API Owner/Responsible contacts	Email Address/phone number of the owner of the Data Asset / API	This information won't be publicly available, but they will be used for internal communications.	e.g., mario.rossi@email.com	MANDATORY, MULTIPLE	INTERNAL

Figure 4: Contacts Sheet

- Technical Documentation: This sheet contains the information about the documentation of the data (composed of Technology, Name of the System, Data Asset data model/ API Interface, Data Model - Standard Glossaries, Data Identifier - Standard used, Data Model - Specific Data Model, Data volume, Update frequency and Data Archiving and preservation). Some of the rows are very important to be filled, especially, the information related to “Data Archiving and preservation”, including the storage, amount of data allowed and back up.

ATTRIBUTE	DESCRIPTION	NOTES FOR GUIDANCE	ALLOWED VALUES / EXAMPLES	MANDATORY / OPTIONAL / CARDINALITY	VISIBILITY (PRIVATE / INTERNAL / PUBLIC)
Technology	Technology adopted to expose the Data Asset / API		e.g., REST, WEB SERVICE SOAP, CSV, Data stream, …	MANDATORY	PUBLIC
Name of the System	Name of the system that is the main source of the data asset	The name should include only alphanumeric characters and the maximum length of the field is 50 characters, spaces included.	eParking, BCBP (Bar Coded Boarding Passes), …	OPTIONAL, MULTIPLE	INTERNAL
Data Asset data model/ API Interface	Link to a document defining the data model of the Data Asset / the interface of the API.		e.g., ER diagram, excel table swagger.json/swagger.yaml	MANDATORY	PUBLIC
Data Model - Standard Glossaries	References to standard (ecosystem) Glossaries, Taxonomies, Data Models, etc. adopted in this Data Asset / API	Report the links to the Glossaries/Taxonomies/Data Models adopted.		OPTIONAL, MULTIPLE	PUBLIC
Data Identifier - Standard used	Plain text, contains if any standard is used	Outline if the data produced or used in the project are identifiable and locatable by means of a standard identification mechanism.	e.g., "Digital Object Identifier"	OPTIONAL	INTERNAL
Data Model - Specific Data Model	Links to non-standard Glossaries defined specifically for this Data Asset / API	The files describing the Data Model must include all the information necessary to understand the data shared by the Data Asset / API.		OPTIONAL	PUBLIC
Data volume	The tipical/expected volume of data			OPTIONAL	INTERNAL
Update frequency	The tipical/expected velocity of data			OPTIONAL	INTERNAL
Data Archiving and preservation	Plain text, contains the information about the location where the Data Asset / API will be archived	A data storage and back-up strategy must be defined. Consider whether the data is safely stored in certified repositories. Storing data on laptops, computer hard drives or external storage is not reccomended. Write down if exists any limitation in the preservation time according with the policy applied by the storage provider	e.g., "Zenodo"	MANDATORY	INTERNAL

Figure 5: Technical Documentation Sheet

- Access and Security: (composed of Access Level and Access Mechanism)

ATTRIBUTE	DESCRIPTION	NOTES FOR GUIDANCE	ALLOWED VALUES / EXAMPLES	MANDATORY / OPTIONAL / CARDINALITY	VISIBILITY (INTERNAL / PUBLIC)
Access Level	Definition of the Access Level of the Data Asset / API, i.e., whether the Data Asset / API will be open to all the applications of the TT or whether it is limited to one Pilot	The meaning of the the possible values is the following: - WITH APPROVAL means that whoever has access to the digital ecosystem (catalogue) has to ask directly to the data set's owner the permission to access the data - WITHOUT APPROVAL means that whoever has access to the digital ecosystem (catalogue) may access the data set without having to get the permission to do so. - OPEN DATA means that the dataset is published on the internet and available to everyone	"WITH APPROVAL" | "WITHOUT APPROVAL" | "OPEN"	MANDATORY	PUBLIC
Access Mechanism	Link to one or more files attached to the Data Asset / API Id Card documenting specific security aspects related to the Access Mechanism		e.g., security_specification.pdf	OPTIONAL, MULTIPLE	INTERNAL

Figure 6: Access and Security Sheet

- Data Quality KPI's: To be developed

This template is available on basecamp (Docs & Files -> WP2 Tech mgnt, T2.3 – Work in Progress)

3 Data Management Plan

3.1 WP4 – SMART HIGHWAYS - AUSOL LOAD BALANCING PILOT

3.1.1 Context

Starting from the previous phase and having that practice analysing data, in this case all the experience will be applied in Ausol´s case where the predictive modelling will be covered. Different approaches will be defined in order to face the short/medium/long period. DGT will also support partners as Public Road Authority and owner of data sources that will be considered in the project. [2]

The Sol Highway consists in two sectors: Malaga-Estepona and Estepona-Guadiaro, both are part of the E15 European itinerary. They gather both the short and medium-haul traffic from Málaga and its airport and the long-haul traffic with origin and destination in the Bay of Algeciras. It has 105,2km length and 17076 vehicles a day. [3]

3.1.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **Yahoo Weather API**: Yahoo Weather API allows you to get current weather information for your location. It makes use of YQL (Query Language) Query a SQL-like language that allows to obtain the meteorological information, the API is exposed like a service REST and returns the information in a data structure JSON. The data are updated every 2 seconds.
- **APPmobile**: APP for mobile phones developed ad-hoc for the Transforming Transport project with the aim of collecting data about origin and destinations plus travel time. Technology used: REST. Real time frequency.
- **CCTVcameras**: Last pictures collected by the CCTV cameras deployed on the roads that shows the current status of the traffic flows on real time. Technology used: Web service - DATEX II. Real time frequency.
- **CCTV – Highway**: Videos stored by AUSOL's CCTV cameras deployed on the roads. They show status traffic flows offline. MP4 videos.
- **Speed Radar**: Location of the speed radars (both, fixed and mobile locations) deployed by the Authorities (National, Regional and Local) along the road network. Technology used: Web service - DATEX II. Real time frequency.

- **Traffic Events**: Traffic events on the road that have an impact on the road traffic, such as road works, accidents, traffic congestions, etc. Technology used: Web service - DATEX II. Real time frequency.
- **Twitter4j**: Twitter4J is an unofficial Java library for the Twitter API that can easily integrate your Java application with the Twitter service. Technology used: JAVA. The data are updated every 2 seconds.
- **VariableMessageSign**: Information showed on the Variable Message Signs which can have an impact on the road traffic. Technology used: Web service - DATEX II. Real time frequency.

3.1.3 Standards and Metadata

No specific standards or metadata have been identified for the time being for the proposed datasets except for CCTV – Highway dataset that is using mp4 files standard.

3.1.4 Data Sharing

Users will be able to access data in two different manners, using our own data portal from M6 or using the URI of the datasets:

- http://infocar.dgt.es/datex2/ (CCTVcameras, SpeedRadar, TrafficEvents and VariableMessageSign datasets)
- https://query.yahooapis.com/v1/public/yql?q for YahooWeatherAPI dataset.

Other considerations or each dataset are describing as following:

- **Yahoo Weather API** dataset is open access and terms of use could be checked at https://policies.yahoo.com/us/en/yahoo/terms/product-atos/apiforydn/index.htm.
- **APPmobile** dataset, which rightsholder is CINTRA, is open access and terms of use will be defined (license under development).
- **CCTVcameras, Speed Radar, Traffic Events and VariableMessageSign** datasets, which rightsholder is DGT, are open access and terms of use could be checked at http://infocar.dgt.es/datex2/dgt/.
- **CCTV – Highway** dataset, which rightsholder is CINTRA, has access only with approval.
- **Twitter4j** dataset terms of use could be checked [8].

3.1.5 Archiving and Preservation

During the lifecycle of the project, each dataset will be archived into Indra's infrastructure (Stage 1 and 2) and Road operator's infrastructure (Stage 3) except YahooWeatherAPI which will offer the information in real time, then, no storage is required.

The amount and preservation of the data will depend on the policy applied by INDRA.

3.1.6 Ethical Aspects

All the dataset (except APPmobile, CCTV – Highway and Twitter4j datasets) don't include personal data. No other ethical aspect has been reported.

3.2 WP4 – SMART HIGHWAYS - NORTE LITORAL LOAD BALANCING PILOT

3.2.1 Context

This phase is the replication pilot. It will be conducted in conjunction with the main partner of the road (or transport) work package in order to share the maximum of information and best practices learnt during the pilot in the first case study. The methodology that has been developed in the main pilot will be used in this phase to be sure that this way of working is correct. At this point will be involved all the data collected over time. [2]

The Norte Litoral Highway has 119km length and runs along the northwest coast of Portugal, from Oporto to Caminha, near the Spanish border, with a branch towards the interior between Viana do Castelo and Ponte da Lima. [3]

3.2.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **Yahoo Weather API**: Yahoo Weather API allows you to get current weather information for your location. It makes use of YQL (Query Language) Query a SQL-like language that allows to obtain the meteorological information, the API is exposed like a service REST and returns the information in a data structure JSON. Technology used: not defined. The data are updated every 2 seconds.
- **CCTV – Highway**: Videos stored by NORTE LITORAL's CCTV cameras deployed on the roads. They show status traffic flows offline. MP4 videos. Technology used: not defined.
- **Speed Radar**: Location of the speed radars (both, fixed and mobile locations) deployed by the Authorities (National, Regional and Local) along the road network. Technology used: Web service - DATEX II. Real time frequency.
- **Traffic Events**: Traffic events on the road that have an impact on the road traffic, such as road works, accidents, traffic congestions. Technology used: Web service - DATEX II. Real time frequency.
- **Twitter4j**: Twitter4J is an unofficial Java library for the Twitter API that can easily integrate your Java application with the Twitter service. Technology used: JAVA. The data are updated every 2 seconds.
- **Traffic Data**: Traffic Data obtained with various counters along highway. Technology used: SQL Server updated in packets of last 5 minutes.

- **Socioeconomic**: National, regional and local socio-economic inputs from the influence area of the pilot test. Technology used: CSV and/or XLS updated every year.
- **PMV Messages**: Messages sent to and received from highway panels. Technology used: not defined.
- **Meteo Information**: Meteorology information of the highway. Technology used: SQL Server updated in packets of last 5 minutes.

3.2.3 Standards and Metadata

No specific standards or metadata have been identified for the time being for the proposed datasets except for CCTV – Highway dataset that is using mp4 files standard.

3.2.4 Data Sharing

Data generated by Speed Radar and Traffic Events datasets are Open Access and we do not consider any restriction, also, Socioeconomic, Yahoo Weather API and Twitter4j datasets are open access but we must check their licenses. [7]

Data generated by PMV Messages and Traffic Data datasets are public but access with approval is needed.

3.2.5 Archiving and Preservation

Yahoo Weather API and Twitter4j datasets will offer the information in real time, then, no storage is required.

Meteo information, Traffic Data and PMV Messages datasets will be archived and preserved in a SQL Server Cluster and Backup Storage. The amount and preservation of the data will depend on the policy applied by the owner.

Socioeconomic, Speed Radar, Traffic Events and CCTV – Highway datasets will be archived into Indra's infrastructure (Stage 1 and 2) and Road operator's infrastructure (Stage 3). The amount and preservation of the data will depend on the policy applied by INDRA.

3.2.6 Ethical Aspects

All the dataset (except CCTV – Highway and Twitter4j datasets) don’t include personal data. No other ethical aspects has been reported.

3.3 WP5 – SUSTANINABLE CONNECTED VEHICLES - SUSTAINABLE CONNECTED CARS PILOT

3.3.1 Context

This task will carry out the pilot to enhance the management of car fleets analysing information that can be extracted from the aggregated analysis of the complete car dataset (its position, CO2 emissions, engine status, engine revolutions, light sensors, temperature, ESP interventions, incidents suffered, strong decelerations, etc.). Having this information (updated every few seconds) for a large number of cars will allow to perform extremely interesting analysis on citizen's mobility, CO2 real emissions and fine dust concentration, traffic incidents in real time, environment/road status, predictive analysis on mobility or traffic incidents, etc. And finally, combining information gathered from cars with data coming for other external sources such as weather stations, highways (use case 'Smart Highways'), other urban transport (use case 'Multi-modal urban mobility') important conclusions can be extracted for the benefit and modernization of EU transport and mobility. [2]

3.3.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **Autoaid API & Enhanced DTC**: Autoaid provides the infrastructure to access telematics web services to various functionality required to read data like identifier, sensor values or fault codes from electronic control units (ECU) build into a vehicle and translates their output for users. Technology used: REST and JSON.
- **Car sensor**: The data is collected directly from vehicles in real time (except for some data such as travelled distance which is sent once at the end of the trip...). Technology used: REST and JSON.
- **Referential**: To have more precision on each vehicle, we use a service (SIV) which is available for French cars only. Another part of the data can be filled by users. Technology used: REST and JSON.
- **Geo-tracking**: The dataset contains GPS coordinates of each trip. Technology used: REST and JSON.

3.3.3 Standards and Metadata

No specific standards or metadata have been identified for the time being for the proposed datasets.

3.3.4 Data Sharing

No open dataset identified for the time being, users has to ask directly to the dataset's owner the permission to access the data.

3.3.5 Archiving and Preservation

Autoaid API & Enhanced DTC dataset has two replicas. Car sensor and Geo-tracking datasets are archived in cluster mode InfluxDB and Referential dataset is archived in a PostgreSQL transactional data base. There is a full backup of the whole datasets nightly.

The amount and preservation of the data will depend on the policy applied by MIT and PostgreeSQL.

3.3.6 Ethical Aspects

All the datasets don't include personal data. No other ethical aspect has been reported for the time being.

3.4 WP5 – SUSTANINABLE CONNECTED VEHICLES - SUSTAINABLE CONNECTED TRUCKS PILOT

3.4.1 Context

Resilience of supply chains depends to a large extent on the reliability of the transport links between production and logistics sites. Traffic disturbances, extreme weather situations or site specific events cause deviations to planned operations leading to delays in delivery time. Time buffers or more frequent deliveries splitting delivery are measures applied to guarantee cargo availability at production and logistics sites just in time. The sensing truck pilot is to address the problem of reliability of transport services by including big data sources from road users, satellite images and traffic events. The aim of the sensing truck pilots is to set up and validate added value services derived from a big data platform. Core is the processing of big data volumes from various sources and to generate aggregated data layers to be deployed in transport and logistics application for dynamic planning and execution. [2]

3.4.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **Traffic Data**: Traffic Data from truck fleets. Technology used: XML.

3.4.3 Standards and Metadata

No specific standards or metadata have been identified for the time being for the proposed datasets.

3.4.4 Data Sharing

No open dataset identified for the time being, users has to ask directly to the dataset's owner the permission to access the data.

3.4.5 Archiving and Preservation

To be refined further in the following versions of the DMP.

3.4.6 Ethical Aspects

The dataset doesn't include personal data. No other ethical aspect has been reported for the time being.

3.5 WP6 – PROACTIVE RAIL INFRASTRUCTURES - PREDICTIVE RAIL ASSET MANAGEMENT PILOT

3.5.1 Context

The pilot will be delivered to overcome the following major barriers to moving to a predict-and-prevent rail maintenance approach. Rail infrastructures involve a complex supply chain between equipment manufacturers, maintainers and operators. Acquiring long term, performance (maintenance, fault) data, understanding the quality, accuracy, and provenance of the largely unstructured data, processing it to identify emerging faults (diagnostics) and disseminating useful, timely prognosis information with known confidence levels for preventive and predictive maintenance has proved historically problematic. This pilot will apply Big Data predictive and prescriptive analytics to a UK national rail route, to reduce the long term cost of maintenance and increase network availability through the facilitation of focused short and medium term proactive interventions. The Pilot will be run on historical and real-time data sources. [2]

3.5.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- Asset Maintenance and Renewal Analysis for Overhead Line Equipment (OLE). Technology used: not defined.
- Asset Failure Diagnosis and Prognosis for Points and Track Circuits. Technology used: not defined.
- Track Maintenance using On-Train Monitoring. Technology used: not defined
- Geo-spatial location of assets, tunnels, cuttings, embankments. Technology used: not defined.
- Localized weather and environmental data: humidity, temperature, precipitation, ice, wind speed & direction, lightning/electrical storm activity. Technology used: not defined.
- Usage: km, number of passes by train type / frequency, passenger / freight, speed, electric/diesel, performance characteristics. Technology used: not defined.
- Power feed, voltage, current draw, current leakage. Technology used: not defined.

- Component manufacturer e.g. contract strips, manufacturer name, model number, serial number. Technology used: not defined.
- Age, wear %, estimated remaining life of asset of new vs. current. Technology used: not defined.
- Temporary Speed Restriction (TSR) data. Technology used: not defined.
- OLE maintenance history, schedule, and renewal history. Technology used: not defined.
- Interaction between pantograph and OLE: pantograph force impact, pantograph electrode wear, maintenance history, pantograph instrumentation. Technology used: not defined.
- Yellow train measurement data. Technology used: not defined.
- Fault (FMS), maintenance (Ellipse), and Inspection records. Technology used: not defined.
- Track geometry, quality, & age. Technology used: not defined.
- Schedule 8 payments (delays minutes/incident, fault types). Technology used: not defined.
- UGMS (Unmanned Track Geometry Measurement Systems) from passenger trains. Technology used: not defined.
- Bump box (Accelerometer) ride quality data. Technology used: not defined.
- Tamping history. Technology used: not defined.
- Access to Intelligent Infrastructure Historian and its route database. Technology used: not defined.

3.5.3 Standards and Metadata

Naming of objects will follow the Thales RDBMS code standards document (83512046-DDQ-GBR-EN-001). Where the convention changes, this will be documented.

The software tools used will be easily used and exportable to an open format e.g. CSV, text, JSON, or equivalent, also WP6 will adopt the OSA-CBM (ISO 13774) data model.

3.5.4 Data Sharing

Data will be reused in the sense that existing repositories of data will be sent to WP6 for ingestion. Permission from the customer is required as the data will be securely stored and processed in the cloud. The customer/provider of the data remains the owner, permission is granted to allow Thales to use the data. Permission to use the raw data can be withdrawn by

the provider at any time. Any insights, trends, aggregations, or other data generated by WP6 remains the IP of Thales.

3.5.5 Archiving and Preservation

The data, metadata, documentation, and code are all deposited in a cloud infrastructure, likely to be Microsoft Azure. The data, metadata, documentation, and code will be secured by means of encryption and either certificate or key/password combination such that only authorised users may access the required assets. This does not preclude subsets from being in the public domain. The resource group selected within the Cloud provider chosen shall, where possible, be within the EU.

3.5.6 Ethical Aspects

No ethical issues have been identified with long term preservation of data as:

- The customer can withdraw their own data at any time.
- The data is encrypted with the customer's own encryption keys.
- No personally identifiable information is stored, only asset information.
- The information can, and has been for certain customers, obfuscated or anonymised at the customer's request. Although anonymised data is of limited value and obfuscated is by far the preferred method.
- At all times when handling data WP6 contributors shall abide by Thales Ethics Policy and the ethics policy agreed by the Transforming Transportation working group.

Legal issues with sharing data are as follows:

- The raw data belongs to the customer and is their information to share, not ours.
- The information generated by WP6 contains Thales IP. This information may be shared subject to Thales' written approval.
- Due to the data processing laws of various countries, certain data may not leave the country in which it was created. This will be handled on a case by case basis. Where possible resource groups will be created within the country of origin. Where this is not possible, and where necessary, the data will be anonymised, obfuscated, or not stored or processed.

3.6 WP6 – PROACTIVE RAIL INFRASTRUCTURES - PREDICTIVE HIGH SPEED NETWORK MAINTENANCE PILOT

3.6.1 Context

The objective of this pilot is to improve the reliability of high speed rail networks by optimising operator´s performance and maintenance of the rail infrastructure. The pilot will consist on the application of Big Data technologies to process heterogeneous data collected in order to understand the variables that have impact on the operators performance and model the nature of the maintenance incidents occurred in the infrastructure (tracks, tunnels, bridges) based on rail traffic, rolling-stock flows, maintenance data, planning & control Data and other information sources. This analysis will also include external variables such as weather forecasts or specific events (summer holidays) to anticipate the maintenance activities on the rail network and therefore to improve the operations of the maintenance of the whole rail infrastructure. Moreover, this solution might also allow rail operators to predict in real-time the impact of certain events on traffic management. [2]

3.6.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP: The list is provided below:

- **AEMET**: Observing networks used for meteorological and climatological studies due to their good spatial distribution throughout the Andalusian territory and their long series of observations. Technology used: CSV, PDF or XML.
- **CGRH4:** It is a multidisciplinary platform on which all of the organization's operating areas are represented (traffic, infrastructure maintenance, stations, civil defence and communication), the incidents are solved globally.
- **CLIMA:** The Environmental Climatology Information Subsystem (CLIMA) consists of an extensive network of meteorological stations belonging to different organisms and the computer application that allows the integration of data, the control of their quality, and their joint exploitation. Technology used: XLS.
- **DAVINCI:** The DaVinci platform is an evolution in the multidisciplinary integration of railways, designed to provide all-round management of processes, systems and users by grouping into a single system all the previously independent subsystems.
- **MDT05/MDT05-LIDAR**: Digital model of the terrain related to the National territory with a mesh of 5x5m. Technology used: Web Service with an estimated volume of 1107x60 MB.

- **DYNAMIC INSPECTION**: The high-speed lines have the periodic dynamic inspection of track. Main data from these inspections are saved in databases to evaluate the problems in the track.
- **European Railway Data**: The Agency is responsible for developing and maintaining several registers and databases in order to ensure transparency and equal access to documents for all railway market actors. These datasets include aspects such as incidents, safety, interoperability or rolling stock. Technology used: Web Service / API REST.
- **Railway Design Projects [digital support]**: Full construction designs including all the elements of the infrastructure and the as-built info collected during the construction stage. Technology used: PDF, CAD, JPG, XLS, DOC, TXT, MOV, MPEG4, DWG … with an estimated volume over 100TB.
- **Railway Design Projects [paper support]**: Full construction designs including all the elements of the infrastructure and the as-built info collected during the construction stage. Technology used: Paper support with an estimated volume of 260m3, 36 million A4 sheets.
- **Maintenance Operation Drainage Clearing [digital support]**: Maintenance reports regarding the operations carried out for drainage clearing. Technology used: PDF, CAD, JPG, XLS, DOC, TXT, MOV, MPEG4, DWG … with an estimated volume of 1GB a year and daily update frequency.
- **Maintenance Operation Drainage Clearing [paper support]**: Maintenance reports regarding the operations carried out for drainage clearing. Technology used: Paper support.
- **Ferrovial Drone flights**: The drone will provide topographic data from the flights to be carried out in the Pilot area. These datasets comprise cloud point of the terrain with their coordinates, data from the drone (location, altitude) as well as aerial photos. Technology used: LAS format and aerial photos with an estimated volume of 2GB per flight. The update frequency depends on the flight planning.
- **Maintenance Operation Line Fencing Preservation [digital support]**: Maintenance reports regarding the operations carried out for clearing of embankment slopes. Technology used: PDF, CAD, JPG, XLS, DOC, TXT, MOV, MPEG4, DWG… with an estimated volume of 1GB a year and daily update frequency.
- **Maintenance Operation Line Fencing Preservation [paper support]**: Maintenance reports regarding the operations carried out for drainage clearing. Technology used: Paper support.
- **Machinery Engine Work**: Data related the machinery engine when is working (standstill periods, working stages, transit). Technology used: XLS and CSV with an estimated volume of 10MB a year and monthly update frequency.

- **Machinery Automatic Engine Work**: Data related the machinery engine when is working (standstill periods, working stages, transit) (automatically obtained). Technology used: XLS and CSV with an estimated volume of 200MB per machine and 5 minutes of update frequency.
- **Machinery General Engine Data**: Data related to engine data of the machinery focused on its operation [automatically obtained]. Technology used: XLS and CSV with an estimated volume of 200MB per machine and 5 minutes of update frequency.
- **Machinery Fuel Consumption**: Data regarding machinery fuel consumption during the maintenance operations. Technology used: XLS and CSV with an estimated volume of 10MB a year and monthly update frequency.
- **Machinery Automatic Fuel Consumption**: Data regarding machinery fuel consumption during the maintenance operations (automatically obtained). Technology used: XLS and CSV with an estimated volume of 200MB a year per machine and 5 minutes of update frequency.
- **Machinery GPS location**: Data related to the location of the machinery according to their GPS coordinates and the theoretical distances to reach the maintenance worksite. Technology used: XLS and CSV with an estimated volume of 200MB a year per machine and 5 minutes of update frequency.
- **Machinery Work Mode**: Data related the machinery work mode (standstill periods, working stages, transit). Technology used: XLS and CSV with an estimated volume of 10MB a year and monthly update frequency.
- **Machinery Automatic Work Mode**: Data related the machinery work mode (standstill periods, working stages, transit) [automatically obtained]. Technology used: XLS and CSV with an estimated volume of 200MB a year per machine and 5 minutes of update frequency.
- **Machinery Technical Specifications**: Technical specification data (type, manufacturer, plate (national and international), max speed, year of acquisition…). Technology used: PDF, XLS, JPG, MPEG4 and DOC with an estimated volume over 1GB.
- **Machinery Number of Tamping Insertions**: Data related the number of tamping insertions carried out by the tamping machine [automatically obtained]. Technology used: XLS and CSV with an estimated volume of 200MB a year per machine and 5 minutes of update frequency.
- **Machinery Tamping Device Temperature**: Data related to the monitoring of the temperature of the tamping devices while the machine is working [automatically obtained]. Technology used: XLS and CSV with an estimated volume of 200MB a year per machine and 5 minutes of update frequency.

- **Maintenance Operation Embankment Slopes Clearing [digital support]**: Maintenance reports regarding the operations carried out for clearing of embankment slopes. Technology used: PDF, CAD, JPG, XLS, DOC, TXT, MOV, MPEG4, DWG… with an estimated volume of 1GB a year and daily update frequency.
- **Maintenance Embankment Slope Clearing [paper support]**: Maintenance reports regarding the operations carried out for drainage clearing. Technology used: paper support.
- **Track Geometry**: Data related to the track geometry before and after of the maintenance activities used to control the track parameters are within the required tolerances. Technology used: XLS, CSV and paper with an estimated volume over 10MB and daily update frequency.
- **Maintenance Operation Track Bed Profiling [digital support]**: Maintenance reports regarding the operations carried out for clearing of embankment slopes. Technology used: PDF, CAD, JPG, XLS, DOC, TXT, MOV, MPEG4, DWG… with an estimated volume of 1GB a year and daily update frequency.
- **Maintenance Operation Track Bed Profiling [paper support]**: Maintenance reports regarding the operations carried out for drainage clearing. Technology used: paper support.
- **GEOMETRICAL INSPECTION**: The parameters of the track geometry are registered by the Vehicle Track Geometric Control. These parameters are evaluated the excel file used to calculate quality index with the data obtained in the geometric inspection vehicle.
- **Hydrogeologic map**: Hydrogeological GIS information of the National territory. Technology used: Web Service.
- **ICECOF**: ICECOF is a monitoring and control system of the punctuality commitments compliance of the railway operation. ICECOF manages the full workflow of all incidents, the system registers and analyses causes, solutions, impact and responsibilities. It includes all the related documentation regarding an incident. Technology used: XML.
- **BTN100**: Topography base of National territory map in scale 1:100.000. Technology used: Web Service with an estimated volume of 1107x2GB.
- **PNOA MÁXIMA ACTUALIDAD**: Recent orthophotographs mosaic of the national territory in scale 1:50.000.
- **PIDAME**: This maintenance tool is used in all the maintenance bases of the high speed lines in Spain. All the maintenance tasks have to be listed in the database before taking place. With this tool the works that have to be done in the track are planned. Technology used: XML.

- **Recent Seismicity**: Recent seismic activity and earthquakes within National territory and boundaries. Technology used: Web Service with daily update frequency.
- **S&C INSPECTIONS**: The aim of S&C inspection is to prevent the values of track gauge, dimension of deflection and groove width in the diamond area from being exceeded by timely maintenance. Technology used: XML.
- **Seismogenic Zones**: Seismic related zones and elements within National territory and boundaries. Technology used: Web Service.
- **SIOS**: This platform gathers all the information related to works, projects and maintenance tasks. It is used every day by ADIF to plan maintenance jobs. Technology used: XML.
- **TAMPING TASK**: Tamping works information. Technology used: XML.
- **MTN25 RASTER**: Recent raster file of the National territory map in scale 1:25.001. Technology used: Web Service with an estimated volume of 1107x200 MB.
- **Noise map of Malaga**: GIS information of the noise distribution in the city of Malaga related to train traffic. Technology used: Web Service.
- **TRAMIFICACION**: The Common Tramification contains alphanumeric information on the links and operational points of ADIF and ADIF Alta Velocidad, associating to these entities their spatial representation at a macro topological level, without reaching the level of detail to identify. Technology used: XML.
- **TRAMIFICACION IDEADIF**: IDEADIF contains alphanumeric information on the links and operational points of ADIF and ADIF Alta Velocidad, associating to these entities their spatial representation at a macro topological level, without reaching the level of detail to identify the track geometry, with the railway lines being represented by the line Longitudinal (centreline). Technology used: SHP.
- **BTN25**: Recent vectorial file of the National territory map in scale 1:25.000. Technology used: Web Service with an estimated volume of 1107x10 MB.
- **Weather Information**: Weather information gathered by the ADIF's weather stations deployed through its facilities and infrastructures. Technology used: Web Service.
- **Weather Information**: Weather information gathered by the AEMET network. Technology used: API REST and daily update frequency.

3.6.3 Standards and Metadata

Three different types of standard have been notified (internal standard):

- Infrastructure design characterization for Railway Design Projects dataset.
- Maintenance operation for Maintenance Operation Drainage Clearing [digital and paper support], Maintenance Operation Embankment Slopes Clearing [digital and

paper support], Maintenance Operation Track Bed Profiling [digital and paper support] datasets.

- Machinery characterization for Machinery Engine Work, Machinery Automatic Engine Work, Machinery General Engine Data, Machinery Fuel Consumption, Machinery Automatic Fuel Consumption, Machinery GPS location, Machinery Work Mode, Machinery Automatic Work Mode, Machinery Technical Specifications, Machinery Number of Tamping Insertions, Machinery Tamping Device Temperature datasets.

3.6.4 Data Sharing

For open datasets or without approval, users will be able to access data through the URI:

Dataset	URI
AEMET	http://www.aemet.es/es/datos_abiertos/catalogo
CLIMA	http://www.juntadeandalucia.es/medioambiente/servtc5/WebClima/
European Railway Data	http://www.era.europa.eu/Core-Activities/Pages/Registers.aspx
Hydrogeologic map	http://info.igme.es/cartografiadigital/geologica/mapa.aspx?parent=../tematica/tematicossingulares.aspx&Id=19
PNOA MÁXIMA ACTUALIDAD	http://centrodedescargas.cnig.es/CentroDescargas/catalogo.do;jsessionid=4CA918ED89D670E68F9A520DB166B2B8#selectedSerie
Noise map of Malaga	http://datosabiertos.malaga.eu/dataset?q=ferroviario
MTN25 RASTER	http://centrodedescargas.cnig.es/CentroDescargas/buscadorCatalogo.do?codFamilia=02102
Recent Seismicity	No link available
Seismogenic Zones	http://info.igme.es/zesis/
BTN25	http://centrodedescargas.cnig.es/CentroDescargas/buscadorCatalogo.do?codFamilia=02101
BTN100	http://centrodedescargas.cnig.es/CentroDescargas/buscadorCatalogo.do?codFamilia=BT100

For datasets with approval access (DAVINCI, DYNAMIC INSPECTION, Railway Design Projects [digital and paper support], Maintenance Operation Drainage Clearing [digital and paper support], Ferrovial Drone flights, Maintenance Operation Line Fencing Preservation [digital and paper support], Machinery Engine Work, Machinery Automatic Engine Work, Machinery General Engine Data, Machinery Automatic Fuel Consumption, Machinery GPS location, Machinery Work Mode, Machinery Automatic Work Mode, Machinery Technical Specifications,

Machinery Number of Tamping Insertions, Machinery Tamping Device Temperature, Maintenance Operation Embankment Slopes Clearing [digital and paper support], Track Geometry, Maintenance Operation Track Bed Profiling [digital and paper support], GEOMETRICAL INSPECTION, MONITORING AND CONTROL SYSTEM OF RAILWAYS OPERATION, PIDAME, S&C INSPECTIONS, SIOS, TAMPING TASK, TRAMIFICACION, TRAMIFICACION_IDEADIF, Weather Information (ADIF and AEMET), users has to ask directly to the dataset's owner the permission to access the data.

Open Data datasets are under creative commons license [10] and can be freely used, re-used and redistributed by anyone with the requirement to attribute and share alike (datasets MDT05/MDT05-LIDAR, Hydrogeologic map, BTN100, BTN25, PNOA MÁXIMA ACTUALIDAD, Recent Seismicity, Seismogenic Zones, MTN25 RASTER, Noise map of Malaga)

3.6.5 Archiving and Preservation

The data of the project will be stored in several places:

Dataset	Storage
CGRH4, DAVINCI, MONITORING AND CONTROL SYSTEM OF RAILWAYS OPERATION, PIDAME, TRAMIFICACION, TRAMIFICACION_IDEADIF	DCSI-DELICIAS
European Railway Data, Noise map of Malaga	External Database
Seismogenic Zones	ZESIS Database
Hydrogeologic map	IGME
Recent Seismicity	IGN Database
BTN100, BTN25, PNOA MÁXIMA ACTUALIDAD, MTN25 RASTER	CNIG Download Centre
SIOS	INECO
Weather Information	ADIF Database
Weather Information	AEMET Database
Railway Design Projects [digital support], Maintenance Operation Drainage Clearing [digital support], Ferrovial Drone flights, Maintenance Operation Line Fencing Preservation [digital support], Machinery Engine Work, Machinery General Engine Data, Machinery Fuel Consumption, Machinery Automatic Fuel Consumption, Machinery GPS location, Machinery Work Mode, Machinery Automatic Work Mode,	Ferrovial Server

Machinery Technical Specifications, Machinery Number of Tamping Insertions, Machinery Tamping Device Temperature, Maintenance Operation Embankment Slopes Clearing [digital support], Track Geometry, Maintenance Operation Track Bed Profiling [digital support]	
Railway Design Projects [paper support], Maintenance Operation Drainage Clearing [paper support], Maintenance Operation Line Fencing Preservation [paper support], Maintenance Operation Embankment Slopes Clearing [paper support], Maintenance Operation Track Bed Profiling [paper support]	Ferrovial Warehouse

The data will be available throughout the whole life of the project.

3.6.6 Ethical Aspects

All the datasets (except MONITORING AND CONTROL SYSTEM OF RAILWAYS OPERATION, PIDAME and SIOS) don´t include personal data.

For other rights in open datasets (MDT05/MDT05-LIDAR, Hydrogeologic map, BTN100, BTN25, PNOA MÁXIMA ACTUALIDAD, Recent Seismicity, Seismogenic Zones, MTN25 RASTER, Noise map of Malaga) you could check the link https://www.boe.es/boe/dias/2015/12/26/pdfs/BOE-A-2015-14129.pdf.

3.7 WP7 – PORTS AS INTELLIGENT LOGISTICS HUB - VALENCIA SEA PORT PILOT

3.7.1 Context

Valencia Port is the largest container main sea port in the Mediterranean. The Valencia Port pilot will focus on a sea port, one of the most important in Europe and will use intelligent predictive and optimization modules based on Big Data technologies to improve performance, efficiency, productivity and competitiveness in this hub and create an optimized ecosystem on which all its agents (warehouses, road hauliers, railways, port authorities, customs, border protection agencies, port terminals and vessels) can interact seamlessly and efficiently. The pilot is built on existing databases (Terminal Operating System (TOS) or Port Community System (ValenciaportPCS) and ongoing projects focused on the IoT layer (SEA TERMINALS, INTER-IoT) that provide real time, reliable and precise information of all the movements taking place in the container yard, together with a tight control of the equipment fleet as a whole as well as individual machines and street-trucks coming to the terminal. [2]

3.7.2 Data Description

The project partner has identified the datasets that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **CATOS**: CATOS has a full suite of planning, operation and management system modules covering the total operations of the terminal and utilizing terminal resources fully. These three activity types share one and the same database within CATOS, therefore, the integrity of your data is guaranteed. Technology used: ORACLE 11gr2 with an estimated volume of 80GB.
- **SEAMS Platform**: Each one of the machines has a Siemens 1200 PLC that is collecting data from the sensors like GPS, encoders, inductive or photocells, using an internal polling that stores the values of all the sensors every second, with timestamp. The average number is 82 variables per machine.
  The central server polls all the machines every second asking for updated data. If the machine polled is available, the machine answers with the last values of the 82 variables and mark then as “sent”. If the machines are not available in the poll moment, each machine keeps the data with timestamp every second and is stored in a SD card in the PLC. When the machine is available again, it will send CSV files with 15 minutes of information each. The total capacity of the SD Card is 6 hours of data for the 12 MB and 3 hours for the 6 MB. Technology used: SQL Server with an estimated volume of 1,3GB and weekly update frequency.

- **Valenciaport AIS**: The automatic identification system (AIS) is an automatic tracking system used for collision avoidance on ships and by vessel traffic services (VTS). AIS equipment provides information such as unique identification, position, course, and speed. Technology used: Data stream.
- **ValenciaportPCS**: ValenciaPortPCS (http://www.valenciaportpcs.net) is an information system that makes available logistical information among the actors involved in port-related freight distribution.
- **Valenciaport Scada**: Supervisory control and data acquisition (SCADA) is a control system architecture. SCADA is a category of software application program for process control, the gathering of data in real time from remote locations in order to control equipment and conditions. SCADA system connects to all the sensors deployed in the Port of Valencia: energy sensors, security, doors, lighthouses, fire hydrants). In SCADA architecture, information from sensors or manual inputs is sent to PLCs (programmable logic controllers) or RTUs (remote terminal units), which then send that information to computers with SCADA software. SCADA software analyses and displays the data in order to help operators and other workers to decision making. Technology used: SQL Server.

3.7.3 Standards and Metadata

Valenciaport Scada uses C37.1-2007 standard (IEEE standard for SCADA and Automation Systems).

3.7.4 Data Sharing

Internal CUSTOMER shall have the non-exclusive, non-transferable License only for use CATOS at the Site under the terms and conditions of its License Agreement.

The rest of the dataset are under approval access, then users has to ask directly to the dataset's owner the permission to access the data.

3.7.5 Archiving and Preservation

Valencia AIS and Valenciaport Scada datasets will be stored into local devices with local backups of the dataset whilst ValenciaportPCS dataset will be stored into local devices but with local and external backups. No information about storage and backup of the other datasets.

3.7.6 Ethical Aspects

All the datasets don't include personal data. No other ethical aspect has been reported for the time being.

3.8 WP7 – PORTS AS INTELLIGENT LOGISTICS HUB - DUISPORT INLAND PORT PILOT

3.8.1 Context

Duisport is the world's largest inland port. Com-pared to Valencia port, the added complexity in duisport stems from the fact that the port is situated in the middle of large city and at the centre of Germany's largest metropolitan area. This means that duisport has a multitude of roads, tracks and water ways that serve as entry and exit points for containers to and from the actual terminals and ports. In addition, roads need to be shared with many other cars within the metropolitan area. This task will extend the results of a large national innovation project on logistics control towers (LoFIP) and enhances it with SoftwareAG's advanced Big Data analytics and visualization capabilities that integrate the various relevant data sources from the port and terminals. This will allow operators to better manage terminal and port operations in real-time by offering advanced decision support, including predictive and prescriptive analytics for proactive decision making. As part of this task, potential data sources that will be explored include video streams of trucks and trains (from video gates), road traffic management data, information about the early registrations of trucks (incl. their weight), collected transport fees, vessel tracking data, dwell time of vessels, UIC-number of rail wagons, container numbers, and electric wagon lists of container trains. [2]

3.8.2 Data Description

To be refined further in the following versions of the DMP.

3.8.3 Standards and Metadata

To be refined further in the following versions of the DMP.

3.8.4 Data Sharing

To be refined further in the following versions of the DMP.

3.8.5 Archiving and Preservation

To be refined further in the following versions of the DMP.

3.8.6 Ethical Aspects

To be refined further in the following versions of the DMP.

3.9 WP8 – SMART AIRPORT TURNAROUND - SMART PASSENGER FLOW PILOT

3.9.1 Context

Smart Passenger Flow Pilot will carry out the pilot to enhance the passenger process, to analyse the different types of passengers and their behaviours, to improve the understanding of when the passengers arrive to the airport, how they go through the different airport zones, their commercial preferences, etc. in order to adapt the airport resources to provide better and personalized services. The pilot will ultimately be deployed at the Athens International Air-port and the solution will be connected with real-time and historical data. Tests and additional adjustments will be carried out in this stage when the solution is deployed at its operational environment. [2]

3.9.2 Data Description

The project partners have identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **Airport Slot**: Airport Slot provides arrival and departure information for flights operated at the airport according to flight schedule. Technology used: Web Services.
- **Baggage**: Baggage checked by AEGEAN passengers, including connecting passengers. Technology used: “e-mail”.
- **Boarding Pass Reading**: Passenger boarding pass read at the checking points before the security screening lines. Technology used: not defined.
- **Flight Plan**: Flight Plans emitted by the airlines/pilots and received from Eurocontrol for flights operated at Athens International Airport. Technology used: Web Services.
- **Mobile Phone Location**: Location at the airport of mobile phones (MAC address) connected to the airport Wi-Fi. Technology used: REST.
- **Passenger**: Non-personal data related to (AEGEAN) passengers, including connecting passengers. Technology used: “e-mail”.

3.9.3 Standards and Metadata

No specific standards or metadata have been identified for the time being for the proposed datasets.

Flight Plan dataset uses eurocontrol glossary model [5] and Mobile Phone Location uses Cisco MSE REST API Guide [6] for data model standard glossaries.

3.9.4 Data Sharing

Flight plans terms of use are under terms and conditions of the NM Access Services. No information about the terms of use of the rest of the dataset (to be defined) for the time being.

3.9.5 Archiving and Preservation

No information yet about archiving and preservation. To be refined further in the following versions of the DMP.

3.9.6 Ethical Aspects

All the datasets don't include personal data. No other ethical aspect has been reported for the time being.

3.10 WP8 – SMART AIRPORT TURNAROUND - SMART TURNAROUND, ETA PREDICTION AND PASSENGER FLOW PILOT

3.10.1 Context

Smart Turnaround, ETA Prediction and Passenger Flow Pilot will synchronize the turnaround process and flight scheduling for all aircraft of an airline at one of their hubs. Based on a holistic view of the overall fleet, priorities for incoming and outgoing flights will be defined and the ongoing and planned turn around processes will be adapted to realize a holistic solution. By analysing the huge amount of real-time data about resources on the ground, their availability, position and task lists, the ground operations will be managed to support the holistic plan. A connection to ATC will be established to communicate preferences within the own airline in order to influence sequencing of own aircraft. Moreover, models will be created and evaluated for a more accurate estimation of ETA predictions. By means of such models, the turnaround process as well as the overall flight planning can be significantly improved. Further, passenger flow models from the Task 8.2 will be verified, evaluated and replicated. Last, the passenger flow as well as ETA prediction models will be integrated in the overall smart turn around process. This feeds into a holistic approach to optimize the turnaround process. The Big Data solutions will ultimately be deployed at Munich/Frankfurt airport. [2]

3.10.2 Data Description

To be refined further in the following versions of the DMP.

3.10.3 Standards and Metadata

To be refined further in the following versions of the DMP.

3.10.4 Data Sharing

To be refined further in the following versions of the DMP.

3.10.5 Archiving and Preservation

To be refined further in the following versions of the DMP.

3.10.6 Ethical Aspects

To be refined further in the following versions of the DMP.

3.11 WP9 – INTEGRATED URBAN MOBILITY - TAMPERE INTEGRATED URBAN MOBILITY AND LOGISTICS PILOT

3.11.1 Context

Through integration and processes of the different data sources more accurate information can be gained on traffic jams and exceptional situations. Tools will be provided to the Traffic Management Centre (TMC) for improved situation awareness, resting on the collection of data and real-time analysis of data for TMC. Methods to use stored data to assess the impact of traffic management measures on congestion will be explored. For the improvement of freight delivery, a system to reserve loading bays will be integrated. [2]

3.11.2 Data Description

To be refined further in the following versions of the DMP.

3.11.3 Standards and Metadata

To be refined further in the following versions of the DMP.

3.11.4 Data Sharing

To be refined further in the following versions of the DMP.

3.11.5 Archiving and Preservation

To be refined further in the following versions of the DMP.

3.11.6 Ethical Aspects

To be refined further in the following versions of the DMP.

3.12 WP9 – INTEGRATED URBAN MOBILITY - VALLADOLID INTEGRATED URBAN MOBILITY AND FREIGHT PILOT

3.12.1 Context

The approach for this task will be to focus on trip building models within urban areas related to urban goods and services supply. Big data on real life trip patterns will be used – provided by Grupo Lince (logistics freight and services provider). This data will mainly consist of depot, stops, service time, etc. and will be based on time stamp and geographical information. The spatial execution of the trip (from stop to stop) can be given, but also missing. Trips will be mapped with an existing road network. The trip building model will finally be adapted to the observed parameters, meaning that model parameters will be derived allowing near to reality reproduction of actual transport processes in urban delivery. A common platform will be established for transport service providers and for city planners. The infrastructure will make use of cloud storage for GPS traces to be processed towards origin/destination matrixes for Valladolid, as well as of traffic-related data. This task includes the development of heuristic optimization procedures (e.g. Simulated Annealing, genetic algorithms, etc.). In order to describe and develop an optimization scenario it is needed that the trip building model will be evaluated for several parameter settings. The big data approach will allow the processing of a large number of trip patterns and to provide the needed parameter (necessary for any model assessment) in the short time. Moreover, this optimization scenario can be used to assign order to routes to complete the most stops in the least amount of time, taking into consideration traffic conditions, local regulations including pollution control policy, delivery windows and vehicle type. [2]

3.12.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **LINCE GPX TRACES**: Time (Date-Hour) and GPS position (longitude, latitude and elevation) related to freight and services supply in Valladolid city. Technology used: GPX with an estimated volume of 1Mb a day.
- **VALLADOLID MAGENETIC LOOPS:** Number of vehicles going over a sensor every 15 minutes. Technology used: XLS with an estimated volume of 8Mb a month.
- **VALLADOLID WEATHER DATA**: It contains the temperature, rainfall and visibility in Valladolid city. Technology used: XLS with an estimated volume of 0,8Mb a year.

3.12.3 Standards and Metadata

No specific standards or metadata have been identified for the time being for the proposed datasets.

3.12.4 Data Sharing

All the datasets are under approval access, then users has to ask directly to the dataset's owner the permission to access the data.

3.12.5 Archiving and Preservation

LINCE has their own copies for LINCE GPX TRACES dataset and VALL has their own copies of VALLADOLID MAGENETIC LOOPS dataset.

CARTIF will store copies to be used during the project (all datasets). Backup copies will be stored in a NAS server which is also connected to the internal network of the company, with physical access control and restricted access to authorised personnel only.

3.12.6 Ethical Aspects

No datasets include personal data. Valladolid Weather Data dataset rights are affected by the following document:

- https://sede.aemet.gob.es/documentos/es/servicios/publicos/AEMET/L1_reverso_general_AEMET.doc

3.13 WP10 – DYNAMIC SUPPLY NETWORKS

3.13.1 Context

Dynamic Supply Chain Networks pilot will be focused on e-commerce logistics and its basic part is to provide shared logistics scenarios that take into consideration various routing and customer preference characteristics such as maximum travel time diversion, multiple destinations or customer journey. In its scope, advanced computational logic based on Big Data technologies will be used in order to improve performance, efficiency, productivity and competitiveness in logistics and increase consumers satisfaction. In order to implement and validate the scenarios described above, a demanding information sharing infrastructure will be used. Its main purpose will be to collect data by different e-shops, and via a specific algorithm procedure to be able to combine them with the alternative courier prices. All these information / data can be used to forecast customer demands and upcoming orders, to detect potential route for shared transportation and missing logistics capabilities, which can be address with the shared logistic scenarios. They further require that an advanced inventory routing algorithm is executed over a multitude of data collected from various nodes in the supply chain and provide updated availability and delivery time information to the end customer also in real time. [2]

3.13.2 Data Description

The project partner has identified the dataset that will be produced during different phases of the project. This list is indicative and will be adapted (addition/removal/modification of datasets) in the following versions of the DMP. The list is provided below:

- **Mantis Lvision**: Mantis' Logistics Vision™ flagship software product suite is an integrated 'state of the art' family of 3rd generation software products that allows modern large enterprises to plan, manage and optimize their entire supply chain. It incorporates best practices of top multinational corporations in most markets. The latest barcode, Radio Frequency, Pick-by-light, Voice logistics and material handling technologies are being utilized by Logistics Vision™ to provide integrated 'turn-key' solutions, while its open architecture allows easy integration with other information systems. 100GB of data are expected. Technology used: not defined.

3.13.3 Standards and Metadata

No specific standards or metadata have been identified for the time being for the proposed datasets.

3.13.4 Data Sharing

Internal Customer shall have the non-exclusive, non-transferable License only for the project. The data access must be with approval, that is, whoever who wants to access to the digital

ecosystem (catalogue) has to ask directly to the dataset's owner the permission to access the data.

3.13.5 Archiving and Preservation

Dataset is archived to external storage for one year with a total size of 100GB. Data backup is being taken on daily basis.

3.13.6 Ethical Aspects

The dataset includes personal data but all personal data will be filtered out by logika. No other ethical aspect has been reported for the time being.

4 Conclusions and Next Steps

In this deliverable, a DMP for the TT project has been presented, structured around the TT domains and their corresponding pilots. A collection of datasets has been put together during the creation of this document. It provides an overview about the amount of data that will be handled in the project. However, as aforementioned, this is a living document and, as time goes by, new datasets will appear and some existing ones may not be relevant any more, what will require this document to be constantly updated.

In the future, we will keep on developing the template ID card for clearing up some aspects of the information that we have noticed at this stage. In this first version, some factors could be considered as ambiguous, incomplete or inconsistent.

The next steps in this WP will be focused on the development of an open data portal for TT, where the metadata that has been gathered as a result of this initial exercise will be registered, so that these datasets will be searchable in the portal. In those cases where the datasets are considered to be open, then the datasets will be uploaded and made available in the portal, so that they can be downloaded by potential users. In those cases where datasets are not open, only metadata will be made available. This should enable search, with the corresponding contact information to allow anybody interested to gain access if considered appropriate by the data owner.

References

Below, a list of documents used for creating this deliverable is shown:

[1] - H2020 Programme (Guidelines on FAIR Data Management in Horizon 2020, v3.0): http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] - Grant Agreement-731932-TT.

[3] - Ausol and Norte Litoral information: http://www.ferrovial.com/en/projects/.

[4] - https://www.securosis.com/blog/data-security-lifecycle-2.0

[5] - https://www.eurocontrol.int/sites/default/files/article/attachments/eurocontrol-glossary-for-flight-statistics-and-forecasts.pdf

[6] – Cisco MSE REST API Guide: http://www.cisco.com/c/en/us/td/docs/wireless/mse/8-0/MSE_REST_API/Guide/Cisco_MSE_REST_API_Guide.html

[7] – Yahoo license: https://policies.yahoo.com/us/en/yahoo/terms/product-atos/apiforydn/index.htm

[8] - Tweeter license: http://twitter4j.org/en/index.html#license.

[9] - INE license: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE.

[10] – Creative commons license: https://creativecommons.org/licenses/by-sa/3.0/