Grant Agreement Number: 731993

Project acronym: AUTOPILOT

Project full title: AUTOmated driving Progressed by Internet Of Things

D6.9

INTERIM REVISION OF THE DATA MANAGEMENT PLAN

Due delivery date: 30/04/2018

Actual delivery date: 03/07/2018

Organization name of lead participant for this deliverable: AKKA

Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the GSA)	
RE	Restricted to a group specified by the consortium (including the GSA)	
CO	Confidential , only for members of the consortium (including the GSA)	

Project funded by the European Union's Horizon 2020 Research and Innovation Programme (2014 – 2020)

Document Control Sheet

Deliverable number:	D6.9
Deliverable responsible:	AKKA
Work package:	6
Editor:	AIT AAZIZI Amine – ZOUGARI Sadeq

Author(s) – in alphabetical order		
Name	**Organisation**	**E-mail**
AIT-AAZIZI Amine	AKKA	Amine.ait-aazizi@akka.eu
FISCHER François	ERTICO	f.fischer@mail.ertico.com
ZOUGARI Sadeq	AKKA	Sadeq.zougari@akka.eu

Document Revision History			
Version	**Date**	**Modifications Introduced**	
		Modification Reason	**Modified by**
V0.1	05/04/2018	First Draft	AIT AAZIZI, Amine
V0.2	13/04/2018	Draft ready for peer review	AIT AAZIZI, Amine ZOUGARI, Sadeq
V1.1	16/05/2018	Peer review	Solmaz, Gurkan Alesiani, Francesco KOVACS, Ernoe SIMEON, Jean-François
V2.0	25/05/2018	Final version for upload	AIT AAZIZI, Amine ZOUGARI, Sadeq BHANDARI, Rita

Abstract

This document presents the second version of the data management plan. It identifies and describes all data that will be collected and generated during the AUTOPILOT project and defines the data management principles that shall be followed by the Consortium.

This includes the elicitation of the different data categories, how data will be created, stored and backed-up, who owns it and who is responsible for the different data, which data will be preserved and shared.

This version is the first update of deliverable D6.7. It takes into account the progress of the work done in WP2, WP3 and WP4. In particular, the definition of the next version of the datasets will ensure the sharing and dissemination of relevant information to all project partners for data management purposes.

Abbreviations and Acronyms

Acronym	Definition
ADAS	Advanced Driver-Assistance Systems
AVP	Automated Valet Parking
CAN	Controller Area Network
DMP	Data Management Plan
EC	European Commission
FAIR	Findable – Accessible – Interoperable-Reusable
FESTA	Field opErational teSt supporT Action
FOT	Field Operational Test
GA	Grant Agreement
GPS	Global Positioning System
IoT	Internet Of Things
ITS	Intelligent Transportation Systems
LIDAR	LIght Detection And Ranging
NDS	Naturalistic Driving Studies
ORDP	Open Research Data Pilot
PO	Project Officer
RADAR	RAdio Detection And Ranging
RPM	Revolutions Per Minute
WP	Work Package
WPL	Work Package Leader

Table of Contents

List of Figures

List of Tables

AUTOPILOT

Executive Summary

In Horizon 2020 a limited pilot action on open access to research data has been implemented. Participating projects are required to develop a Data Management Plan (DMP).

This deliverable provides the second version of the DMP elaborated by the AUTOPILOT project. The purpose of this document is to provide an overview of the main elements of the data management policy. It outlines how research data will be handled during the AUTOPILOT project and describes what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved. Besides a data types list, metadata and global data collection processes are also defined in this document.

The AUTOPILOT data management plan refers to the latest EC DMP guidelines[1]. This version has explicit recommendations for full lifecycle management through the implementation of the FAIR principles, which state that the data produced shall be Findable, Accessible, Interoperable and Reusable (FAIR).

Since the data management plan is expected to mature during the project while taking into account the progress of the work, the last version will be produced as additional deliverable by the end of the project.

1 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

1. Introduction

1.2 Objectives of the project

Automated driving is expected to increase safety, to provide more comfort and to create many new business opportunities for mobility services. The Internet of Things (IoT) is about enabling connections between objects or "things"; it is about connecting anything, anytime, anyplace, using any service over any network.

AUTOmated Driving **P**rogressed by **I**nternet **O**f **T**hings” (AUTOPILOT) project will especially focus on utilizing the IoT potential for automated driving.

The overall objective of AUTOPILOT is to bring together relevant knowledge and technology from the automotive and the IoT value chains in order to develop IoT-architectures and platforms which will bring Automated Driving towards a new dimension. This will be realized through the following main objectives:

- Use, adapt and innovate current advanced technologies to define and implement an IoT approach for autonomous and connected vehicles
- Deploy, test and demonstrate IoT-based automated driving use cases at several permanent pilot sites, in real traffic situations with: Urban driving, Highway pilot, Automated Valet Parking, Platooning and Real-time car sharing
- Create and deploy new business products and services for fully automated driving vehicles, used at the pilot sites: by combining stakeholders’ skills and solutions, from the supply and demand side
- Evaluate, with the involvement of users, public services and business players at the pilot sites:
    - The suitability of the AUTOPILOT business products and services as well as the ability to create new business opportunities
    - The user acceptance related to using the Internet of Things for highly or fully automated driving
    - The impact on the citizens’ quality of life
- Contribute actively to standardization activities as well as to consensus building in the areas of Internet of Things and communication technologies

Automated vehicles largely rely on on-board sensors (LiDAR, radar, cameras, etc.) to detect the environment and make reliable decisions. However, the possibility of interconnecting surrounding sensors (cameras, traffic light radars, road sensors, etc.) exchanging reliably redundant data may lead to new ways to design automated vehicle systems potentially reducing cost and adding detection robustness.

Indeed, many types of connected objects may act as an additional source of data, which will very likely contribute to improving the efficiency of automated driving functions and enabling new automated driving scenarios. This will also improve the safety of the automated driving functions while providing driving data redundancy and reducing implementation costs. These benefits will help push the SAE level of driving automation to full automation, keeping the driver out of the loop. Furthermore, by making autonomous cars a full entity in the IoT, the AUTOPILOT project enables developers to create IoT/AD services as easy as accessing any entity in the IoT.

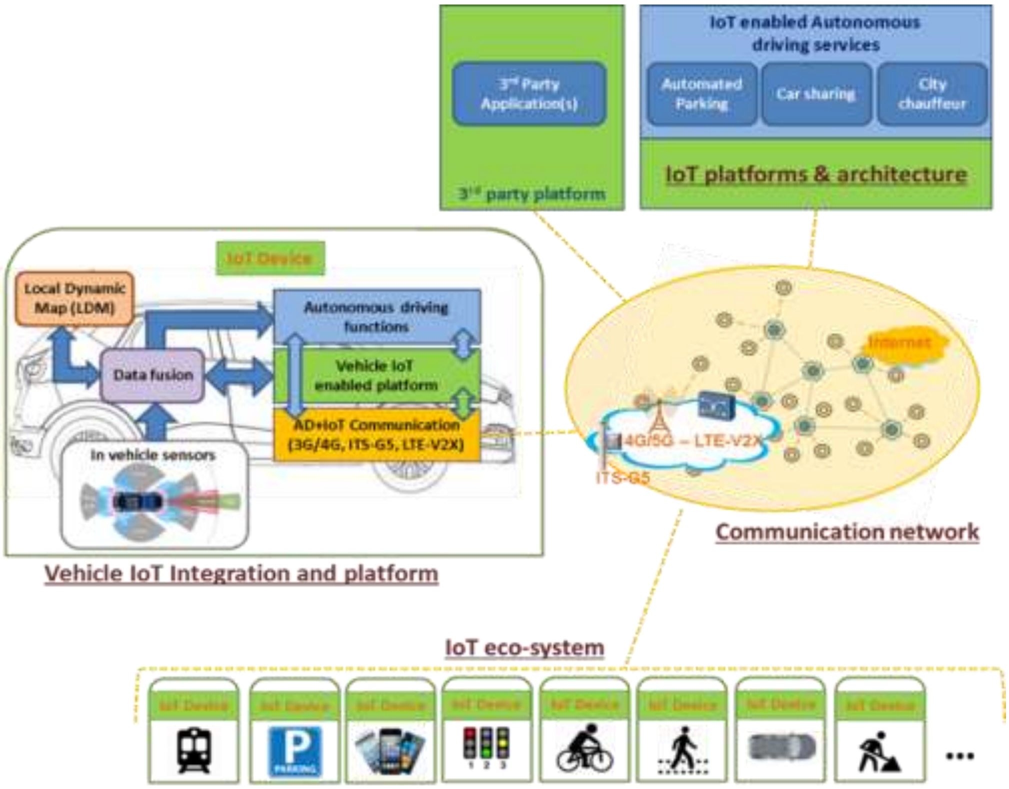

Figure 1 – The AUTOPILOT overall concept

The Figure above depicts AUTOPILOT's overall concept. The main ingredients needed to apply IoT to autonomous driving as represented in the image are:

- The overall IoT platforms and architecture, allowing the use of IoT capabilities for autonomous driving.
- The Vehicle IoT integration and platform to make the vehicle an IoT device, using and contributing to the IoT.
- The Automated Driving relevant sources of information (pedestrians, traffic lights, etc.) becoming IoT devices and extending the IoT eco-systems to allow enhanced perception of the driving environment on the vehicle.
- The communication network using appropriate and advanced connectivity technology for the vehicle as well as for the other IoT devices.

1.3 Purpose of the document

This deliverable presents the second version of the data management plan elaborated for the AUTOPILOT project. The purpose of this document is to provide an overview of the dataset types present in the project and to define the main data management policy adopted by the Consortium.

The data management plan defines how data in general and research data in particular will be handled during the research project and will make suggestions for data management after the project. It describes what data will be collected, processed or generated by the IoT devices and by the whole IoT ecosystem, what methodologies and standards shall be followed during the collection process, whether and how this data shall be shared and/or made open not only for the evaluation needs but also to comply with the ORDP requirements[2], and how it shall be curated and preserved. Besides, the data management plan identifies the four (4) key requirements that define the data collection process and provides first recommendations to be applied.

In comparison to the first version provided at M06, this **second version (M16)** of the data management plan includes more detailed dataset descriptions according to the progress of the work done in the WP2, WP3 and WP4. The descriptions will be filled following the template provided in chapter 5.

The AUTOPILOT data management plan will be updated by the end of the project. The **M32 upcoming version** will outline the details of all datasets involved in the AUTOPILOT project. These datasets include acquired or derived data and aggregated data (IoT data, evaluation data, test data and research data). These dataset types are explained in detail in chapter 5.

This document is structured as follows: **Chapter 2** outlines a data overview in the AUTOPILOT project. It details AUTOPILOT data categories, data types and metadata, then the data collection processes to be followed and finally the test data flow and test data architecture environment.

Chapter 3 gives a global vision of the test data management methodology developed in WP3 across pilot sites.

Chapter 4 gives insights into the Open Research Data Pilot under H2020 guidelines.

Chapter 5 provides a detailed description of the datasets used in the AUTOPILOT project with focus on used methodologies, standard and data sharing policies.

Chapter 6 gives insights into the FAIR Data Management principle under H2020 guidelines and the steps taken by AUTOPILOT in order to be FAIR compliant.

Finally, the chapters 7 and 8 outline the necessary roles, responsibilities and ethical issues.

1.4 Intended audience

The AUTOPILOT project addresses highly innovative concepts. As such, the intended audience of the project is the scientific community interested in IoT and/or automotive technologies. In addition, due to the strong expected impact of the project on their respective domains, the other expected audience consists of automotive industrial communities, telecom operators and standardization organizations.

[2] https://www.openaire.eu/ordp/

2 Data in AUTOPILOT: an overview

The aim of this chapter is:

- To provide a first categorization of the data;
- To identify a list of the data types that will be generated;
- To provide a list of metadata that will be used to describe generated data and enable data re-use;
- To provide recommendations on data collection and sharing processes during the project and beyond.

The AUTOPILOT project will collect a large amount of raw data to measure the benefit of IoT for automated driving with multiple automated driving use cases and services, at different pilot locations.

Data from vehicles and sensors will be collected and managed through a hierarchy of IoT platforms as illustrated in the architectural diagram [3] of

Figure 2.

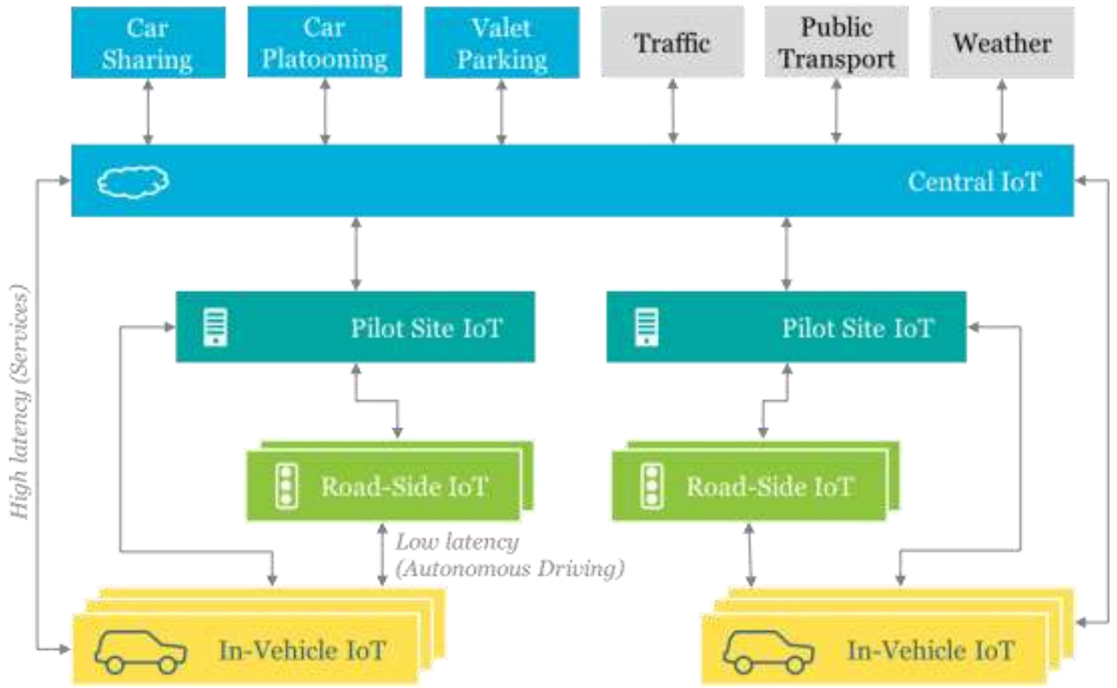

Figure 2 – The AUTOPILOT hierarchical IoT architecture

The diagram above shows a federated architecture with the following four layers:

- **In-vehicle IoT Platforms:** Here is everything that is mounted inside the vehicle, i.e., components responsible for AD, positioning, navigation, real-time sensor data analysis, and communication with the outside world. All mission critical autonomous driving functions should typically reside in this layer.

[3] AUTOPILOT D2.3 - Report on the Implementation of the IoT Platform

- **Road-side IoT Platforms:** Road-side and infrastructure devices, such as cameras, traffic light sensors, etc., are integrated and managed as part of road-side IoT platforms covering different road segments and using local low latency communication networks and protocols as required by the devices and their usage.
- **Pilot Site IoT Platforms:** This layer constitutes the first integration level. It is responsible for collecting, processing and managing data at the pilot site level.
- **Central IoT Platform:** This is a Cloud-based top layer that integrates and aggregates data from the various pilot sites as well as external services (weather, transport, etc.). This is where the common AD services such as car sharing, platooning, etc. will reside. Data, at this level, are standardized using common formats, structures and semantics. The central IoT platform will be hosted on IBM infrastructure.

The data analysis will be performed according to Field Operational Test studies (FOT[4]) and using FESTA[5] methodology. The FESTA project funded by the European Commission developed a handbook on FOT methodology which gives general guidance on organizational issues, methodology and procedures, data acquisition and storage, and evaluation.

From raw data a large amount of derived data will be produced to address multiple research needs. Derived data will follow a set of transformations: cleaning, verification, conversion, aggregation, summarization or reduction.

In any case, data must be well documented and referenced using rich metadata in order to facilitate and foster sharing, to enable validity assessments and to enable its usage in an efficient way.

Thus, each piece of data must be described using additional information called metadata. The latter must provide information about the data source, the data transformations and the conditions in which the data has been produced. More details about the metadata in AUTOPILOT are described in section 2.2.

2.1 Dataset categories

The AUTOPILOT project will produce different categories of datasets:

- **Context data**: data that describe the context of an experiment (e.g. metadata);
- **Acquired and derived data**: data that contain all the collected information from measurements and sensors related to an experiment;
- **Aggregated data**: data summary obtained by reduction of acquired data and generally used for data analysis.

2.1.1 Context data

Context data is any information that helps to explain observation during a study. Context data can be collected, generated or retrieved from existing data. For example, it contains information such as vehicle, road or driver characteristics.

2.1.2 Acquired and derived data

Acquired data is all data collected to be analysed during the course of the study. Derived data is created by different types of transformations including data fusion, filtering, classification and reduction. Derived data are easy to use and they contain derived measures and performance indicators referring to a time period when specific conditions are met. This

[4] http://fot-net.eu/
[5] http://wiki.fot-net.eu/index.php/FESTA_Handbook

category includes measures from sensors coming from vehicles or IoT and subjective data collected from either the users or the environment.

The following list outlines the data types and sources that will be collected:

<table>
<tr><td colspan="2">In-vehicle measures are the collected data from vehicles, either using their original in-car sensors or sensors added for AUTOPILOT purposes. These measures can be divided into different types:</td></tr>
<tr><td></td><td>Vehicle dynamics are measurements that describe the mobility of the vehicle. Measurements can be for example longitudinal speed, longitudinal and lateral acceleration, yaw rate, and slip angle.</td></tr>
<tr><td></td><td>Driver actions on the vehicle commands that can be measured are, for instance, steering wheel angle, pedal activation or HMI button press variables, face monitoring indicators characterizing the state of the driver, either physical or emotional.</td></tr>
<tr><td></td><td>In-vehicle systems state can be accessed by connecting to the embedded controllers. It includes continuous measures like engine RPM or categorical values like ADAS and active safety systems activation.</td></tr>
<tr><td></td><td>Environment detection is the environment data that can be obtained by advanced sensors like RADARs, LIDARs, cameras and computer vision, or more simple optical sensors. For instance, luminosity or presence of rain, but also characteristics and dynamics of the infrastructure (lane width, road curvature) and surrounding objects (type, relative distances and speeds) can be measured from within a vehicle.</td></tr>
<tr><td></td><td>Vehicle positioning is the geographical location of a vehicle determined with satellite navigation systems (e.g. GPS) and the aforementioned advanced sensors.</td></tr>
<tr><td></td><td>Media mostly consist of video. The data consist of media data but also index files used to synchronize the other data categories. They are also often collected from the road side.</td></tr>
<tr><td colspan="2">Continuous subjective measures: Complimentary to sensors and instrumentation, some continuous measures can also be built in a more subjective way, by analysts or annotators, notably using video data.</td></tr>
<tr><td colspan="2">Road-side measures are the vehicle speed measurement and positioning, using radar, rangefinders, inductive loops or pressure hose. In ITS systems, it may also contain more complex information remotely transferred from vehicles to road-side units.</td></tr>
<tr><td colspan="2">Experimental conditions are the external factors which may have an impact on participants' behaviour. They may be directly collected during the experiment, or integrated from external sources. Typical examples are traffic density and weather conditions.</td></tr>
<tr><td colspan="2">IoT data are the external sources of data that will be collected/shared through IoT services.</td></tr>
<tr><td></td><td>Users Data can be generated by smartphones or wearables. The users can be the pedestrians or the car drivers. These data helps the user experience for the usage of services by vehicle or infrastructure. The privacy aspects are well explained in chapter 4.</td></tr>
<tr><td></td><td>Infrastructure Data are all the data giving additional information about the environment. Typical examples are the traffic status, road-works, accidents and road conditions. They can also be directly collected from Road-side cameras or traffic light control units and then transferred to IoT Platforms. For instance, the infrastructure data can transfer hazard warnings or expected occupancy of busses on bus lanes to vehicles using communication networks.</td></tr>
</table>

	In-Car data defines the connected devices or sensors in vehicles. Typical examples are navigation status, time distance computations, real-time pick-up / drop-off information for customers, and events detected by car to be communicated to other vehicles or GPS data to be transferred to maps.
Surveys data are data resulting from the answers of surveys and questionnaires for user acceptance evaluation	

2.1.3 Aggregated data

Aggregated data is generally created in order to answer the initial research question. They are supposed to be verified and cleaned, thus facilitating their usage for analysis purposes.

Aggregated data contains a specific part of the acquired or derived data (e.g. the average speed during a trip or the number of passes through a specific intersection). Its smaller size allows a simple storage in e.g. database tables and an easy usage suitable for data analysis. To obtain aggregated data, several data reduction processes are performed. The reduction process summarizes the most important aspects in the data into a list of relevant parameters or events, through one or all of the following processes: validation, curation, conversion, annotation.

Besides helping in answering new research questions, aggregated data may be re-used with different statistical algorithms without the need to use raw data. For AUTOPILOT, aggregated data will represent the most important data types that will be shared by the project. It does not allow potentially problematic re-uses because it does not contain instantaneous values that would highlight illegal behaviour of a vehicle, a driver or another subsystem.

2.2 Metadata

2.2.1 General principles

This section reviews the relevant metadata standards developed or used in the previous and ongoing FOTs and naturalistic driving studies (NDS) as a basis for the development of the metadata specifications of the pilot data. Such standards will help the analysis and re-use of the collected data within the AUTOPILOT project and beyond.

The text in this section is derived from the work done in the FOT-Net Data project[6] for sharing data from field operational tests. The results of this work are described in the Data Sharing Framework[7]. The CARTRE project[8] is currently updating this document to specifically addressing road automation pilots and FOTs.

As described in the previous sections, the pilots will generate and collect a large amount of raw and processed data from continuous data-logging, event-based data collection, and surveys. The collected data will be analysed and used for various purposes in the project including the impact assessment carried out by partners who are not involved in the pilots. This is a typical issue encountered in many FOT/NDS projects in which the data analyst (or re-user) needs to know how the raw data was collected and processed in order to perform data analysis, modelling and interpretation.

Therefore, good metadata is vital. The Data Sharing Framework defines metadata as '**any information that is necessary in order to use or properly interpret data**'. The aim of this

[6] http://fot-net.eu
[7] http://fot-net.eu/Documents/d3-1-data-sharing-framework/
[8] http://connectedautomateddriving.eu

section is to provide methods to efficiently describe a dataset and its associated metadata. The result will serve as suggestions for good practices in documenting a data collection and datasets in a structured way.

Following the definition of metadata by the data sharing framework, we divide the AUTOPILOT metadata into four different categories as follows:

- **AUTOPILOT pilot design and execution** documentation, which corresponds to a high-level description of data collection: its initial objectives and how they were met, description of the test site, etc.
- **Descriptive** metadata, which describes precisely each component of the dataset, including information about its origin and quality;
- **Structural** metadata, which describes how the data is being organized;
- **Administrative** metadata, which sets the conditions for how the data can be accessed and how this is being implemented.

Full details of these metadata categories can be found in the Deliverables of the FOT-Net Data project such as D4.1 Data Catalogue and D4.3 Application of Data Sharing Framework in Selected Cases which can be found on the project website[9].

FOTs have been carried out worldwide and have adopted different metadata formats to manage the collected data. One good example is the ITS Public Data Hub hosted by the US Department of Transport[10]. There are over 100 datasets created using ITS technologies. The datasets contain various types of information --such as highway detector data, travel times, traffic signal timing data, incident data, weather data, and connected vehicle data -- many of which will also be collected as AUTOPILOT data. The ITS Public Data Hub uses ASTM 2468-05 standard format for metadata to support archived data management systems. This standard would be a good starting point to design metadata formats for various types of operational data collected by the IoT devices and connected vehicles in AUTOPILOT.

In a broader context of metadata standardisation, there are a large number of metadata standards available which address the needs of particular user communities. The Digital Curation Centre (DCC) provides a comprehensive list of metadata standards[11] for various disciplines such as general research data, physical science as well as social science and humanities. It also lists software tools that have been developed to capture or store metadata conforming to a specific standard.

2.2.2 IoT metadata

The metadata describing IoT data are specified in the context of OneM2M standard[12]. In such a context "data" signifies digital representations of anything. In practice, that digital representation is associated with a "container" resource having specific attributes. Those attributes are both metadata describing the digital object itself, and the values of the variables of that object, which are called "content".

Every time an IoT device publishes new data on the OneM2M platform a new "content instance" is generated, representing the actual status of that device.

All the "content instances" are stored in the internal database with a unique resource ID.

[9] http://fot-net.eu/Documents/fot-net-data-final-deliverables/
[10] https://catalog.data.gov/dataset
[11] http://www.dcc.ac.uk/resources/metadata-standards/list
[12] http://www.onem2m.org/

The IoT metadata describe the structure of the information, according to the OneM2M standard. The IoT metadata are described in the table below.

Table 1 – OneM2M Metadata for IoT data [13]

Metadata Element	Extended name	Description
pi	parentID	ResourceID of the parent of this resource.
ty	resourceType	Resource Type attribute identifies the type of the resource as specified in clause. E.g. "4 (contentInstance)".
ct	creationTime	Time/date of creation of the resource. This attribute is mandatory for all resources and the value is assigned by the system at the time when the resource is locally created. Such an attribute cannot be changed.
ri	resourceID	This attribute is an identifier for the resource that is used for 'non-hierarchical addressing method', i.e. this attribute contains the 'Unstructured-CSE-relative-Resource-ID' format of a resource ID as defined in table 7.2-1 of [5]. This attribute is provided by the Hosting CSE when it accepts a resource creation procedure. The Hosting CSE assigns a resourceID which is unique in that CSE.
rn	resourceName	This attribute is the name for the resource that is used for 'hierarchical addressing method' to represent the parent-child relationships of resources. See clause 7.2 in [5] for more details.
lt	lastModifiedTime	Last modification time/date of the resource. The lastModifiedTime value is updated when the resource is updated.
et	expirationTime	Time/date after which the resource will be deleted by the Hosting CSE.
acpi	accessControlPolicyIDs	The attribute contains a list of identifiers of an <accessControlPolicy> resource. The privileges defined in the <accessControlPolicy> resource that are referenced determine who is allowed to access the resource containing this attribute for a specific purpose (e.g. Retrieve, Update, Delete, etc.).
lbl	label	Tokens used to add meta-information to resources. This attribute is optional. The value of the labels attribute is a list of individual labels, that can be used for example for discovery purposes when looking for particular resources that one can "tag" using that label-key.
st	stateTag	An incremental counter of modification on the resource. When a resource is created, this counter

[13] AUTOPILOT D3.6 - Data collection and integration methodology

		is set to 0, and it will be incremented on every modification of the resource.
cs	contentSize	Size in bytes of the content attribute.
cr	creator	The ID of the entity (Application Entity or Common Services Entity) which created the resource containing this attribute.
cnf	contentinfo	Information that is needed to understand the content. This attribute is a composite attribute. It is composed first of an Internet Media Type (as defined in the IETF RFC 6838) describing the type of the data, and second of an encoding information that specifies how to first decode the received content. Both elements of information are separated by a separator defined in OneM2M TS-0004 [3].
or	ontologyRef	This attribute is optional. A reference (URI) of the ontology used to represent the information that is stored in the contentInstances resources of the <container> resource. If this attribute is not present, the contentInstance resource inherits the ontologyRef from the parent <container> resource if present.

3 Data management methodology in AUTOPILOT

The AUTOPILOT data collection process and data management is built upon requirements coming from 4 processes:

- **The evaluation requirement** defines the minimum data that must be collected in order to perform the evaluation process at the end of the project
- **The test specification** provides details about the data to be collected on the basis of the evaluation requirements and according to use cases specifications
- **The test data management** defines the data collection, harmonization, storage and sharing requirements using the above two processes and the ORDP process
- **The Open Research Data Pilot[14] (ORDP)** defines the requirement related to sharing of research data

3.1 Evaluation process requirements

The evaluation process is defined in task 4.1 which develops the evaluation methodology. Named FESTA (Field opErational teSt supporT Action), this methodology must be implemented thoroughly and incorporated into the planning to guarantee that all pilots are collecting the required information needed for the evaluation.

The following figure shows a high-level view of the data that will be collected and integrated in the evaluation process. Different types of data (in blue) are collected, stored and analysed by different processes. The workflow will be defined per pilot site but in a homogeneous way. The data types and requested formats will be defined in the evaluation task deliverable D4.1.

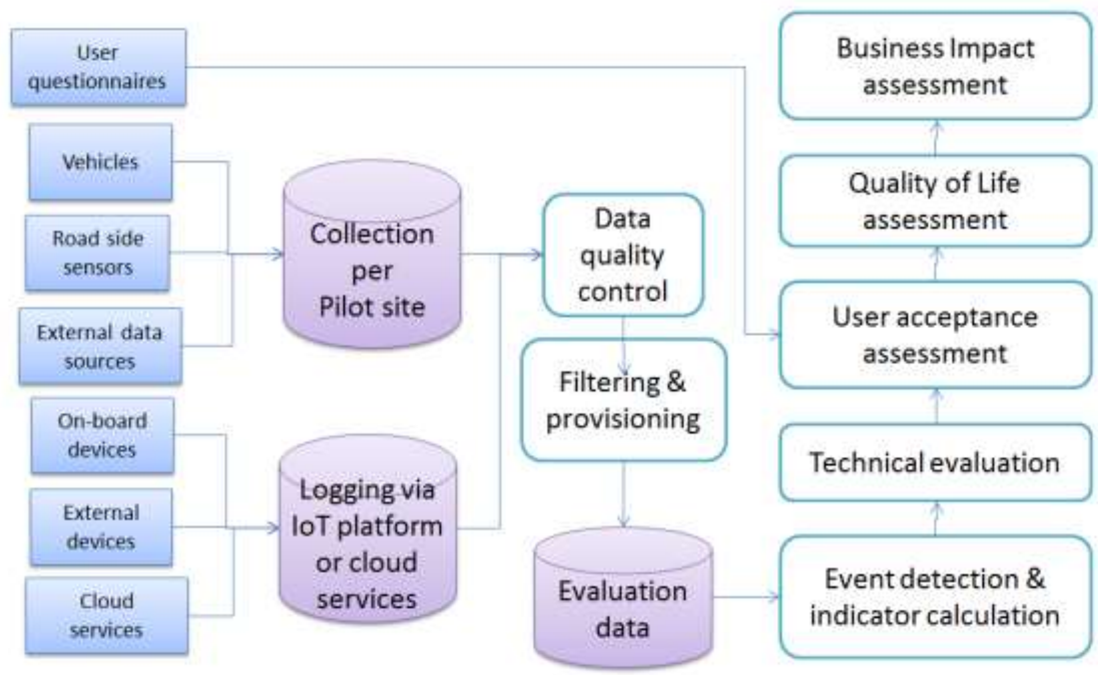

Figure 3 – Evaluation methodology

To fulfil the project objectives, a design of experiment is performed during the evaluation task. This design creates requirements that define the number of scenarios and test cases, the duration of tests and test runs, the number of situations per specific event, the number

[14] https://www.openaire.eu/opendatapilot

of test vehicles, the variation in users, the variation in situations (weather, traffic, etc.). Each pilot site must comply with this design of experiment and provide sufficient and meaningful data with the required quality level to enable technical evaluation. Refer to D1.1 for additional information regarding design of experiment and data quality (Time synchronization of devices & logging, accuracy & frequency of logging, alternative data sources, cross-checking from automated vehicles, on-board devices, road side detectors, detection of failures in systems and logging).

3.2 Tests specification process requirements

The pilot tests specification Task T3.1 plays a major role that must be thoroughly followed. Indeed, this task will convert the high-level requirements defined in the evaluation process into precise and detailed specifications of data formats, data size, data currencies, data units, data files, and storage. The list of requirements will be defined for each of the following items: Pilot sites, Scenarios, Test Cases, Measures, Parameters, Data quality, etc. and will be described in deliverable D3.1. All the development tasks of WP2 must implement completely, if impacted, the requirement defined in D3.1 in order to provide all the data (test data) as expected by the technical evaluation.

3.3 Open research data pilot requirement process

Additional requirements related to ORDP are defined in this document to guarantee that the collected data will be provided in compliance to European Commission Guidelines[15] on Data Management in Horizon 2020. Those requirements are clearly defined and explained in chapter 4.

3.4 Test data management methodology

The main objective of the data management plan is to define the methodology to be applied in AUTOPILOT across all pilot sites, in particular test data management. This includes the explanation of the common data collection and integration methodology.

One of the main objectives within T3.4 “Test Data Management” is to ensure the comparability and consistency of collected data across pilot sites. In this context, the methodology is highly impacted by the pilot site specifications of Task 3.1 and compliant with the evaluation methodologies developed in Task 4.1. In particular, technical evaluation primarily needs log data from the vehicles, IoT platforms, cloud services and situational data from pilot sites to detect situations and events, and to calculate indicators.

The log data parameters that are needed for technical evaluation are organized by data sources (vehicle sources, vehicle data, derived data, positioning, V2X messages, IoT messages, events, situations, surveys and questionnaires).

For IoT data, some pilot sites use proprietary IoT platforms in order to collect data from specific devices or vehicles (e.g. the Brainport car sharing service and automated valet parking service use Watson IoT Platform™ to collect data from their vehicles).

On the top of that, we have a OneM2M interoperability platform in each pilot site. This is the interoperability IoT platform for exchanging IoT messages relevant to all autonomous driving (AD) vehicles at pilot site level. Then, the test data will be stored in pilot site test server storage that will contain mainly the vehicle data, IoT data and survey data. Further, the test data will be packaged and sent to the AUTOPILOT central storage that will allow evaluators to access all the pilot site data in a common format. This includes the input from all pilot

[15] Guidelines on FAIR Data Management in Horizon 2020 - Version 3.0 - 26 July 2016

sites and use cases and for all test scenarios and test runs.

Every pilot site has its own test storage server for data collection (distributed data management). In addition, there is a central storage server where data from all pilot sites will be stored for evaluation and analysis.

The following figure represents the data management methodology and architecture used in AUTOPILOT across all pilot sites.

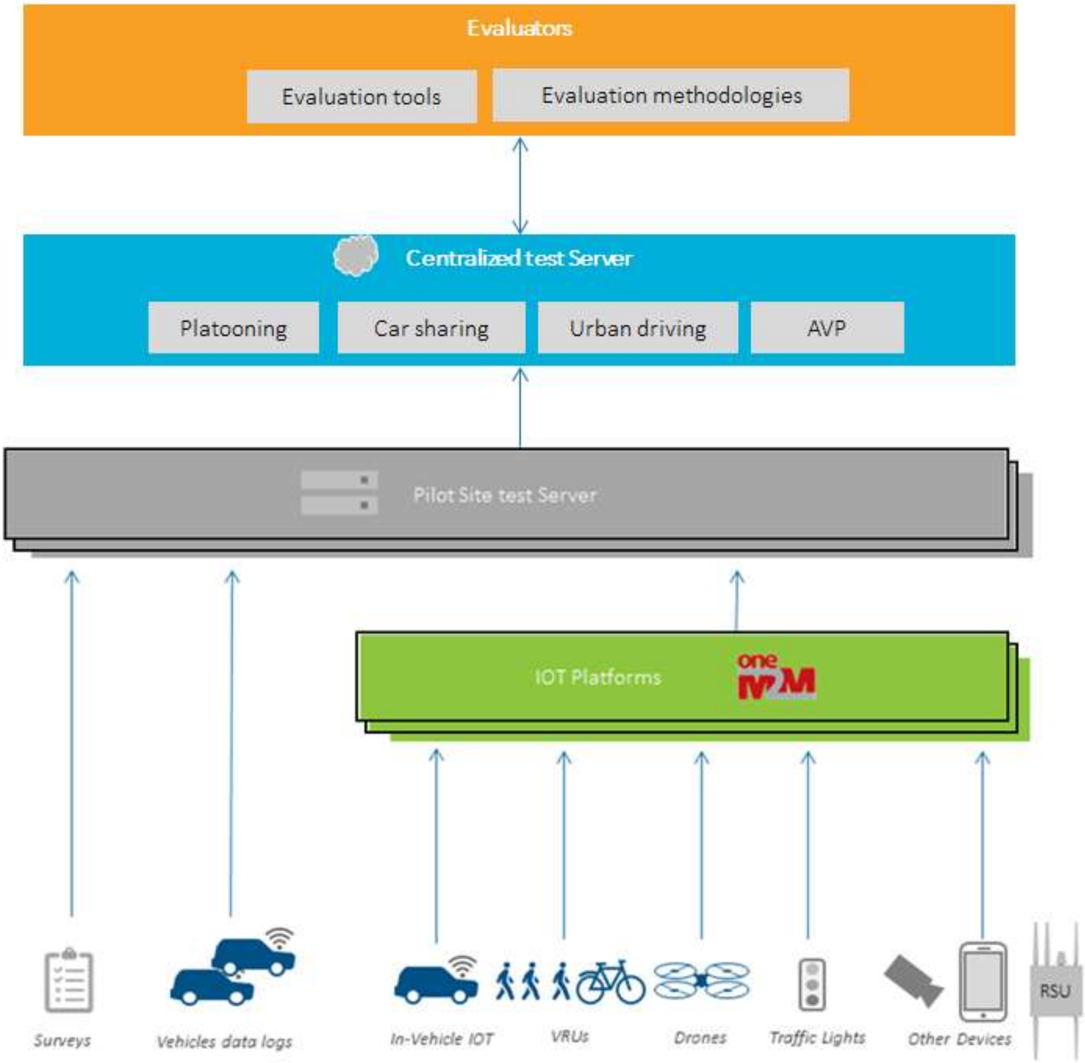

Figure 4 – Generic scheme of data architecture in AUTOPILOT

4 Participation in the open research data pilot

The AUTOPILOT project has agreed to participate in the Pilot on Open Research Data in Horizon 2020[16]. The project uses specific Horizon 2020 guidelines associated with 'open' access to ensure that the project results provide the greatest impact possible.

AUTOPILOT will ensure open access[17] to all peer-reviewed scientific publications relating to its results and will provide access to the research data needed to validate the results presented in deposited scientific publications.

The following lists the minimum fields of metadata that should come with an AUTOPILOT project-generated scientific publication in a repository:

- The terms: "European Union (EU)", "Horizon 2020"
- Name of the action (Research and Innovation Action)
- Acronym and grant number (AUTOPILOT, 731993)
- Publication date
- Length of embargo period if applicable
- Persistent identifier

When referencing Open access data, AUTOPILOT will include at a minimum the following statement demonstrating EU support (with relevant information included into the repository metadata): "This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 731993".

The AUTOPILOT consortium will strive to make many of the collected datasets open access. When this is not the case, the data sharing section for that particular dataset will describe why access has been restricted (See Chapter 5).

A number of AUTOPILOT project partners maintain institutional repositories that will be listed in the following DMP version, where the project's scientific publications and in some instances, research data will be deposited. The use of a specific repository will depend primarily on the primary creator of the publication and on the data in question.

Some other project partners will not operate publically accessible institutional repositories. When depositing scientific publications they will use either a domain specific repository or use the EU recommended service OpenAIRE (http://www.openaire.eu) as an initial step to finding resources to determine relevant repositories.

Project research data will be deposited in the online data repository ZENODO[18]. It is a free service developed by CERN under the EU FP7 project OpenAIREplus (grant agreement no.283595).

The repository will also include information regarding the software, tools and instruments that were used by the dataset creator(s) so that secondary data users can access and then validate the results.

The AUTOPILOT data collection can be accessed in the ZENODO repository at an address such as the following link: https://zenodo.org/collection/<<autopilot>>

[16] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
[17] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm
18 https://zenodo.org/

In summary, as a baseline AUTOPILOT partners will deposit:

- Scientific publications – on their respective institute repositories in addition (when relevant) to the AUTOPILOT ZENODO repository
- Research data – to the AUTOPILOT ZENODO collection (when possible)
- Other project output files – to the AUTOPILOT ZENODO collection (when relevant)

This version of the DMP does not include the actual metadata about the research data being produced in AUTOPILOT. Details about technical means and services for building repositories and accessing this metadata will be provided in the next version of the DMP. A template table is defined in section 5.2 and will be used by project partners to provide all requested information.

5 AUTOPILOT dataset description

5.1 General Description

This section provides an explanation of the different types of datasets to be produced or collected in AUTOPILOT, which have been identified at this stage of the project. As the nature and extent of these datasets can evolve during the project, more detailed descriptions will be provided in the next version of the DMP towards the end of the project (M32).

The descriptions of the different datasets, including their reference, file format, standards, methodologies and metadata and repository to be used are given below. These descriptions are collected using the pilot site requirements and specifications.

It is important to note that the dataset bellow will be reproduced by each use case in all the pilot sites. The dataset categories are:

- IoT dataset
- Vehicle dataset
- V2X messages dataset
- Survey dataset

5.2 Template used in dataset description

This table is a template that will be used to describe the datasets.

Table 2 – Dataset description template

Dataset Reference	**AUTOPILOT_PS_UC_datatype_ID** Each dataset will have a reference that will be generated by the combination of the name of the project, the pilot site (PS) and the use case in which it is generated. **Example**: AUTOPILOT_Versailles_Platooning_IoT_01
Dataset Name	Name of the dataset
Dataset Description	Each dataset will have a full data description explaining the data provenance, origin and usefulness. Reference may be made to existing data that could be reused.
Standards and metadata	The metadata attributes list and standards. The used methodologies.
File format	All the format that defines data
Data Sharing	Explanation of the sharing policies related to the dataset between the next options: **Open**: Open for public disposal **Embargo**: It will become public when the embargo period applied by the publisher is over. In case it is categorized as embargo the end date of the embargo period must be written in DD/MM/YYYY format. **Restricted**: Only for project internal use. Each dataset must have its distribution license. Provide information about personal data and mention if the data is anonymized or not. Tell if the dataset entails personal data and how this issue is taken into account.
Archiving and Preservation	The preservation guarantee and the data storage during and after the project

	Example: databases, institutional repositories, public repositories.

5.3 IoT dataset

This pro-forma table is a description of the IoT Dataset used in AUTOPILOT.

Table 3 – IoT dataset description

Dataset Reference	**AUTOPILOT_PS_UC_IoT_ID**
Dataset Name	IoT data generated from connected devices
Dataset Description	This dataset refer to the IoT datasets that will be generated from IoT devices within use cases. This includes the data coming from VRUs, RSUs, smartphones, Vehicles, drones, etc.
Standards and metadata	During the project, the metadata related to the IoT data are based on OneM2M standard. The OneM2M IoT platforms are implemented across pilot sites to cover the interoperability feature. More details are provided in section 2.2.2. In addition, the data model of these data is inspired by the DMAG (data management activity group) work done in T2.3. The DMAG defined a unified data model that standardizes all the IoT messages across pilot sites. The AUTOPILOT common IoT data model is based on different standards: SENSORIS, DATEX II. After the project, the metadata will be enriched with ZENODO's metadata, including the title, creator, date, contributor, pilot site, use case, description, keywords, format, resource type, etc.
File format	JSON
Data Sharing	This dataset will be openly available for use by 3rd party applications and will be deposited in the ZENODO repository.
Archiving and Preservation	During the project, the data will first be stored in the IoT platform. Then, the data will be transferred to the pilot site test server before finishing up in the centralized test server. At the end of the project, the dataset will be archived and preserved in ZENODO repositories.

5.4 Vehicles dataset

Table 4 – Vehicles dataset description

Dataset Reference	**AUTOPILOT_PS_UC_VEHICLES_ID**
Dataset Name	Data generated from the vehicle sensors.
Dataset Description	This dataset refers to the vehicle datasets that will be generated from the vehicle sensors within use cases. This includes the data coming from the CAN bus, cameras, RADARs, LIDARs and GPS.
Standards and metadata	The vehicle data standards used in AUTOPILOT are developed in task 2.1. The pilot site implementations are based on well-known standards for common data formats: CAN, ROS, etc.

	More details are provided in D2.1. After the project, the metadata will be based on ZENODO's metadata, including the title, creator, date, contributor, pilot site, use case, description, keywords, format, resource type, etc.
File format	XML, CSV, SQL, JSON, Protobuf
Data Sharing	This dataset will be openly available for use by 3rd party applications and will be deposited in the ZENODO repository.
Archiving and Preservation	During the project, the data will first be stored in pilot site test servers before finishing up in the centralized test server. At the end of the project, the dataset will be archived and preserved in ZENODO repositories.

5.5 V2X messages dataset

Table 5 – V2X messages dataset description

Dataset Reference	**AUTOPILOT_PS_UC_V2X_ID**
Dataset Name	V2X messages communicated during test sessions
Dataset Description	This dataset refer to the V2X messages that will be generated from the communication between the vehicles and any other party that could affect the vehicle. This includes the other vehicles and the pilot site infrastructure.
Standards and metadata	The V2X messages are mainly generated from the ITS-G5 communication standard. After the project, the metadata will be enriched by ZENODO's metadata, including the title, creator, date, contributor, pilot site, use case, description, keywords, format, resource type, etc.
File format	CAM, DEMN, IVI, SPAT, MAP
Data Sharing	This dataset will be openly available for use by 3rd party applications and will be deposited in the ZENODO repository.
Archiving and Preservation	During the project, the data will first be stored in pilot site test servers before finishing up in the centralized test server. At the end of the project, the dataset will be archived and preserved in ZENODO repositories.

5.6 Surveys dataset

Table 6 – Surveys dataset description

Dataset Reference	**AUTOPILOT_PS_UC_SURVEYS_ID**
Dataset Name	Survey data collected during test sessions
Dataset Description	This data refers to the data resulting from the answers of surveys and questionnaires for user acceptance evaluation.
Standards and metadata	Survey data will use some well-known tools (Google Forms, Survey Monkey, etc.) The work of defining a common format for survey data is still in progress by the user acceptance evaluation team.

	After the project, the metadata will be enriched by ZENODO's metadata, including the title, creator, date, contributor, pilot site, use case, description, keywords, format, resource type, etc.
File format	CSV, PDF, XLS
Data Sharing	This dataset will be openly available for use by 3rd party applications and will be deposited in the ZENODO repository. It is important to note that these data will be **anonymized** before data sharing.
Archiving and Preservation	During the project, the data will first be stored in pilot site test servers before finishing up in the centralized test server. At the end of the project, the dataset will be archived and preserved in ZENODO repositories.

6 FAIR data management principles

The data that will be generated during and after the project should be **FAIR**[19], that is Findable, Accessible, Interoperable and Reusable. These requirements do not affect implementation choices and don't necessarily suggest any specific technology, standard, or implementation solution.

The FAIR principles were generated to improve the practices for data management and data curation. FAIR principles aim to be applicable to a wide range of data management purposes, whether it is data collection or data management of larger research projects regardless of scientific disciplines.

With the endorsement of the FAIR principles by H2020 and their implementation in the guidelines for H2020, the FAIR principles serve as a template for lifecycle data management and ensure that the most important components for the lifecycle are covered.

The intention is to target the implementation of the FAIR concept rather than a strict technical implementation of the FAIR principles. AUTOPILOT project has undertaken several actions, described below, to carry on the FAIR principles.

Making data findable, including provisions for metadata

- The datasets will have very rich metadata to facilitate the findability. In particular for IoT data, the metadata are based on the OneM2M standard.
- All the datasets will have a Digital Object Identifiers provided by the public repository (ZENODO).
- The reference used for the dataset will follow this format: **AUTOPILOT_PS_UC_Datatype_XX.**
- The standards for metadata are defined in chapter 5 tables and explained in section 2.2.

Making data openly accessible

- All the datasets that are openly available are described in the chapter 5.
- The datasets for evaluation will be accessible via AUTOPILOT's centralized server.
- The datasets will be made available using a public repository (e.g. ZENODO) after the project.
- The data sharing in chapter 5 explains the methods or software used to access the data. Basically, no software is needed to access the data.
- The data and their associated metadata will be deposed in a public repository or in an institutional repository.
- The data sharing in the section 5 tables will outline the rules to access the data if restrictions exist

Making data interoperable

- The metadata vocabularies, standards and methodologies will depend on the public repository and are mentioned in the chapter 5 0tables.
- The AUTOPILOT WP2 took several steps in order to define common data formats. This work was developed in task 2.1 for vehicle data and task 2.3 for IoT data. The goal is to have the same structure across pilot sites and enable evaluators dealing with the same format for all pilot sites.

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

- AUTOPILOT pilot sites use IoT platforms based on OneM2M standards to enable data interoperability across pilot sites.

Increase data re-use (through clarifying licenses)

- All the data producers will license their data to allow the widest reuse possible. More details about license types and rules will be provided in the next version (M32).
- By default, the data will be made available for reuse. If any constrains exist, an embargo period will be mentioned in the section 4 tables to keep the data for only a period of time
- The data producers will make their data for third-parties within public repositories. They will be reused for the purpose of validating scientific publications.

7 Responsibilities

In order to face the data management challenges efficiently, all AUTOPILOT partners have to respect the policies set out in this DMP and datasets have to be created, managed and stored appropriately.

The data controller role within AUTOPILOT will be undertaken by Francois Fischer (ERTICO) who will directly report to the AUTOPILOT Ethics Board.

Each data producer and WPL is responsible for the integrity and compatibility of its data during the project lifetime. The data producer is responsible for sharing its datasets through open access repositories, and for providing the latest version.

Regarding ethical issues, the deliverable D7.1 details all the measures that AUTOPILOT will use to comply with the H2020 Ethics requirements.

The data manager role within AUTOPILOT will directly report to the Technical Meeting Team (TMT). The data manager will coordinate the actions related to data management and in particular the compliance to Open Research Data Pilot guidelines. The data manager is responsible for implementing the data management plan and he ensures it is reviewed and revised.

8 Ethical issues and legal compliance

Ethical issues related to the AUTOPILOT project will be addressed in the D7.1

As explained in chapter 2, the IoT platform is a cloud platform that will be hosted on IBM infrastructure, and maintained by IBM IE. It will integrate and aggregate data from the various vehicles and pilot sites.

All data transfers to the IBM hosted IoT platform are subject to and conditional upon compliance with the following requirements:

- Prior to any transfer of data to the IBM hosted central IoT platform, all partners must execute an agreement as provided for in Attachment 6 of the AUTOPILOT Collaboration Agreement.
- All the partners must agree to commit not to provide personal data to the central IoT platform and to ensure that they secures all necessary authorizations and consents before sharing data or any other type of information ("background, results, confidential information and/or any data") with other parties.
- Every partner that needs to send and store data in the central IoT platform has to request access to the servers, and inform IBM IE what type of data they will send.
- IBM IE will review all data sources BEFORE approving them and allowing them into the central IoT platform, to ensure they are transformed into data that cannot be traced back to personal information.
- No raw videos/images or private information can be sent to the central IoT platform. The partners who will send data to the platform must anonymize data first. Only anonymized information that will be extracted from the raw images/videos (e.g., distance between cars, presence of pedestrians, etc.) will be accepted and stored.
- The central IoT platform will only be made available to the consortium partners, and not to external entities.
- IBM IE reserves the right to suspend partners' access in case of any suspicious activities detected or non-compliant data received. IBM IE may re-grant access to the platform if a solution demonstrating how to prevent such sharing of personal data and sensitive personal data is reached and implemented.
- IBM IE may implement validation procedures to check that the submitted data structures and types are compliant with what the partners promised to send to the central IoT platform.
- All the data will be deleted at the end of the project from all servers of the central IoT platform.

The privacy and security issues related to the AUTOPILOT project will be outlined in deliverable D7.1 and addressed in the WP1 Task 1.5 for Security, Privacy and Data Specification issues.

9 Conclusion

This deliverable provides an overview of the data that AUTOPILOT project will produce together with related data processes and requirements that need to be taken into consideration.

The document outlines an overview of the dataset types with detailed description and explains the processes that will be followed for test sites and evaluation within high-level representations.

Chapter 5, which describes the datasets, has been updated from the previous version of the DMP (D6.7) to reflect the current progress of the project. . This includes a detailed description of the standards, methodologies, sharing policies and storage methods.

The Data Management Plan is a living document. The final version of the DMP will be available at the end of the project and will provide all the details concerning the datasets. These datasets are the result of the test sessions performed at pilot site. The data will be stored in public repository after the project.

10 Annexes

10.1 Annex 1: Open research data pilot

Open access refers to the online provision of scientific information that is free of charge to the end-user and reusable. This scientific information handles the peer-reviewed scientific research articles/publications and the research data underlying publications.

Under the H2020 terms, the project must also aim to deposit the research data needed to validate the results presented in the deposited scientific publications, known as “underlying data”. In order to effectively supply this data, project participants need to consider at an early stage how they are going to manage and share the data they create or generate under H2020 guidelines on data management and with respect to the AUTOPILOT Grant Agreement.

“The Commission is running a flexible pilot under Horizon 2020 called the **Open Research Data Pilot** (ORD pilot). The ORD pilot aims to improve and maximize access to and re-use of research data generated by Horizon 2020 projects and takes into account the need to balance openness and protection of scientific information, commercialization and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions.”

“By extending the pilot, open access becomes the default setting for research data generated in Horizon 2020.

However, not all data can be open. Projects can therefore opt out at any stage (either before or after signing the grant) and so free themselves retroactively from the obligations associated with the conditions if:

- participation is incompatible with the obligation to protect results that can reasonably be expected to be commercially or industrially exploited
- participation is incompatible with the need for confidentiality in connection with security issues
- participation is incompatible with rules on protecting personal data
- participation would mean that the project's main aim might not be achieved
- the project will not generate / collect any research data or
- there are other legitimate reasons (you can enter these in a free-text box at the proposal stage).”

After depositing publications beneficiaries must ensure open access to those publications via the chosen repository.

“The two main routes to open access are:

- Self-archiving / 'green' open access – the author, or a representative, archives (deposits) the published article or the final peer-reviewed manuscript in an online repository before, at the same time as, or after publication. Some publishers request that open access be granted only after an embargo period has elapsed.

Open access publishing / 'gold' open access – an article is immediately published in open access mode. In this model, the payment of publication costs is shifted away from subscribing readers.”

In the research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital format and stored in a public repository. Normally, users can access, mine, exploit, reproduce, and disseminate openly accessible research data free of charge as explained in the following figure.

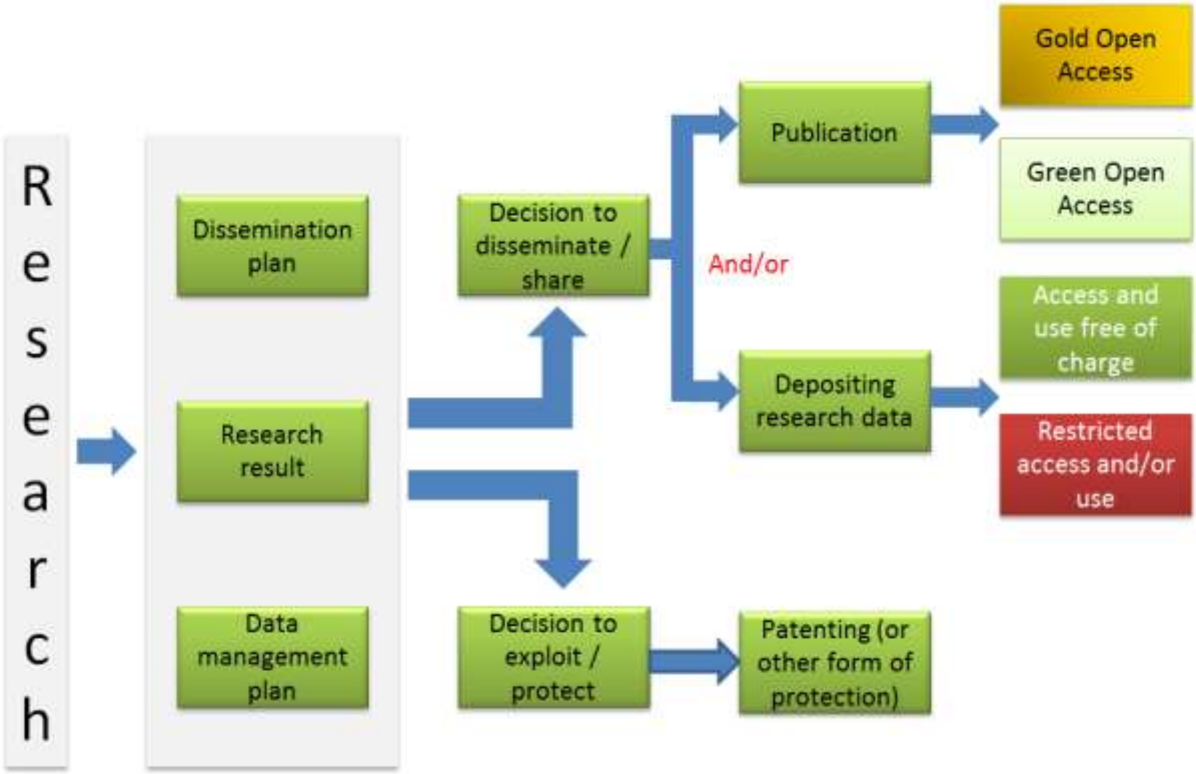

Figure 5 – Principles of H2020 open access to research data