# [L7.1] SAFE AND SECURE PROCESS FOR DATA COLLECTION

**Main authors: M. Kaczmarek, M Oubaiabra, M Ndoye (APSYS)**

**Update: 2021-10-29**

**Keywords:** Data collection, Big Data, Data management, Artificial Intelligence, GDPR

**Abstract :** Many publications and study projects have outlined the data ecosystem implemented at different levels and at the interfaces of the different actors involved in autonomous mobility.

This deliverable provides a representative (and not exhaustive) panorama of the state of the art in data management. It also opens the way for the various constraints and requirements in terms of security, cybersecurity, performance of autonomous mobility especially from the point of view of the use and the generation of mass data.

**Résumé :** De nombreuses publications et projets d'étude ont esquissé l'écosystème des données mises en œuvre à différents niveaux et aux interfaces des différents acteurs de la mobilité autonome.

Ce livrable donne un panorama représentatif (et non exhaustif) de l'état de l'art dans la gestion des données. Il ouvre aussi la voie aux différentes contraintes et exigences en termes de sécurité, de cyber-sécurité, de performance de la mobilité autonome sur le registre de l'utilisation et de la génération de données en masse.

**TABLE OF CONTENTS**



**FIGURES**

**TABLES**

**ACRONYMS**

| | |
|---|---|
| ADAS | Advanced Driver Assistance Systems |
| AI | Artificial Intelligence |
| ANSP | Access Network System Planning |
| AV | Automotive Vehicle |
| CAN | Control Area Network |
| CNIL | Commission Nationale Informatique & Libertés |
| CSU | Channel Service Unit |
| DoD | Department of Defence |
| eMaas | Ethernet Mobility as a service |
| FAA | Federal Aviation Administration |
| GDPR | General Data Protection Regulation |
| GNSS | General Navigation Satellite System |
| GPS | Global Positioning System |
| HAD | Hardware Acquisition Data |
| HLP | High Level Parameters |
| IA | Intelligence Artific"ielle |
| IAAS | Infrastructure As A Service |
| ICST | Information Communication and Space Technology |
| ICT | Information and Communication Technologies |
| IoT | Internet of Things |
| ITS | Information Technology Service |
| LiDAR | Light Detection And Ranging |
| M2M | Man to Machine |
| Maas | Mobility as a service |
| MAC | Multiple Access Control |
| MaTeLo | Markov Test Logic (tool) |
| MOOVE | MOnitoring Outillé du Véhicule dans son Environnement |
| OBD | On Board Diagnostic |
| ODD | Operation Design Domain |
| OEM | Original Equipment Manufacturer |
| PAAS | Platform As A Service |
| R&D | Recherche & Développement |
| RDBMS | Relational Data Base Management System |
| REX | Retour d'EXpérience |

| | |
|---|---|
| RGPD | Règlement Général sur la Protection des Données |
| SAAS | Soft As A Service |
| SCS | Safety Critical Scenario |
| SGAM | Smart Grid Architecture Model |
| SMIS | Seamless Mobile Information Sharing |
| SNA | Smart Network Architecture |
| SSD | Solid State Drive |
| SVA | Safety of Vehicle Autonomous |
| TAAS | Transportation As A Service |
| TBD | To Be Defined |
| USB | Universal Serial Bus |
| UTM | Unmanned Traffic Management |
| V2I | Vehicle to IoT |
| V2V | Vehicle to Vehicle |
| V2X | Vehicle to X |
| VA | Véhicule Autonome |
| VIN | Vehicle Identification Number |
| WP | Work Package |
| XiL | XML Imaging Language |
| XML | eXtensible Markup Language |

# 1 INTRODUCTION

This deliverable makes it possible to take the real dimension of autonomous mobility.

The autonomous vehicle communicates, speaks, listens, examines, analyzes, decides and acts.

The various WPs provided an overview of what already exists, both from a regulatory and normative point of view, trials and tests, and information security technologies.

This document will focus on the concrete data that make it possible to operate the autonomous vehicle in an operational manner. This data inventory will hardly be exhaustive, it will be enriched over time with the zooms that will be carried out on all the functionalities of each subsystem making up the overall autonomous mobility system.

It will also be enriched with technological developments which will generate increasingly reliable and massive volumes of data that can be used by algorithms derived from artificial intelligence.

This document will also be fully consistent with the nomenclatures, glossaries and taxonomies validated within the general framework of the project.

The precision of the descriptions and characteristics listed in this deliverable, will be a bridge that will have a great interest in the realization of the deliverable Cf. [17].

These two documents may however be self-supporting.

# 2 SCALABLE INVENTORY OF DATA INVOLVED IN AN AUTONOMUS TRANSPORTATION SYSTEM

This cartographic representation aims to list the main data used during the autonomous vehicle operations. This list is not exhaustive, because it depends on different technologies involved for each functional need.

This description splits the data used during:

- the learning phase,
- the autonomous driving on a dedicated route,
- the autonomous driving in an open road environment.

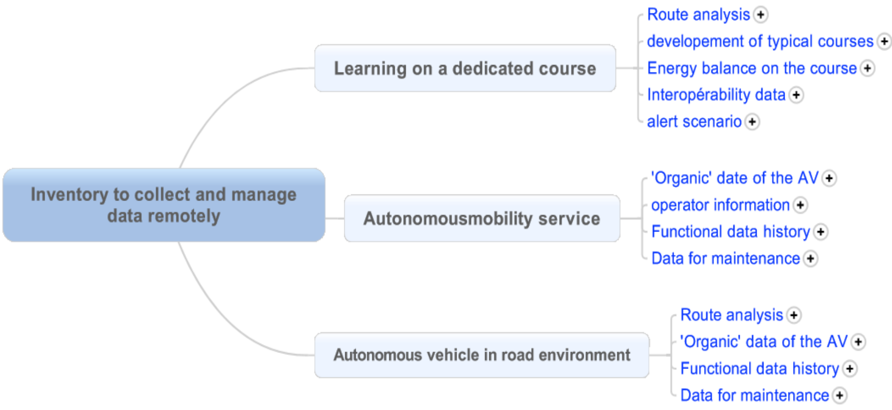


**Figure 1 : Splitting data used**

## 2.1 The learning phase: preliminary inventory descriptive list

This inventory shall be completed during the main steps of the project, to become more and more exhaustive and precise.

It describes the main ‘categories’ of data which are needed to process the learning sequence of an autonomous shuttle or vehicle.

Some data are specific to:

- The operator internal processes,
- The autonomous vehicle decision system,
- The autonomous vehicle maintenance requirements,
- The autonomous vehicle communication and data exchange processes.

For each category or family of data, the technical solution used to assume a functional need, can be designed in a specific architecture or framework format. The challenge for autonomous transportation is to set requirements of interoperability in data exchange and (re)-use them in real time in the most thrifty way. All vehicles operated by different companies or owners will be constrained to share the most precise information and data about traffic conditions (congestion, density, unexpected event, weather…).

The figure below provides an example.

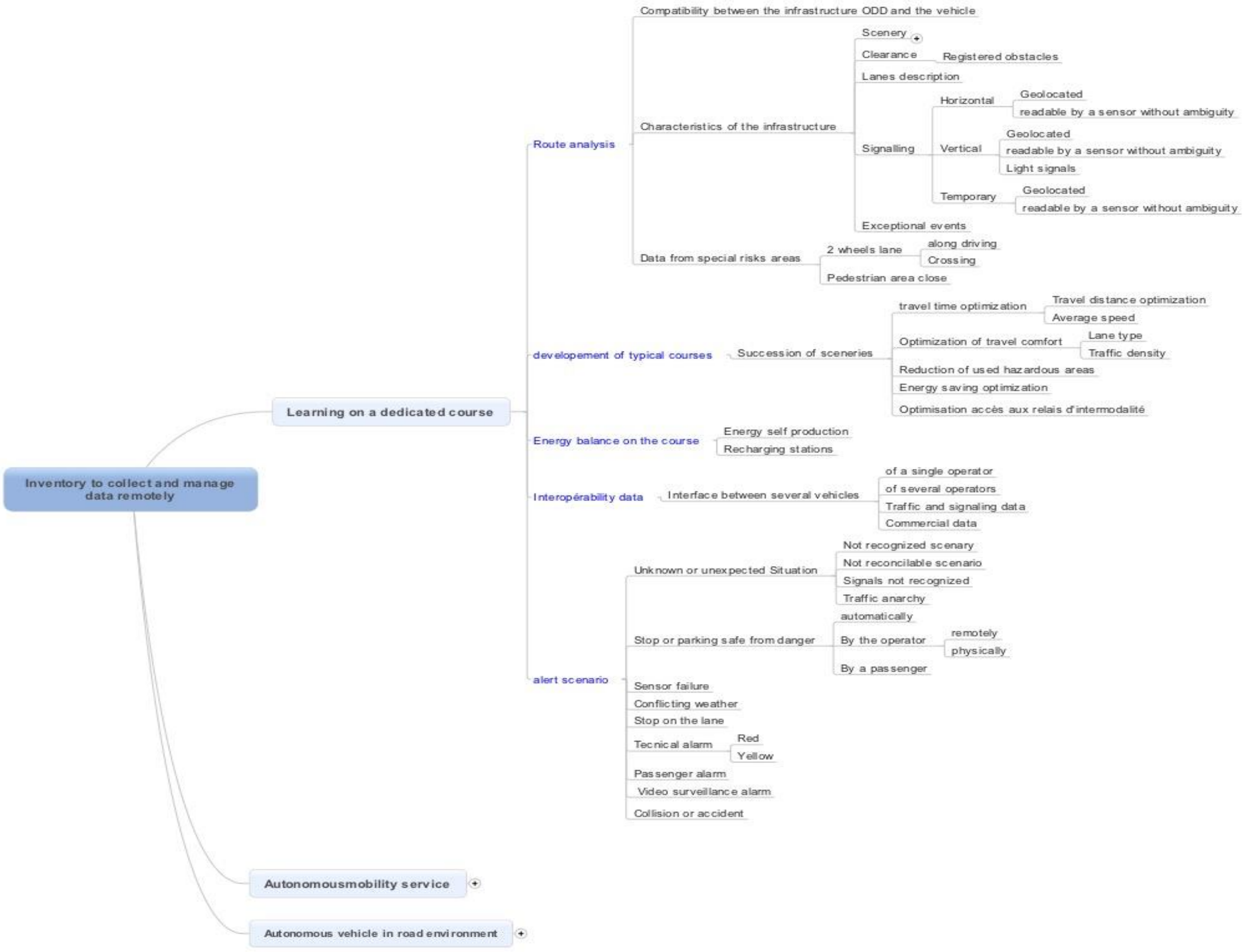


**Figure 2 : Example for data categories in learning phase**

## 2.2 The autonomous driving on a dedicated route: preliminary inventory descriptive list

This inventory complements the previous learning phase with the data collected and involved during the autonomous driving on a dedicated way.

The autonomous vehicle generates and transmits its own data captured and processed along its trajectory.

Some of this data are used in the IA algorithms, and many other information are stored to be processed by the operators in digital twins models for examples or upgraded learning algorithms.

The figure below provides an example.

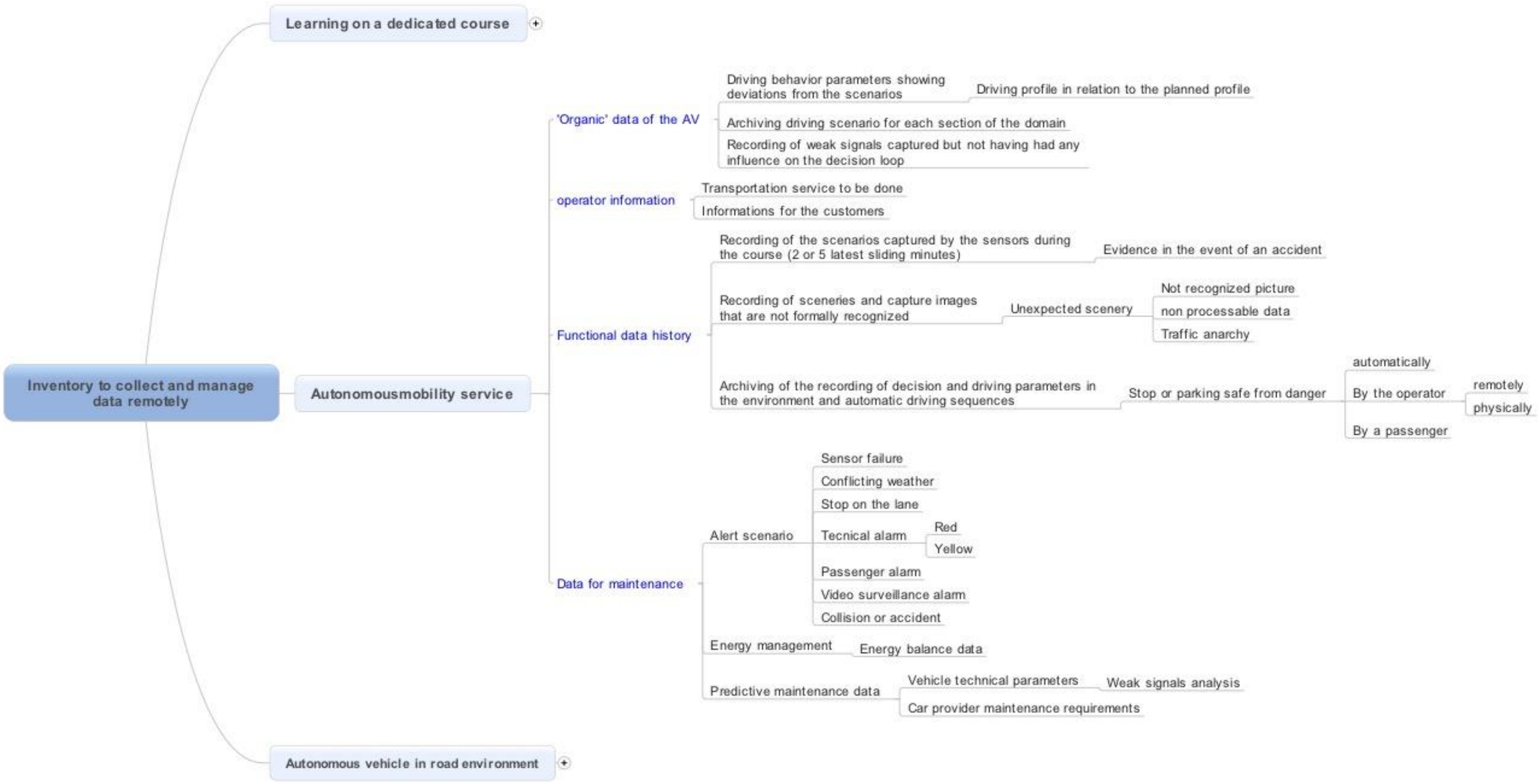


**Figure 3 : Example for data categories in autonomous driving on a dedicated road**

## 2.3 The autonomous driving on a road environment: preliminary inventory descriptive list

This description aims to highlight the difference between the operational data involved in a dedicated way and a road in an open environment.

The figure below provides an example.

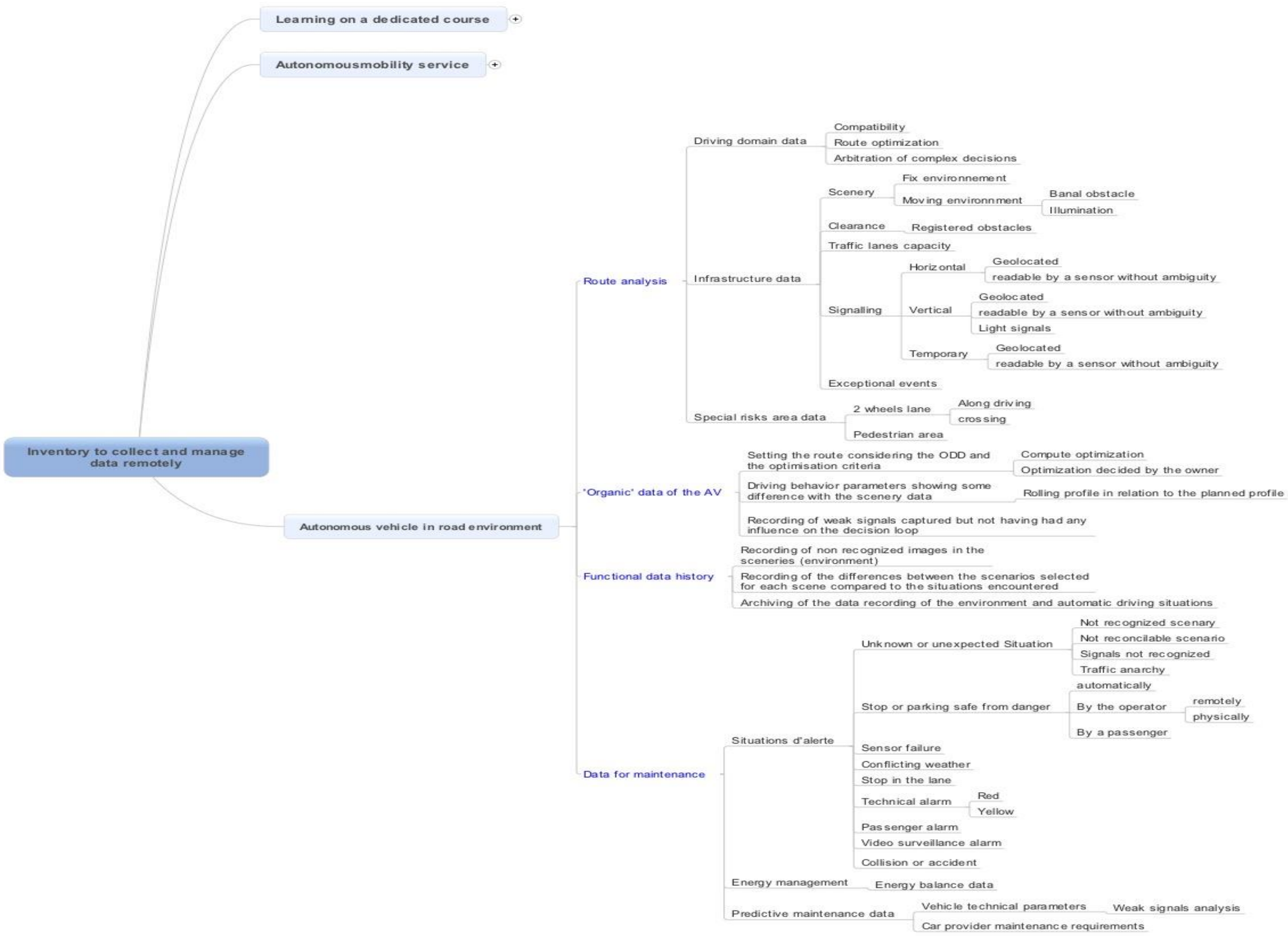


**Figure 4: Example for data categories in autonomous driving on a road environment**

# 3 DATA CLASSIFICATION, CHARACTERISTICS, MANAGEMENT MODALITIES, ONTOLOGIES, AND SEMANTICS (MATRIX DESCRIPTION)

The data is classified by type. Each data is characterized by descriptors:

- Data type ("organic" data for the AV, operation data, maintenance data, events and functioning recordings…),
- Sensor or device which generates the data (camera, radar, LiDAR, sensor, satellite, GPS…),
- Item description,
- Operator or owner (data referring to industrial property, shareable or interoperate data…),
- Format or technology,
- Data volume (rate of flow…),
- Location (embedded storing, cloud…),
- Transfer mode (late or live transmission…),
- GPRD constrained,
- Big Data,
- Ontology link,
- Comments,
- These excel files are available in a separate root file.

The table below is an example.

| Type de donnée | Item | Description de l'item (si besoin) | Owner | Format / Techno | Volume | Hébergement | Transfert | RGPD | Big data | Lien ontologie | Commentaires |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Données 'organiques' du VA | Capteur | Capteur photographique | | Caméra | 20 à 60 Mo/s | Interne car besoin en temps réel | Transfert vers le Cloud dans certains cas | | Oui -> Données sécuritaires hautement prioritaires, voire en temps réel (situations d'urgence, événements fortuits, évitement de collision…) | | Les volumes cités pour les capteurs sont issues du site : https://pro.largus.fr/actualites/les-vehicules-autonomes-a-lepreuve-du-stockage-des-donnees-9503394.html |
| Données 'organiques' du VA | Capteur | Capteur par ondes sonores inaudibles | | Sonar | 10 à 100 Ko/s | Interne car besoin en temps réel | Transfert vers le Cloud dans certains cas | | Oui -> Les données sécuritaires hautement prioritaires, voire en temps réel (situations d'urgence, événements fortuits, évitement de collision…) | | |
| Données 'organiques' du VA | Capteur | Capteur de temps de vol | | Lidar | 10 à 70 Mo/s | Interne car besoin en temps réel | Transfert vers le Cloud dans certains cas | | Oui -> Les données sécuritaires hautement prioritaires, voire en temps réel (situations d'urgence, événements fortuits, évitement de collision…) | | |
| Données 'organiques' du VA | Capteur | Capteur par ondes électromagnétique | | Radar | 10 Ko/s | Interne car besoin en temps réel | Transfert vers le Cloud dans certains cas | | Oui -> Les données sécuritaires hautement prioritaires, voire en temps réel (situations d'urgence, événements fortuits, évitement de collision…) | | |
| Données 'organiques' du VA | Récepteur de navigation par satellite | / | | GNSS GPS | GPS -> 50 Ko/s | Interne car besoin en temps réel | Oui si besoin d'indiquer sa position à un tiers. | Oui dès lors que le traitement de donnée nécessite un envoi à un tiers ou de réaliser une fonction à distance. | Oui -> Les données sécuritaires hautement prioritaires, voire en temps réel (situations d'urgence, événements fortuits, évitement de collision…) | | Voir site de la CNIL pour ce qui concerne la RGPD : https://www.cnil.fr/fr/reglement-europeen-protection-donnees |

| Type de donnée | Item | Description de l'item (si besoin) | Owner | Format / Techno | Volume | Hébergement | Transfert | RGPD | Big data | Lien ontologie | Commentaires |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Données 'organiques' du VA | Software | / | | / | 1 Go (ex: Model Tesla) | Interne car besoin en temps réel | ? | | | | Le volume est issu du site : https://blocksandfiles.com/2020/01/17/connected-car-data-storage-estimates-vary-widely/ |
| Données 'organiques' du VA | Autres capteurs | Capteur de vitesse véhicule<br>Capteur de vitesse moteur<br>Capteur de température (eau/huile)<br>Capteur de pression (Collecteur / Pneu)<br>Débit d'air<br>Etc... | | / | Faible par rapport aux autres capteurs | Interne car besoin en temps réel | | | | | |
| Données exploitation | Gestionnaire, exploitant | / | | | | Externe (Stockage objets ou cloud) | | | | | |
| Données exploitation | Trafic | / | | Utilisation du récepteur de navigation | | Externe (Stockage objets ou cloud) | | Oui dès lors que le traitement de donnée nécessite un envoi à un tiers ou de réaliser une fonction à distance. | | | Données utilisé par Waze / Google maps |
| Données exploitation | Parcours | / | | Définition du parcours à partir d'un assistant de navigation + utilisation de l'IA pour le pilotage sur le parcours | | Pour le pilotage hébergement en interne car besoin en temps réel (Intelligence artificiel) mais la définition du parcours peut-être éventuellement réalisé | | | Oui | | |

| Type de donnée | Item | Description de l'item (si besoin) | Owner | Format / Techno | Volume | Hébergement | Transfert | RGPD | Big data | Lien ontologie | Commentaires |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | à partir des données externe. | | | | | |
| Données exploitation | Données de l'infrastructure | Signalisation | | V2I (Vehicle to Infrastructure Communication) | | Interne car besoin en temps réel | | | | | https://www.unitec.fr/wp-content/uploads/2018/11/Dossier-de-veille-Le-v%C3%A9hicule-autonome-en-milieu-urbain.pdf |
| Données exploitation | Données Géographiques | Position du véhicule | | GNSS (localisation par satellite), d'INS (navigation inertielle) et d'odométrie (mesure de déplacement des roues) ou des capteurs embarqués | ≈ 6 Mo/h -> 1,67 Ko/s (Ex : Waze / GOOGLE MAPS) | Interne / Externe ? | | | | | Données utilisé par Waze / Google maps |
| Données exploitation | Interopérabilité | Cohabitation de plusieurs véhicules: Un même exploitant | | Le V2V (Vehicle to Vehicle Communication) | | | Oui car il s'agit de communiquer des informations à un tiers. | | Oui si il y a de nombreux véhicules. | | https://www.unitec.fr/wp-content/uploads/2018/11/Dossier-de-veille-Le-v%C3%A9hicule-autonome-en-milieu-urbain.pdf |
| Données exploitation | Interopérabilité | Cohabitation de plusieurs véhicules: Plusieurs exploitants | | Le V2V (Vehicle to Vehicle Communication) | | | Oui car il s'agit de communiquer des informations à un tiers. | | Oui si il y a de nombreux véhicules. | | https://www.unitec.fr/wp-content/uploads/2018/11/Dossier-de-veille-Le-v%C3%A9hicule-autonome-en-milieu-urbain.pdf |

| Type de donnée | Item | Description de l'item (si besoin) | Ow ner | Format / Techno | Volume | Hébergement | Transfert | RGPD | Big data | Lien ontologie | Commentaires |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Données exploi-tation | Intero-pérabi-lité | Cohabita-tion de plu-sieurs véhicules: Données commer-ciales | | | | Externe | Oui car il s'agit de communiquer des informations à un tiers dans un but commer-cial. | Oui dès lors que le traitement de don-née nécessite un en-voi à un tiers ou de réaliser une fonction à distance. | Oui si il y a de nom-breux véhicules. | | |
| Données histo-riques de fonc-tionnement | Bases donnée | / | | | | | | Oui si les données sont associées à un véhicule / une per-sonne. | | | |
| Données histo-riques de fonc-tionnement | Scénario | Scénario d'alerte | | | | Interne / Ex-terne | Oui possible (ex communication avec les services d'urgence en cas d'accident) | Oui si les données sont associées à un véhicule / une per-sonne. | | Utilisation de terme commun pour la défi-nition de scénario et donc l'ontologie est un bon moyen pour définir correctement ces scénarios. Utiliser essentiellement pen-dant les phases de testing et pour incré-menter la base de données. | Voir les articles suivant :<br>- Jean-Louis BOULANGER, 2011, Reducing the Gap Between For-mal and Informal Worlds in Au-tomotive Safety-Critical Systems<br>- IRT System X, Reducing the Gap Between Formal and Infor-mal Worlds in Automotive Safety-Critical Systems<br>- Chen WEI, 2018, Formal Mod-els for the Conceptualization and Characterization of Use Cases for the Autonomous Vehi-cle Activity Report |
| Données histo-riques de fonc-tionnement | Scénario | Scénario de fonctionne-ment | | | 2TB of data is col-lected / day | Externe | | Oui si les données sont associées à un véhicule / une per-sonne. | | Utilisation de terme commun pour la défi-nition de scénario et donc l'ontologie est un bon moyen pour définir correctement ces scénarios. Utiliser essentiellement pen-dant les phases de testing et pour incré-menter la base de données. | Voir commentaire ci-dessus + Projet MOOVE |
| Données histo-riques de fonc-tionnement | Com-porte-ment | / | | | | | | Oui si les données sont associées à un véhicule / une per-sonne. | | | |

| Type de donnée | Item | Description de l'item (si besoin) | Ow ner | Format / Techno | Volume | Hébergement | Transfert | RGPD | Big data | Lien ontologie | Commentaires |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Données de maintenance | Alarmes | / | | Voyant lumineux + message sur le tableau de bords | Intégré au données 'software' | Interne / Externe (Remonté base de donnée exploitant) | Oui possible | Oui si les données sont associées à un véhicule / une personne. | | | |
| Données de maintenance | Défaillance | / | | | Intégré au données 'software' | Interne / Externe (Remonté base de donnée exploitant + garage de la marque) | Oui possible | Oui si les données sont associées à un véhicule / une personne. | | | |
| Données de maintenance | Energie | / | | Interne au software | | | Si besoin pour proposition commerciale (borne de recharge etc...) | Oui dès lors que le traitement de donnée nécessite un envoi à un tiers ou de réaliser une fonction à distance. | | | |
| Données de maintenance | Steps de maintenance | Maintenance prédictive Programmation Changement de pneu | | | Intégré au données 'software' | Interne / Externe (Remonté base de donnée exploitant + garage de la marque) | Oui possible | Oui si les données sont associées à un véhicule / une personne. | | | |
| Données de maintenance | etc | / | | | | | | | | | |

**Tableau 1 : Example to data classification**

## 4 REPRESENTATION OF INFORMATIONS FLOWS

To operate with a high level of range, the autonomous vehicle requires the acquisition and transfer of a large amount of data. As a result, the value of these vehicles will no longer depend on the equipment but on the software part composed of an artificial intelligence part allowing an optimal management of these data. We then find the flows specific to the vehicle, namely the data from sensors such as a camera, a sonar or a LiDAR.

These data allow the vehicle to understand its environment. The vehicle will also have to register in a company and therefore have to communicate with both other vehicles but also infrastructures, the manager of the car park, insurance companies and much more.

To manage the complexity and interoperability of mobility systems with a large number of connected vehicles, smart charging stations and information systems, a Smart Network Architecture (SNA) model was proposed. It has quickly become the benchmark for intelligent systems engineering and analysis Cf. [1].

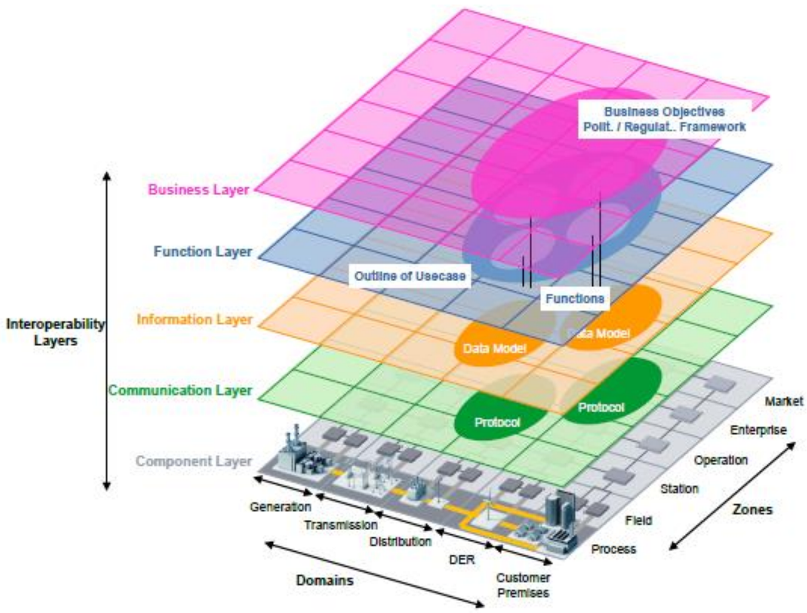


**Figure 5 : Smart Grid Architecture Model (SGAM) Cf. [2]**

The Figure 5 consists of three dimensions (interoperability layers, zones and fields).

The interoperability of the components is ensured by the different functional (Business, Function) and technical (communication, Information, Component) layers.

The fields represent the energy conversion chain.

Zones are control systems based on Information and Communication Technologies (ICT).

Various projects have experimented with the use of SGAM in architecture management, and new uses and areas of application continue to emerge but still no common architecture.

The mobility ecosystem below is an example of data structuration as related with the dependency actors and objects.

- Mobility ecosystem

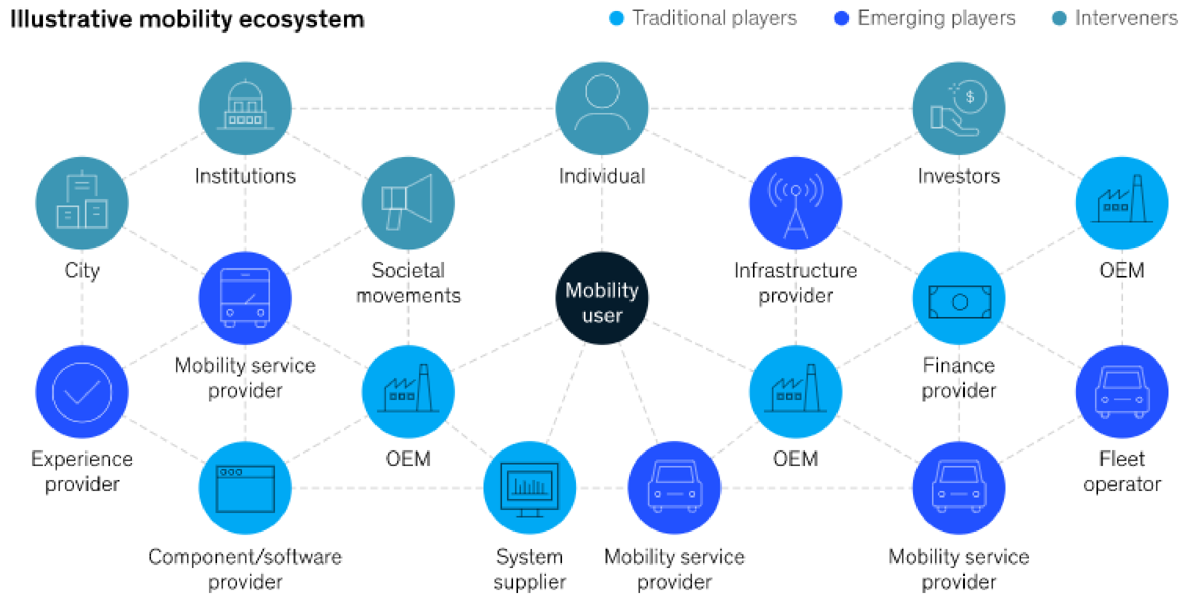


**Figure 6 : Source Mckinsey Center for future mobility Cf. [7]**

The deployment of autonomous mobility services will have an impact on the mobility of our society and the beneficiaries are varied (Cf. Figure 6).

- Passengers (citizens and businesses) contribute data through their use of city infrastructure, mobile technologies and knowledge services. They also enjoy the utilization of those facilities by gaining improved access to city services, employment, friends and family.
- Regional, national and international organizations provide for standardization of procedures, data production and interaction, so as to make sure best practice, interoperability and production, aimed toward improving efficiency and reducing costs.
- City governments set policy and regulations around data collection and data sharing, invest in digital infrastructure, collect data from citizens and city activities, and use that data to set urban planning, design and operation.
- Transport operators including public sector agencies and their suppliers, who install and operate the digital infrastructure, collect and analyze data and use it to guide planning and service delivery.
- Private businesses include people who own and operate digital assets on their own property and gather data from those assets, also as organizations that make value from their activities in processing, analytics and knowledge service development, both international corporations and start-ups.

To support embedded systems, the opportunities lie in the contribution of infrastructure. It can rely on the deployment of perception equipment for example extending the perception of the autonomous driving system compared to a solution that would be based on vehicle sensors. It can also rely on the installation of connectivity equipment in particular to help the vehicle to navigate or to serve the on-board user experience as physical road facilities that includes signs or service center construction for the installation of charging stations for autonomous shuttles.

- Mobility as a service

This sophisticated information system is the corner stone of a new business model: “Maas”.

The Maas is presented by some specialists as one of the keys to the success of the new sustainable mobility offer in the territories. According to document Cf. [3], Maas is a user-centric, intelligent mobility distribution model in which all mobility service providers’ offerings are aggregated by a sole mobility provider (the Maas provider) and supplied to users through a single digital platform.

Concretely the Maas is an application available on smartphone that will provide both information on the services available that will allow to define the optimal path according to the criteria that have been selected by the user and that will also allow to pay and validate access to transport.

The Maas is promising, but today faces a number of obstacles including the economic model which is not yet found and which is based on the exploitation of data. The authors Cf. [4] argue that although the number of travelers continues to increase, the list of suppliers is not long and this can be explained by the lack of a common architecture of the Maas which facilitates the complex integration of all actors involved in the ecosystem.

In addition to the traditional actors of mobility, new operators and of digital services are emerging and becoming increasingly complex. This raises the thorny question of whether the Maas should be managed by a public authority or at the initiative of private operators. The answer to this question is not yet found.

The multitude of data generated by the Maas offers a broad opening of the data which leads to a number of reticence and legal problems, hence the demand to protect or anonymize the most sensitive of them.

Figure 7 shows an example of a Maas ecosystem presented by the Finnish Minister of Transport and Communications which is a benchmark in the field.

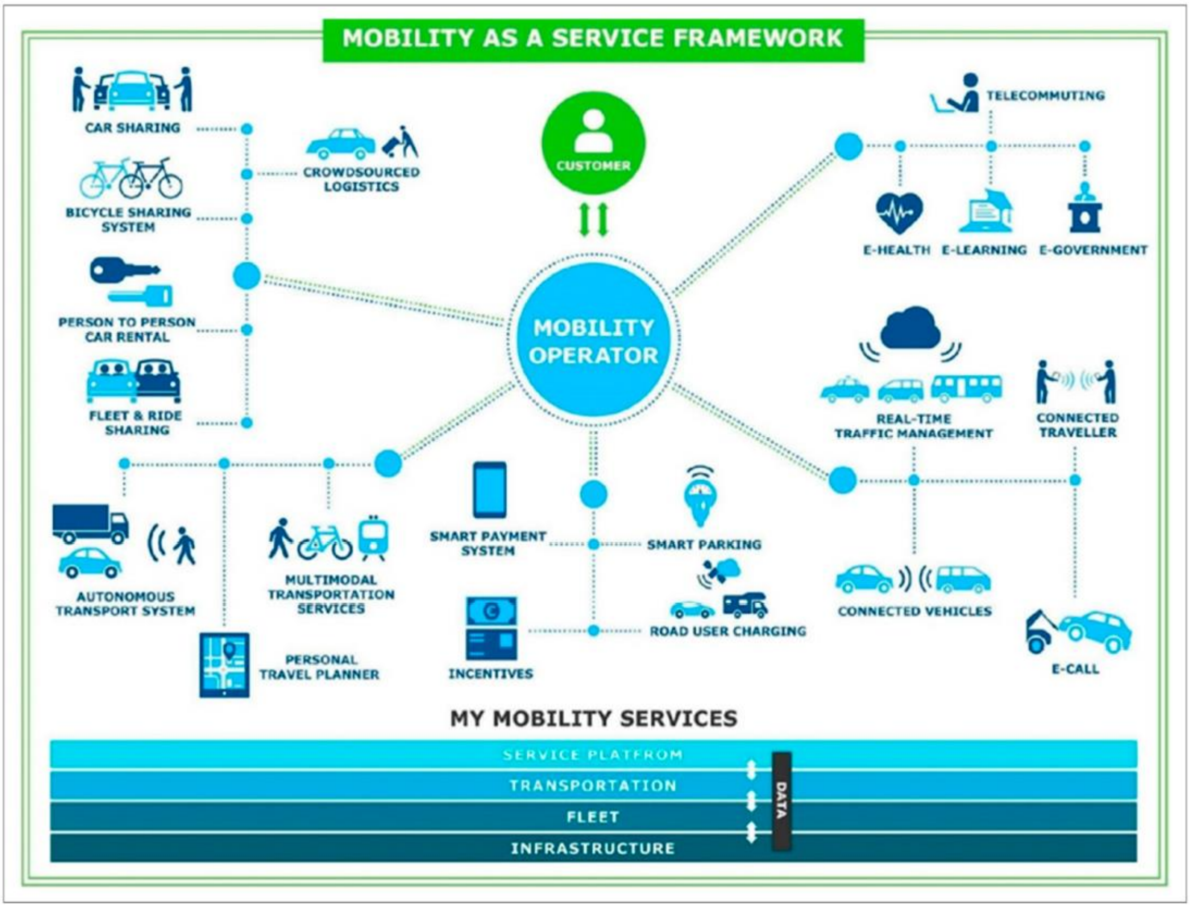


**Figure 7 : The Mobility as a Service framework. Source [5]**

The traveler interacts with many products and services while heading to some destination. Indeed, the traveler could also be the user of a smartphone application to program the journey or check the schedule of the bus for instance, getting the metro station and interacts with the tickets machine, then boards the metro using information panels.

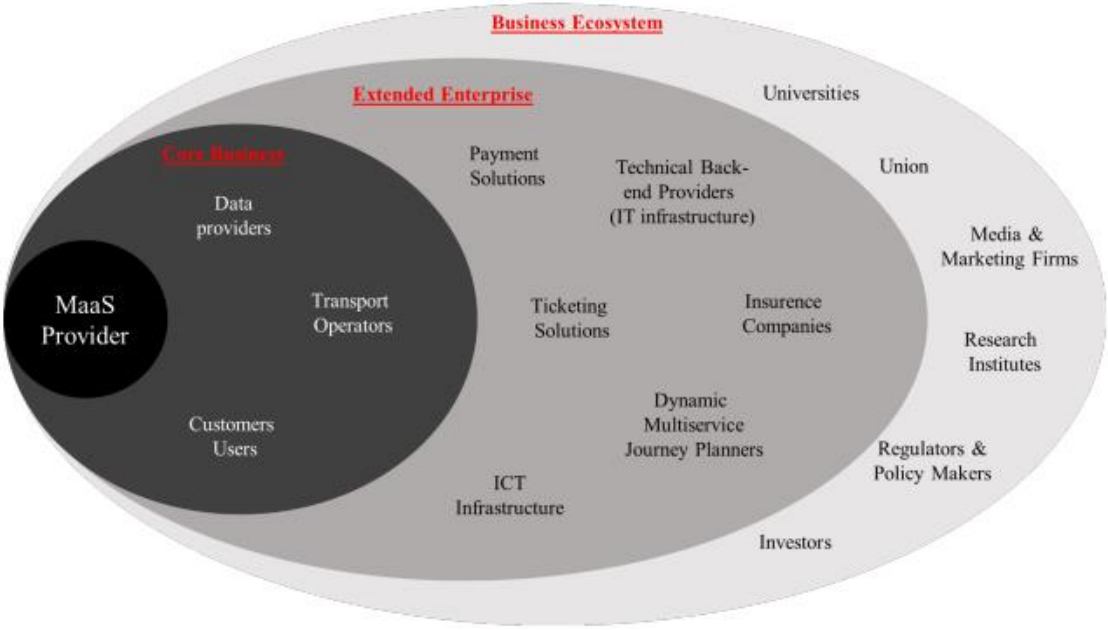


**Figure 8 : The proposal of Mobility as a Service Ecosystem Source: Adapted from [3]**

An additional perspective of the Maas ecosystem is the Maas business ecosystem.

Figure 8 shows the layers of the business ecosystem proposed by authors of documents Cf. [3] and Cf. [4].

There is also the eMaas concept, which is built on the Maas model and which refers to ecology in order to reduce the number of vehicles on the roads, thus reducing the effects related to pollution.

In the field of aviation, author Cf. [6] proposed the Seamless Mobile Information Sharing (SMIS) architecture, which is initially the concept of integrating business and urban aviation into MaaS. The architecture is based on a system-wide information management program that allows many different users to subscribe to the data at the same time. The aim of the plan is to support the sharing of information between stakeholders in the national airspace system by providing communication infrastructure and architectural solutions. An overview of the SMIS architecture and its layers is shown in Figure 9.

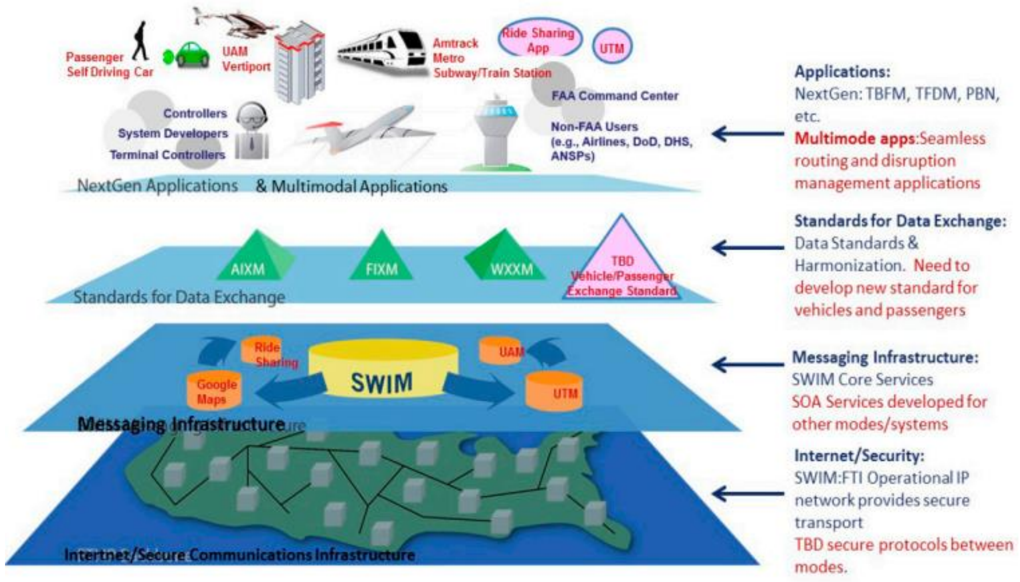


**Figure 9 : Seamless Mobility Information Sharing Architecture Cf. [6]**

# 5 TECHNOLOGY THAT IS NOW BEING USED, AS WELL AS TECHNOLOGICAL ADVANCEMENTS THAT ARE BEING WATCHED

## 5.1 Safe and secure processes

According to document Cf. [8], the authors define a scene as «Big Data blends together the collection of large volumes of high velocity, heterogeneous, evolving domain data and the use of advanced techniques and models to store, retrieve, manage, process and analyse the captured information.»

The life cycle of data and services according to document Cf. [8] can be the following:

- data collection,
- infrastructure provisioning,
- data analysis,
- data retrieval and security.

Below different method used for each step:

- Data collection -> Io9 paradigm
- Infrastructure provisioning -> Cloud Computing and SAAS/PAAS/IAAS
- Data retrieval -> Open Data
- Data analysis -> Big Data Analytics (Deep Learning, Computational Intelligence, Soft Computing …)
- Security -> Cybersecurity

Below, a diagram of the intersection between the paradigms. It is an extract of document Cf. [8].

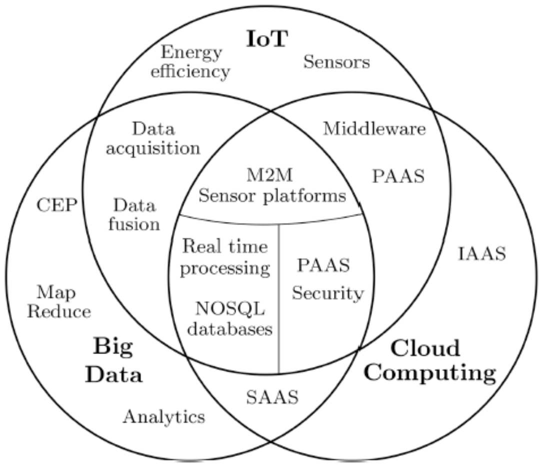


**Figure 10 : Schematic diagram showing the intersection between IoT, Cloud Computing and Big Data paradigms Cf. [8]**

The Figure 10 illustrates the intersection of three paradigms, which have recently become essential technological components of intelligent data services and applications in transport and mobility: IoT, Cloud Computing and Big Data.

It is possible to analyze the data with Big Data technology without stored in database thanks to data stream.

Big Data can be define with four features:

- Volume: Amount of data captured, managed and analyzed,
- Variety: Variety of data sources,
- Velocity: « Rate at which data instances are produced; and the speed at which such data samples are received and processed »,
- Veracity: Authenticity and security of data.

There are different types of Big Data:

- Structured data -> a formal structure,
- Unstructured data -> no predefined schema or data model as example a sensor or social media,
- Semi-structured data, it is a mix of structured and unstructured data as example the language XML.

The main Big Data sources are:

- Social Media,
- Sensor data (IoT paradigm have accelerated the proliferation of physical devices embedded with electronics, software, sensors and actuators Cf. [8]),
- Open Data (Anyone can freely access and use, modify and share this data).

Below a table of the main data sources. It is an extract of document Cf. [8].

| | Source | Type | Format | Example |
|---|---|---|---|---|
| traditional sources | ITS | structured | CSV, XML etc. | car navigation or traffic signal control data |
| | legacy systems (RDBMS) | structured | ERM, XML, JSON | database of electronic toll users |
| | GIS data and maps | semi-structured | GIS | road map of a region |
| new potential data sources | Open Data | non-structured | text, CSV | Smart Cities open data |
| | Linked Data | semi-structured | RDF | semantic urban mobility enrichment data |
| | Sensor Data | raw data | text, CSV, sensor formats | new sensors on roadways to measure pavement conditions |
| | Social Media | semi-structured | text, CSV, XML, RDF etc. | crowdfunding, collaborative apps data or user mobile devices data |

**Tableau 2 : Main data source related to transport and mobility domains**

Techniques and methods used for Big Data analytics can be classified as following:

- Descriptive analytics: Unsupervised machine learning, pattern recognition and statistics to discover regular patterns in data,
- Predictive analytics: Learning algorithms,
- Prescriptive analytics: Suggest the best action or decision thanks to optimization techniques, expert systems and other elements from Computational intelligence, Mathematical Programming and Operations Research. (Cf. [8]).

The authors emphasize the importance of ‘Deep Learning models noted lately in the literature, mainly due to their capacity to learn complex data representations as opposed to task-specific feature engineering approaches. Deep Learning aims to capture and understand human mobility and transportation patterns at larger scales than traditional implementations of other machine learning approaches’.

Today, there are several Big Data platform.

| Purpose | Technique, s | Vertical/horizontal approach | Big Data competitive advantage | Real time? |
|---|---|---|---|---|
| *Descriptive* | integration models | horizontal (ITS) | Big Data fusion | yes |
| | statistical functions | horizontal (ubiquitous cities) | Smart Cities | yes |
| | massive data mining | horizontal (space, air and ground data) | Big Data Analytics platform | no |
| | massive data mining | horizontal (real-time datasets) | Big Data Analytics platform | yes |
| | data integration and data mining | horizontal (ITS) | Big Data Analytics platform | no |
| | data mining | vertical (intelligent monitoring) | Big Data Analytics platform | no |
| | trajectory analysis | horizontal (urban and trajectory data) | Big Data Analytics platform | no |
| | scheduling | vertical (logistics) | Big Data sources | no |
| | statistical and econometric models | horizontal (weather conditions, drivers features and road data) | Big Data sources | no |
| *Predictive* | neural networks | vertical (time series) | collaborative applications | no |
| | short-term prediction models | horizontal (ITS) | Big Data sources | no |
| | prediction models | vertical (urban roads) | big data sources | no |
| | prediction models | vertical (bus and transit) | Big Data sources | yes |
| | classification | vertical (bus transport) | Big Data sources | no |
| | classification and statistical models | vertical (*t*-axis licences) | Big Data sources | no |
| *Prescriptive* | heuristic algorithm | vertical (vehicular routing) | Big Data sources | no |
| | deep learning | vertical (vehicle license plate recognition) | deep learning | no |
| | deep learning | vertical (traffic accidents detection) | deep learning and social media data sources | no |
| | optimisation and route planning | vertical (transit routes) | real-time optimisation routing | yes |

**Tableau 3 : Summary of reference works illustrating each of the purposes for which Big Data analytics can be used in transportation and mobility**

This table (Cf. [8]) summarizes the different studies concerning the integration of Big Data analytics in the transport industry.

Big Data is classified into three categories of descriptive, predictive and prescriptive analysis.

Descriptive analysis encompasses unsupervised machine learning which is a means of experimenting with artificial intelligence for data analysis.

Predictive analysis focuses on the learning algorithm that is applied to examples of supervised data in order to link observed dataset with a target value. This analysis works by training.

Prescriptive analysis points to the best decision to choose among several possibilities with optimization technique.

Deep learning, neural networks such as other methods are adapted to descriptive, predictive and prescriptive analysis.

The table below provides overview of applications and related projects presented in the article Cf. [8]:

| Big Data approach | Problem/application (year) |
|---|---|
| *Driver assistance and instrumented vehicles* | |
| multi-sensor data fusion for instrumented vehicles (2012) | Big Data fusion |
| efficient vehicle design (2015) | Big Data Analytics |
| driving data fusion techniques (2016) | Big Data fusion |
| RDMP framework for ADAS (2016) | Big Data platform |
| driving tendency recognition method (2016) | Big Data Analytics |
| behaviour and vehicle dynamics risk analysis (2016) | Big Data framework and policies |
| mobile agents for data management in vehicular networks (2017) | vehicular networks |
| *Traveller information* | |
| Big Data schemes in social transportation systems (2016) | Big Data social transportation |
| guidelines to pioneer public transport (2016) | Big Data services |
| *Roadway operation and management* | |
| roadway control environmental footprint (2016) | Big Data Analytics |
| traffic congestion on limited access roadways (2016) | Big Data Analytics |
| road traffic operation (2016) | Big Data |
| *Traffic management* | |
| traffic flow prediction based on deep learning (2015) | Big Data predictive analysis |
| RC evolution patterns (2015) | Big Data real-time analysis |
| OD matrix generation (2017) | Big Data Analytics |
| *Transit management* | |
| route planning services optimisation (2016) | Big Data Analytics |
| bus planning (2016) | Big Data Analytics |
| general traffic planning using IoT (2016) | Big Data Analytics |
| safety analysis based on simulation (2016) | Big Data |
| pedestrian planning (2017) | Big Data Analytics |
| measuring and monitoring transit system performance (2017) | Big Data fusion and analysis |
| *Emergency and incident management* | |
| detection of incidents into public infrastructure (2016) | Big Data social sensor data |
| predict safety risks over rail incident data (2016) | Big Data Analytics |
| resilience of taxi and subway trips (2016) | Big Data |
| *Transport modes* | |
| assessment of external force acting on ship (2013) | Big Data fusion and analytics |
| condition-based maintenance in railway systems (2015) | Big Data streaming analysis |
| railway risk analysis (2015) | Big Data visualisation |

**Tableau 4 : Big Data approaches used in ITS application area**

You will find below different examples of project in smart mobility information system:

- MOOVE Project

The MOOVE (MOnitoring Outillé du Véhicule dans son Environnement) project is an example of real world driving scenarios carried on by VEDECOM. The aim is to collect a large amount of driving related data on non-automated vehicles Cf. [9].

The first phase of the project is data acquisition. These data are collected on each of the MOOVE vehicles and are synchronized, stored on an SSD mass storage and then transferred to the data center for analysis: finally, the results can be used by all partners sharing the project show Figure 11.

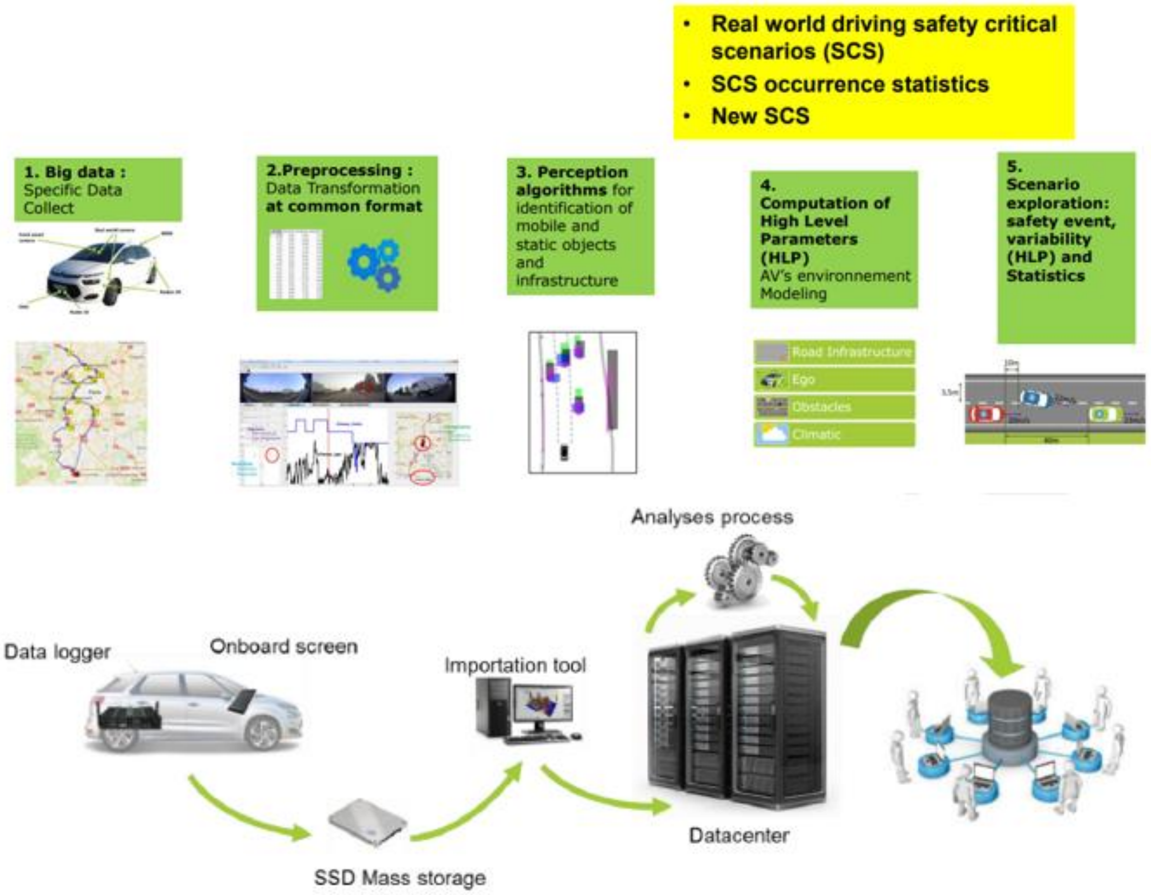


**Figure 11 : Data collection synopsis in MOOVE Cf. [9]**

Data collected from smart sensors and live cameras is directly synchronized using an integrated data logger connected to 8 CAN networks, 3 independent Ethernet networks and various inputs. Several types of watchdog have been put in place to ensure a high quality transfer of data recorded from the data logger. The first is in real time and allows the driver to start recording. For each entry, the number of images per second is compared to the expected rate values. Thanks to the records reports sent to the data center, further verification is carried out by examining the image identifier and frame rates. Finally, a final check is performed by quantifying or each parameter the significant values.

- PEGASUS Project

PEGASUS have proposed a new method: the circuit of relevant scenario.

This project allow to integrate today's available methods and tools. The central element of the circuit of relevant scenario is a data base and its processing toolchain. Furthermore, the toolchain « must be capable to include and use different data sources and therefore heterogenic ».

During the PEGASUS project, the researchers developed a program that guarantees a unified evaluation and verification of the training function in the most efficient way. Pegasus makes a significant contribution to the certification of future autonomous vehicles through development requirements, to the processes, the overall method allows a continuous test sequence through a set of all requirements for the driving function and a collection of relevant traffic situations. Collection is based on field test, simulator and accident data. The data are processed uniformly and made available via a central database for simulation applications, at the actual traffic test site.

The data collection is based on field test, simulator and accident data. The data are processed uniformly and made available via a central database for simulation applications, at the actual traffic test site. The result is a process recommendation and safety assessment. Below is a schematic representation of the project:

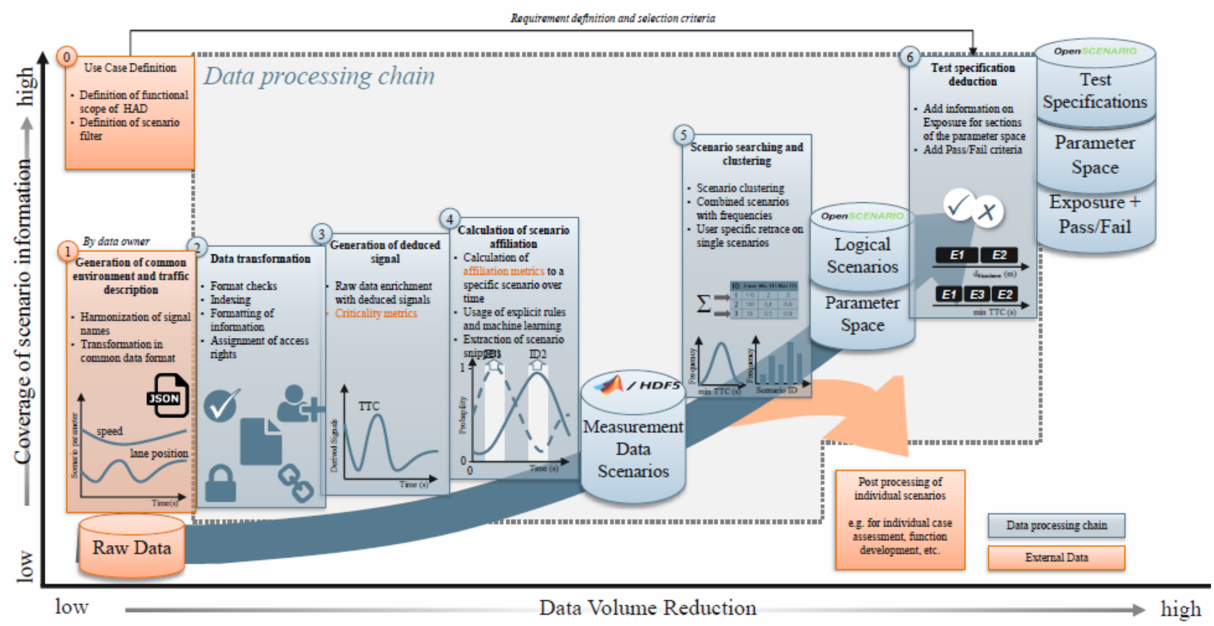


**Figure 12 : Data Base + Data Process Chain**

Figure 12 deals with the coverage of scenario information based on the reduction in data volume. Testing and validating automated vehicles requires new methods and tools for an effective protection process.

Available and known methods and tools can be used throughout the data processing process. The relevant scenarios, which are currently established in the project are:

- the data processing chain must be able to process different data sources and heterogeneous data quality in order to provide a common specification test.
- the database concept presented allows efficient processing of high data volumes through a flexible tool chain.

- SVA Project (Simulation for the Safety of the Autonomous Vehicle)

Autonomous cars interpret their environment through sensors such as cameras and radars, and their design must ensure the safety of their occupants and other road users. Therefore, its design must be able to cope with technical failures that can affect the efficiency of its sensors, external disturbances (such as rain, fog, glare, etc.), and difficulties in interpreting its environment (such as erasing ground marks). The project was launched in 2015. Its objective is to address the issue of the verification of autonomous vehicles by numerical simulation by developing methods and tools to assist in design and verification.

- HEADSTART Project (Harmonized European Solutions for Testing Automated Road Transport)

With multiple different projects in different countries on autonomous vehicles, the European Commission has funded the HEADSTART project in order to identify and harmonize a methodology for validation and verification of automated driving functions, taking into account the needs of different stakeholders.

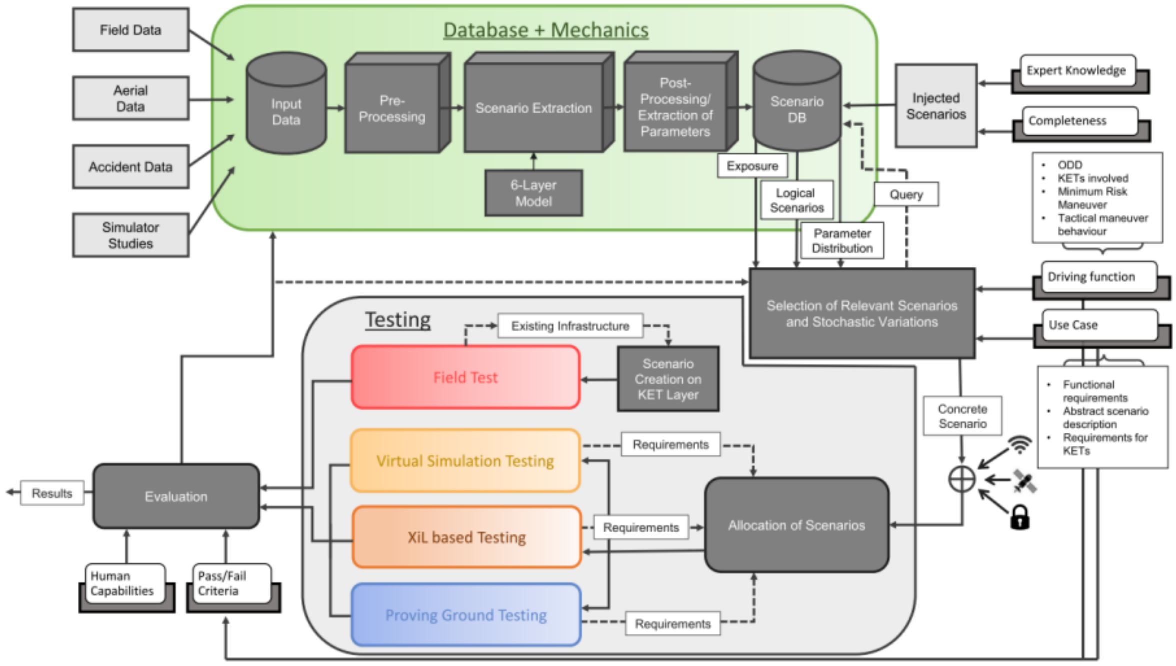


**Figure 13 : Overview on the HEADSTART Methodology from [10]**

An overview of the overall methodology is presented in Figure 13 with scenarios representing input data from different sources. After extracting the scenario from the database, check the scenarios covering all the events listed in the performed table. The next step is to define the relevance of the parameters and to combine them. The test methods targeted by HEADSTART for the simulation allocation process are:

- Ground testing,
- Virtual Testing / Simulation,
- Test X in the Loop (XiL).

- COVADEC Project (Design and Validation of On-board Driver Assistance Systems)

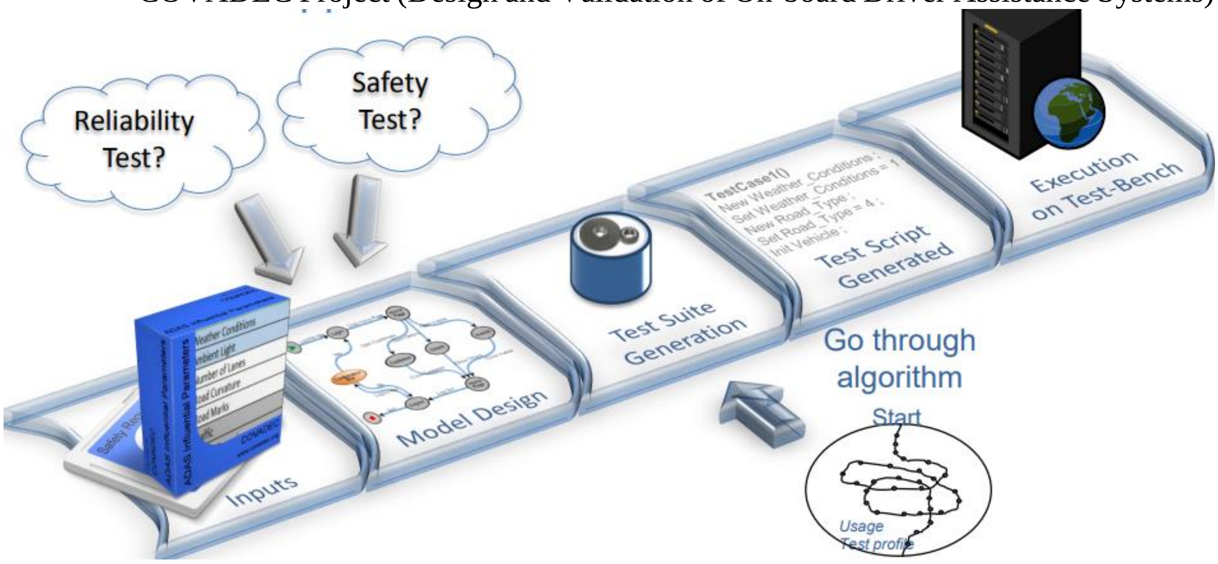


**Figure 14 : Model-Based Testing and Test Automation applied to Advanced Driver Assistance Systems Validation source: covadec**

The objective of this project is to demonstrate that ADAS behavior meets safety objectives and to verify the reliability of ADAS functions by using test activities to measure whether safety objectives are met. The solution is based on MaTeLo which is a Model-based Testing tools. By designing a model with states and transitions, MaTeLo automatically generates test scenarios, test suites, data and scripts for automation. MaTeLo can be used to test computers, embedded systems or artificial intelligence. It generates hundreds of thousands of combinatorics for each unit test maneuver, the tests are graphical, reproducible and can be stored in any ALM or database.

Two types of tests are needed to validate ADAS dependability:

- Safety oriented
- Reliability oriented
- Test Cases

Verify that the ADAS behavior is compliant with safety requirements.

## 5.2 Projection of the impacts of these technological developments*

Following elements enumerate such impact on technology:

- Large amount of data to process (Big Data)
    - How the autonomous vehicle can take advantage of technological developments without practicing an excessive technological over dimensioning trend? The functions of the AV which will benefit from these technological developments in priority are:
        - The level of security
        - The level of performance and autonomy (sensors, decision and fusion algorithms, etc.)
        - Storage of non-essential data in real time (REX of scenarios, traceability of decisions, possibility of replay, etc.)
    - System engineering must ensure that ease of technology does not lead to overconsumption or redundancy of data without demonstrated benefits for safety and intrinsic performance.
    - Non-essential data in real time can be processed in time offset. What about their use, exploitation and pooling of results for the benefit of AV operators.
- Impact on transportation, infrastructure

  The emergence of autonomous vehicles will lead to changes in road and highway infrastructure to adapt to this new mode of transport. Further work and reconstruction of the road will be carried out in accordance with the new standards. Manufacturers work with cartographers to integrate ultra-accurate high-definition maps into cars and automatically update them when changing panels and speeds.
    - Communications infrastructure will evolve with global technology. Only its level of solicitation and transfer must be synchronous with the rationalization of data needs solely for performance and security purposes.
- Autonomous Vehicle Legal Status (Accident Liability)
    - The possibility of recording driving and behavior events is an imperative to provide the necessary evidence for the division of responsibilities in the event of an accident. This data must remain complete and safe until it is collected by the authorities. Data interoperability (Autonomous shuttle).
    - This notion of the possibility of exchange and use of data essential to the management of traffic and behavior of AVs between different operators or vehicle generations is a need to limit the redundancy of data representing the same state of the ecosystem by example.
- Ecological impact

- Big Data has an ecological footprint that is increasingly quantified. Data from the AV ecosystem is no exception.

## 5.3 Protection of confidential "industry" data and personal data

The collection of data is a principle that aims to obtain and measure information from several equipment (sensors, camera ...). These data are stored and allow to have a precise idea about the behavior of the system. It is necessary to be able to collect a large amount of data in the different possible situations.

As mentioned above, this method involves the management of a large amount of data. This management has advantages but also presents risks.

Data collection involves data protection and privacy. The respect of privacy is very much regulated by the law. The collection of data must be carried out for a specific and legitimate purpose. Each collection must stipulate how long it will be kept and offers the possibility of being informed about the data collected. This data protection has been completed by the RGPD. In addition to the protection issues, there is the security of stored data. The access to these data must be controlled to avoid any bad intention.

The collection system is based on the exchange and processing of data between certain elements:

- Inside the car (driver warning sensor),
- Between the connected object in the vehicle and the servers,
- Between vehicle (V2V), vehicle and infrastructure (V2I) - an operation that requires large amounts of data to be collected,
- Vehicle condition, event description, driving mode, etc.,
- Potential identifying data, directly or indirectly (MAC address, VIN number, geographic location, data matching, etc.).

Transmission in real or delayed time, storage for a longer or shorter period

- External activation in the vehicle and in case of alarms, repairs, maintenance, emergency calls, etc.
- In different connection technologies for external media,
- Conventional communication network (Wi-Fi, Bluetooth, 3G/4G, etc.),
- Collaborative Systems Frequency Band (ITS-G5),
- Internal (OBD, USB).

To facilitate compliance with the “European Framework Directive 2010/40/EU” on the deployment of ICST and the “General Data Protection Regulation 2016/679” (GDPR), the CNIL issued recommendations on the protection of connected cars and personal data. In October 2017, it published a compliance report on privacy and connected cars, describing some basic aspects. In these recommendations on the protection of data collected by these types of cars, it states that “all data that may be linked to an identified or identifiable natural person, including the number of the registration plate or the serial number of the vehicle, are protected personal data”.

The rights of connected vehicle users:

- The right not to be known or recognized on travel,
    - In addition to the purpose and conditions of authorization (according to the law, by the regulatory agency CNIL).
- The right to know,
    - Complete, clear and accessible,
    - Concerning the purpose, the method of processing, the recipient of the data,
    - Conditions for exercising access rights, right of objection, etc.

- The right to accept collection and processing, unless legally restricted,
    - "Informed" consent: free, informed (informed) and specific.

Vehicle safety: a major objective

Rules that do not cover the IT security of connected vehicles

- no certification process,
- no direct vehicle regulations yet,
- Automotive manufacturers need to integrate new skills to ensure system safety,
- Regulations relating to the protection of personal data and privacy that require measures to be taken to prevent the security breach.

## 6 POINTING OF DATA SUBJECT TO AI PROCESSING

Artificial intelligence systems have learned to predict the best trajectory from large amounts of data. Then they combined that data with data from real-time information. The driver assistance system warns the driver of possible dangers. For example, it could be an audible warning when the car suddenly deviates from the track.

The vehicle is equipped with physical sensors such as lasers (LiDAR), stereo vision CAMERAS and RADARS. Then use algorithms to merge different sources of information. To help the VA follow the right path and observe dynamic obstacles (pedestrians and vehicles), sensor fusion and estimation that allow detection of different conditions must be able to distinguish very dangerous situations Cf. [11].

The location function determines the location of the vehicle by estimating the location and direction of the vehicle. Common positioning systems use satellite navigation systems (GNSS), such as the Global Positioning System (GPS) or Galileo. The decision-making and control system is responsible for generating and executing the vehicle's behavior. Given the path that vehicles must follow at the road level to accomplish strategic tasks in order to achieve their high-level objectives, this task-level planning process involves different functions. Pathway planning or behavior planning is used in the short term to manage interactions with others road users and infrastructure, and to monitor local driving.

Vehicle-to-Vehicle (V2V), Infrastructure Vehicle (V2I) and Vehicle-to-All (V2X) communications broadcast the current position of the vehicle to nearby traffic, reminding traffic of upcoming operations, traffic jams, accidents, and road construction. V2I technology sends traffic light information to autonomous vehicle acceleration and braking systems to help plan routes based on traffic conditions and frequency of traffic light changes Cf. [12].

European action on the development of the automated and connected vehicle is particularly important in the field of access and data exchange, to ensure interoperability of services and interfaces, ensuring security and privacy. Indeed, the European Commission has not hesitated to consider interoperability as one of the key success factors of the digital transformation, the objective of the AUTOCITS project is to contribute to the deployment of C-ITS in Europe by improving the feasibility of autonomous vehicles and by strengthening the role of C-ITS as a catalyst for the implementation of autonomous driving. To this end, three pilot projects will be implemented in the main European cities like Paris, Madrid… Cf. [13].

# 7 PROJECTION OF THE EVOLUTION OF AI TECHNIQUES IN R&D ON THE USE AND STORAGE OF DATA

The automotive industry is undergoing a unique transformation. Road vehicles are becoming more and more autonomous. Consumers, regulators and industries are preparing for autonomous cars. One of the technical evolutions of artificial intelligence in R&D is the digital twin. In recent years, a large number of Digital Twins concepts and repositories have been proposed and combine artificial intelligence, machine learning and data analysis with data to create a digital simulation model that updates and changes as its physical counterpart changes. Digital twins use multiple sources to learn and constantly update themselves to represent their status, working conditions or location in near real time. The learning system uses the following methods to learn on its own and transmit sensor data from all aspects of its operating state; Human experts, such as engineers with in-depth and relevant industrial knowledge; other similar machines; Other machinery fleets of similar machines; And the wider system and the environment to which it belongs.

Author Cf. [14] research on the application of digital twin technology in intelligent electric vehicles. Intelligent electric vehicles that have been systematically divided into specific areas within intelligent vehicle systems, such as autonomous navigation control, advanced driver assistance system, vehicle condition monitoring, battery management systems and vehicle power and electrical drive system. Author Cf. [15] maintains that the vision of the digital twin refers to the complete physical and functional description of the component, product or system, which includes more or less all information that may be useful at the current stage and later life cycle.

The Digital Twin uses its high-fidelity virtual model, massive twin data, and real-time bi-directional dynamic interaction to perform functions such as simulation, diagnosis, prediction, and optimization control. AI analyzes, integrates and deeply exploits the twin data by matching the best intelligent algorithms to complement services with different needs. With the support of AI, Digital Twin can greatly improve the value of data as well as the responsiveness and accuracy of various functions Cf. [16].

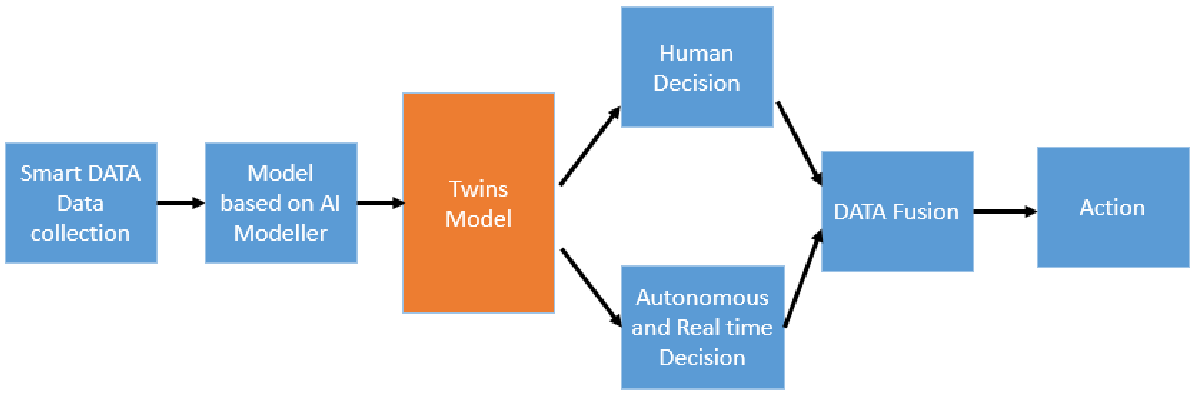


**Figure 15 : Example of Digital Twins process**

AI systems that will evaluate AI to improve data and differentiate between reliable and flawed data.

Reliably assessing the confidence of a deep neural network and predicting its failures is of paramount importance for the practical deployment of these models.

# 8 SUMMARY TABLE OF UPDATES AND CHANGES TO THE DOCUMENT

# 9 APPENDICES

## 9.1 Archives of chapters that became obsolete during the project

## REFERENCES


[1] Benedikt Kirpes, P. D. (2019). Architecture des systèmes de mobilité électrique : un cadre basé sur des modèles pour gérer la complexité et l'interopérabilité. *Information sur l'énergie 2, https://doi.org/10.1186/s42162-019-0072-4.*

[2] Coordination, C.-C.-E. S. (2012). Smart Grid Reference Architecture. Bruxelles: https://ec.europa.eu/energy/sites/ener/files/documents/xpert_group1_reference_architecture.pdf.

[3] Kamargianni, M. M. (2017). *The Business Ecosystem of Mobility as a service.* Washington DC, 8-12: 96th Transportation Research Board (TRB) Annual Meeting.

[4] Reyes García, J., Lenz, G., Haveman, S., & Bonnema, G. (2020). State of the Art of Mobility as a Service (MaaS) Ecosystems and Architectures—An Overview of, and a Definition, Ecosystem and System Architecture for Electric Mobility as a Service (eMaaS). *World Electric Vehicle Journal, https://doi.org/10.3390/wevj11010007*, 1-19.

[5] Kivimäki, M. (2014). MaaS-Finland on the leading edge. *In Proceedings of the Mobility as A Service Seminar and Networking Event; Ministry of Transport and Communications.* Ventaa, Finlande.

[6] Tuchen Sarasina. (2018). TuchenRole of Aviation in Seamless End-to-End Mobility. *In Proceedings of the 2018 IEEE/AIAA 37th*, 1-8.

[7] Kersten Heineke, T. H. (2021). Récupéré sur www.mckinsey.com: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/defining-and-seizing-the-mobility-ecosystem-opportunity

[8] Ana Isabel Torre-Bastida, J. D.-C. (2018). *Big Data for transportation and mobility: recent advances, trends and challenges.*

[9] VEDECOM. (2017). *Identification of real world driving scenarios for the functional safety of autonomous vehicles.*

[10] Wagener, N. (2020). WageCommon Methodology for Data-Driven Scenario-Based Safety Assurance in the HEADSTART Project. *Virtual ITS European Congress.*

[11] Khayyam H., J. B. (2020). *Artificial Intelligence and Internet of Things for Autonomous Vehicles.* Nonlinear Approaches in Engineering Applications. Springer, Cham. https://doi.org/10.1007/978-3-030-1896.

[12] Abhishek Gupta, A. A. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues, *https://www.sciencedirect.com/science/article/pii/S2590005621000059*.

[13] Rodrigo Castiñeira, J. N. (2018). AUTOCITS – Regulation study for interoperability in the adoption of autonomous driving in European urban nodes. *Transport Research Arena, Vienna, Austria. ffhal-01898256*.

[14] Ghanishtha Bhatti, H. M. (2021). Towards the future of smart electric vehicles: Digital twin technology,. *Renewable and Sustainable Energy Reviews,.*

[15] Boschert S, R. R. (2016). Digital Twin—The Simulation Aspect. In: Hehenberger P.,Mechatronic Futures. Springer, Cham. https://doi.org/10.1007/978-3-319-32156-1_5.

[16] Wang, W. W. (2021). Application of Digital Twin in Smart Battery Management Systems. Chinese Journal of Mechanical Engineering, https://doi.org/10.1186/s10033-021-00577-0.

[17] WP7.2 (TBD) Process of data analysis and identification of singular situations by BPI france