This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 814761

D3.5
Standard protocol for handling big data

Safe tolerance zone calculation and interventions for driver-vehicle-environment interactions under challenging conditions

iDREAMS

Project identification

Grant Agreement No	814761
Acronym	i-DREAMS
Project Title	Safety tolerance zone calculation and interventions for driver-vehicle-environment interactions under challenging conditions
Start Date	01/05/2019
End-Date	30/04/2022
Project URL	www.idreamsproject.eu

Document summary

Deliverable No.	3.5
Deliverable Title	Standard protocol for handling big data
Work Package	WP3
Contractual due date	30 September, 2020
Actual submission date	30 September, 2020
Nature	Report
Dissemination level	Public
Lead Beneficiary	TUM
Responsible Author	Christelle Al Haddad, Md Rakibul Alam
Contributions from	Kui Yang, Constantinos Antoniou (Technical University of Munich), Rachel Talbot, Ashleigh Filtness, Graham Hancox (Loughborough University), Muhammad Adnan, Yves Vanrompay (University Hasselt), Thomas Stieglitz, Bart de Vos (DSS), Amir Pooyan Afghari (Delft University of Technology), Philipp Blass, Martin Winkelbauer (KFV, Austrian Road Safety Board), André Lourenço (CardioID), Eva Michelaraki, Christos Katrakazas (National Technical University of Athens), Rodrigo Taveira, João Vieira (BARRA), Petros Fortsakis (Oseven).

Please refer to the document as:

Al Haddad, C.; Alam, M. R.; Yang, K.; Antoniou, C.; Talbot, R.; Fitness, A; Hancox, G.; Adnan, M.; Vanrompay, Y; Stieglitz, T.; de Vos, B; Afghari, A. P.; Blass, P.; Winkelbauer, M.; Lourenço, A.; Michelaraki, E.; Katrakazas, C.; Taveira, R.; Vieira, J.; Fortsakis, P. (2020). *Standard protocol for the handling of big data.* Deliverable 3.5 of the EC H2020 project *i*-DREAMS.

Revision history (including peer review & quality control)

Version	Issue date	% Complete	Changes	Contributor(s)
V1.0	11-04-2020	10%	Contents	Kui Yang (TUM)
V1.1	06-07-2020	60%	First draft	See responsible authors and contributions from above
V1.2	22-07-2020	100%	Full draft	As above
V1.3	30-07-2020	100%	Full draft for review	As above
V1.6	30-09-2020	100%	Revised according to internal and external review	As above

Disclaimer

Copyright

Table of contents

List of Figures

List of Tables

Glossary and Acronyms

Word / Abbreviation	Description
API	Application programming interface
CAN	Controller area network
CardioID GW	CardioID gateway
CDC	Central data centre
CF	CompactFlash
DAS	Data acquisition system
DMP	Data management plan
DSS	DriveSimSolutions
DOF	Degrees of freedom
DPO	Data protection officer
EU	European Union
FAIR	Findable, accessible, interoperable and re-usable
FESTA	Field opErational teSt supporT Action
FMS	Fleet Management System
FOT	Field operational tests
GDPR	General data protection regulation
GEOS	Geospatial data
GPRS	General Packet Radio Services
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HDD	Hard disk drive
HTTPS	Hypertext Transfer Protocol Secure
i-DREAMS	Smart- Driver and Road Environment Assessment and Monitoring System
IMU	Inertial measurement unit
JSON	JavaScript Object Notation
LDC	Local data centre
NDS	Naturalistic driving study
NTUA	National Technical University of Athens
OBD II	On-board diagnostics 2nd generation
OBU	On-board unit
OEM	Original equipment manufacturer
RDBMS	Relational database management systems
REST	Representational state transfer

SHRP2	The Strategic Highway Research Program 2
SQL	Structured Query Language
STZ	Safety tolerance zone
UK	United Kingdom
UMTS	Universal Mobile Telecommunications System
VMC	Vehicle management centre

Executive Summary

The *i*-DREAMS project intends to develop a framework for the definition, development, testing and validation of a context-aware safety envelope for driving called the 'Safety Tolerance Zone'. Taking into account driver background factors and real-time risk indicators associated with the driving performance as well as the driver state and driving task complexity indicators, a continuous real-time assessment will be made to monitor and determine if a driver is within acceptable boundaries of safe operation. Moreover, safety-oriented interventions will be developed to inform or warn the driver in real-time as well as on an aggregated level after driving, through an app-and web-based gamified coaching platform (post-trip intervention). Furthermore, a user-license Human Factors database with anonymised data from the simulator and field experiments will be developed.

The conceptual framework of the *i*-DREAMS platform integrates aspects of monitoring (such as context, operator, vehicle, task complexity and coping capacity), to develop a Safety Tolerance Zone for driving. In-vehicle interventions and post-trip interventions will help to maintain the safety tolerance zone as well as provide feedback to the driver. This conceptual framework will be tested in simulator studies and three stages of on-road trials in Belgium, Germany, Greece, Portugal, and the United Kingdom (UK) with a total of 600 participants representing car, bus, truck, rail drivers.

During the experiments, large amounts of data will be generated, from the different data collection tools, originating from different modes and countries. This "Big Data" will inevitably need to follow guidelines that would specify how they could be best handled, and how they would pass from one entity to another while complying with legal and ethical regulations set out at a national, and EU level; these would be compliant with the GDPR regulations (2016/679), aiming to protect personal data, and ensuring that the proper framework is set out in case of agreement infringement.

The aim of this deliverable is to therefore provide the necessary protocols for handling the generated big data, passing through the different steps from data collection, to data storage. Looking at the different collection phases, a specific indication would be given on special considerations that would be needed for other modes, where available.

The specific objectives of this deliverable are therefore:

- Provide a methodology for the handling of big data, based on learnings from previous studies/projects; particularly naturalistic driving studies (NDS) studies in Europe
- Provide standard protocols for the handling of big data, informing i-DREAMS experiments on the procedures to be followed to best handle collected data, while complying with the necessary regulations and ethical considerations

The standard protocols are to be continuously controlled throughout the i-DREAMS project, and would be in accordance to the data management plan (DMP), and the data and knowledge management committee; partners at different countries would be responsible for their own data collection, and obliged to follow the proper standards, while consulting with their national and local authorities.

The protocols will checked for feasibility and maintainability and at the end of the project, a report will be drawn to identify issues that have or could have been improved, to serve as guidelines for future projects/research involving similar data collection.

1 Introduction

1.1 The i-DREAMS project

The overall objective of the *i*-DREAMS project is to setup a framework for the definition, development, testing and validation of a context-aware safety envelope for driving ('Safety Tolerance Zone'), within a smart Driver, Vehicle & Environment Assessment and Monitoring System (*i*-DREAMS). Taking into account driver background factors and real-time risk indicators associated with the driving performance as well as the driver state and driving task complexity indicators, a continuous real-time assessment will be made to monitor and determine if a driver is within acceptable boundaries of safe operation. Moreover, safety-oriented interventions will be developed to inform or warn the driver real-time in an effective way as well as on an aggregated level after driving through an app- and web-based gamified coaching platform. Figure 1 summarises the conceptual framework, which will be tested in a simulator study and three stages of on-road trials in Belgium, Germany, Greece, Portugal and the United Kingdom (UK) on a total of 600 participants representing car, bus, truck and rail drivers, respectively. Specifically, the Safety Tolerance Zone (STZ) is subdivided in three phases, i.e. 'Normal driving phase', the 'Danger phase', and the 'Avoidable accident phase'. For the real-time determination of this STZ, the monitoring module in the *i*-DREAMS platform will continuously register and process data for all the variables related to the context and to the vehicle. Regarding the operator however, continuous data registration and processing will be limited to mental state and behaviour. Finally, it is worth mentioning that data related to operator competence, personality, socio-demographic background, and health status, will be collected via survey questionnaires.

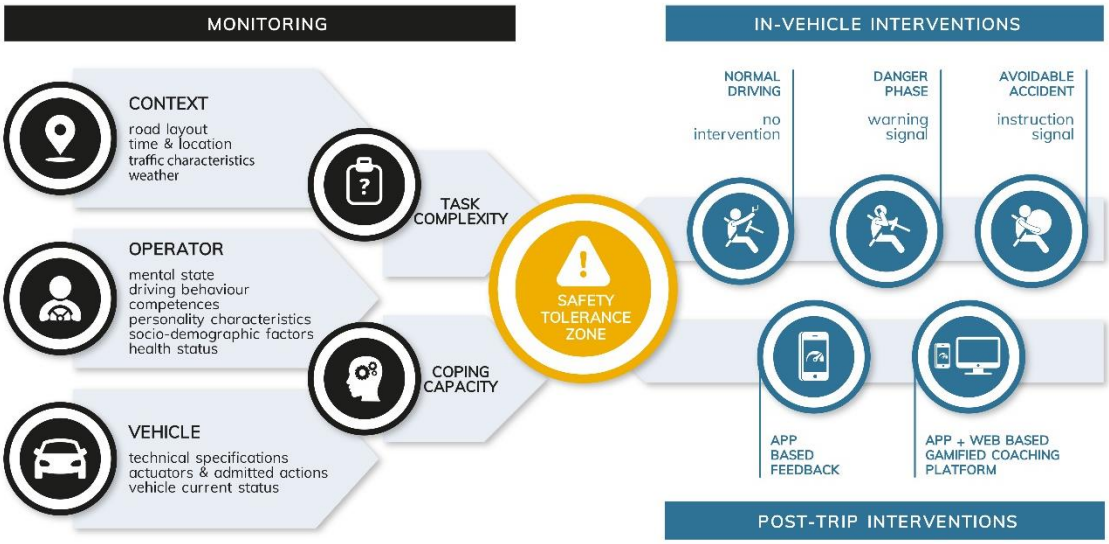

Figure 1: Conceptual framework of the i-DREAMS platform

The key output of the project will be an integrated set of monitoring and communication tools for intervention and support, including in-vehicle assistance and feedback and notification tools as well as a gamified platform for self-determined goal setting working with incentive schemes, training and community building tools. Finally, a user-license Human Factors database with anonymised data from the simulator and field experiments will be developed.[1]

[2] Further general project information can be found on the website: https://idreamsproject.eu

1.2 Deliverable

The applied nature of the project will result in large data sets. This deliverable is concerned with the logistics of handling such data and the process for making it useable. Standard i-DREAMS procedures will be written to meet the legal and ethical requirements of collecting, handling, and storing such data.

1.2.1 Aims and objectives

- Develop Standard *i-Dreams* protocols for the project to meet the legal and ethical requirements of collecting, storing, and handling data (data formats, standards, and communication protocols). These would cover all data handling processes from in-vehicle data collection to a data storage for future analysis.
- Map out the data flow between different partners and servers and highlight the arising legal and ethical implications.
- Link to the technology development work package to ensure that the standards set out in this deliverable are feasible to implement.

1.2.2 Structure

The deliverable is structured as follows. In the next section, a literature review is presented, based on the review of previous naturalistic driving studies (NDS) in Europe, focusing on lessons learned on big data handling, storage, and ethical and legal considertaions. This section ends with a gap identification, and best practices descriptions and implications for i-DREAMS. In the third section, methods are proposed for big data management in NDS, comprising data collection, processing, storage, and access. After that, the project data is presented in detail for the data collected in both simulator and on-road trials across different modes and countries. Finally, standard protocols are given, highlighting the steps needed for the proper handling of data in the context of i-DREAMS.

2 Literature Review

Driving simulator experiments and naturalistic driving studies have recently increased in popularity, due to technological advances allowing the unobtrusive measurements of drivers' behaviour. In this process, several components are crucial for the handling of data. The FESTA handbook defines the guidelines for data acquisition, including storage and analysis tools (Section 7), emphasizing the importance of laws and regulations in such protocols (FESTA Handbook, Version 7, 2018). Defined data upload schemes include picking up the data (by taking a storage device out of the vehicle and physically sending it to a database), data transmission via wired, cellular, or WIFI connection. Based on these recommendations, the following subsections present previous NDS projects conducted in Europe, by looking at aspects of data handling (collection, processing), storage, and their corresponding legal and ethical concerns. A summary of findings is then given with a special focus on implications for the i-DREAMS project.

The selected projects to be discussed are: AOS[2] (2007-2009), SeMiFOT[3] (2008-2009), euroFOT[4] (2008-2011), TeleFOT[5] (2008-2012), INTERACTION[6] (2008-2012), 2BeSAFE[7] (2009-2011), PROLOGUE[8] (2009-2011), UDRIVE[9] (2012-2016), Track & Know[10] (2019-Ongoing). For most projects, insights were provided by the authors or colleagues who previously worked in one of the projects (no associated references in the literature).

2.1 Big data collection

AOS: A Dutch project which investigated for 8 months trucks (2400 trucks) and collected over a total of around 77 million kilometres of driving. On-board units collected data and performed an initial data reduction. Warning systems were used for the following: Forward Collision Warning/Headway, Monitoring and Warning Systems, Lane Departure Warning Assist, Adaptive Cruise Control, Directional Control, Black Box Feedback System, Data loggers. Quantitative data collected included distance to vehicles in front, accelerations, decelerations, intensity of braking, rollover risks, line crossings, speed, average speed, speed variation, use of cruise control, headway time, distance to lane marking, warning events, number of incidents, total km driven, average km per day. Qualitative data included questionnaires and interviews with truck drivers. Data was transferred wirelessly to transmit data to the data centres.

SeMiFOT: This project included partners in Sweden and the US. Data was collected for cars, and trucks. Equipment used included GPS (1 Hz), accelerometer (100 Hz), Mobileye (image processing, distance, position of other road users, traffic sign recognition, etc), CAN access, eye-tracking (at 50/60 Hz), 7 Video channels (MPEG400GPL). Data collected included 2944 hours of vehicle data. In addition to quantitative data, personal data was collected: socio-

[2] Information can be found under: http://wiki.fot-net.eu/index.php/AOS
[3] Final project report available under: https://document.chalmers.se/download?docid=1773834060
[4] Information on the project can be found under: www.eurofot-ip.eu
[5] Information can be found under: http://wiki.fot-net.eu/index.php/TeleFOT
[6] Information can be found under: http://wiki.fot-net.eu/index.php/INTERACTION
[7] Final deliverable reports are given under: https://cordis.europa.eu/project/id/218703/reporting/it
More information can be found on the project website under: http://www.2besafe.eu/
[8] Project deliverables can be found under: https://prologue.kfv.at/prologue/deliverables/
[9] Final summary report is available under: https://cordis.europa.eu/docs/results/314/314050/final1-udrive-final-publishable-summary-report.pdf
[10] More information is available on the project website under: https://trackandknowproject.eu/

demographics, permanent or temporary driver impairments, driving experience, attitudes, accident history and the usual time of driving and roads used.

euroFOT: In this project, 28 partners across Europe collaborated including: vehicle manufacturers, automotive suppliers, universities, and research centres. The field operational trial (FOT) was organised by four Vehicle Management Centres (VMC) across Europe: in France, Germany, Italy, and Sweden. This project investigated cars (Over 600 cars from French, German Italian and Swedish VMCs) and Trucks (Over 130 from German and Swedish VMCs). The FOT equipment included: CAN data, sensors (eye and head tracker, GPS, accelerometer and yaw-rate sensor, radar/headway position sensing, lateral positioning sensing), video data (video of driver's face, forward view, driver's feet and cabin view). Collected data was mainly quantitative from the vehicle devices and sensors. Sensitive data included geolocation GPS data, video data containing driver's face, but also demographics (age, gender) from the questionnaire data.

TeleFOT: Data was collected through CAN data/ data logger, travel diaries, for cars. Collected data included: duration of journey, time of journey, speed, TTC, DTC, date of journey. If there was more than one participating driver in each vehicle, the travel diary will be consulted to see who did which trip. Travel diary included the mode of travel selected, type of driving environment (highway, rural, urban). The project had several test sites across Europe but each one effectively ran their own unique trials in silo, testing different nomadic devices. This made data collection and processing easier as most of it was carried locally without the need to collect all data on one big server but did limit the participant numbers/ scalability as it was not possible to directly compare data across sites. The need to collect data from an SD card in the satellite navigation device for UK trials was said to lead to a spike in data after this change over as participants were reminded the device was in the car and therefore turned it on more, which was required for data to start recording. This spike in activity was advantageous as it allowed for more data to be generated than if the device turns on automatically (as in i-Dreams) but could also be considered disadvantageous as the prompt of manually operating the device may also change driving behaviour as participants are reminded they are being studied. It is therefore recommended that participants do not need to physically turn on the device as this may alter true driving behaviour, if possible, collecting data automatically has both logistical and experimental advantages.

INTERACTION: The study included 30 vehicles comprising thirteen different specific vehicle models, all of which were passenger cars (rentals), and included countries in Europe (France, Czech Republic, Spain, Austria, Portugal, the Netherlands, the United Kingdom, Finland, Australia). The DAS used an array of different monitoring devices including cameras, GPS, accelerometers, infra-red sensors, GSM sensors and pressure sensors to monitor driver activity. Data collected also included questionnaire data.

Each partner was responsible of their own data collection, coding, and analysis. Tools to be distributed and optimised to work on computers used in each partner organisation. Video coding by staff trained in each partner organisation, which might lead to inconsistency.

Findings from this project:

- Data differs for different vehicles and the respective sensor setup.
- Recommended to centralise responsibilities in terms of coding, processing, and analysis to create a consistent dataset.

2BeSAFE: The study included 27 partners from 11 countries, i.e. France, UK, Austria, Greece, Italy, Spain, Portugal, Israel, Australia, Belgium, Germany (European Commission, 2017), and investigated Powered Two-Wheeler (PTW) rider behaviour and the interaction between PTWs and other road users (Espié, Bekiaris, & Nikolaou, 2010) (Espié, Boubezoul, Aupetit, & Bouaziz, 2013) (Barnard, Utesch, van Nes, Eenink, & Baumann, 2016). Collected data included motorcyclists' profile questionnaire translated in 6 languages (Baldanzini, et al., 2009). PTWs (different across countries) were fitted with special equipment and sensors in the UK, France, Italy, and Greece. The equipment used for the experiment included a variety of different devices and sensors, such as two data-loggers, one GPS, three front cameras, one camera recording the rider's face, one 3-axes accelerometer/ gyroscope, one steering wheel position sensor, a throttle position sensor, blinker and brake contacts sensors and two wheel displacement sensors (European Commission, 2017). Miniature sensors recorded throttle and accelerometer position, handlebar rotation, operation of hand and foot brake levers, foot peg pressure and turn signal operation. In addition, a gyroscope and GPS equipment recorded position and orientation, and a video camera recorded the visual context. Innovative tools were designed and included: six instrumented PTWs, an instrumented car, two riding simulators, a driving simulator, and a video-based tool for investigating motorcyclists' risk awareness (Vlahogianni, Yannis, & Golias, 2014). All instruments were installed discretely. Quantitative data that was available also included additional road safety data, and contextual data (road infrastructure, etc.).

PROLOGUE: This project included partners from the Netherlands, Greece, Israel, Norway, Spain, the United Kingdom, and Austria. The DAS technology was not specifically designed for the project; rather off-the-shelf technologies were used, and different partners used different DAS units. Among these were the "pdrive"® (recording 4 continuous video channels, g-forces, speed and GPS, voice recording as optional), a combination of it with two stationary ("site-based") cameras , and a highly instrumented experimental vehicle ("ARGOS") that included 7 video channels and eye-tracking and several environmental conditions (climate, noise). Data collected included vehicle signals, dynamics, GPS, and lane position and time headway.

The off-the-shelf systems often had their own software and proprietary data and video format (like pdrive), which however remain confidential.

The pdrive system was used at different resolution between 10 and 100 Hz. It compiled four channels to one VGA stream. If GPS was used, it was limited to 1 Hz in all DAS units. The off-the-shelf DAS units were all not particular designed for scientific purposes, but for driver training, for observation of novice drivers or as driver assistance systems. The experimental vehicles used also were already there before the study and designed to capture comprehensive information. The site-based observation only used video streams. Sensitive data (personal data) was hardly collected (van Schagen, et al., 2011) (Sagberg, et al., 2011).

Findings from this project:

- A European large-scale experiment would have to use a common DAS instead of using different DAS units creating data according to a common standard.
- A basic and relatively simple and cheap off-the-shelf G-based DAS provides very useful data for many research questions.
- Attention must be paid to the reliability and validity of the identified safety-related events, i.e. avoiding false alarms, and missed events.

UDRIVE: Data was collected at seven operation sites for cars, trucks, and scooters. This data was pre-processed at three local data centres (LDC). During the pre-processing, the data was enriched with e.g. map data (i.e., the addition of road and infrastructure attributes from digital

maps). A smart camera measured the distance to potential crash objects and pedal positions, vehicle speed, longitudinal and lateral accelerometers will be used to identify driver reactions such as avoidance manoeuvres. Positioning data was used to determine road type and other environmental variables. For cars, 7 cameras, inertial measurement unit (IMU) sensors; GPS; Mobileye smart camera, CAN data, and sound level were used. Trucks had the same equipment with 8 cameras, while scooters just had 5 cameras, IMU sensors and GPS. Measured data was mostly quantitative, from sensors. Additional data was enabled by video cameras like number of passengers, seatbelt use, traffic conditions, temporal factors (day/night, rush hours). The positioning data was used to identify the environment road type (single carriageway, motorway), road geometry (intersection, link, bend, junction), locality (urban, rural). Sensitive data including questionnaire data such as driver characteristics such as age, gender, driving experience, sensation seeking and locus of control, other attitude assessments, but also crash and traffic violations history, and video-based hazard perception test. Contact details of participants, and their home and work addresses were also collected (Eenink, Barnard, Baumann, Augros, & Utesch, 2014)(van Nes, Bärgman, Christoph, & van Schagen, 2019).

Track & Know: The project aims to investigate cars, trucks, and buses and partners cover countries in the UK, Belgium, Greece, Italy, Luxembourg, Ireland, Switzerland, France, Germany. Data was obtained from Systemetica Inc ®, based on sensors installed in the vehicles (mainly GPS position data of vehicles and other risk variable as events like harsh braking, harsh cornering), Vodafone Innovus ® sensors installed in the fleets (this is mainly GPS position data of vehicles), Oximeter data for sleep apnoea patients, some other contextual data are also collected such as road network information, weather information and point of interests etc. Collected data were mostly quantitative and included: GPS- based spatio-temporal positional data of vehicles, real time streams and historical data (for the past year).
A summary of data collection practices in the above projects is summarised in Table 1.

Table 1: Summary of data collection practices in previous EU-NDS projects

	Project name	AOS	SeMiFOT	euroFOT	TeleFOT	INTERACTION	2BeSafe	PROLOGUE	UDRIVE	Track & Know
Transport mode	Passenger car		X	X	X	X	X	X	X	X
	Truck	X	X	X					X	X
	Motorcycle						X			
	Other						X		X	X
Data scale	Sample size/ sampling frequency	Over 77 M kilometres mileage in 8 months	39 drivers, 3K hrs of vehicle data, 170K km mileage	More than 25TB data over a period of one year		92 drivers, 3K hours of driving data, 138K km mileage		Resolution between 10 Hz and 100 Hz		More than 1TB for the previous 1 year
Data collection tools	CAN		X	X	X				X	
	Sensors		X	X		X	X		X	X
	GPS		X	X		X	X	X		X
	Accelerometer		X	X		X	X	X	X	
	Video camera		X	X		X	X	X	X	
	Other	Data logger			Data logger, travel diaries	Infra-red	Data logger	Voice recorder	Sound recorder	Oximeter for Sleep Apnoea Patients
Collected data	Quantitative	X		X	X		X		X	X
	Qualitative	X							X	
	Video data			X				X		
Sensitive / Personal information	Image/video			X					X	
	Name / address / contact			X					X	
	Demographics		X	X					X	
	Attitudes and psychological characteristics		X							
	Other		Permanent or temporary driver impairment				Stakeholder information			Positioning data of individuals

2.2 Big data storage

Naturalistic driving studies usually require a double storage system comprising a local, on-vehicle medium, and a central database, which collects data from all participants. Regarding on-vehicle storage, some popular means are hard drives mounted on an on-board DAS (European Commission, 2017) (Dozza, Bärgmann, & Lee, 2013) (Wu, Aguero-Valverde, & Jovanis, 2014), or flash storages such as SD cards plugged into the on-board diagnostics interface (OBS II) or on an on-board computer (Fridman, et al., 2019) (Knoefel, Wallace, Goubran, & Marshall, 2018) (Wallace, et al., 2015). Transfer of data to central servers can be done manually, for example in the Candrive study participants met regularly with the study team in order to move data into the central database and empty the on-board storage media (Wallace, et al., 2015), or through wireless networks such as Bluetooth, WLAN or cell network transmission.

The selection of storage architecture for the central database can have considerable implications for the research project. Key features such as processing speed, file servers that deliver data to researcher computers, and links to database servers for hosting large datasets are critical components. Reasonable quality video data rates are 6-8 megabytes of video per minute, resulting in a total of 20 gigabytes of data per vehicle per month. The video data typically comprise 80-95% of the total data collection compared to 5-15% for the vehicles naturalistic data (Klauer, Perez, & McClafferty, 2011). (Tselentis, 2018) mentions that data can be stored in a central database, where they are managed and processed for example by calculating indicators based on machine learning and data mining techniques such as filtering and aggregation using complex python scripts.

The Structured Query Language (SQL) has proven a popular programming language for developing central databases. PostgreSQL is an open-source relational database management system, which has been used very often in naturalistic driving studies (Fridman, et al., 2019) (Bender, et al., 2016) (Dozza, Bärgmann, & Lee, 2013). According to (van Schagen, et al., 2011) databases should support SQL programming, as it has been proven by the large-scale SHRP2 NDS project in the US, which resulted in 1 petabyte of data. Databases have also been developed based on MATLAB (Machiani & Abbas, 2016) (European Commission, 2017) and stored in a secure Amazon S3 folder (Babulal, et al., 2016). Finally, (Oussous, Benjelloun, Lahcen, & Belfkih, 2018) supported Apache Hadoop technology, which can store, process and analyse large volumes of data. Hadoop executes tasks where data are stored i.e. it does not copy in memory the whole distant data to execute computations, which ensures better processing speed.

All in all, different studies suggest a multitude of different ways of storing and structuring data within databases. A common observation is that studies including videos require vast amounts of storage space, according to video duration and quality. Another important observation is that data should be stored in a way that enables easy access of different participants and researchers and allows for resource-efficient processing.

In the below sub-section, a summary of data storage schemes for the previously mentioned EU projects are detailed.

SeMiFOT: Hard drives (HDDs) were chosen as the primary media for data storage in the vehicles. The data was stored locally in the vehicle, while summary/status information was uploaded remotely via wireless 3G/GPRS. When the status information was uploaded, it was transferred into a database and displayed in a web interface for quality and status checks (for example hard drive space left and data source up-time). After the monitoring system had sent an overflow warning for the in-vehicle HDD (or there was some other problem with the data), the hard drives were switched manually by OEM personnel. The OEM then separated

proprietary and project data and copied this to a transfer disk. When some disks had been copied to the transfer disk, it was moved to SAFER for data upload. A custom developed software was then used for uploading and synchronizing the data into the main database. Video was stored separately, with a possibility to extract individual frames on a per time stamp basis (synchronised with the database data). The analysts could then access both the database and the video data synchronously, either via the analysis viewer or directly from MATLAB. Security and access procedures were implemented according to what was decided in the Consortium Agreement. As for the installed data acquisition units, different vehicles had different hardware configurations. An Oracle database was chosen together with the University of Michigan Transportation Research Institute (UMTRI) inspired database model of 10Hz sample frequency, but with a few separate tables with high frequency data. At the very end of data handling map data attributes were added based on map matching of GPS positions.

Data has been stored in its genuine resolution. For some data, multiple resampling processes were performed to harmonise resolution of different types of data for different purposes.

The central database was arranged as a part of the Chalmers University computer system (as it was done in UDRIVE es well). They did some data enrichment int terms of GPS and map data. GPS was exported, matched with the (public available) geographical information of the Swedish road system. This included geometrical data as well as legal information, e.g. speed limits. The Swedish administration publishes an updated digital map every two week, which is fully available on the internet for free. SAFER used an Oracle database and had a graphical user interface and processing based on MATLAB.

For SeMiFOT the working solution is to grade data access depending on both data type and company/institute. They used Oracle built in authentication and access to data is determined by roles.

SeMiFOT had the problem that data had to be protected even among partners, since some CAN data should not be disclosed to other partners. This problem was solved by XML storage (configuration files were not kept in a central data storage).

euroFOT: Data was stored in oracle and file servers at each VMC: more than 25TB data have been stored across all VMCs over one year. Data from the vehicles were obtained via transfer hard disk. A base software was developed to enable users to browse and select data from a database. It allowed users to visualise and present per trip data in a user-friendly way, by means of graph, table view and/or video. Furthermore, it allowed users who had access to insert new data to the database. The software modules work with time-based signals (CAN data, GPS, other external sensors, and video), so they display and work on continuously recorded sequences of data. Data can be accessed by using SQL queries.

TeleFOT: Video data was stored as files in file management systems. The non-video data were generally stored in relational databases; Oracle, MySQL, and MS Access were used. For some partners, the non-video data were also stored as MATLAB files in the file management systems, to allow analysis through database and/or direct with MATLAB. Raw trips were converted to a standard readable format (MATLAB file) initially only containing unsynchronised time-history signals. The signals were then synchronised and harmonised according to a uniform data model. The trips were then enriched with aggregated data, by applying user-defined scripts. Trips data from the MATLAB files were then stored in relational database. The analysis was performed by directly querying aggregated data and its description from the database using standard statistical tools, such as SAS (Statistical Analysis Software, SAS/STAT 2011).

INTERACTION: On-board Micro-SD interface for data storage centre console/storage. The Enterprise Application Platform's (EAP) storage was separated into a dedicated EAP storage and "Sensor Data" storage. EAP storage managed user data and access rights (read, modify, delete), as well as application configuration and project reference data. It is implemented in a traditional RDBMS database. Sensor data storage on the other hand has specific requirements of big data that RDBMS either cannot support or supports them with the drawback of performance. For Sensor data implementation it is considered the use of a distributed file system (Hadoop) for offline processing and a data streaming framework for real-time data stream processing (Apache Strom).

2BeSAFE: Storage included subjective data from participant questionnaires, logbooks, and interviews. These were converted to electronic format by partner institutions for transfer to the communal server. The original paper copies of these materials were stored by the relevant partner institution in a secure cabinet, to which only the project leader has access. In order to ensure that electronic data was kept securely and under the same conditions, data from all the partners were transferred to a single server using a File Transfer Protocol (FTP) to be set up and maintained by Transport Research Laboratory (TRL). The survey was developed in LimeSurvey, an-open source online survey application. Designed to be user-friendly, it enabled users to develop and publish surveys, and collect responses, without having to write any coding.

PROLOGUE: All operations were organised at local level, there was no common storage, nor compatible data. For most of the application, the proprietary software tools of the systems were used to execute these tasks, as far as they were executed at all. Data storage policies were not set up in PROLOGUE regarding the limited size of the experiment. For the experimental vehicles, their usual system of storing data was used; pdrive data was transferred from the CF-cards to password-protected file server folders, protected only by the IDs and passwords of the scientific staff and data protection training they had to go through and agreements they had to sign.

UDRIVE: The pre-processed data was to be stored at a central data centre (CDC). The partners had remote access to the data at the CDC. Analysis and annotation were done directly on the central data set at the CDC through the remote access. The project required physically transferring data using hard drives, nothing was sent over the air and this introduced many issues. The process required scanning QR codes on the vehicle, data logger and hard drive to try and ensure data was correctly matched but this allowed room for error to occur. The hard drives were sent from project partners across Europe to be collated and processed on a central database in France so there was a risk of data being lost in shipping. Currently though, with existing cloud storage methods, the faster, more accurate, less risky means of collecting, storing, and transferring data should be used. The data was collated at seven operation sites. This data was pre-processed at three LDCs. During the pre-processing, the data was enriched with e.g. map data (i.e., the addition of road and infrastructure attributes from digital maps). The pre-processed data was to be stored at a CDC. To facilitate processing, annotation and analysis of the data, a common software toolset was developed within the project. For all trips, the driver ID will be confirmed based on the first video frames. In this step the trip is matched to a driver, as more than one driver per vehicle can participate in the study.

Track & Know: Kafka clusters-based technology was deployed to handle data, data were stored in servers using Mongo DB, HBase databases. The servers can store and process data

of about 1 TB of data. Kafka connectors are developed to communicate between the data storage and processing components within the Track & Know platform.

MongoDB as scalable NoSQL storage system is used for storing data and integrated within Kafka Apache Cluster platform

Kafka Apache Cluster -based communication and storage platform is used to communicate and further processing of the data

A variety of big data processing tools were developed; e.g, NoDA, sub-trajectory joining operations, data enrichment pipeline (having cleansing, map matching, weather info, & POI

Various analytics are developed on the data, e.g. Machine learning based accident risk prediction, Individual mobility network clustering and its use in improving accident risk prediction, future location prediction model, sub-trajectory clustering and its processing to facilitate various mobility applications such as Carpooling.

Historical GPS data is stored in local servers and machines (mostly by the partner who owned the data). Kafka Cluster is used for real-time stream data and when a certain process is required to be done with historical data.

A summary of data collection practices in the above projects is summarised in Table 2.

Table 2: Summary of data storage practices in previous EU-NDS projects.

	Project name	AOS	SeMiFOT	euroFOT	TeleFOT	INTERACTION
Server	Local server			X		
	Central server		X			X
	In-vehicle hard drive		X			X
Data query	Tools / protocols / software packages		Customised developed software	Oracle, MySQL, MS Access, MATLAB		Hadoop
Data handling	Pre-processing	X		X	X	
	Data enrichment		X	X		
	Synchronisation / indentification					

	Project name	2BeSafe	PROLOGUE	UDRIVE	Track & Know
Server	Local server	X		X	X
	Central server	X		X	
	In-vehicle hard drive		X		
Data query	Tools / protocols / software packages	FTP			Kafka Cluster, MongoDB, Kafka Apache
Data handling	Pre-processing	X	X	X	X
	Data enrichment			X	X
	Synchronisation / indentification			X	

2.3 Legal and ethical considerations

Besides data collection, storage, and access, legal and ethical considerations are crucial for successful NDS experiments, and for their viability in the first place. This section will present practices that were conducted in previous projects, focusing on data sharing, data protection, data maintenance after the project lifetime, along with ethical and legal considerations, where applicable.

euroFOT: Video, geolocation, and questionnaires data, as collected in an FOT with real drivers, represent personal data, and "personal data" was subject to European Directives (e.g. D95/46/EC [2]) as well as national laws. In the contract with the driver, though, each FOT-centre had stipulated that personal data will not be disclosed without permission of the driver. Instead of allowing limited access to data for public, the project followed a different path. The access was to be granted based on submission of research proposal that will be reviewed by owners of data. The research proposal needed to be submitted to the Management of euroFOT. Data protection and maintenance was the responsibility of VMC.

TeleFOT: Data stored in lockable filing systems – no personal data stored on database. Participant identification shredded shortly after use.

2BESAFE: To protect participant anonymity, partners were required to pseudonymise/ anonymise data files (using some specific codes) before transferring the files to the communal server. Partners therefore remained responsible for managing participant confidentiality and for retaining the ability to identify participants if required and only the project leader had access to the relevant information. The survey was developed with the help of LimeSurvey, an open-source online survey application. Designed to be user-friendly, it enabled users to develop and publish surveys, and collect responses, without having to write any coding.

The survey was anonymous; thus, the records of the survey responses did not contain any identifying information about the participants. The only “personal” information asked was the opening question (“Which groups of stakeholders do you belong to?”). Ethical and legal considerations for naturalistic riding experiments were particularly challenging.

All personal data were destroyed once it no longer became necessary for analysis, if this was defined; and if not, the data were destroyed at the end of the study period at the latest. Non-personal data were kept beyond the period of the study if it is felt that it was of value to partner institutions in further work and if it was agreed that it could be held. In such case the data were made available to all partners, who shared equal intellectual property rights, regardless of the nationality of the source data.

UDRIVE: It was an ethical requirement that the first and last minute or so of each journey was deleted so participants could not be identified by data alone. This procedure implied some data loss yet ensured that addresses such as home were not shared. Legal, ethical, and logistical issues could also arise if participants drive their equipped car to another country for holiday etc. Due to ethical reasons, this was reported to be an issue in UDRIVE and let to the loss of big portions of the data of such trips due to complications with the equipment when crossing borders. Finally, if the driver in question was not a participant, but maybe someone that drives this vehicle incidentally, the trip was deleted to protect the privacy of non-participants.

The data collected within the project will be available for further research after the project for partners and for third parties, with certain limitations to adhere to the privacy of the participants as agreed in the informed consent forms.

Track & Know: All project activities conformed to International, European, and national laws. Particular attention is being paid to EU Regulation 2016/679, or the General Data Protection Regulation, which came in effect from 25 May 2018. Moreover, the project had its own Ethical committee (based on external experts) that looked after the ethical issues in relation to preparation of deliverables. Each deliverable also contained a proforma to be submitted and reviewed continuously during the process of deliverable activities. Finally, pseudonymisation operations were in place (such as tokenisation or data encryption) or complete anonymisation where possible.

A Data Protection Officer for the project ensured that data collection and processing within the scope of the project, was carried out according to EU and national legislation. As the project is still ongoing, it is not clear what data would be available to public/open access.

Table 3 Summary of legal and ethical considerations in previous EU-NDS projectss

	Project name	AOS	SeMiFOT	euroFOT	TeleFOT	INTERACTION
Legal considerations	Legal protocols		X			
Ethical considerations	Anonymisation					
	Permission for disclosure		X	X		
Data protection considerations	Access restriction		X	X		
	Deletion of data				X	

	Project name	2BeSafe	PROLOGUE	UDRIVE	Track & Know
Legal considerations	Legal protocols				
Ethical considerations	Anonymisation			X	X
	Permission for disclosure				X
Data protection considerations	Access restriction			X	X
	Deletion of data	X			

2.4 Summary from previous EU-NDS projects

The state-of-the-practice in big data handling within several European naturalistic driving projects was reviewed in this section including AOS (2007-2009), SeMiFOT (2008-2009), euroFOT (2008-2011), TeleFOT (2008-2012), INTERACTION (2008-2012), 2BeSAFE (2009-2011), PROLOGUE (2009-2011), DaCoTa (2010-2012), UDRIVE (2012-2016), Track & Know (2019-Ongoing). While the focus in most of these projects has been on passenger cars and trucks, other transport modes such as motorcycles and scooters have also been investigated in some cases. The data collected in these projects comprise of both quantitative and qualitative data. Big data handling in these projects has been reviewed from three main perspectives: data collection, data storage, and legal and ethical issues related to big data. A summary of this review is presented in Table 4. It is important to note that the analyzed projects were constrained by the technology that was available at the time. This might mean that in the i-DREAMS project, other approaches might be used that were not accessible at the time.

Finally, it is worth mentioning that the EU has recently issues new guidelines for the processing of personal data in the context of connected vehicles and mobility-related applications[11].

[11] These guidelines can be found under: https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_202001_connectedvehicles.pdf

Table 4: Summary of the state of big data handling practice in previous EU-NDS projects[12]

		Project name	AOS (2007-2009)	SeMiFOT (2008-2009)	euroFOT (2008-2011)	TeleFOT (2008-2012)	INTERACTION (2008-2012)	2BeSafe (2009-2011)	PROLOGUE (2009-2011)	UDRIVE (2012-2016)	Track & Know (2019-Ongoing)
Data collection	Transport mode	Passenger car		X	X	X	X	X	X	X	X
		Truck	X	X	X					X	X
		Motorcycle						X			
		Other						X		X	X
	Data scale	Sample size / sampling frequency	Over 77M kilometres mileage in 8 months	39 drivers, 3K hrs of vehicle data, 170K km mileage	More than 25TB data over a period of one year		92 drivers, 3K hours of driving data, 138K km mileage		Resolution between 10 Hz and 100 Hz		More than 1TB for the previous 1 year
	Data collection tools	CAN		X	X	X				X	
		Sensors		X	X		X	X		X	X
		GPS		X	X		X	X	X		X
		Accelerometer		X	X		X	X	X	X	
		Video camera		X	X		X	X	X	X	
		Other	Data logger			Data logger, travel diaries	Infra-red	Data logger	Voice recorder	Sound recorder	Oximeter for Sleep Apnoea Patients
	Collected data	Quantitative	X		X	X		X		X	X
		Qualitative	X							X	
		Video data			X				X		
	Sensitive / Personal information	Image/video			X					X	
		Name / address / contact			X					X	
		Demographics		X	X					X	
		Attitudes and psychological characteristics		X							
		Others		Permanent or temporary driver impairment				Stakeholder information			Positioning data of individuals
Data storage	Server	Local server			X			X		X	X
		Central serve		X			X	X		X	
		In-vehicle hard drive		X			X		X		
	Data query	Tools / protocols / software packages		Customised developed software	Oracle, MySQL, MS Access, MATLAB		Hadoop	FTP			Kafka Cluster, MongoDB, Kafka Apache
	Data handling	Pre-processing	X		X	X		X	X	X	X
		Data enrichment		X	X					X	X
		Synchronisation/identification								X	
Legal and Ethical Requirements / Policies	Legal considerations	Legal protocols		X							
	Ethical considerations	Anonymisation								X	X
		Permission for disclosure		X	X						X
	Data protection considerations	Access restriction		X	X					X	X
		Deletion of data				X		X			

[12] Provided details are only summaries of the obtained and publicly available references. The authors are not responsible for discrepancies resulting from non-published protocols or results.

2.5 Implications for i-DREAMS

This section focuses on the key implications from the previously reviewed practices of big data handling within EU-ND projects and aims to outline some practices could be applied within the i-DREAMS project in the light of the existing gaps. These implications are presented in the same order as before: data collection, data storage and legal and ethical considerations.

Data collection:

- A basic, relatively simple, and cheap off-the-shelf G-based data acquisition system (e.g., accelerometer) provides very useful data for many research questions. However, the reliability and validity of the identified safety-related events (e.g., false alarms, missed events) must be carefully checked.
- For the projects that are implemented in multiple European countries, data should be collected using a common data acquisition system for all partners and following the same protocols and standards.
- As the collected data differs for different vehicle models and the respective sensor setup, the number of vehicle models to be employed for data collection should be minimsed. This also reduces the burden of installation and de-installation of sensors in each vehicle.
- In general, it is recommended to centralise the responsibilities in terms of coding, processing, and analysis to create a consistent dataset.

Data storage:

- The collected data should be pre-processed prior to storage. Advanced video processing techniques increase the efficiency of the data pre-processing task since the video data comprise most of the total data within naturalistic driving studies.
- The collected data may be enriched by external data sources prior to storage. Digital maps, roadway engineering attributes, traffic characteristics, climatic data and questionnaires are useful external data sources that could enrich the collected data.
- Ease of access to data is a more important aspect than storage capacity within the data storage phase. The data should be stored in open formats so that all partners can have access to it. Moreover, the data should be well-structured for analysis to prevent complexities for different partners.
- Partners who are involved in data collection should adopt appropriate privacy protocols so that the ethical considerations are not violated.
- It is necessary to set up a systematic back-up scheme to prevent loss of data. The Festa Handbook (2018) postulates: *"a backup strategy should be based on "acceptable downtime". Off-site backups are mandatory for managing a disaster scenario. The majority of the data is never edited (video and raw data in the database) and data mirroring should be sufficient. The backup policy must be based on the time it takes to recover data and the acceptable loss of data. Even though some studies may use the original logger data as backup, any private or published data created afterwards must have valid continuous backups."*
- The data should be well defined. As there are usually many partners involved in European projects, the data collected and stored by each partner (either in a unique data server or in multiple serves) should be understandable by all partners. This is in general achieved through a clear Data Management Plan, which defines datasets, variables within each dataset, and who generates each dataset.
- The video files should be stored separately but linked with the rest of the data so that the videos can be retrieved in synchronisation with the database. To do so, the video

data should be stored as files in file management systems while the non-video data should be stored in relational databases.

- Transferring the data from the local devices (e.g., hard drives, USB drivers, or SD cards) to central servers should be done electronically for example using wireless networks, Bluetooth, WLAN or cell network transmission. Manual transmission should be minimal because it imposes extra burden on participants resulting in reluctance to participate and data loss.
- For those files that still need manual extraction and\or transfer (e.g., data coming from questionnaires), the hard copy of the data should be stored after transforming it to the electronic version.
- Processing speed and links to database servers are important data storage architecture considerations to be aware of. The data could be stored in an open standard file format such as JSON (JavaScript Object Notation). JSON is widely used by popular programming languages which would make the processing the data to gain insights. JSON is a data interchange format which is easily readable for human and consists of attribute–value pairs and array data types. Key-value pairs allows flexibility of extending the data structure. JSON data could be converted to other standard format if required since converter libraries exist in all the popular programming languages. In the case where data is highly relational in nature, relational database management systems such as SQL, MySQL, PostgreSQL etc. should be used. Wherever the data structure is not fully known and fixed, NoSQL databases like MongoDB, CouchDB, DynamoDB etc. should be used. Advantages of using NoSQL databases include directly storing JSON formatted data, faster schema on writing mechanism etc. Python,MATLAB, apache Hadoop, Oracle, and MS Access can be used to make queries on the stored data.

Legal and ethical issues:

National data protection regulations and European-GDPR should be adopted for protecting data in naturalistic driving studies.

The data should not be collected, stored, and/or used without the informed consent of participants (in the different experiments). In addition, directly sharing the data to third-parties (parties that are not in the consortium) during or after the experiments would require the consent of consortium partners, and can only be granted upon approval of relevant ethical committees, and upon consultation with the responsible data protection officer (DPO). This might require a submission of proposals elaborating on the need of the external parties in question of the data, and a plan of how they intend to use the data and goals they plan to achieve.

Data should be pseudonymised before being shared to other partners within the consortium (i.e. uploaded to a common central server); meaning participant names need to be replaced with a unique identifier that is available only to the partners collecting that data in the first place. This is done in consultation with relevant DPOs.

To preserve participants' privacy, first and the last minute(s) of driving in each trip may be deleted to avoid any possible relation with sensitive[13] information such as destinations with religious or political implications.

The legal and ethical considerations should be carefully considered for all possible countries of involvement as driving in Europe may occur across multiple countries. This also includes

[13] Sensitive information according to GDPR regulations can be found under: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en#references

cases in which the driver in question is not a participant, but someone that drives the vehicle incidentally, as they did not consent to their data being collected.

Careful considerations and protocols must be defined for treating personal and sensitive data after the project timeline. In the context of i-DREAMS, sensitive data include health-related data (on heart rate and variability and health-related questions in questionnaires[14]) and ethnicity (not directly ethnicity, but rather nationality).

[14] Questionnaires in i-DREAMS are given in Deliverable 3.4 (Pilkington-Cheney, et al., 2020).

3 Methods for Big Data Management in NDS

Naturalistic Driving Studies inevitably result in large amounts of data that need to be managed with specific standards. Previous projects indicate the need of good data management practices, comprising different steps including data collection, data pre-processing, data processing, data storage, but also external considerations like ethical, legal, and data protection. All different steps are to be conducted with high quality standards; in Europe these need to comply to the GDPR data protection regulation (2016/679), that came into effect in Europe in May, 2018 (European Parliament and Council of European Union, 2016). Accordingly, this section presents methods for handling of Big Data in the context of Naturalistic Driving Studies, by proposing an overall methodology (summarised in Figure 2).

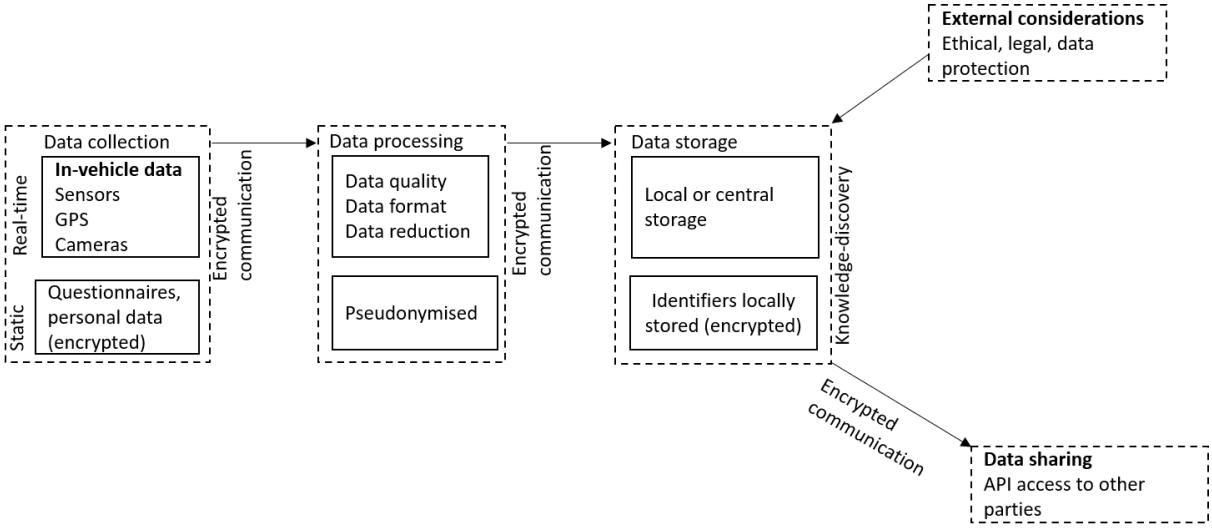

Figure 2: Generalised data handling framework (own illustration)

3.1 Data collection

Simulator and NDS studies collect data using different equipment, often as a collaboration of different partners or countries, which can make things rather complex, mostly when different standards are used. Data is often collected by sensors that measure different metrics in-vehicle, GPS data, video cameras, all of which are often centrally integrated through a communication component; these can be specifically designed for the NDS or off-the-shelf equipment. Ideally, different partners would use the same DASs, to reduce issues of data incompatibility, as learned from the UDRIVE project. Instrumenting vehicles can vary depending on the mode of interest (cars, buses, trucks, or even driving simulators); yet, for all of these the data would mostly be quantitative and of a big size: could result eventually in several terabytes of data. In addition to quantitative data, qualitative data is often collected by means of surveys and questionnaires, which include personal information like name, health status, age, gender. Additional personal information would come from GPS measurements, as they pertain to the home and/or work addresses of participants.

A main concern and challenge would therefore be to properly dispose of this data; for personal data, the way to correctly anonymise it before linking it to the experiments data, and for geo-location data, how to best handle it. Besides, raw GPS signals could be better managed when cleaned and simplified, using for instance the Ramen-Douglas-Peucker algorithm (Muckell, Hwang, Lawson, & Ravi, 2010). Since certain events (near-collision warnings) need to be

mapped to an exact location, the two nearest-GPS points could be added per event to the simplified trace.

Data collection should only be conducted after receiving the approval from the respective ethical committees, having consulted the outstanding legal departments, and being in accordance with the responsible data protection officers. Approvals are accompanied with agreements on data sharing, but also participants' consent, where details on how their data will be handled are explained.

3.2 Data pre-processing and processing

This includes aspects of data preparation including data quality procedures like data cleaning (attempts to best smooth the noise of the data) and dealing with missing, inconsistent, or erroneous values (due to calibration issues); missing values can be imputed. Data pre-processing can include data aggregation, and formatting the data to facilitate future processing, such as resampling processes to harmonise resolutions of different data types. Tools used to process the data would depend on the usability and purpose of analysis. These can be chosen according to the partners' preferences, who can develop scripts in MATLAB, Python, etc.

To process survey data, coding is required according to categorical values, and ensuring consistency across different surveys. Having obtained the consent of participants to collect data, this data is stored locally (where the experiments are conducted), and assigned a unique participant identifier; the only cross-reference between the identifier and the identity of the person would be the consent form which is to be secured, with limited access (to be defined; normally the DPO and assigned points of contact to the experiments). During the project, only pseudonymised (at least) version of personal data should be handled and linked to the rest of the data (vehicle data); protocols and specifications are to be discussed with relevant DPOs.

3.3 Data storage

Several types of storage can be of use: one would be local, and the other central. In a central storage architecture, all partners would have access to a central database. Depending on the ethical and legal regulations, and on the data sharing agreements, personal information may need to be locally anonymised or pseudonymised, before being transferred to a central server. This data would then be pre-processed.

On the other hand, in a local process, one approach could be to transfer vehicle data through hard disks, then store it in encrypted local servers at each data collection point. This however would be an obtrusive way, high demanding, and require an effort from the participants, as they would be reminded of the nature of the data collection experiment. Alternatively, data could be automatically transmitted (WIFI, wireless, Bluetooth), and locally downloaded (from the servers of the data collection equipment providers). For example, in case the technology is provided from a technology provider, this data would be automatically stored in the database of such systems (off-the-shelf systems). Then the data could be saved to a central back-end server, which saves different components of the integrated and processed data, which is then copied and shared with different partners based on defined access levels and following safe protocols with access points encryption. Transferring data would take place over HTTPS and hence secured with public/private key encryption mechanism. Local partners would get a copy of their data directly from the central server (for efficiency and consistency purposes); alternatively, they also can access the data directly from the technology providers' databases.

Subjective data or rather questionnaire and qualitative data, though not directly resulting in big datasets, needs to be uniformly re-coded and pseudonymised (at least), prior to being added

by the responsible parties to the central server. Accordingly, the only cross-reference between the trip data and participant personal data (name, address, etc.) would be through this identifier; this link would be securely stored at the premises of partners who collect this data, ideally in a separate computer (password-protected), with no other data and accessible only to a limited number of people (to be defined in a separate agreement or protocol at each point of data collection according to the regulations of each entity collecting this data).

3.4 Data access

Access to the data needs to be controlled and defined according to different levels, with different rights. A joint data agreement for processing personal data is to be signed between different partners and would aim at specifying different roles in accordance with GDPR regulations, including processors, controllers, joint controllers, the categories of personal data, and a protocol to report breach incidents. Different granted permissions would give different roles for access to different people. Personal data would not be stored on any database, should be anonymised, and stored securely in a separate computer. Only the anonymised or pseudonymised data can be linked to the experimental data in the databases.

When no longer needed, or at most at the end of the project lifetime, personal data shall be deleted. Pseudonymised data, shall be stored only until deemed necessary; a provisional duration shall be determined. Only the local partners collecting data shall have access in a secure location to the link to the identifier which enables the personal data identification. Typically, personal data would be destroyed or fully anonymised, after which it is no longer possible to link the person with the data; this is usually added as a disclaimer in the participant consent form.

The access of external parties to the data needs to be defined through a protocol according to which partners would agree to grant or not the data to an external party, and also specify the terms and duration of data access and use, until deletion. External access can be granted according to data agreement, where only the fully anonymised data can be made publicly available for research purposes. For this, a platform shall be chosen to ensure sustainability of the data to be archived; data shall remain accessible long after the end of the project. This would be to ensure the FAIR principles (data should be Findable, Accessible, Interoperable, Re-usable) as indicated in the DMP.

4 Collected Data of the Project

Methods described in Section 3 looked at a range of methods for NDS, proposing best practices based on lessons learned. To apply these in the context of i-DREAMS, it is essential to first introduce the data collected in the project. This section aims therefore to provide an overview of the collected data of the project. A detailed description of the dataset types was provided in the DMP. In addition, data collected is also mentioned in Deliverable 5.1 (Hancox, et al., 2020) and Deliverable D3.4 (Pilkington-Cheney, et al., 2020).

At the heart of data collection are technologies provided by the CardioID system (integrating different components including the Mobileye, CardioWheel/wristband, OBDII devices, and GPS), the smartphone app OSeven, the application from IMOB for heavy vehicles, and the driving simulator technologies-namely DSS. Besides the quantitative data, qualitative data is collected through questionnaires and surveys, as outlined in Deliverable D3.4 (Pilkington-Cheney, et al., 2020). In the following, an overview on these different data sources and types are given, and where relevant differences from the previous updates on data collection is highlighted.

4.1 Quantitative data generated using the CardioID system

To collect a range of vehicle and driver-related driving attributes, the project uses a system that is part of the i-DREAMS platform; here referred to as the CardioID system developed and developed by CardioID[15] (a technology company, founded in 2014). The CardioID system is available to consortium partners who are responsible for conducting on-field trials.

Based on the *CardioID System API document*[16], this system is composed of several devices that are marketed by CardioID and used for collecting data and then implementing real-time interventions. The different devices are connected with a gateway (CardioID GW) that gathers and centralises information from other components and handles data connectivity and transmission. These components are:

- Mobileye in conjunction with DashCam (installed on the windshield)
- CardioWheel system (installed on the steering wheel) for buses and trucks; for cars, trains, and trams, this is a wearable solution (wristband)
- Connection with On-board vehicle diagnostic units
- GPS device (inside the gateway, and the antenna outside)

Additionally, the CardioID GW is capable of local, real-time edge processing, and provides an output for a buzzer or a haptic engine; this could also be another sound via a speaker and is still under discussion. The gateway also has a capability of Ethernet, wifi and mobile data connectivity. Each gateway is conceptually tied to a vehicle (not a specific user), with data being acquired within a trip session. A trip session is defined from the moment the vehicle is turned on until it is turned off. There is a grace period (5 minutes) during which a quick turn off and back on is considered the same trip. The collected data goes through some usual automated processes to remove data noise (i.e. detection of outliers and their removal) and stored in CardioID cloud servers (subject to availability of required connectivity protocol). CardioID also provides a web API to support data access within the i-DREAMS project. The API follows the REST style, based on JSON data (available only for consortium partners). REST – representational state transfer – is a software architectural style for designing web APIs. The REST architecture for APIs creates a clear separation between the client from the servers (e.g. the data storage server) which improves the portability, makes the APIs highly

[15] https://www.cardio-id.com/about
[16] Confidential Document (Only available for Consortium Partners)

extendable and scalable. Further developments become faster due to the inherent modularity. APIs, if following REST style, is independent of the type of platform or languages.

The DMP datasets should be updated to account for:

- Wristband datasets for cars, and rail, if relevant
- Vehicle inertial information collected by accelerometer, gyroscope, magnetometer (9 DOF motion sensors installed on the Gateway)
- Driving Behaviour data collected by accelerometer, gyroscope, magnetometer, to detect harsh acceleration, braking/cornering

For heavy vehicles (trucks, buses, rail modes), FMS logging will be implemented but may not be available to all heavy vehicles; therefore a more generic term such as (vehicle) CAN interface can be used.

4.2 Video data generated via dashboard camera

Apart from the quantitative data, the CardioID system also provides video data generated from its dash cam which works in support of the Mobileye component. This dash cam is currently triggered by events from Mobileye, in order to better understand, in a post- trip analysis, what led to the event in consideration (e.g. a forward collision warning), but it will be triggered also by other events (e.g. harsh braking, CardioWheel sleepiness event). The prime objective of collecting such video data is to visualise the real scenario on road because of which a warning (i.e. event) is generated by the Mobileye system. This is important because Mobileye can sense situations causing the system to generate a warning where surrounding vehicles are the main cause of the occurrence of such events. For example, the driver has kept a safe distance from the forward vehicle; however, another vehicle suddenly came in between the subject vehicle and forward vehicle by changing its lane which results in the reduction of the time headway. For developing post-trip interventions in gamification environment such types of situations need to be identified so that appropriate information is sent back to the driver to make the post-trip intervention more effective and avoid false positives which would undermine the acceptance of the system. Video clips are recorded when specific events occur and provide the situation just before, during and after the event. Each video clip size is around 3 MB (10 seconds before the event, the time of the event, and 5 seconds after the event), and it is known from previous experience of CardioID that on average around 50 clips are recorded per 100 km driven by the vehicle. These video clips can be accessed from the CardioID cloud servers via provided web API (only available to consortium partners).

4.3 Driving simulator data

The driver and road environment assessment and monitoring system that will be developed in this system will undergo a testing phase using driving simulators. Special risk scenarios will be developed within a simulated environment to test a variety of situations to assess the performance of the preliminary system. Along with the usual driving simulator system, the CardioID system will also be integrated within a driving simulator for real-time intervention testing. In Germany, Belgium, and Portugal, simulators will be provided by DSS, using the driving simulator software STISIM3. In Greece, NTUA will use the FOERST simulator, with F10 as a driving simulator. Specifications of the different simulators and software are provided in Annexes 3 to 6 of Deliverable 5.1 (Hancox, et al., 2020). For the rail simulators in the UK, the simulators are located within the operator's premises. Three types of simulator are available for rail. The first uses software by Sydac, an Australian company, and replicates Bombardier cabs. The second uses software by TransUrb, a Belgium company, to replicate Stadler cabs. The third simulator was designed by Ian Rowe Associates with software

developed by its sister company, Avansim LTD. The latter simulator is designed as a training tool and therefore the sampling rate is 1-3 Hz.

Updates of the DMP should therefore make the distinction between different dataset types for simulator data coming from different simulator types (DSS, FOERST, and the rail simulators).

The STISIM3 simulator software automatically collects driving parameters at frame rate (+/-60 Hz). These parameters are linked to time and include travelled distance, speed, acceleration, steering inputs, brake input, lateral positions etc. Following parameters can be measured by the DSS simulator, but not by the FOERST simulator, namely:

- Current roadway curvature (1/foot or 1/meter)
- Vehicle yaw rate (radians/second)
- Minimum range (feet or meters) between the driver's vehicle and all vehicles in the driver's direction.
- Minimum range (feet or meters) between the driver's vehicle and all vehicles opposing the driver's direction.

For simulator trials, optional considerations include eye-tracking equipment (glasses) and video recording, for all simulators (cars, buses, trucks, and rails).

4.4 Quantitative data generated from smartphone applications

OSeven is another technology provider within the i-DREAMS consortium. OSeven will provide a state-of-the-art android and iOS-based smartphone application that also monitors and collect driving behaviour of individuals using a variety of parameters. The app will be used by drivers recruited for on-field trials and will use different smartphone sensors to collect such data. Drivers recruited for the experiments will download the app on their personal smartphones. Raw data collected by the app includes date and time, GPS data, angles formed by the local axes of the phone to the North and horizontal planes, rate of change of these angles, accelerometer data, gyroscope data, activity data (walking, stopping, driving), screen state (for mobile use), smartphone device data, while processed trip data includes (but not limited to) number of trips, distance travelled, trip duration, number of harsh brakes, number of harsh accelerations, driving over the speed limit, average speed, mobile phone use, and distance travelled. The datasets collected from this app and the CardioID technology will be fused together for more accurate prediction of driving behaviour and also to test and validate the performance of i-DREAMS real-time platform. OSeven will provide an API to i-DREAMS partners to access these datasets. The smartphone app will also have a functionality to provide post-trip intervention via a feedback mechanism.

The OSeven app will be used in the field trials solely-not in simulator contexts-and only for cars since motion sensing is based on GPS data; the app is not able to provide driving behaviour information for a stationary simulator trial. For other modes (trucks, buses, rail), another app would be used (IMOB); the exact specification of which are not shared at this stage. The main idea would be to address driving behaviour by means of a performance score and interventions based on gamification techniques. For more details on the interventions, please refer to Deliverable 3.3 (Brijs, et al., 2020).

4.5 Qualitative and quantitative data on levels of participation and user experience/opinions

Apart from data generated by technological equipment, pre- and post-experiment questionnaire surveys will also be conducted to obtain driver socio-demographic information

and drivers' driving attitudes and feedback in relation to their experience during simulator and on-field experiments. These data provide a meaningful base to conduct analysis to test and investigate the performance of the developed system and then improve it prior to conducting on-field trials. In addition to this contextual information interview-based survey, additional surveys are also planned within the i-DREAMS project. These are online questionnaire-based surveys to get expert/stakeholders opinion to get insight on current state-of-the-practice and the possible advancements suggested by an informed community; these can be incorporated within the development of the i-DREAMS platform. Questionnaire details are provided in Deliverable 3.4 (Pilkington-Cheney, et al., 2020).

5 Standard Protocols for Big Data Handling

Based on previous learnings (Section 2), proposed methods (Section 3), and data collected in i-DREAMS (Section 4), this section aims to provide protocols for the handling of big data resulting from the conducted experiments. Legal and ethical considerations throughout the process are to be taken into consideration by all partners involved in data collection. The latest updates on these are provided in Section 7 of Deliverable 3.4 (Pilkington-Cheney, et al., 2020) and Section 7 of Deliverable 5.1 (Hancox, et al., 2020).

5.1 Setting the scene

When describing protocols for the handling of big data, it is crucial to differentiate between parties with different interests and responsibilities in the i-DREAMS project.

- Technology providers: these are namely CardioID, OSeven. These directly collect data through the equipment they provide. These are the source of data collection, ensure proper and consistent data collection, and make it accessible to the rest of the partners.
- Simulator partners: this is namely DSS, which will integrate the different technologies, and provide the simulators which will log the data (integrated simulator and technology data) and store it locally (at the simulator PC); other simulators (different car simulator at NTUA, rail simulator(s)) will need to follow similar integration procedures and would also fall under this category for simulator experiments.
- Field trial partners: this includes partners who are responsible for the logistics of setting up the experiments (simulators, and field trials). These are responsible for correctly managing the experiments at their premises and collect the questionnaire data and are responsible for correctly linking it to the experimental data, after having adequately pseudonymised it (link to a unique and securely stored identifier; details on this are given in Section 5.2.3).
- Data processors: include partners who will access data to analyze it and test hypotheses derived from the project's research questions. In addition, UH also processes data by providing the post-trip intervention framework, which does not really collect data, but generates scores and interventions based on the collected data at the end of a trip/time duration.

Partners and their distinct roles are vital to the correct and efficient data collection, storage, and access, which would in turn facilitate data analysis and usability. The overall i-DREAMS architecture and its different components is depicted in Figure 3 and comprises various components including data sources, user interfaces, in-vehicle systems (presented in Section 4 and will be updated in the next DMP version), data processor components, the post-trip intervention framework, and a back-office component (that is developed under Deliverable 4.3-Confidential).

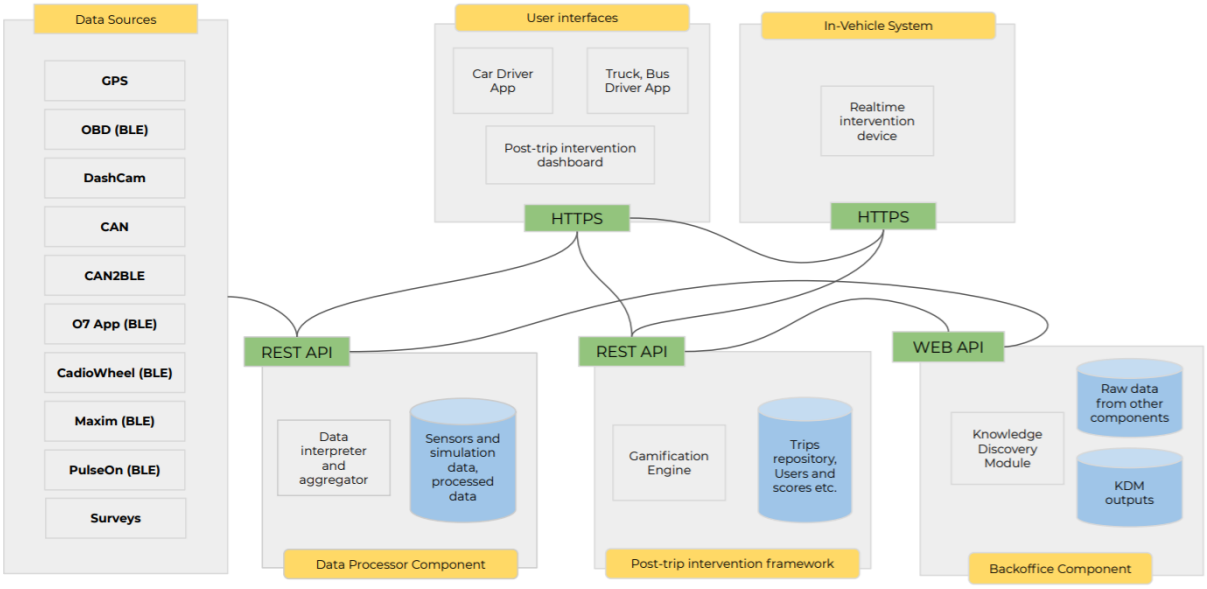

Figure 3: i-Dreams system components (own illustration)

5.2 Protocols for handling data

This section outlines protocols for handling the big data generated in i-DREAMS, with a focus on data collection, data storage and backup, data sharing, and data access. Where relevant, details are given regarding specific protocols in place used by partners.

5.2.1 Data collection

For country-specific trials, local partners from each country are responsible for the logistics of setting up the scenarios, leading to the collected data. In the context of i-DREAMS, based on the defined methodology (Section 3), and to fulfil best practices, data collection should cover the following aspects.

- Means of collection: To ensure consistency of processes and quality of data, data acquisition (vehicle data) should be done through the same mechanisms (servers, communication protocol, code etc. should be same) even for country specific scenarios. Data collection components in the i-DREAMS context are listed in detail in Section 4 of this deliverable.
- Data availability: for the support of the experiments and trials, data from external sensors (Mobileye, CardioWheel, Wristband) needs to be made available.
- Frequency of collection: given the fact that each sensor has a different frequency rate; each sample has an associated time stamp to it for appropriate synchronisation. Detail information on sampling frequency of CardioID system measurements is provided in detail in Table 2 of Deliverable 3.2 (Katrakazas, et al., 2020). For DSS simulator scenarios, data will be collected based on simulator frame updates (60 frames per second). Simulators will communicate with the Cardio ID GW at each time frame update and register data from the gateway when available.
- Communication protocol:
  - Realtime: ZMQ. This messaging library was chosen since it is very performative – via sockets that carry atomic messages across various transports like in-process, inter-process, TCP, and multicast. Moreover, it allows integration in various programming languages, which makes software framework more modular. This is

very important when there are several research groups implementing over the same software.
    - Non-Realtime: HTTPS through API
- Data integrity: Each data-collecting system (i.e. the Gateway and associated sensors) should be conceptually tied to a vehicle, not a specific user. Data is acquired within a trip session, which is defined from the moment the vehicle is turned on until it is turned off. A grace period (5 minutes) during which a quick turn off and back on should be considered as the same trip.
- Data pre-processing:
  This is primarily done locally at the gateway and in the tech partners' databases.
    - Handling missing data (sensor and communication failure): With sensor failure, a trigger and alarm can be sent to the user/ field trial supervisor to ensure that nothing was disconnected (equipment in-vehicle). For communication failure however, all the data is logged so off-line synchronisation is possible even without any real-time communication. Finally, missing data can occur by the non-collaboration of the user – for example, hands off the CardioWheel or not wearing correctly the wearable; while for the latter no specific protocol has been created at the time this deliverable was written, it could be if necessary.
        - Ensure temporal order in case of time-series data
        - Handling the time zone information carefully
        - Rectify GPS data: sensor can sometimes report incorrect latitudes and longitudes when there are momentary losses of GPS signals. A filtering procedure may be implemented to remove these positional jumps.
        - Video data: may be pre-processed in a way to reduce data volume without compromising the quality of the video. Metadata of the videos (event, timestamps, trip info etc.) should also be attached with each video for ease of future analysis.
        - Outliers and anomaly detection: Outliers and anomalies are needed to be detected to ensure quality of data. Detection processes should be done at the source of collection when possible.
    - Data verification: Due to the potentially huge amounts of data collected, data verification is important since the probability of errors during the communication process is high. The data aggregated in CardioID GW should be validated at the end of a trip session, ensuring temporal order of the data points, and verifying that repeating sample points are filtered out. This is needed to deal with possible network requests from the collection end to the cloud server that do not arrive in the correct order, or when data is received by the server, but its acknowledgement does not reach the data collection end (and data is resent). This procedure should apply to all suitable data types. For example: *GPS, LOD, IBI, DriverChange, CAN etc.*
    - Data loss minimisation: This often happens at the retrieval/upload. To prevent data loss during the data upload/retrieval procedure it is thus critical to verify that data is consistent before deleting it from the vehicle. In case data is picked up and inconsistencies are identified, the vehicle data logger should be checked as soon as possible lso that any issues can be recognised and fixed.
    - Vehicle data deletion: Data from the vehicle should be deleted after the data has been backed-up and verified.
    - Data format: A description of the data variables should be provided by the technical partners generating the data and should be sufficient for future reference.
        - Simulation data format: ensure consistent formats between different simulation formats (from DSS, NTUA, rail simulators).

- Field Trials data format: consistency ensured through same technology use (integrated CardioID gateway).
- Questionnaire data format: ensure similar format by using same surveys (translated); similar survey platform would be encouraged. Otherwise surveys can be exported to a similar CSV format, and the answers to categorical questions consistently coded (protocol or agreement to be separately defined). For survey data to be stored, the following related information should be attached to each instance:
  - Date and time (hh:mm) of start
  - Date and time (hh:mm) of end
  - Unique identifier; the link between this identifier and the personal data (name, address, etc.) is only stored at the local partners' premises, so that they only could cross-reference the data with the participant profiles.
  - If applicable, reference to objective data (file name, location).
  - Although survey data is static, it would still be a good practice to have the above-mentioned information. These data entries might however change for field trial experiments, during which participants' attitudes may be measured half-way through the trials.

- Data privacy and ownership:
  - Prior to the start of the experiments, trial partners should have received the approval of their respective ethical committee, established contact with their respective DPO (for compliance with both GDPR and national regulations), and have signed a joint data agreement for the processing of personal data (confidential). This agreement, which remains confidential among consortium partners, distinguishes between data processors, data controllers, categories of personal data, and specifies how data leakage or breaches of the agreements are to be reported.
  - A data protection officer in each country where data is collected is appointed to approve the correct handling of the data.
  - Prior to participation in the experiment, participants sign an informed consent, where they give the partners permission to collect and process their data during experiments, including details of which data is to be shared. Only then, their personal information may be collected, and they are given a unique identifier which is a cross-reference between the experiment data and their personal data. This personal data should be encrypted to ensure security and be placed in an offline file system. .and only accessible to limited people (to be defined, typically the DPO and persons of contact assigned for the experiments). Only the local partner has the access to the unique identifier which can point to a participants' personal data.
  - Servers and hard drive encryption (following the GDPR recommendation: article 34, recital 83) should ensure that all data (including non-personal) is protected, as a mitigation against breaches, even if the data is pseudonymised, but specifically for personal data.
  - In case of sharing among partners residing in different countries and to assure compliance with privacy regulations, the local partner needs to clear the simulation hardware of collected data before handing over the hardware to another partner.

5.2.2 Data storage and backup

Data storage and backup protocols should cover following aspects.

- Partners are free to choose the storage engines (databases, filesystems) for local storage facilities. For storing of json formatted data, alternatives such as MongoDB, CouchDB etc would be ideal candidates. For relational data, MySQL, MariaDB,

PostGreSQL have been quite popular. For a mixture of relational and non-relational data, PostGreSQL would be suitable.

- Data should be automatically stored locally (preferably, in the same machine). Only anonymised or at least pseudo-anonymised personal data should stored together with the vehicle data.
- Data storage can be categorised in two types- local and central storage.

Local storage: Local storage refers to storage systems which are not accessible through standard API to external world (other partners and/or third parties) To ensure proper handling of the data, a local storage should fulfil these requirements:

- Persistence: data should remain stored locally for at least the end of the experiments.
- Reliability: To achieve reliability:
  - Periodic backups should be taken. Cardio ID databases perform automatic backup at AWS (daily backup, storing last 7 days). For simulator experiments and the baseline stage of field trails, it is suggested to back up by the end of each day the collected data.
  - Delete/Modify operations should be handled carefully to make sure data consistency and validity.
  - Availability: Wherever the necessity of sharing local data among partners arises, data could be sent to the central data storage by uploading through the available API of the central storage system. Data access rights at the central server would be necessary to define to control rightful access to it.

Publicly accessible storage: Publicly accessible storage refers to storage systems which are accessible through standard API to the external world (other partners and/or third parties) This may also include third party cloud storage. A publicly available storage should fulfil these requirements:

- Persistence: data should remain stored at least until the end of the project.
- Reliability: To achieve reliability these steps are required:
  - Periodic backups should be performed.
  - No delete/modify permission is given to any user of the storage (other than the local storages), only read permission would be provided to the appropriate user. Exceptions can take place in extraordinary circumstances and contingent upon approval of the *superadmin*.
- Availability: Once available in the storage, data should be immediately available to the authorised user, preferably via an application programming interface (API).
- Serviceability: In the time of storage server maintenance (server downtime), data may not be available up to a certain period.

For CardioID data (most of the vehicle data collection), the data that is available using CardioID API is built on top of the database structure. The details for its implementation and access are provided in a documentation that to date remains confidential between project partners. As advised by the FESTA Handbook, data should be backed-up and stored in a safe place as soon as it is available. Ideally, the backed-up data and the main copy of the data should be in two different safe places. If possible, to minimse data loss, it is recommended to use direct on vehicle data backups (FESTA Handbook, Version 7, 2018).

Within i-DREAMS, a back-end database or a back-office component (shown in Figure 3) is developed to store centrally raw and processed data from other components. Partners would

then comfortably access data from the back-office through a WEB API. The specifications of this back-office are given in detail in Deliverable 4.3 (confidential).

5.2.3 Data sharing

Data sharing should follow following procedures.

- Partners will upload the data to the back-office server in a specific format.
- Consortium partners and authorised third parties can get access to the data through the developed server web API. API specifications will be part of deliverable 4.3, which remains confidential between consortium partners.
- A joint agreement between partners-for sharing personal data- provides the details of use of personal data among consortium partners until the end of the project; after the project end, a fully anonymised subset of the data may be made accessible for research, and following GDPR regulations.

Data preparation: Before sharing the data, the responsible partner (generating data through collection, surveys, etc.) should create an understandable data format along with description of variables which would help other partners to understand easily. Data preparation should be done to ensure:

- Consistency: The types of data being shared should reflect the expected versions of the data being collected.
- Completeness: Shared data should contain complete information that is intended to be shared.
- Integrity: Data being shared should be correct, relevant and accurately represent what it should.
- Timeliness: The data should be received at the expected time wherever needed in order for the information to be utilised efficiently.

Data transfer: To transfer data efficiently, each partner should either provide API on their own or upload the data to a back-office server from where other partners can collect the data.

- If an API is exposed to transfer data from the responsible partner's side, an API specification is also expected from the partner. These APIs should also need to be secured through an authentication mechanism.
- Similarly, the data back-office should also provide an API specification listing out how to access data which are available through its API.

Data pseudonymisation: Before sharing among consortium partners or uploading to a cloud storage service, the local partner should first make sure the data is anonymised or at least pseudonymised. This would ensure the privacy of participants and comply with the national and EU regulations (GDPR) and would be reviewed and approved by the respective DPOs.

- Data format:
  - Expected data format is JSON for both simulator and field trial data where possible.
  - Descriptions of the data variables are attached and should be sufficient for future reference.
- Data processing tool:
  Processing tools are tools that partners would use to analyze collected and generated data; partners have the freedom to select their own tools (MATLAB, Python…) and develop their own scripts, according to usability purposes.

Data anonymisation: Following procedures can be applied for anonymisation after the agreed time after the end of the project (5 years) and for making the data accessible in an open-source platform according to the project objectives.

- The unique identifier that connects the data in the partners' databases (internal servers, central databases) with the personal data of the user is replaced with a random number. The process would then be irreversible and there would no longer be any possibility of connecting the data in any database with the personal data of the user.
- Where primary data (including location data) relates to the DriverID, the DriverID is replaced by a random code for each trip. This process is irreversible and there is (i) no longer any possibility of linking the primary data of the trips (including location data) to the personal data of the user and to any DriverID and (ii) no longer any correlation between the trips of a user which is anonymised.
  Following the above procedures, the data of the user would be fully anonymised since it is impossible to connect this data with a natural person. The extent to which these exact procedures would be applied (whether the first point or both mentioned points) would depend on the approval of the respective DPOs and would need to be agreed between consortium partners.

5.2.4 Data access

Following procedures should be followed for the access of data.

- Different user types should be defined with different rights of access (e.g.superadmin, admin, user etc.). A list of roles shall be made between data access during the project lifetime, and after the project end.
- During the project lifetime, data access should follow joint agreements set out between partners. Pseudonymised data shall be accessible to consortium partners, according to joint data agreements. Personal data shall be only accessibly locally by authorised personnel.; personal data shall not be stored longer than necessary. A duration of five years is provisionally foreseen and should be agreed by respective parties (for the personal data).
- In accordance with principles defined in the DMP, research data shall follow FAIR guidelines to be findable, accessible, interoperable, and reusable.
- The data aggregated by CardioID can be accessed using a web API which follows the REST style, based on JSON data (URL: https://api.cardio-id.com). The access needs to be authenticated, and the chosen mechanism is based on an access token attributed by CardioID to each data client, with a finite validity period. The token should be included in every request to the API, as a header. The API provides GET and POST methods.
- An anonymised portion of the data (a few datasets) will be made available and offered to third-parties at the end of the project; in according with GDPR, these should excluse personal data. This will be made available on the Zenodo[17] digital repository, which will guarantee archiving and sustainability. This has been indicated in the DMP as the best way to ensure access to the generated data remains long after the project ends.

[17] Zenodo was launched at the CERN Data Centre in May 2013 with a grant from the European Commission with a special commitment to sharing, citing and preserving data and code. As a digital repository, Zenodo registers DOIs for all submissions through DataCite. The platform is based on the Invenio open-source software, Zenodo profits from and contributes to the foundation of code used to provide Open Data services to CERN and other initiatives around the world.

5.3 Special considerations

Simulators:

Within the i-DREAMS project, several driving simulators will be used for experiments that cannot be performed on the road, either because concepts or designs are in an early development stage or because of practical or safety concerns.

- Since data collected in the simulator results directly from the calculation of variables that make up the simulation itself, there is no risk of faulty sensors, noise or incorrect data. This means that additional validation or processing of the simulation data is not necessary before it can be used for analysis.
- Output data from the simulator studies is treated differently compared to data from on-road experiments. Data from the simulator studies will not be automatically uploaded to the cloud nor will it be made available through an API like data from the on-road experiments would be (through the CardioID API for instance). The purpose of the data acquired by simulator studies is to analyse the data and answer questions based on what was initially set out during the design of the simulator experiment.
- The CSV data file that is created during the simulator experiment is automatically stored locally on the simulator PC. Video files and data from eye tracking are recorded on a removable storage medium, like an SD-card. The partner that is organizing the simulation experiment is responsible for creating a daily back-up of the collected data files, either locally or on a protected, non-shared cloud environment and to remove video files from the SD-card, store them locally and also create a daily back-up.
- When exchanging simulators or simulator PC's between multiple partners, all collected data should be removed from the simulator PC before handing over the PC to the other partner. Any other data storage devices, like memory cards for cameras should also be cleared of all data before being exchanged by partners.
- For NTUA simulators, the process will follow the same procedure as for the DSS simulator to have as much consistency as possible (simulator specifications will be provided in Deliverable 5.2).
- Where simulator data is collected at the operator site (rail), data will be securely transferred to the responsible partner via secure data connection or by an external hard drive in a secure case.

A summary of the handling protocols defined above with respect to the lessons learned from previous projects (as presented in Section 2.5) is given in Table 5.

Table 5: Lessons learned implications to i-DREAMS

	Previous findings	Implementation in iDREAMS	Remarks
Data collection	Reliability and validity checks	☑	For both the local and publicly accessible storage, period backups are performed. Additionally, no delete/modify permissions are given to any users of the storage (besides the local storage)
	Common DAS	☑	Achievable through the i-DREAMS system which is used in all trial countries
	Minimse number of vehicle models	☑	This is a recruitment strategy that would be followed to help reducing data collection burdens, but also installation and de-installation efforts.
	Centralise responsibilities for coding, processing, and analysis	☑	This can be part of the central back-end system, where processed data and specific metrics can be added for the processing and analysis of collected data.
Data storage	Data pre-processing prior to storage	☑	Primarily done at the gateway and the tech partners' databases and includes handling missing data from sensor and communication failure.
	Advanced video processing techniques		This has not been discussed in the scope of this deliverable but would need to be discussed in the next steps and depending on the needed analysis. Advanced techniques would be recommended for this purpose.
	External data sources		Data enrichment using external data is possible and would need to be discussed in the next steps.
	Ease of access of data	☑	Ensured through the central back-end API
	Systematic back-ups	☑	At each storage system
	Data well defined and understandable	☑	Role of the DMP
	Video files stored separately but linked with the rest of the data in file management systems		This has not been discussed as video data processing has not been discussed in this deliverable but should be possible in i-DREAMS.
	Transferring the data should be done electronically, avoiding manual transmissions	☑	Through the gateway (Except for the simulators)
	Manual extraction files like questionnaires: store hard copies after electronic storage		Questionnaire setups have not been finalised but this should be possible within i-DREAMS.

	Ease of access of data. Recommended architectures: open standard format like json for relational data. For non-relational or not fully known data, storage systems like MongoDB are recommended also for fast access	☑	
Legal, ethical, and data protection considerations	Informed consent of participants	☑	Revised forms by relevant ethical committees.
	Agreements for third-party sharing	☑	Needs to be discussed between partners to draft such an agreement protocol, and in accordance with the relevant ethical and DPO committees.
	GDPR for Europe	☑	Supervised by the relevant DPOs.
	Data pseudonymisation before sharing to partners in the consortium	☑	Unique identifier available only to partners collecting the data (consultation with ethical/DPO entities)
	First and last minutes of driving may be deleted to avoid any possible relation with sensitive information		Recommended for i-DREAMS.
	Driving across multiple countries		Need to draft a protocol for cross-country crossing. What happens to the data?
	Non-participant driving the vehicles incidentally		Need to be able to stop recording as the driver did not give consent to their data being collected. To avoid this scenario, this could be well explained in the recruitment (to ask if possible, to prevent driving by non-participants).
	Data use after project lifetime (personal and sensitive data)	☑	Defined within national ethical and DPO committees, for the use by local partners.

6 Conclusions and Next Steps

This deliverable aimed at providing protocols for the handling of data generated throughout the i-DREAMS project. The experiments to be conducted will collect data of different types, from different countries, and therefore result in “Big Data”. Guidelines to correctly handling the data and the logistics of doing so are therefore key for i-DREAMS, in its different phases. By first looking into lessons learned from previous projects of similar nature, findings were drawn that could be key to i-DREAMS. Accordingly, these paved the way to a methodology for the handling of generated data, looking at different aspects, and starting from data collection, to storage, and considering legal, ethical, and data protection. This resulted in the protocols for the handling of this generated data in the context of i-DREAMS; special considerations to modes like simulators, trucks, buses, rail are given, when specific information was available.

The following was decided:

- Partners who collect data (where trials are running; in their premises) are responsible for the proper collection of data, including the specified protocols. This can be distinguished between partners who provide the technology (CardioID-OSeven), partners who conduct experiments (different trial partners), and partners who process the data (Intervention framework, and the ones who access the central backend). The collection should deal with communication issues, loss of signals, at the source of data collection.
- Storage should be done locally (in-vehicle devices: gateways, simulations: in local computers, and personal data: in local partners’ systems), and in cloud servers (where data is stored from the different collection points).
- Storage should be done locally (directly from the data collection equipment to the technology’s sever), and centrally (where data is stored from the different collection points).
- Personal data should remain where it was locally collected. Before being uploaded to a central back-end server, it should be at least pseudonymised. The identifier, which is the unique key between the personal data and the corresponding collected data, shall be encrypted and placed in an offline file system with limited access (to the DPO and persons of contact who are responsible for the experiments).

Next steps:

In the next steps of the project, the next version of the DMP should be updated to take into account the recent additions to the data collection (wristbands, potential eye-tracking equipment, video recording, and other simulator types like the one in NTUA, or the rail simulators in the UK). Details of the simulator protocols will be detailed in the upcoming Deliverable 5.2.

The next steps in the organisation of experiments include:

- Identifying issues in data collection from the pilots.
- Identifying issues in data storage from the pilots.
- Starting the implementation of the back end as soon as data is available; giving access to partners and identifying issues.
- Drafting an agreement for data access, where access rights are defined by partners, with different rights (accessing, editing).
- Drafting a protocol for the handling of questionnaire data, detailing the platform to use (to collect the data), ensuring uniformity consistency, also taking translation into consideration, and taking into account issues that should be handled at the source

(cleaning, pre-processing). Once questionnaires are finalised, this is to be addressed for operationalisation.

- Drafting the data sharing agreement for beyond the project end: are proposal plans needed for accessing portion of the anonymised data? Will it be available for all researchers? What exact anonymisation procedures will be followed?
- Drafting data handling documentation, which has to be filled by partners involved with data collection and storage and processing, to ensure proper quality control of data at different ends, and which could be used at the end of the project, to monitor achieved goals vs. what was actually proposed or expected.

Overall, this document should be a living one, updated where applicable into the necessary steps, and serving as a guideline for how to best handle the big data generated throughout the project. This deliverable will of course be enriched through adequate consultation of the next DMP version, and with the upcoming D5.2, for driving simulator experiments in particular, which will be a first learning lesson for not only understanding better the system, risk scenarios, etc., but also for how to best handle the data, and how we can further improve the process and make it more efficient. Updates on handling and processing the data will also be reflected in Deliverable D5.4, around the end of the project, where a unified big data fusion framework will be presented for exploiting driving performance data.

7 References

Adnan, M., Brijs, T., Donders, E., & Hermans, E. (2019). *Data Management Plan. Deliverable 1.2 of the EC H2020 project i-DREAMS.*

Babulal, G., Addison, A., Ghoshal, n., Stout, S., Vernon, E., Sellan, M., & Roe, C. (2016). Development and interval testing of a naturalistic driving methodology to evaluate driving behavior in clinical research. *F1000Research, 5*, 1716.

Baldanzini, N., Hurth, V., Regan, M., Spyropoulou, I., Eliou, N., Lemonakis, P., & Galanis, P. (2009). *Deliverable 4, 2BESAFE project. Literature review of data analysis for naturalistic driving study. .* Belgium: European Commission.

Bärgmann, J. (2016). Methods for Analysis of Naturalistic Driving Data in Driver Behavior Research. Gothenburg, Sweden.

Barnard, Y., Utesch, F., van Nes, N., Eenink, R., & Baumann, M. (2016). The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters. *European Transport Research Review, 8(2)*, 14.

Bender, A., Ward, J. R., Worrall, S., Moreyra, M., Konrad, S., Masson, F., & Nebot, E. (2016). A Flexible System Architecture for Acquisition and Storage of Naturalistic Driving Data. *IEEE Transactions on Intelligent Transportation Systems, 17*(6), 1748-1761.

Brijs, K., Brijs, T., Ross, V., Donders, E., Yves Vanrompay, Y., Geert Wets, G., . . . Gaspar, C. (2020). *Toolbox of recommended interventions to assist drivers in maintaining a safety tolerance zone. Deliverable 3.1of the EC H2020 project i-DREAMS.*

Doerzaph, Z., Dingus, T. A., & Hankey, J. (2010). Improving Driver Safety through Naturalistic Data Collection and Analysis Methods. *SAE International Journal of Passenger Cars - Electronic and Electrical Systems, 3*, 162-169.

Dozza, M., Bärgmann, J., & Lee, J. D. (2013). Chunking: A procedure to improve naturalistic data analysis. *Accident Analysis and Prevention, 58*, 309-317.

Eenink, R., Barnard, Y., Baumann, M., Augros, X., & Utesch, F. (2014). UDRIVE: the European naturalistic driving study. . *Proceedings of Transport Research Arena. IFSTTAR.*

Espié, S., Bekiaris, E., & Nikolaou, S. (2010). Naturalistic rider studies for the analysis of riders' behavior and safety:" 2BESAFE". . *Proceedings of the Road Safety on Four Continents Conference (Vol. 15, pp. 194-199).*

Espié, S., Boubezoul, A., Aupetit, S., & Bouaziz, S. (2013). Data collection and processing tools for naturalistic study of powered two-wheelers users' behaviours. *Accident Analysis & Prevention, 58*, 330-339.

European Commission. (2017, May 25). *2-WHEELER BEHAVIOUR AND SAFETY*. Retrieved from CORDIS EU research results: https://cordis.europa.eu/project/id/218703

European Commission. (2017, March 29). *Final Report Summary - 2-BE-SAFE (2-WHEELER BEHAVIOUR AND SAFETY)*. Retrieved from CORDIS EU research results: https://cordis.europa.eu/project/id/218703/reporting

European Commission. (2017). *UDRIVE: eUropean naturalistic Driving and Riding for Infrastructure and Vehicle safety and Environment.* European Commission.

European Parliament and Council of European Union. (2016). *Regulation (EU) 2016/679.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

FESTA Handbook, Version 7. (2018). Retrieved from https://connectedautomateddriving.eu/wp-content/uploads/2019/01/FESTA-Handbook-Version-7.pdf

Fridman, L., Brown, D. E., Glazer, M., Angell, W., Dodd, S., Jenik, B., . . . Seppelt, B. (2019). MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction With Automation. *IEEE Access, 7*, 102021-102038.

Giustiniani, G., Carrocia, R., & Robibaro, M. (2010). *DaCoTA Deliverable 2.2. Specification of Data System.* Loughborough, United Kingdom: European Commission Directorate-General for Mobility and Transport.

Group, i.-S. (n.d.). Retrieved from https://i-sense.iccs.gr/projects/finished-projects/item/116-2besafe

Hancox, G., Talbot, R., Pilkington-Cheney, F., Filtness, A., Brijs, K., Brijs, T., . . . Yang, K. (2020). *Simulator & Field Study Organisation & Support. Deliverable 5.1 of the EC H2020 project i-DREAMS.*

Hart, E., Barmby, P., LeBauer, D., Michonneau, F., & Mount, S. (2016). Ten Simple Rules for Digital Data Storage. *PLOS Computational Biology, 12*(10).

i-Sense Group. (n.d.). Retrieved from https://i-sense.iccs.gr/projects/finished-projects/item/116-2besafe

Katrakazas, C., Michelaraki, E., Yannis, G., Kaiser, S., Senitschnig, N., Ross, V., . . . Taveira, R. (2020). *Toolbox of recommended data collection tools and monitoring methods and a conceptual definition of the safety tolerance zone. Deliverable 3.2 of the EC H2020 project i-DREAMS.*

Klauer, S. G., Perez, M., & McClafferty, J. (2011). Naturalistic Driving Studies and Data Coding and Analysis Techniques. In B. Porter, *Handbook of Traffic Psychology* (pp. 73-85). Norfolk, VA, USA: Elsevier.

Knoefel, F., Wallace, B., Goubran, R., & Marshall, S. (2018). Naturalistic Driving: A Framework and Advances in Using Big Data. *Geriatrics, 3*(16).

Machiani, S. G., & Abbas, M. (2016). Safety surrogate histograms (SSH): A novel real-time safety assessment of dilemma zone related conflicts at signalized intersections. *Accident Analysis and Prevention, 96*, 361-370.

Muckell, J., Hwang, J. H., Lawson, C. T., & Ravi, S. S. (2010). Algorithms for compressing GPS trajectory data: an empirical evaluation. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University – Computer and Information Sciences, 30*, 431-448.

Pilkington-Cheney, F., Talbot, R., Hancox, G., Filtness, A., Cuenen, A., Polders, E., . . . & Al Haddad, C. (2020). *Experimental protocol. Deliverable 3.4 of the EC H2020 project i-DREAMS.*

Sagberg, F., Eenink, R., Hoedemaeker, M., Lotan, T., van Nes, N., Smokers, R., & Winkelbauer, M. (2011). *Recommendations for a large-scale European naturalistic driving observation study. PROLOGUE Deliverable D4.1.*

Tselentis, D. (2018). Benchmarking Driving Efficiency using Data Science Techniques applied on Large-Scale Smartphone Data. Athens, Greece.

van Nes, N., Bärgman, J., Christoph, M., & van Schagen, I. (2019). The potential of naturalistic driving for in-depth understanding of driver behavior: UDRIVE results and beyond. *Safety Science, 119*, 11-20.

van Schagen, I., Ruth Welsh, A., Backer-Grondahl, M., Hoedemaeker, T., Lotan, A., Morris, F. S., & Martin, W. (2011). *Towards a large scale European Naturalistic Driving study: final report of PROLOGUE: deliverable D4.2.* Loughborough, United Kingdom: University of Loughborough.

van Schagen, I., Welsh, R., Backer-Grondahl, A., Hoedemaeker, M., Lotan, T., Morris, A., & Winkelbauer, M. (2011). *Towards a large scale European Naturalistic Driving study: final report of PROLOGUE: deliverable D4.2.*

Vlahogianni, E. I., Yannis, G., & Golias, J. C. (2014). Detecting powered-two-wheeler incidents from high resolution naturalistic data. *Transportation research part F: traffic psychology and behaviour, 22*, 86-95.

Wallace, B., Goubran, R., Knoefel, F., Marshall, S., Porter, M., Harlow, M., & Puli, A. (2015). Automation of the Validation, Anonymization and Augmentation of Big Data from a Multi-year Driving Study. *2015 IEEE International Congress on Big Data*, (pp. 608-614). Santa Clara, CA, USA.

Wu, K.-F., Aguero-Valverde, J., & Jovanis, P. P. (2014). Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level. *Accident Analysis and Prevention, 72*, 210-218.