D3.5 Standard protocol for handling big data

Interview with Christelle Al Haddad

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 (STZ) for driving. In-vehicle interventions and post-trip interventions will help to maintain the STZ as well as provide feedback to the driver. This conceptual framework will be tested in simulator studies and on-road trials in Belgium, Germany, Greece, Portugal, and the United Kingdom (UK) with around 600 participants representing car, bus, truck and rail drivers.

During the experiments, large amounts of data will be generated, from the different data collection tools, originating from different modes and countries. The aim of this deliverable is to therefore provide the necessary protocols for handling this "Big Data". More specifically, this deliverable will:

- Provide a methodology for the handling of big data, based on learnings from previous studies/projects; particularly naturalistic driving studies (NDS) 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 need to be continuously controlled throughout the project and in accordance with the Data Management Plan (DMP)[1] and the Data and Knowledge Management Committee[2]. Project partners at different countries are responsible for their own data collection, and obliged to follow the proper standards, while consulting with their national and local authorities. Overall, deliverable 3.5 should be a living document, updated where applicable into the necessary steps, and serving as a guideline for how to best handle the big data generated throughout the project.

Christelle, you are the author of this deliverable. You started your work with reviewing previous naturalistic driving studies (NDS) in Europe? Which projects did you review exactly and are they comparable to what you do in i-DREAMS?

CHRISTELLE AL HADDAD: *"We reviewed 10 European naturalistic driving projects[3] and focused on how they handled data collection, data preparation, data storage, and the legal and ethical considerations. It is important to mention that most of these studies focused on passenger cars and trucks and in some cases on powered two-wheelers (motorcycles and scooters). Although the reviewed projects do not always have the same scope as in i-DREAMS (cars, trucks, buses & rail), it goes without saying that it is crucial to take the lessons learned from these projects into account."*

[1] The second and third update of the DMP are confidential deliverables.

[2] The Data and Knowledge Management Committee is a supervisory body with a particular focus on data flow management and protection.

[3] The following naturalistic driving projects were evaluated: AOS (2007-2009), SeMiFOT (2008-2009), euroFOT (2008-2011), TeleFOT (2008-2012), INTERACTION (2008-20212), 2BeSAFE (2009-2011), PROLOGUE (2009-2011), UDRIVE (2012-2016), Track & Know (2019-2022).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761

In the report you devote an entire chapter to methods for managing Big Data. What methods are you talking about exactly?

CHRISTELLE AL HADDAD: *"It comprises all the methods for data collection, preparation (including data processing), storage, access and sharing) in naturalistic driving studies. Many of the methods that we describe in the deliverable often reflect common sense, but of course they are more than that. They are based on what we learned from previous projects and the goals we want to achieve in our project. We developed methods to streamline data collection, (pre-)processing, storage, access and sharing, while of course keeping in mind the different ethical and legal considerations."*

Can you give us some examples of the methods you will apply in i-DREAMS related to data collection for example?

CHRISTELLE AL HADDAD: *"Of course, these methods have to do with how we collect data and which rules we need to follow. We will use the same Data Acquisition System for example in all experiments to reduce issues with data incompatibility. Since most of our data, collected with sensors, is quantitative, supplementing it with additional layers of data including survey and GPS data, it becomes a challenge to dispose of it properly. We use specific techniques to correctly safeguard personal information before linking it to the experiment. And of course, we only start collecting data after consulting and receiving approval from the respective ethical committees and data protection officers in the different data collection sites, and after receiving the signed consent of participants to take part in the study, allowing us to process their data for the purpose of research. Another important aspect to mention for instance is the data sharing agreement that was drafted for the exchange of data between partners."*

This raises many new questions in my mind. You refer to a specific Data Acquisition System … which one is that?

CHRISTELLE AL HADDAD: *"For collecting data, we need to distinguish between simulator data and data from the on-road trials. When it comes to the latter, we use our so-called i-DREAMS-system. This system is composed of several devices that are marketed by CardioID (one of our tech partners in i-DREAMS) and used for collecting data and then implementing real-time interventions in the vehicle. The different devices are connected with a gateway that gathers and centralizes information from other components and handles data connectivity and transmission.*

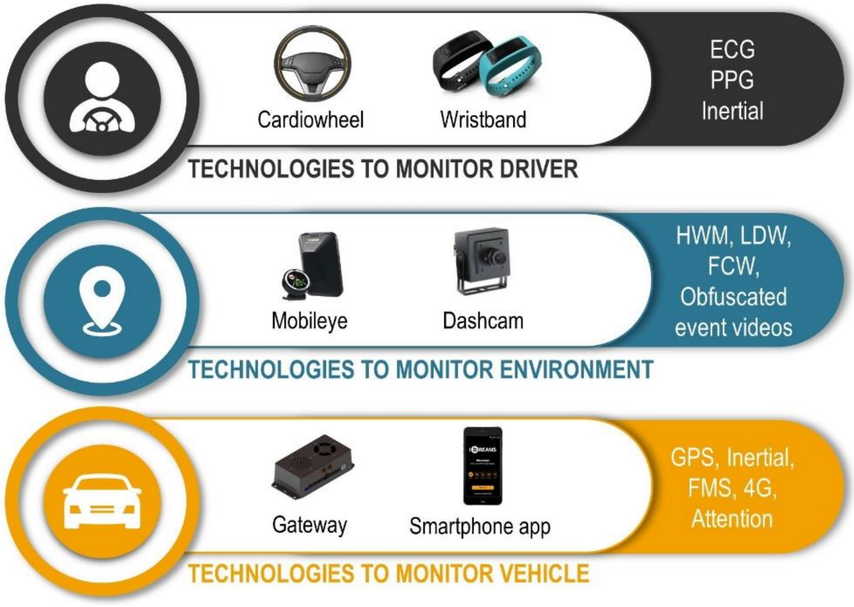
Cardiowheel
Wristband
ECG
PPG
Inertial
TECHNOLOGIES TO MONITOR DRIVER
Mobileye
Dashcam
HWM, LDW, FCW, Obfuscated event videos
TECHNOLOGIES TO MONITOR ENVIRONMENT
Gateway
Smartphone app
GPS, Inertial, FMS, 4G, Attention
TECHNOLOGIES TO MONITOR VEHICLE

Figure 1: i-DREAMS Monitoring technologies

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761

For the post-trip interventions, we use technology from OSeven PC (another tech partner in i-DREAMS) which is a state-of-the-art Android and iOS-based smartphone application that also monitors driving behaviour of individuals using a variety of parameters. Data from the Oseven PC app is fused with CardioID technology for more accurate prediction of driving behaviour. Furthermore, other app technology comes from UHasselt, the project coordinator, that provides the driver with feedback on their personal performance and challenges and motivates the driver via gamification features."

What practical lessons did you learn from the ND-Studies you evaluated?

Christelle Al Haddad: "The most important lessons regarding data collection are that it is advisable to minimise the number of vehicle models that we equip with our Data Acquisition System to reduce the complexity of installation and de-installation. Furthermore, we learned the importance of centralizing responsibilities in terms of coding, processing and analysis to create dataset consistency.

With respect to data storage, we learned quite a few practical things such as the importance of pre-processing and enriching the data and storing data in open formats to increase accessibility. Of course, there is also the importance of a systematic back-up system, and the clear definition of data so everyone speaks the same language. Transferring data from local to central servers should be done electronically and without manual intervention. Files that do require manual intervention (e.g. paper questionnaires) should be stored after transforming them to the electronic version. And of course, it is also important to consider processing speed and links to databases when thinking about the data architecture.

Then lastly, regarding the legal and ethical issues, there is GDPR to consider, in addition to all the national and local regulations, which might differ from one country to another. Participants are therefore asked to sign an informed consent before they start the experiment and we ensure that the necessary procedures are in place to safeguard personal information."

I can imagine that when working out all the necessary procedures, it is important that all partners involved have a clear idea of what their own responsibilities are in this big picture. Can you briefly elaborate on these responsibilities?

Christelle Al Haddad: "It is important to distinguish between technology partners and data collection partners, but also between simulator experiments and field trial experiments.

CardioID and Oseven PC are our technology partners. DriveSimSolutions (DSS) is our simulator partner. DSS integrates technologies and provides simulators that log data, which is locally stored on the simulator PC. Partners using another driving simulator (NTUA and LOUGH) need to follow a similar approach.

The field trial partners are UHasselt, TUM, NTUA, BARRA and LOUGH. They are responsible for the logistics to set up and follow-up experiments taking place at their respective sites. The data processors (UHasselt, LOUGH, TUM, TUD and NTUA) will access the collected data to analyse and test hypotheses derived from the research questions. UHasselt also processes data via the post-trip intervention framework by generating scores and interventions based on collected data. I can of course provide more in-depth information on all these responsibilities, but this kind of gives you an idea of the big picture."

This project received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 814761

When all that data is collected, how do you proceed from there?

CHRISTELLE AL HADDAD: *"For the simulator tests, data is automatically stored locally. Partners are free to choose their own storage engines (databases, filesystems) for local storage. Local refers to systems that are not accessible through API. Collected data is pseudonymized (with a plan to completely anonymize the data after completing the experiments) so that the personal information is safeguarded."*

What does API mean?

CHRISTELLE AL HADDAD: *"API is the acronym for Application Programming Interface, which is a software intermediary that allows two applications to talk to each other. APIs are mechanisms that enable two software components to communicate with each other using a set of definitions and protocols. For example, the weather bureau's software system contains daily weather data. The weather app on your phone "talks" to this system via APIs and shows you daily weather updates on your phone. The same principle is applied to the collected i-DREAMS data. Within i-DREAMS, a back-end database or a back-office component is developed to store centrally raw and processed data from other components. Partners can then comfortably access data from the back-office through a web API."*

How does locally stored data end up in that back-office database?

CHRISTELLE AL HADDAD: *"As mentioned, the vehicle data is automatically made available to partners, via the developed centralized back-office. For that, partners do not need to upload anything manually. However, additional information pertaining to the experiments, which partners might need to provide to aid the analysis, would need to be manually uploaded to the back-office server in a specific format. This is also the case for the questionnaire data, but also the simulator data files. Other partners can get access to the data through the developed server web API. A joint agreement between partners on 'sharing personal data' arranges how personal data is used among consortium partners until the end of the project. Moreover, after the project end, the aim is to make accessible an anonymised portion of the data (complying with GDPR and other regulations of course), so that other researchers could benefit from it."*

You already mentioned that personal data is safeguarded at all times. What do you do precisely to ensure this?

CHRISTELLE AL HADDAD: *"Data is first pseudonymised. This means that a unique identifier is appointed to each participant. The document linking the identifier to the participant is not saved on a (central cloud) server, but locally. Access to this document is strictly protected and personal data is not stored longer than necessary (max 5 years after project end). After these 5 years data is anonymised. This means that the previously appointed unique identifier that connects data in the partners' databases with the personal data of the participant, is replaced by a random number. So, there is no longer any link with the data and the personal information. When this is done, (part of) the data is ready to be made available in an open-source platform which is one of the project objectives."*

This project received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 814761

My last question relates to data accessibility. What is included in your data access procedure?

CHRISTELLE AL HADDAD: “There are a couple of things we pay special attention to. We first define different user types, each with different access rights (e.g. superadmin, admin, user…). The pseudonymised data is made accessible to the consortium partners according to the stipulations set out in joint data agreements. Our research data follow the FAIR[4] guidelines. Aggregated data from CardioID is made available via a web API and an anonymised portion of the data (a few datasets) will be made available and offered to third parties at the end op the project (complying with GDPR) on a digital repository.”

What mainly stays with me from our conversation is that there is a tremendous amount of details you need to take into account to handle all the aspects related to big data. I am impressed with how thoughtful you have been at every step. I wish you the best of luck with all the experiments!

Edith Donders
Discom Manager

[4] FAIR stands for Findable, Accessible, Interoperable and Reusable.

This project received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 814761

Report 3.5 is part of WP3:
Operational design of i-DREAMS
Download the report here

Researcher in the spotlight

CHRISTELLE AL HADDAD

Graduated as *a Civil Engineer (BEng.)* from the *American University of Beirut (2015)*, and as a *Transportation Engineer (MSc)* from the *Technical University of Munich (2018)*.

Employed at the *Technical University of Munich*, the *Chair of Transportation Systems Engineering since 2019*.

Passionate about *culture, food, people, languages, and music (singing in particular).*

Tasks in i-DREAMS: *Leading the operational work package within i-DREAMS (WP5), in particular the data collection in Germany, but also the coordination of the different tasks TUM is responsible for (back-end, algorithm development, etc.)*.