Building a Data Management Toolchain for a Level 3 Vehicle Automation Pilot

Ben Nagy [1], Johannes Hiller [2], Nisrine Osman [3], Sami Koskinen [4], Erik Svanberg [5], Francesco Bellotti [3*], Riccardo Berta[3], Ahmad Kobeissi [3], Alessandro De Gloria [3]

[1] Assisted & Automated Driving, Jaguar Land Rover, Banbury Road, Gaydon, CV35 0RR, United Kingdom
[2] Institute for Automotive Engineering, RWTH Aachen University, Steinbachstr. 7, 52074 Aachen, Germany
[3] DITEN, Università degli Studi di Genova, Via Opera Pia 11/A, 16145, Genova, Italy
[4] Automated vehicles, VTT, Visiokatu 4, 33720 Tampere, Finland
[5] SAFER, Chalmers University of Technology, Lindholmspiren 3A, 417 56 Göteborg, Sweden
[*]franz@elios.unige.it

Abstract: L3Pilot is the first comprehensive test of ADFs with hands-off the wheel on public roads across Europe. L3Pilot will test ADFs in 100 cars with 1,000 drivers across 10 different countries in Europe. The tested functions will be mainly of SAE automation level 3, some of them of level 4. This paper describes the data management toolchain we have designed and developed in order to exploit pilot data for answering a set of research questions about evaluation of such aspects as: technical and traffic, user acceptance, impact, socio-economic impact. The toolchain, supporting various confidentiality levels (prototype vehicle owner, consortium, public), has been designed to allow cross-vehicle owner data management, with efficient storage of data and its iterative processing with a variety of analysis and evaluation tools. Most of the tools in the data processing chain have been developed to a prototype version, tested in lab and are ready to be deployed for the pre-pilots in the various sites.

1. Introduction

Analyzing road-test data is of paramount importance for developing autonomous vehicles Not only can test data reveal performance values and issues [1], but also indicate user behaviour and acceptance, which is a key factor for designing useful and successful products.

The L3Pilot research project aims at testing the viability of automated driving as a safe and efficient means of transportation on public roads. It will focus on large-scale piloting of SAE Level 3 functions, with additional assessment of some Level 4 functions [2]. The functionality of the systems will be exposed to variable conditions with 1,000 test subjects, 100 cars, by 12 vehicle owners (either Original Equipment Manufacturers, or suppliers), across 10 European countries, including cross-border routes. The technologies being tested cover a wide range of driving scenarios, including parking, overtaking on highways and driving through urban intersections. The tests will provide valuable data for evaluating technical aspects, user acceptance, driving and travel behaviour, as well as impact on traffic and safety.

From a methodological viewpoint, the project uses the FESTA method [3, 4], driven by a set of research questions and hypotheses that have been published recently in Deliverable 3.1 [5]. Research questions will be revised as more information on the tested automated driving functions is available, and as further work is completed on matching the experimental procedures and evaluation methods to the research questions.

Targeted research questions are grouped in four clusters. Each research question is further divided in sub-questions, that are directly mapped to data analysis:

- **Technical and traffic evaluation.** E.g. what is the impact of Automated Driving Function (ADF) on the interaction with other road users?
- **User and acceptance evaluation.** E.g. what is the user experience?
- **Impact evaluation.** E.g. what is the impact of ADF on travel behaviour? (Exposure)
- **Socio-economic impact evaluation.** E.g. what are the socio-economic impacts of ADF?

[5] also defines the performance indicators needed to answer the research sub-questions. The definition of performance indicators required for the evaluation supports the selection of piloting tools and testing of the data processing and analysis chain, before full-scale, on-road testing starts.

In such an extensive pilot, data come from a multitude of heterogeneous sources, and have to be processed by a variety of tools for analysis and evaluation. Recorded data concerns different in-vehicle sensors/systems and are enriched with external data sources. Signal time-series are also supplemented with video streams recording the driver activity.

In this extended abstract we describe the toolchain we have designed and are developing in order to deal with the data management throughout the whole pilot project.

2. Related work

Data acquisition and telemetry are, since long time, key factors for quality and performance in vehicle development and management [6]. Recently, connectivity has been introduced in automotive production series, making vehicles highly mobile Internet of Things (IoT) nodes. In this context, [7] presents the Common Vehicle Information Model as a harmonized data model allowing a common understanding and generic representation, brand-independent throughout the whole data value and processing chain. Since the volume of data collected from fleets of vehicles using telematics services can be very high, it is important to design the systems and frameworks in a way that is highly scalable and efficient. [8] explore the opportunities of leveraging Big Automotive Data for knowledge driven product development, and present a technological framework for capturing and online analyzing data from connected vehicles.

Concerning the data format, [9] describes an approach to combine standards specified for the automotive test data management with the widely used Unified Modeling Language (UML).

Over the past decade a large number of Field Operational Tests (FOT) have been conducted to test Intelligent Transport Systems in real traffic conditions with thousands of drivers (e.g., euroFOT [10], that also investigated whether automatic video processing could improve naturalistic data analyses [11]). In order to ensure scientifically sound studies a FOT methodology was developed in the FESTA project, mainly centred on three focuses: the user, the vehicle or the context [4]. L3Pilot follows this methodology.

Several collaborative industrial research projects have been conducted in Europe concerning the first levels of automation. The Adaptive project developed several automated functions offering different levels of assistance, including partial, conditional, and high automation [12]. Drive C2X investigated cooperative awareness enabled by periodic message exchange between vehicles and roadside infrastructure [13]. The FOT-Net Data project has prepared a Data Sharing Framework including hands-on recommendations on how to prepare for and perform data sharing of transport research data [14]. TEAM focused on apps for collaborative mobility [15].

3. Overall system data architecture

The data processing flow involves several different layers, beginning with raw data acquisition from the pilot vehicles and ending with the analysis and display of the results. Fig. 1 shows an outlook of this process, which will be described in the reminder of this article.

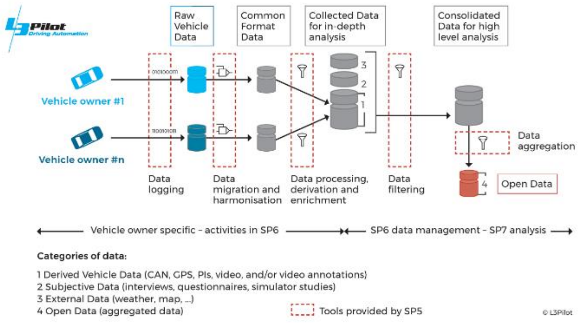

L3Pilot Driving Automation
Raw Vehicle Data
Common Format Data
Collected Data for in-depth analysis
Consolidated Data for high level analysis
Vehicle owner #1
Vehicle owner #n
3
2
1
4
Data aggregation
Open Data
Data logging
Data migration and harmonisation
Data processing, derivation and enrichment
Data filtering
Vehicle owner specific - activities in SP6
SP6 data management - SP7 analysis
Categories of data:
1 Derived Vehicle Data (CAN, GPS, PIs, video, and/or video annotations)
2 Subjective Data (interviews, questionnaires, simulator studies)
3 External Data (weather, map, ...)
4 Open Data (aggregated data)
Tools provided by SP5
© L3Pilot

Fig. 1. *High level schema of the overall data management architecture*

Fig. 1 highlights four layers which correspond to different levels of data access and confidentiality. The left-most one is the vehicle owner layer, involving proprietary data logging from the vehicle (for vehicle owner we intend the company who set up the vehicle prototype. Which is either an original equipment manufacturer, OEM or a supplier). Filtered data, according to the vehicle owner confidentiality requirements and policies, is then converted to a common data format [16], in order to allow usage of the same tools across all the analysis and evaluation partners in the project, as described in the next sections. These data will be made available to a "selected partner", a trusted project partner, one or more for each vehicle owner, who is responsible for a detailed analysis. The third step consists of aggregating data for higher level analyses, and feeds the de-identified consortium and public databases, as described in section 7.

4. Data logging, conversion to the common format and storage in HDF5 files

The first step consists of logging data, in a proprietary format, from the vehicular buses. Logged data is then converted in the L3Pilot Common Data Format (CDF), which we designed in order to allow a common processing of heterogeneous source data [16]. Conversion is made through MATLAB, Python or C++ scripts which produce HDF5 files. One HDF5 file is produced for every pilot vehicle trip. HDF5 is a binary file format that we decided to adopt given its ability to contain and compress large and complex structured data. HDF5 also involves a data model and software libraries for storing and managing data. Portability is also important for exploiting the rising potential within artificial intelligence (e.g., automatic scene detection and video annotation).

In L3Pilot, an HDF5 file is organized in an extensible set of datasets, as reported in the following bullet list. For each section, data is recorded row by row in a timeline with a 10 Hz sampling frequency.

- **Ego vehicle.** This dataset contains the main signals of the vehicle. Examples include: speed, acceleration, brake pressure, etc.

- **Positioning.** This dataset contains information about the position and heading of the vehicle.
- **Objects.** This dataset contains information about the objects that surround the vehicle (e.g., cars, bicycles, pedestrians). Info includes classification, longitudinal and lateral positions, velocity, angular rates, etc.
- **Lane lines.** This dataset contains information about the lane markings. Info includes lateral position, curvature, etc.
- **External.** This dataset contains information about the external environment. Info includes number of lanes, road type, speed limit, etc.
- **Metadata.** Descriptive information about the trip is stored for each file. It includes information about the driver, vehicle, timing, type of experiment, etc.

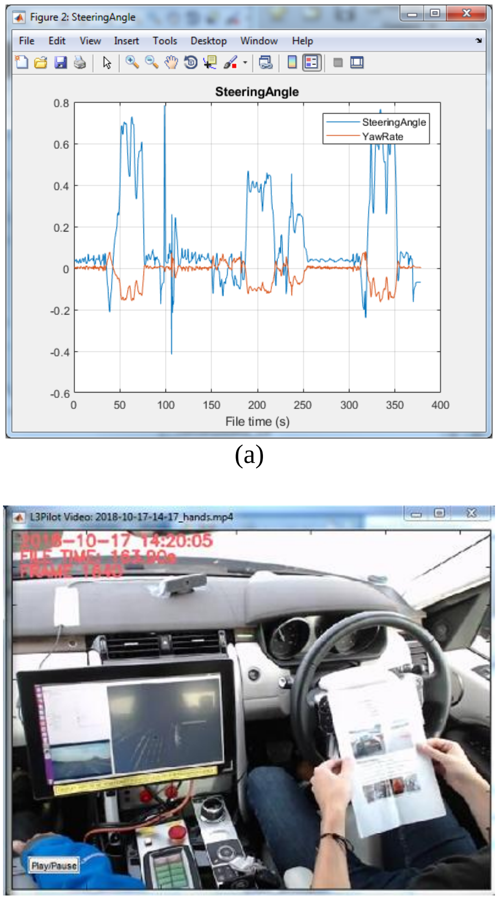
Data converted in the common data format is next checked for correctness and consistency. This step is performed through a data quality check tool that parses the HDF5 files in order to guarantee that all the defined data structures are present and correct, so that they are compatible with the post-processing and evaluation tools.

De-identification (or pseudonymization) is necessary for pilot driver privacy and, when sharing data, for protecting vehicle owner confidentiality. It is managed by the vehicle owners and first concerns the trip ID and the user ID. The ID is an 8-character string, obtained through a simple procedure, based on a SHA-256 hashing. Source information (e.g., driver name, date of birth, trip place, vehicle owner, etc.), integrated with a secret word for “salting”, is processed through a hash function (e.g., SHA-256), that generates a 64 character identifier. For the purposes of this project, it is sufficient to extract the left-most 8 characters to have the driver/trip ID. Using this ID, vehicle manufacturers can then track their participants and trips through the HDF5 files, subjective questionnaires and consortium database which are described later. When sharing data beyond selected partners, further sensor-level data anonymization steps may be deemed necessary to protect product IPR.

5. Data post-processing and enrichment

The second main step is represented by data enrichment. This involves the processing – through MATLAB or python scripts – of the data present in an HDF5 file in order to obtain additional information, particularly related to the methodological requirements and the project research questions [5], such as:

- **Derived Measures (DMs).** These are quantities computed from the above-mentioned raw signals. Examples of derived measures include: time headway, time-to-collision, etc. The calculated derived measures are then stored in the derivedMeasures HDF5 struct.
- **Performance Indicators (PIs).** PIs aggregate information from one or more timelines in an HDF5 file. In general, performance indicators are single values and no longer time-series data. PIs are specified in the methodology document [5]. Examples include: distribution of difference between speed and speed limit, time spent in automated mode, etc. The calculated performance indicators are then stored in the HDF5 performance Indicators struct.
- **Events and scenarios.** These are stereotypical situations that may happen during a drive. They are recognized by analysing the raw signal timelines. Scenarios are saved in the scenario timeline data structure. Examples of detected scenarios include: uninfluenced driving, lane change, merge, cut-in, etc. Events are unforeseen incidents that may occur during a trip. Examples include near-crashes and accidents.



(a)

(b)

Fig. 2. *Snapshots from the viewer tool for video annotation*

Another important post-processing step is given by the video annotation. To this end, a MATLAB-based viewer tool has been developed. Through this software interface a user can display one or more video signals stored in an HDF5 file, watching the current signal evolution and inserting annotations about the recording. Examples of annotations include the tasks being performed by the driver, e.g., eating/drinking, reading/writing, smoking, mobile phone, no task, etc. The project will also test automatic feature extraction from video by machine learning algorithms.

All this additional information is incrementally added in the HDF5 file, so that a new version is enriched with computed data, beside the original signal values.

6. Subjective data tool

Until now, we have dealt with data about vehicles and environment. Another important aspect is given by subjective data collection. This is achieved with an online survey tool which pilot vehicle users will use in order to reflect and report about their test experience, mainly addressing the impact evaluation and socio-economic impact evaluation research questions. In order to support execution and management of the subjective data collection, we chose the LimeSurvey commercial tool [17]. Three contextualized questionnaires were developed: Urban Pilot, Traffic Jam/Motorway and Parking Functions. The questionnaires are translated in all the pilot sites' local languages. Survey results are exported to feed the "selected partner" and for the Consortium and Public database.

7. Consortium and public database

The final step of the data processing involves the preparation of information for the consortium and public databases. The goal is to collect aggregated and anonymized information in these databases from all the HDF5 files (i.e., pilot trips) and make it available to the whole consortium (while the "selected partner" analysis is restricted to each specific vehicle owner) and the general public, respectively.

From an architectural point of view, we have developed the prototype of a software platform for supporting project-level data aggregation and presentation. The platform relies on a MongoDB non-relational database, which is accessed through Node.js. The platform exposes a set of RESTful APIs [18, 19] for inserting and retrieving data. A Graphical User Interface (GUI) is being implemented in order to allow a user-friendly access to data. Different user roles have been defined for administrators, site managers, vehicle owners, consortium and public users. Such roles implement different data read/write rights, in order to meet the project information confidentiality requirements.

A MATLAB script has been implemented, which allows vehicle owners and test site managers to extract relevant data from HDF5 files and post them to the database. Subjective data is also inserted, to keep into account the driver profile. Information from HDF5 files is of the following kind:

- **Statistical data** from the single timeline signals. Statistical descriptors include: average, stdev, median, min, max, histogram, correlations (between two signals). In order to support a finer grain analysis, the MATLAB script allows extracting information related to different events (e.g., automatic drive function ON, Hands on Steering Wheel, etc.) and scenarios (e.g., uninfluenced driving, cut-in, etc.). A combination of events/scenarios (e.g., automated driving function ON & uninfluenced driving) is also supported.
- **Performance Indicators.** These are the PIs presented in section 5, and already included in the enriched version of the HDF5 files.

8. Status and results

As of April 2019, most of the tools in the data chain have been developed to a prototype version, tested in lab and are ready to be deployed for the pre-pilots in the various sites.

9. Conclusions

The L3Pilot industrial research project will pilot 100 L3 SAE automation vehicles, by 12 vehicle owners, with 1,000 test subjects, across 10 European countries. This extended abstract has described the data management toolchain we have designed and developed in order to exploit pilot data for answering a set of research questions about evaluation of such aspects as: technical and traffic, user acceptance, impact, socio-economic impact.

The toolchain, supporting various confidentiality levels (vehicle owner, consortium, public), has been designed to allow cross-vehicle owner data management, with efficient storage of data and its iterative processing with a variety of analysis and evaluation tools. Most of the tools in the data processing chain have been developed to a prototype version, tested in lab and are ready to be deployed for the pre-pilots in the various sites.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 723051. The sole responsibility of this publication lies with the authors. The authors would like to thank all partners within L3Pilot for their cooperation and valuable contribution.

References

[1] Li, L., Huang, W., Liu, Y., Zheng, N. and Wang, F.,: 'Intelligence Testing for Autonomous Vehicles:

A New Approach', in *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 2, pp. 158-166, June 2016.

[2] SAE,: 'Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems (J3016)'. Technical Report, SAE International (2016)

[3] 'Updated version of the FESTA Handbook', https://connectedautomateddriving.eu/wp-content/uploads/2017/08/2017-01-31_FOT-NET_FESTA-Handbook.pdf, accessed 8 January 2019

[4] Barnard, Y., Innamaa, S., Koskinen, S., Gellerman, H., Svanberg, E., Chen, H.,: 'Methodology for Field Operational Tests of Automated Vehicles', Transportation Research Procedia, Volume 14, 2016, Pages 2188-2196,

[5] L3Pilot Deliverable D3.1: 'From Research Questions to Logging Requirements', https://l3pilot.eu/detail/news/l3pilot-research-questions-and-hypotheses-available-for-download-1/ , accessed 8 January 2019

[6] Chandiramani, J. R., Bhandari, S., and Hariprasad, S. A.: 'Vehicle Data Acquisition and Telemetry', 2014 Fifth International Conference on Signal and Image Processing, Bangalore, India, 2014, pp. 187-191.

[7] Pillmann, J., Sliwa, B. and Wietfeld, C.,: 'The AutoMat CVIM - A Scalable Data Model for Automotive Big Data Marketplaces', 2018 19th IEEE International Conference on Mobile Data Management (MDM), Aalborg, 2018, pp. 284-285.

[8] Johanson, M., Belenki, S., Jalminger, J., Fant, M., and Gjertz, M.,: 'Big Automotive Data: Leveraging large volumes of data for knowledge-driven product development', 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, 2014, pp. 736-741.

[9] Bartz, R.,: 'Mapping data models of the standardized automotive testing data storage to the unified modeling language', 2013 IEEE International Conference on Industrial Technology (ICIT), Cape Town, 2013, pp. 1327-1332.

[10] Burzio, G., Mussino, G., Tadei, R., Perboli, G., Dell'Amico, M. and Guidotti, L.,: 'A subjective field test on lane departure warning function in the framework of the euroFOT project', *2009 2nd Conference on Human System Interactions*, Catania, 2009, pp. 608-610.

[11] Dozza, M., Pañeda González, N.,: 'Recognising safety critical events: Can automatic video processing improve naturalistic data analyses?', Accident Analysis & Prevention, Volume 60, 2013, Pages 298-304

[12] 'Adaptive Project Final Report', http://www.adaptive-ip.eu/files/adaptive/content/downloads/AdaptIVe-SP1-v1-0-DL-D1-0-Final_Report.pdf, accessed 9 January 2019

[13] Boban, M., and d'Orey, P. M.,: 'Measurement-based evaluation of cooperative awareness for V2V and V2I communication', *2014 IEEE Vehicular Networking Conference (VNC)*, Paderborn, 2014, pp. 1-8.

[14] Gellerman, H., Svanberg, R., Barnard, Y.,: 'Data Sharing of Transport Research Data', Transportation Research Procedia, Volume 14, 2016, Pages 2227-2236,

[15] Bellotti, F., Kopetzki, S., Berta, R., Lytrivis, P., Amditis, A., Raffero, M., Aittoniemi, E., Basso, R., Radusch, I., De Gloria, A.,: 'TEAM applications for Collaborative Road Mobility' , IEEE Transactions on Industrial Informatics, DOI: 10.1109/TII.2018.2850005

[16] Hiller, J., Svanberg, E., Koskinen, S., Bellotti, F., Osman, N.,: 'The L3Pilot Common Data Format – Enabling Efficient Autonomous Driving Data Analysis', to appear in Proceedings of NHTSA ESV 2019, Eindhoven

[17] LimeSurvey,: https://www.limesurvey.org/, accessed 8 January 2019

[18] Steiner, T. and Algermissen, J.,: 'Fulfilling the hypermedia constraint via http options, the http vocabulary in rdf, and link headers', in proceedings of the second international workshop on RESTful design. ACM, 2011, pp. 11–14.

[19] Richardson, L., Amundsen, M., and Ruby, S.,: 'RESTful Web APIs', O'Reilly Media, Inc.", 2013