Abstract

This paper describes the content of a proposed Data Sharing Framework for data collected in Field Operational Tests and in Naturalistic Driving Studies. These projects gather data regarding the driver behaviour in relation to the vehicle, ITS and traffic environment during normal driving. Huge amounts of data are stored from these tests and could be re-used in many different fields of road transport research, e.g. safety, automated driving or mobility, to understand the human behaviour in different traffic environments. The framework includes topics such as description of data and metadata, data protection, research support services and topics to address in agreements to make data re-use possible. The purpose of the framework is to facilitate global data sharing and re-use and thereby enhance the availability of data for future research in ITS.

Keywords:
Data sharing, Field Operational Tests, Naturalistic Driving Data
1 Introduction

During the past 15 years, we have seen a fast growth in the number of Field Operational Tests (FOT) and Naturalistic Driving Studies (NDS) performed worldwide. The need to better understand the benefits of safety systems and the factors behind the occurrence of incidents and accidents have been the main driving forces. The availability of technology, such as affordable video capture solutions and data storage, has been a requirement and facilitator for the development of the FOT and NDS methodologies.

The data, which has mainly been collected through naturalistic driving by volunteer drivers, have been used to answer a wide range of research questions. The size of the datasets varies, from gigabytes to several petabytes.

The largest datasets have so far been collected in the US (e.g. IVBSS (1), SHRP2 (2) and the Safety Pilot project (3)) and in Europe (e.g. euroFOT (4), DRIVE C2X (5) and the on-going UDRIVE project (6)). In Japan, large ITS datasets have been collected, e.g. (7), and also Australia has several interesting datasets, e.g. (8). Data collection is on-going in China (9). These datasets provide a huge resource for answering research questions and the interest in data re-use is increasing along with the awareness of the substantial effort and funding needed to do these FOT/NDS.

There are different views on the value of data sharing depending on if you are a data provider or a data user. Substantial efforts have been made to collect data and build up the data infrastructure and tools. It is therefore important to find mutually beneficial business models for both providers and users, for re-use of the data to become a reality. Further, a concept of procedures, agreements and technical structures to make data available would facilitate an increase in the number of data providers who are interested in opening up their datasets. In some countries, publicly funded projects are required to make the data open for further research. Apart from the more general possibilities to share data, there are different constraints that make it difficult to open up datasets. The legal and ethical requirements in each country, where an organisation is involved in collecting, storing, or analysing data, will have an impact on the data sharing constraints. Also, consortium agreements and consent forms signed by the participants may not have had data sharing in focus when they were written and could make data sharing impossible to others than the project partners. Finally, the availability of funding, both for the new research project as well as for the data provider can set considerable constraints of the re-use of the data.

This paper is based on information collected within the European support action FOT-Net 2 (10) and the further developments done by the succeeding FOT-Net Data (11). The knowledge and experiences gained in FOT and NDS are collected through activities at various conferences and through discussions with a variety of stakeholders from Europe, the US, Japan, Australia and China (12).

2 Data Sharing Framework

The availability of a common Data Sharing Framework (DSF) has the potential to substantially contribute to a larger use of the collected FOT/NDS data. Such a framework should include data sharing pre-requisites that could be integrated into project agreements from the start, as well as procedures and templates to facilitate easier data sharing. The organisations setting up new FOT/NDS projects would then not need to develop the specific...
data sharing content for a certain project, and can instead focus on the project specific questions such as research questions, study design, and data acquisition requirements. Also, researchers wanting to re-use already collected datasets could then utilise a more or less standard application procedure, rely on already performed training that is widely accepted, and plan for the costs that using a specific dataset might cause the project.

Seven areas need to be addressed by a DSF:

1. Agreements within the project collecting data, including consortium agreements, participant agreements and agreements with third party data providers
2. Availability of valid data and metadata, including a “standard” description of the documentation of the data
3. Data protection requirements
4. Security and personal integrity training for all personnel involved
5. Support and research services, to facilitate the start-up of projects and offer research capabilities
6. Financial models to provide funding for the data to be maintained and available and for access provision personnel to be available
7. Application procedures and data sharing agreements

Generally, the data could be either managed by partners from the project where the data was collected or by an external data provider. The need for the different elements in the DSF when providing data is strongly depending on the data categories included in the dataset that is provided.

2.1 Project and participant agreements

The initial process of setting up a project is crucial to the possibilities to share data during and after the project. The main documents to focus on are: 1) the grant agreement, if the project has external funding, including the description of the work; 2) the consortium agreement among the project partners; 3) the participant agreement; and 4) agreements with data providers to the project.

In the grant and consortium agreement, it is important to be aware of the topics and issues to be discussed in relation to data sharing and re-use of data and to focus them already during the project application and a possible negotiation phase. It is especially important to pay attention to the possibilities to provide open data after the project, based on the scope of the project and the data to be collected. The topics that should be addressed are: 1) the ownership and access to data and data tools; 2) distribution of data; 3) the access methods; 4) the research and commercial areas where data usage will be allowed; 5) the post-project (re-)use of data; and 6) the post-project financing.

The agreements with the participants of the FOT/NDS have substantial impact on the possibility to re-use data after the project both by the project partners and by external parties. Thus, it is important that these topics are addressed in the agreement. It is difficult to reach participants after the project has been concluded to ask for additional consent. The participant agreement explains the project to the participant and it is vital that the participant understands the use of the data during and after the project. From a data sharing standpoint, it is especially important to describe: 1) what data is collected; 2) where the data will be stored; 3) who is responsible for the data; 4) who will have access to what data and on what conditions; and 5) the access procedures. As each participant allows the project to follow the participant’s private life for a period ranging from a few weeks up to more than a year, it is important that
they have a solid understanding of what the data could be used for. The participant should make an active consent to the most vital topics for data sharing.

Datasets are often enhanced with data collected from sensor systems bought from suppliers and put on the vehicles, and data from external data providers such as companies providing map data, weather data, or other services. Non-disclosure agreements and contracts should be signed and it is important to be aware of the topics that can affect future research, due to possible restrictions in data use.

2.2 Descriptions of data and metadata

The core of data sharing is that the data provided is valid or at least documented to a level where an assessment of the level of validity of the data could be performed. This is especially problematic if one has not been part of the project and does not know the way the validation was performed in detail, which sensor/version was used and how the data was processed (from raw data). The main problem is usually that the data itself is not sufficiently described.

There are different ways of describing the collected data. One is to cluster the data by the category of data or by ownership. The category usually determines the level of protection, whereas the ownership is more related to the readiness to share the data. If a data type already is jointly owned, it is easier to share it with a wider research community.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Data category</th>
<th>Ownership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaires- and interview data</td>
<td>Personal</td>
<td>Jointly</td>
</tr>
<tr>
<td>Video</td>
<td>Personal</td>
<td>Jointly</td>
</tr>
<tr>
<td>GPS</td>
<td>Personal</td>
<td>Jointly</td>
</tr>
<tr>
<td>Vehicle mounted sensors (eye-tracker, radar, etc.)</td>
<td>Sensor</td>
<td>Jointly/supplier</td>
</tr>
<tr>
<td>V2V and V2I data including “activity” data</td>
<td>System/sensor</td>
<td>Jointly</td>
</tr>
<tr>
<td>Enhancing data – road attributes, weather</td>
<td>Infrastructure/sensor</td>
<td>Jointly/supplier</td>
</tr>
<tr>
<td>“Open” and aggregated internal vehicle data</td>
<td>System/sensor</td>
<td>Jointly</td>
</tr>
<tr>
<td>Closed, confidential internal vehicle data</td>
<td>System/sensor</td>
<td>OEM</td>
</tr>
</tbody>
</table>

One of the most important factors in creating a FOT/NDS dataset that can be reused is the simplicity in which the dataset can be understood. The collected data needs to be described with metadata in such a manner, that a person from a research discipline not familiar with this kind of data would be able to understand the data and any issue related to it. At the same time, it needs to be described in such depth that it is possible to verify whether the data is good enough to be used for specific research.

If the collected data in a project shows a large variability in quality or consists of data collected through separate FOT not using the same data format, the description of the metadata is even more important.

Metadata on a higher level includes information such as the experimental protocol used, the subjects and vehicle collecting the data, and video annotations in the form of a code book which states the rules which the annotators had to follow. At a more detailed level however, the metadata involves all information that describes how the data was generated, how it was
derived, and what other properties it has (e.g. resolution, frequency, resampling strategies, details of algorithms, and how quality metrics were calculated).

It is important that data from projects are available in their original forms, with clearly described formats, directly from the data storage source (e.g. database or file storage). That is, both within and after a project, many users would want to access the data in different ways using different tools. Data description formats should preferably be accepted and used by as large part of the community as possible.

2.3 Data protection

Data protection is the key to create the trust needed between the data provider and the research organisation to make the data owners provide access to their data. If the data provider knows that the research organisation have good and proven procedures in place to keep control of who is accessing the data and the researchers have knowledge about the legislation regarding the handling of Personal Integrity Data (PID) and Intellectual Property Right (IPR) data, the more data the data owner will be willing to share.

The data protection level needed depends on the harm the revealed data could do. There are two categories of data that especially need protection: personal data and data that, if revealed, could potentially harm a commercial company. The first category, the personal data, is defined by the European Directive 95/46/EC Art. 2 (13):

‘Personal data’ shall mean any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

And also defines specifically sensitive personal data in Art. 10:

“1. Member States shall prohibit the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life.

2. Paragraph 1 shall not apply where:

(a) the data subject has given his explicit consent to the processing of those data, except where the laws of the Member State provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject's giving his consent”

The personal data, especially if video is collected, is often the data category that sets the highest demands on the data protection due to the need to preserve the personal privacy.

The second category contains several different types of commercial data, such as weather data, radar data and road attributes, which are used to enhance FOT/NDS data. The provision of these data to projects is usually accompanied by agreements, stating the conditions for access and use. Some data might need protection and it is advisable that the parties discuss and agree on a suitable level of protection.

The suggested data protection requirements described in this paper have the aim to guide data providers, here defined as Data Centres (DC), and research organisations, here defined as Analysis Sites (AS), towards setting up a data protection concept that would meet the regulations and agreements and that would guard the will of the participants as stated in the consent forms. Two sets of data protection requirements are suggested below, one for the DC
and one for the AS. Depending on the data type involved in the data sharing, the needed level of protection will vary. The data protection recommendations are related to datasets including both video and proprietary sensor data. If the data to be shared is anonymised, several of the requirements are not applicable.

Requirements for the DC:

DC1: The DC must document its data protection implementation.
DC2: Data stored and processed at a DC must be protected from unauthorized access.
DC3: Data stored and handled at a DC must be protected from accidental deletion or corruption.
DC4: Confidentiality agreements for any involved personnel must be in place.
DC5: Data protection must be ensured by the DC after end of project.
DC6: Data sent between DC and AS must be encrypted.
DC7: Data downloads, where the downloading organisation becomes responsible for the data, are regulated by the Project Agreement(s) and the informed consent of the driver.
DC8: Data extractions for specific purposes, such as results for reports and publications, must be in accordance with the consent forms and project agreement and the extraction must be documented.

Requirements for the AS:

AS1: The AS organisation must document its data protection implementation.
AS2: The analysis work stations must be physically and logically protected.
AS3: Data users must have received relevant training in data protection and integrity issues.
AS4: A confidentiality agreement for any involved AS personnel must be in place.
AS5: An appointed person at the AS administers access requests and forwards those to the DC data access dispatcher.
AS6: Specified procedures for data extraction, stated by the DC, must be used.
AS7: The data user may only extract or re-distribute data if allowed by the agreements with the DC.
AS8: The project data must not be used for research areas not covered by the consent forms.
AS9: All post-project research must investigate the need for approval.

The documentation of the implementation of the requirements above and the exchange of this information between the DC and the AS is a good practice to create a trustful data sharing environment. The introduction of an external part evaluating the implementation could further enhance the trust for sharing the data.

It is recommended to always consider both the DC and AS requirements, even when a single organisation is acting as both DC and AS.

2.4 Training on data protection related to personal data and IPR

It is important for the implementation of the data protection, to train personnel who are going to analyse FOT/NDS data regarding the local implementation of the security precautions,
such as the data protection procedures and the analysis environment set-up together with more
general information and rules following the specific dataset at hand.

For those datasets where video or other personal data is present, training on integrity issues
needs to accompany the general training on data security. There are different kinds of personal
integrity training available, e.g. the US NIH training course (14), where the data user gets a
certificate at the end of a web course.

2.5 Support and research services

Support and research services are an essential part of data sharing. Support services will assist
the researchers during the process, while the researcher is doing the actual work. Research
services are more targeted towards extracting usable datasets for the data user or even perform
the research itself and provide results.

Tools are an integral part of the support services. The tools could consist of a viewing and
annotation tool, scripts that extract useful datasets from the database, MATLAB, and other
licensed software, such as SPSS. It could also include entire frameworks for retrieving,
processing, and pushing data back into the “database”. However, it is important that the
researchers can choose what tools to use and that they are not dependent on complex
frameworks. Support services should impose as few constraints as possible on the processes
researchers use on the data (within the data protection framework).

The research services are beyond the initial start-up provided by the support services. The
data provider takes a larger part in the actual research to be performed, depending on the
needs of the researcher. If the researcher would like to have the FOT/NDS data aggregated to
another format, the research services can assist. The work performed by the research services
could stretch as far as performing a complete package of analysis, answering specific research
questions.

2.6 Financial models for post project funding

Many FOT/NDS datasets have been collected and the issue of post-project funding is a shared
issue. There are several tasks to be performed if a dataset is to be easily accessed. The
following identified cost items are to be funded. The research services are not included, as
they are directly linked to the research and should therefore be paid by the applying project
directly.

<table>
<thead>
<tr>
<th>Research infrastructure for FOT/NDS data</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management &amp; coordination</td>
<td>Management of the infrastructure including data catalogues, and user and data agreements</td>
</tr>
<tr>
<td>Analysis platform support</td>
<td>Data management – Tool support - further develop and adapt the analysis tools to new types of analysis Access management</td>
</tr>
<tr>
<td>Facilities &amp; analysis work stations</td>
<td>Physical secure work space</td>
</tr>
<tr>
<td>IT operations</td>
<td>Database servers, storage and licenses</td>
</tr>
<tr>
<td>Data documentation</td>
<td>Post-project clean-up and structuring of data</td>
</tr>
</tbody>
</table>
There are several ways of funding the cost of maintaining and providing data for re-use:

Per project: The infrastructure gets funding by the projects utilising the data. In conjunction with the data access application, the cost is discussed. The cost is usually a generalized cost split per year, distributed over the estimated amount of projects. However, it is hard to estimate the number of projects. The problem is that the projects often have not planned for these additional data costs. Another drawback with this solution is that if there is a gap between projects, there is no funding to pay for the infrastructure.

Base funding and per project funding: Base funding will cover the basic running costs and gives the opportunity to put some money into marketing the infrastructure in order to attract more projects. As the projects do not get any data cost, they are more willing to re-use the data on a larger scale. It usually includes some paid maintenance work as well, and there is stability in knowing there will be a base funding over a few years.

Base funding with specific purposes: The platform is funded for a specific purpose, where many co-financers split the cost, e.g. through member fees. The funding is sometimes used for assigned research for the members as a whole. These users appreciate the focus on large volume of specific data, e.g. event recorded data. However, most users are not part of such homogenous groups, focusing on a specific matter.

Consulting: As discussed earlier, the support and research services could also perform the complete analysis and provide final results through a consulting agreement with the customer. The cost of maintaining the data would in this case be a part of the agreement and not only the personnel cost.

2.7 Application Procedure

The project collecting the data should at an early stage of the project agree on the conditions for re-using the collected dataset and on an application procedure for re-use. This will facilitate that new research applications which want to utilize the data, will have taken the data application time and potential costs for re-using the data into consideration already during the proposal phase, before the application is sent to the targeted call.

The application procedure shall at least address the following items: 1) where to apply; 2) what information is needed to be provided to be able to evaluate the application; 3) who can approve an application, response times, and conditions to be taken into account in the approval decision; 4) requirements on mandatory training in data protection and integrity issues; 5) information on the data access procedure; 6) requirements on data protection; 7) potential costs for data access, support, and research services; 8) requirements on acknowledgements on publications, reports, and presentations; and 9) documentation of data applications and the related approval decision(s).

3 Main Challenges

There are several large challenges in setting up a common DSF, consisting of the seven areas previously discussed. To make the framework really attractive, it should be usable on a global level, as the datasets are collected in different parts of the world. This raises even more issues.

In a global perspective, the project funding schemes lead to a difference in ownership of the data. In the US, many projects are fully financed by the authorities who thereby claim the
ownership of the data, while in EU-funded projects, participating organisations pay between 25-50% of the cost and as a result, own the data. This leads to different situations when it comes to the possibilities to gather and share the data after the project. Further, the legal setting differs between countries, which put different requirements on the handling of the data depending on where it is collected, stored, and analysed.

Documentation of data and metadata is usually not performed to a sufficient level in the projects. How could this be improved to facilitate and enhance the sharing of data? Another concern related to the data is that data protection procedures need to be reinforced repeatedly because even when procedures are in place, they can be quickly forgotten and overlooked. The people handling and using the data can subsequently perform non-approved disclosures of data, and thus, the awareness of the necessity to protect personal privacy and IPR-related data need to be restored.

The funding to keep the datasets available for research needs to be resolved. The mechanisms for this base funding need to be developed and decided upon, otherwise the data will not be re-used to a larger extent and a tremendous waste of the money, put into the collection of the data, will occur. The money to fund additional projects using existing data is just a minor additional part of the cost already used to collect the data.

The efforts to create and maintain a DSF should not be underestimated. As the research field of collecting and analysing FOT/NDS data is fairly new, there are changes to be expected in the way research will be performed. The framework must be able to incorporate such developments, where the opportunities include new data mining methods, image processing, new data types, continuously larger datasets and thereby the need for new database structures and search methods. Consequently, the DSF need to be flexible and continuously developed, when new requirements are emerging.

Perhaps the largest issue is to persuade the data providers to share their data. They are often more interested in additional or new research projects, than to work on documenting the existing data to permit other researchers to understand and use their data, especially as there are usually no funding left in the project for thorough data documentation. Therefore, maybe the highest priority should be to focus on what motivates a data owner to share the data.

4 Conclusions

This paper discussed the concept of a global DSF for exchanging naturalistic driving data and the seven different topics of the DSF were presented.

Based on the current challenges, the different items of the DSF need further discussions and development to include more of the requirements from different stakeholders and thereby be better suited to be accepted and used globally. The success of the DSF is depending on the interest and participation of the stakeholders, to gain from the on-going collection of knowledge and experiences from projects collecting new data or projects re-using data.

There are certain elements of the DSF, such as the funding to keep datasets available, which are especially important for the success of data sharing. It is also fundamental, that the awareness is enhanced regarding the critical inclusion of possibilities for re-use into the agreements of new projects collecting data. Through project agreements favouring data re-use and the consent for re-use from the participants in the studies, the availability of data will expand.
Finally, the EU project FOT-Net Data (11) will continue to develop the DSF and through the discussions incorporating the stakeholders in FOT/NDS, a further developed DSF will be available in 2016.

Acknowledgements

FOT-Net Data is a Support Action in the seventh Framework Programme Information and Communication Technologies. It is funded by the European Commission (EC), DG Connect, Grant Agreement number 610453.

UDRIVE is co-funded by the European Commission, DG Research and Innovation, in the 7th Framework Programme, Grant Agreement Number 314050.

References

7. ETC 2.0: https://www.go-etc.jp/english/etc2/index.html, retrieved August 2nd, 2015