UNIVERSITY OF LEEDS

This is a repository copy of *Data sharing framework for naturalistic driving data*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/93077/

Version: Accepted Version

Proceedings Paper:
Gellerman, H, Svanberg, E and Barnard, Y (2015) Data sharing framework for naturalistic driving data. In: ITS World Congress 2015 Proceedings. 22nd World Congress on Intelligent Transport Systems and Services, 05-09 Oct 2015, Bordeaux, France. .

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Technical Paper

ITS-2599

Topic: 7. Are Big Data and Open Data Transport's 'Silver Bullets'?

Subtopic: c. Information: Data Access, Service Access; Big Data, Open Data

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

Paper title: Data sharing framework for naturalistic driving data

Author: Helena Gellerman - SAFER at Chalmers, Sweden

Co-Author(s): Erik Svanberg[1], Yvonne Barnard[2][1]SAFER at Chalmers, Sweden, [2]Ertico, Belgium

This paper is submitted for: Conventional Presentation

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 research, i.e. safety, automated driving or mobility, to understand the human behaviour in different traffic environments. The framework includes topics such as data description, data protection and storage, 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.

Draft paper: Data Sharing Framework for Naturalistic Driving Data

1 Introduction

During the past 15 years, we have seen a fast growth in the number of Field Operational Tests (FOTs) 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 a main driving force. The availability of technology, such as affordable video capture solutions and cheap enough data storage, has been a requirement and facilitator for the development of the FOT and NDS.

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, mainly depending on if the data is collected continuously and if it includes video.

The largest datasets have so far been collected in the US (e.g. IVBSS, SHRP2 and the on-going Safety Pilot project) and in Europe (e.g. euroFOT, DriveC2X and the on-going UDRIVE project). In Japan, large data sets based on event recorders have been collected and Australia has several interesting datasets. Data collection has also started in Korea and China. As the number of different datasets has increased and so also the awareness of the substantial effort and funding needed to do these FOT/NDS, the interest in data sharing has increased worldwide.

There are different views on the value of data sharing depending on if you are a data provider or a data user. The owner of the data has spent a large amount of effort 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. Structures and procedures to make data available and access control should be in place. The availability of such procedures would also increase the number of data providers who are interested in opening up their datasets. In some countries, the requirement for making the data open for further research is incorporated into the agreement if public (government) money is funding the project. 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 either data collection and storing or analysis of the data, will have an impact on the data sharing constraints. Also, for example 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 FOT-Net 2 (FOT-Net; FOT-Net 2 Consortium, 2014) and FOT-Net Data activities, at various conferences and through discussions with people from the US, the EU, Japan, Australia and China (Barnard, 2014).

2 Data Sharing Framework

The availability of a common Data Sharing Framework will highly facilitate a larger use of the collected FOT/NDS data. Such a platform 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 researchers setting up new FOT/NDS projects would then not need to develop the data-sharing specific content for a specific 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 Data Sharing Framework:

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 meta data, including a “standard” description of the documentation of the data
3. Data protection requirements both on the data provider and the analysis site
4. Security and personal integrity education 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 (one or more) from the project where the data was collected (original project) or by an external data provider.

In the European support action FOT-Net Data, the Data Sharing Framework is being further developed from the work done in FOT-Net 2 (FOT-Net 2 Consortium, 2014) on the basis of experiences gained in FOTs and NDS, in collaboration with a variety of stakeholders from Europe, the US, Japan and other countries.

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) potential agreements with external 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:

- · Ownership and access to data and data tools
- · Storage and download of data
- · Access methods
- · Research and commercial areas where data usage will be allowed
- · Post-project (re-)use of data
- · Post-project financing

To be able to re-use data after the project and by other parties than the ones who collected the data, participants of the FOT/NDS have to agree to such use. 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 and 3) who is responsible for the data, 4) who will have access to what data and on which conditions, and 5) the access procedures. As the participants allow 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 will or can be used for. The participant should make an active consent to the most vital topics for data sharing.

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 are often used to enhance the data set. 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 are documented to a level where an assessment of the level of validity of the data could be performed. This is potentially 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 or 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 1: Data classification

Data type	Data category	Ownership
Questionnaires- and interview data	Personal	Jointly
Video	Personal	Jointly
GPS	Personal	Jointly
Vehicle mounted sensors (eye-tracker, radar, etc.)	Sensor	Jointly/supplier
V2V and V2I data including "activity" data	System/sensor	Jointly
Enhancing data – road attributes, weather	Infrastructure/sensor	Jointly/supplier
"Open" and aggregated CAN-data	System/sensor	Jointly
Closed, confidential CAN-data	System/sensor	OEM

One of the most important factors in creating a FOT/NDS dataset that can be reused is the simplicity in which the data set can be understood. The collected data need to be described 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 that/if the data is good enough to be used for specific research. That is, 1) if the quality of the collected measures is good enough and 2) if there is video without disruptions accompanying all data of interest.

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

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

It is important that data from projects can be read in a “raw” and clearly described format directly from the data storage source (e.g. database or file storage) regardless of what analysis tools are used in a project (with appropriate access restrictions). That is, both within a project and after it finishes (re-use of data), there are many different types of analyst who will need and want to access the data in different ways. At lowest level the users should be able to get data and metadata in as “raw” data as possible from the data source.

Data description formats and data formats will have to be acceptable and used by as large community as possible.

2.3 Data protection

Data protection is the key to create the trust needed between the data provider and the researcher to make the data owners provide access to their data. If the data provider knows that the researchers have good and proven procedures in place to keep control of who is accessing the data and that the researchers have knowledge about the legislation regarding the handling of personal integrity data and Intellectual Property Right (IPR) data, the more data they are willing to share.

The data protection level needed depends on the harm the revealed data could do. There are especially two categories of data that need protection, personal data and data that, if revealed, could potentially harm a commercial company. The provision of the latter data to projects is usually accompanied by agreements, stating the conditions for access and use.

The European Directive 95/46/EC Art. 2 contains a definition of the term “personal data”:

´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 suggested data protection requirements described in this paper have the aim to guide data centres and analysis sites towards setting up a data protection concept that would meet the regulations and that would guard the will of the participants as stated in the consent form.

There are several types of commercially available data that are used to enhance FOT/NDS data, where some data might need protection. When signing contracts for provision of such data, it is advisable to discuss the foreseen data protection level, so that both parties could agree on a suitable level of protection.

Two sets of data protection requirements are suggested below, one for the Data Centre (DC) and one for the Analysis Site (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 data sets 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: Data stored and processed at a DC must be protected from unauthorized access.

DC2: Data stored and handled at a DC must be protected from accidental deletion or corruption.

DC3: The DC must document its data protection implementation.

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 are regulated by the Project Agreement(s) and the informed consent of the driver.

DC8: Data extractions for specific purposes must be in accordance with the consent forms and project agreement and the extraction must be documented.

Requirements for the AC:

AS-1: The AS organisation must document its data protection implementation AS-2: The analysis work stations must be physically and logically protected.

AS-3: Analysts must have received relevant training in data protection and integrity issues.

AS-4: A confidentiality agreement for any involved AS personnel must be in place.

AS-5: The AS data supervisor administers access requests and forwards those to the DC data access dispatcher.

AS-6: Specified procedures for data extraction must be used.

AS-7: The analyst must not extract or re-distribute data.

AS-8: The project data must not be used for research areas not covered by the consent forms in the project.

AS-9: Visitors/guests to the AS should sign a confidentiality agreement.

AS-10: All post-project research must investigate the need for approval.

2.4 Education on data protection related to personal data and IPR

It is important to educate 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.

The addition of video in an FOT/NDS can add substantial value to a data set. The reason for a sudden brake or steering manoeuvre may then be understood by looking at the video of the event. The inclusion of video in the data set brings at the same time another level of need for protection of the data. For those data sets 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 (http://phrp.nihtraining.com/), where the analyst gets a certificate at the end of a web course.

2.5 Support and research services

Support and research services are one of the cores 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 to perform the research itself or extract usable datasets.

Tools are an integral part of the support services. The tools could consist of a viewing and annotation tool, scripts to extract useful datasets from the database, MATLAB and other licensed software, such as SPSS. It could also include entire frameworks for both retrieving, processing and pushing data back into the “database”. However, it is important that the analysts can choose what tools to use and that they are not dependent on complex frameworks. It is also important that data from all projects can be read in the original and clearly described format directly from the data storage source (e.g. database or file storage) regardless of what analysis tools are used in a project (with appropriate access restrictions). Support services should impose as few constraints as possible on the processes analysts use to analyse 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 analyst. If the analyst 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 2: Items requiring funding

Research infrastructure for FOT/NDS data	**Comments**
Management & coordination	Management of the infrastructure
Analysis platform support	Data management – expert knowledge Tool support - further develop and adapt the

	analysis tools to new types of analysis Access management
Facilities & analysis work stations	Physical secure work space
IT operations	Database servers, storage and licenses
Data documentation	Post-project clean-up and structuring of data

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 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. Most users are though not part of such homogenous groups, focusing on a specific matter.

2.7 Application Procedure

The project collecting the data should agree early on in the project 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:

- · Where to apply
- · Which information is needed to be provided to be able to evaluate the application?
- · Who can approve an application, response times, and conditions to be taken into account in the approval decision?
- · Requirements on mandatory training in data protection and integrity issues
- · Information on the data access procedure
- · Requirements on data protection
- · Potential costs for data access, support and research services
- · Requirements on acknowledgements on publications, reports and presentations
- · Documentation of data applications and the related approval decision(s)

3 Main Challenges

There are several large challenges in setting up a common Data Sharing Framework. 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.
Looking globally, 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.
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? A further related concern raised within projects is that data protection procedures need to be reinforced because even when procedures are in place, they can be quickly forgotten and undermined by those people handling and subsequently exchanging the data.

Funding to keep the datasets available for research needs to be solved. The mechanisms for this base funding needs to be developed and decided upon, otherwise the data will not be re-used and a tremendous waste of money 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 Data Sharing Framework could not be underestimated. As the research field of collecting and analysing FOT/NDS data is fairly new, there are still huge changes to be expected in the way research will be performed and the framework must be able to incorporate such developments. Examples of challenges to address are data mining methods, image processing, new data types, continuously larger data sets and thereby the need for new database structures and search methods.

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

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, GA No. 314050.

References:

FOT-Net, www.fot-net.eu

Barnard, Y. (2014). *Report on FOT Network activities, Deliverable D2.2.* FOT-Net 2

FOT-Net 2 Consortium. (2014). *Working Group Additions to FOT Methodology. Deliverable D3.2.* FOT-Net 2.