EUROPEAN COMMISSION

DG CONNECT

SEVENTH FRAMEWORK PROGRAMME

INFORMATION AND COMMUNICATION TECHNOLOGIES

Coordination and Support Action – Grant Agreement Number 269983

FOT-Net 2

Field Operational Tests Networking and Methodology Promotion

Report from the FOT-Net Data Sharing working group

Deliverable no.	
Dissemination level	Final report
Work Package no.	WP3.2
Author(s)	Helena Gellerman
Co-author(s)	Jonas Bärgman (SAFER), Erik Svanberg (SAFER)
Status (F: final, D: draft)	F
File Name	WG Data sharing report.doxc
Project Start Date and Duration	01 June 2011, 39 months

Document Control Sheet

Main author(s) or editor(s): Helena Gellerman

Work area: FOT-NET WG 3.2 Data Sharing

Document title: Report from the FOT-Net Data Sharing working group

Table of Contents

1 Introduction

During the past 15 years, we have seen a fast growth in the number of Field Operational Tests (FOTs) and Naturalistic Driving Study (NDS) that have been performed worldwide. The need to better understand the causational factors behind incidents and accidents together with the availability of technology with cheap enough storage capabilities and sensors have been the main driving forces for the development of the methodology from the start.

The data which has mainly been collected through naturalistic driving by volunteer drivers have been used to answer the research questions in the original project. The size of the datasets varies, from below 1TB to several PB, 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 on-going DriveCAM and Safety Pilot) and in Europe (e.g. euroFOT, DriveC2X and on-going UDRIVE). In Japan, large data sets based on event recorders have been collected and Australia has several interesting datasets. It is especially interesting that data collection also have started in Korea and China, as the traffic environment and traffic culture is so different from other countries.

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 become more and more in focus worldwide. Data Sharing (including Big Data and Open Data) was also a key theme arising from the latest ITS Congress 2013 in Tokyo. Numerous presentations addressed the issues on all kind of data, not only from FOT/NDS, but the key problems were the same; who should provide the data and how?

Most of the earlier projects focused on learning the FOT/NDS methodology and to answer the research questions set out by the project. That was a major achievement in itself. There was an unawareness of the requirements for data sharing after the project. Many projects do not therefore have the necessary pre-requisites in place in the consortium agreement and the consent form to be able to share the data, at least not outside the former project partners. Due to lack of time and funding, much data is not documented to a proper level which further hampers its re-use. Also, if tools are developed in the project, they are often tailor-made, to suit the needs of the project and the tool requirement sheet did not include the view of a non-partner user. The awareness has increased regarding the personal data and the need for data protection and security measures. And finally, many projects did not discuss the nature of a data sharing procedure and how to approve and assist new projects in re-using the data.

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 large amount of effort (and usually also their own funding) to collect data and build up the data infrastructure and tools. It is also an extra effort to provide data, especially if there is no basic funding for keeping the data up and running. It is therefore important to find win-win situations between the data provider and data user in further re-use of the data, to compensate the effort done to provide easily accessible data. This would also increase the number of data providers who are interested in opening up their datasets.

Apart from the more general possibilities to share, there are different constraints that could 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 conditions. As mentioned earlier, the content in the consortium agreement and the consent forms signed by the participant might not have had data sharing in focus when they were written and could make data sharing impossible after the project. Also 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.

Still, there are several advantages to share collected data, where some of them are pointed out in section 2 “Why data sharing and re-use of data?”

This report is based on information collected during the last three years within the FOT-Net 2 activities, at various conferences and through discussions with people from the US, the EU, Japan, Australia and China. During three FOT-Net workshops in conjunction with ITS World Congress 2010, 2012 and 2013, data sharing has been a topic for a separate session, where the participants from different countries all over the world openly shared their experience from FOT/NDS. During the CS Coordination day 25/05/2012, the seminar on Complementarity of different FOTs and re-use of data 26/11/2012 and the meeting on Lessons learned from Pilots on Cooperative Systems 26/02/2013, the sessions on data sharing gave valuable input to the report. On several occasions, representatives from the EC participated actively in the discussions and expressed their views and expectations as a funding organisation.

Different conferences such as the SHRP2 Safety Research Symposiums 14/07/2011, 12/07/2012, 11/07/2013 and VTTI conference 28-29/08/2012 in the US, the ITS World Congress 14-18/10/2013 and the Fast Zero symposium 22-26/9/2013 in Japan and the DDI conferences 5-7/09/2011 and 4-6/09/2013 in Sweden have all given input to the report. During these seminars and conferences and also on other occasions, separate discussions outside the sessions and workshops with different people have given in-depth knowledge of their views on data sharing.

The foundation of the suggested platform comes from hands-on experience and discussions in many different projects, such as SeMiFOT, euroFOT, DriveC2X, SHRP2 and UDRIVE.

2 Why data sharing and re-use of data?

Performing an FOT/NDS demands considerable amount of time and effort. Sharing the data after the project also requires devoted persons to bring the data and tools to a level, where they are easily understandable for someone not having participated in the project. To stimulate more data providers to take this step, it is essential to understand the possible benefits of sharing the data.

The data provider is usually, at least up until now, also performing research and the chances of getting additional funding to do further analysis is probably the factor that gives the highest motivation to provide data for data sharing. By opening up the access to the dataset, a larger variety of possible research projects would be suggested and the possibility of additional research funding is increased.

The original project usually only performs a small part of the possible research that could be done on the collected dataset. From a funding organisations point of view, utilising the already collected datasets for further analysis is an efficient return on investment. Also for project partners, knowing the data, it is also a good payback on invested efforts, to be able to further explore the data. During this second phase of data use, the funding organisation could require that additional partners are brought in, to open up the use of the data.

Due to the amount of data available from different parts of the world, meta-analysis cross FOTs and NDSs could provide a more quality assured result compared to drawing the conclusions from a single dataset.

Using global datasets for research on the comparisons of specific groups in different contexts and countries, e.g. older drivers, could provide insights in cultural differences in traffic behaviour for the specific group.

If funding for additional research is conditioned by international collaborations and data sharing, the global research community will be strengthened, as the flow of ideas and knowledge will be enhanced.

Research collaborations create trust between organisations and promote thereby an increase in the willingness to share data.

3 Content of a common data sharing platform

The availability of a common data sharing platform, where projects are set up in a similar manner with the data sharing pre-requisites integrated into the project agreements from the start and using procedures/templates with the same content, will highly facilitate a larger use of the collected FOT/NDS data. The researchers setting up new FOT/NDS do not need to go through the content of yet another special framework for a specific project, but can focus on the project specific questions such as research questions and study design. Also, researchers wanting to re-use already collected datasets or maybe several different datasets in the same research can utilise a more or less standard application procedure, rely on already done training that are widely accepted and plan for the costs that using a specific dataset might cause the project.

In the following section, the suggested content of such a platform will be described. On an overall level, the following seven areas need to be addressed by a data sharing platform:

- Project agreements, such as the grant agreement together with the description of the work, the consortium agreement, the participant agreement and external data provider agreements set the pre-requisites and the borders for data sharing together with legal and ethical constraints.
- The availability of valid data and meta data, including a “standard” description of the documentation of the data, e.g. standard format and the related attributes (sampling frequency, accuracy, ….).
- Data protection requirements both on the data provider and the analysis site, including security procedures.
- Security and personal integrity education for all personnel involved.
- Support and research functions, to facilitate the start-up of projects and also e.g. offer processed data for researchers not so familiar with FOT/NDS data. The support also includes the availability of analysis tools.
- Financial models to provide funding for the data to be maintained and available and access services.
- Last, but not least application procedures including content of application form and data sharing agreement.

Another way of describing the common data sharing platform is by its content of documents.

Table 1: Data sharing platform documents and content

Document type	Content
Procedures	Application and approval, support/research functions, data extraction and download
Templates	Application form, Data description, Consent form, Data sharing agreements, Data sharing text for Consortium Agreements, Data security presentation, Approved training certificate Financial models, Data protection implementation, Data extraction request, NDA for

	analysts/visitors, Application to ethical review board, Description of content to be funded
Standards	Data protection - data provider/analysis site, Data extraction format, Data and metadata description, Level of security education

Generally, the data itself could be either managed by the project itself or by an external data provider. A central data provider could also just provide test samples of the different datasets and guide the interested researchers to the organization hosting the complete data set. The general recommendation however, is to let one or more project partner(s) from the original project maintain the data, possibly with test samples as described above, as analysis of the datasets in most cases require a deeper knowledge of the data and the way it was collected.

3.1 Data sharing in project documents

The initial process of setting up a project is crucial to the possibilities to share data during and after the project. Agreements can of course always be renegotiated, but the time and money consumed could be substantial, especially in large consortia, as the partners have entered the consortia on the conditions stated in the agreements, and alterations could lead to reconsiderations. The project agreements cover many different topics, but just a few of them are related to data sharing. Therefore, the time spent during the project application and in the beginning of the project, to agree on the conditions for data access and use including data re-use after the project, are well invested.

The main documents to focus on are the grant agreement, if the project has external funding, including the description of the work, the consortium agreement among the project partners, the participant agreement and potential agreements with external data providers to the project.

3.1.1 Grant Agreement – Description of Work

In the grant agreement and the description of the work, the result of the project and the funding is agreed upon. 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 during the project application and also during 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 description of the work could include most of the topics listed in 3.1.2. One topic that is especially important to address during the project application phase towards the granting organisation, are the possibilities for post-project funding and other conditions for keeping the data available for data sharing after the project, if there is such a requirement on the project.

3.1.2 Consortium Agreement

The consortium agreement is the most important document next to the consent forms in 3.1.3, in setting the conditions and requirements for data sharing and re-use of the data. Numerous topics need to be discussed and decided to set a legal platform for the handling of the data during and after the project. The following table provides guidance to the topics to be handled in the consortium agreement.

Table 2: Data sharing topics within the consortium agreement

Topic	Comments
Ownership and access to data and data tools	Will all partners own all data/part of the data? How could it be used and on what conditions? May the data be licensed to third parties? Will all partners have access to all/part of the data? May third parties have access to the data and on what conditions? Constraints due to personal data, especially video? Are there future agreements with data providers to take into account? Who will own the data tools and on what conditions are they licensed during and after the project? How can data be re-used if the data is owned by one partner and that partner cease as company?
Storage and download of data	Where will the data be stored, centrally or distributed? What are the requirements on data protection and how are they assured? Shall all data/part of the data be downloadable for all partners and if so, under which conditions? Shall all data/part of the data be downloadable for third parties and if so, under which conditions? Is there a time limit to request data for download?
Access methods	Shall a specific access procedure be used and by whom? How will the data be accessed? Can it be remotely accessed, downloaded or only accessed at the premises of any partner? What are the requirements on data protection for partners/third parties analysing the data?
Areas of use	Shall it be possible to use the data for both research and commercial purposes? Are there special conditions for the commercial use? In which research/commercial areas could the data be used? (I.e. safety, mobility etc.)
Post-project re-use of data	Which partner is responsible for maintaining the data after the project? Shall a non-partner be the data provider of the project data during/after the end of the project? Which application procedure shall be used? Who will grant access to data after the project? Are there conditions, such as legal and ethical constraints and availability of funding for data storage and access services to be considered?

Post-project financing	How will the storage and support services for data re-use be financed after the project? How will this funding be distributed?

3.1.3 Participant Agreements including consent forms

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

- What data is collected?
- Where will the data be stored and who is responsible for the data?
- Who are the project partners?
- Who (project partners/third parties) will have access to what data and on which conditions?
- How are the access procedures (overview description)?
- The possibility to consent to the three topics described as YES/NO options below, directly related to data sharing.

As the participants allow the project to follow the participant's private life for a period usually from a few weeks up to a year, it is important to be very clear on the use of the data. A recommendation is to have the participant make an active consent to the most vital topics for data sharing. Common consent needs for data sharing are the following topics, where an example text is provided for European conditions;

I hereby agree to participate in the above described research study. I consent to have the material transferred and shared with research partners in a third county (e.g. country outside EES)

□ Yes □ No

I also consent to have video recordings or pictures being published or shown in public events (e.g. research reports or conferences)

□ Yes □ No

I also consent to have collected data (including video recordings and pictures) to be reused in other research projects focusing on factors regarding:

- *The driver (e.g. drowsiness, distraction, driving style) and/or*
- *Vehicle behaviour (e.g. fuel consumption, system activation) and/or*
- *Environmental factors (e.g. road geometry, weather conditions)and/or*
- *…*

□ Yes □ N0

3.1.4 External data provider agreements

External data providers could be companies providing sensor systems, map data, weather data or other services that the project needs 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. Attention from a data sharing perspective should be given to the following topics.

- What is regarded as confidential information?
- If data is regarded as confidential information, could it be changed/aggregated, to allow for more open access?
- Can the data be accessed by another project partner/third party?
- Can the data be transferred to another project partner/third party?
- Are there special conditions for what the data could be used for?
- Are there special conditions for sharing and re-using the data after the project?
- What happens if the external data provider is bought by another company?

3.2 Valid Data - 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 could be performed. This is potentially problematic if one has not been part of the project and does not know the way the tests were 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.

The FOT/NDS data is often referred to as the data that is actually collected during the tests including questionnaires and interview data. To be able to answer almost any research question, a vast number of other data is needed. In 3.2.1, an overview of data types are provided, which may/may not be part of a projects complete dataset. In 3.2.2, a list of items needed to describe project measures, is provided.

3.2.1 Collected data to share

There are different ways of describing the collected data. One is to cluster the data by the same category of data or ownership. The category usually determines the level of protection, see 3.3.1, 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 3: Data classification

Data type	Data category	Ownership
Questionnaires- and interview data	Personal	Jointly
Video	Personal	Jointly

GPS	Personal	Jointly
Vehicle mounted sensors (eyetracker, lanetracker, 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 CAN-data	System/sensor	OEM

An overview of the variety of possible data to be collected and later shared is seen in the table below.

Table 4: Data that can be collected and shared

Data	**Specification**
Measures	Measures pre-processed and derived from the data collected.
Processed data	Data that has been produced using available data in order to enrich the data set; derived measures, events, locations, situations, performance indicators.
Positional data	GPS positions related to measures and processed data.
Geographical attributes	Data properties attributed to geographical locations. The data is retrieved using GPS positions and can include properties such as traffic situation, speed limits, road information, weather conditions, etc.
Video	Video data collected from cameras covering both exterior and interior environment.
Annotated data	Data produced by annotators.
Questionnaires and interview data	Questionnaires and interview data (with potentially personal data) answered by the participants during the study.
Communication data	Data from monitoring the V2V or V2I communication, e.g. latency.

Participant meta data	Meta data on drivers; driver profile, selected information from questionnaires. De-identified.
Non-participant meta data	Meta data on drivers not selected for participation. De-identified. This data is used for exposure analysis.
Participant consent data	Selected consent form data for participant. May include information on usage restrictions, consent withdrawal, etc.
Vehicle meta data	Meta data on vehicles where data is collected. De-identified.
Test meta data	Meta data on the performed FOT/NDS, such as study design, location, test period etc.
Annotation meta data	Information about the code book, the book stating how the annotators should code the different events.
Created data	New data created by analysts; either private, shared with analyst group, or shared with analysts in organisation.

3.2.2 Description of Data and Metadata

One of the most important factors to make 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, e.g. 1) if the quality of the collected signals are good enough, or 2) if there is video without disruptions accompanying all interesting trips.

Most projects use internal project-specific descriptions and description formats of the collected data. A few have been based on concepts used in earlier projects, but then often somewhat extended or modified. The latter is mainly the case when an organisation is doing an FOT/NDS for the second time and wants to re-use tools, database structures etc. As most projects use their own description of the data, it might take a large effort to re-describe the different, already collected datasets in a common format, and be even harder to re-use available tools with a new dataset.

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. For the individuals involved in data collection and analysis in a specific project (i.e. the projects partners) the metadata is implicit during and shortly after a project, as they have performed the project and for them, the description of the collected data feels more urgent. For the organisations attempting to re-use the data however, the need for description is often the other way around. The metadata is essential to know if the dataset could even be used for the new research purpose. 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).

At the end of the project, usually the last resources are used for the analysis, which means that there is little possibility to do a thorough job of documenting the data and metadata, especially not through the eyes of a re-user of the data. It is therefore important to do the documentation early on in the project.

The attempt to standardize the formats and content of descriptions of the data and metadata presented in section 3.2.1 is a larger effort, and has therefore not been targeted in the work done by the working group on Data Sharing. This is thought to be an important next step, as this standardization task could provide a platform for e.g.

- a harmonized interface for analysis tools reading both data and metadata, making it possible to use the same tools for different datasets (even if the data itself may be different),
- a standard description of the data at a data broker with different datasets available to new data users, and a simplified process for projects when describing their data, i.e. for new projects to state the fulfilment of the data and metadata description as a goal and deliverable of the project.
- a suggestion for a minimum requirements in terms of data characteristics (e.g. sampling frequency, accuracy, delay, age) in order to ensure a reasonable confidence that the data set can be really useful to future project. This standard should take into account technological evolution and consequent more demanding characteristics in the future.

It is important that any description format can handle data protection appropriately. Preferably data description formats should provide a way to describe different layers (or tiers) of processing on top of the originally collected data. That is, to go from several measured metrics to a derived measure using these metrics may include several layers of processing i.e. resampling, filtering and merging algorithms. Metadata description formats should allow for descriptions of all these layers/tiers. Care should be taken to keep units and other metadata consistent through the tiers.

Finally, it is important that data from all 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. Any tool can and should build on these formats and users that require graphical user interfaces can then use those formats, while other users would develop analysis based on direct data access without such graphical interfaces. Examples of different ways to analyse data are:

- to use as close to original data as possible and do on-the-fly processing of all or most derived measures events etc., or
- to calculate all derived measures and events and push them back into the database (or whatever storage is use), and then to apply a more simple set of algorithms.

For the first method, the core is a set of validated functions or algorithms that are consecutively applied for each analysis, while for the latter the derived measures and events pushed back into the database is the core. Data description formats and data formats will have to deal with both approached to be acceptable and used by as large community as possible. Preferably the data formats should be the same across projects, but at least the data description and metadata formats should be the same (as described above).

3.3 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, proven procedures in place to keep control of who is accessing the data and that the researchers have knowledge in the legislation surrounding the handling of personal and IPR data, the more data they are willing to share.

There are different levels of research co-operations, which demands different levels of protection. The example where the data is collected and used within the same organisation is not considered here as the data is not shared externally. As soon as the data is shared between two organisations, the reasoning below becomes valid. How is the data going to be accessed between the partners? Should each partner have a data set? How can the data be transferred? Which demands on the physical/logical access must be in place? etc.

The data could be downloaded via a website, transferred on hard drives to the research organisation, remotely accessed at the data provider or only be accessed from the premises of the data provider. Each method has its own implications and it is usually the data type that has a large impact on the conditions for which method that is used.

In this section, the demands on data protection for different kind of data will be discussed. The section 3.3.2 includes a suggestion for requirements on data protection both at the Data Centres (DC) and at the Analysis Sites (AS), to facilitate the set-up of the necessary framework, to prevent unauthorized access to the collected data.

3.3.1 Data protection level depending on data type

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.

Personal data that needs protection

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 personal data is therefore divided in two categories, sensitive personal data and more general personal data. The suggested data protection requirements stated in 3.3.2 have the aim to guide the data centres and the analysis sites towards setting up a data protection concept that would guard the will of the participants as stated in the consent form. As the FOT/NDS data often is collected including video and GPS, special precautions might be needed if the participants have given their consent to the collection of sensitive data, to guard the anonymity of the sensitive data. These precautions might exceed the requirements in 3.3.2.

Commercial data that needs protection

There are several different kinds of commercial data that might need protection. When signing contracts for provision of such data, it is advisable to discuss the foreseen protection level, so that both parties could agree on a suitable level of protection.

Several things affect the protection level. A way to categorise the commercial data could be:

Table 5: Categorisation of commercial data

Data Category	Access	Ownership
Open	Open for all analysts/all project partners/certain project partners	Owned by all/part of the project consortium
Closed	Open to all project partners/certain project partners during the project. Available on a per project approval by the owner.	Data provider
Proprietary	Data is never shared, as the commercial value of the data is too high for data sharing.	Data provider

The closed data could be made more open via the agreement, through aggregating the signal to a higher level, thus avoiding any commercially harmful misuse.

3.3.2 Data protection at Data Centres and Analysis Sites

Two sets of requirements are suggested below, one for the Data Centre (DC) and one for the Analysis Site (AS). Related documents to both the DC and the AS are listed. Depending on the data type involved in the data sharing, the needed level of protection will vary. The data protection recommendation is related to a data set including both video and proprietary sensor data. If the data to be shared is anonymised, several of the requirements are not applicable.

Data Centres (DC)

List of data protection requirements

DC1: Data stored and processed at a DC must be protected from unauthorized access.
Servers, computing environment (also physical), and network connections must be protected using measures sufficient to prohibit access to unauthorized parties.

DC2: Data stored and handled at a DC must be protected from accidental deletion or corruption.
Sufficient backup and disaster recovery solutions must be in place, and also protected from unauthorized access.

DC3: The DC must document its data protection implementation.
The data protection implementation description must be documented and it is recommended that it should be presented by the DC to the AS.

DC4: Confidentiality agreements for any involved personnel must be in place.
The DC must require signed confidentiality agreements with all involved personnel before they start handling the FOT/NDS data. Agreements can be either specific for the project (for guest researchers, students, etc.) or implicit through means of employment contracts.

DC5: Data protection must be ensured by the DC after end of project.
The data must be stored and protected at the DC after the end of the project, to facilitate data re-use and sharing after the project.

DC6: Data sent between DC and AS must be encrypted.
Data may be transmitted between DC and AS by electronic means, or alternatively transported on physical media. The DC must ensure that the data cannot be accessed during the transfer.

DC7: Data downloads are regulated by the Project Agreement(s) and the informed consent of the driver.
Data sharing could in some cases involve actual downloading of part or the whole set of a project data. The Project Agreement should regulate the possibilities of downloading the data. Also the participants must have given their consent to spread the data outside the project partners.

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

Depending on what the participants have agreed to in the consent forms, different extraction policies can be used. Especially video and GPS extraction is to be treated with special care and the recommendation is to anonymize the personal data content in the videos, especially faces and vehicle number plates. Each extraction must be in accordance with potential content in the project agreement. All extractions must be documented.

Documents

The following specific documents within the context of the Data Center are identified:

- Agreement with external IT infrastructure provider (if applicable)
- Confidentiality or Non-Disclosure Agreement (CDA or NDA), for involved personnel
- Data protection implementation documentation signed by DC leader.

Analysis Sites (AS)

List of data protection requirements

AS-1: The AS organisation must document its data protection implementation if handling data within their organisation.
In order for data access to be granted to the analysts from the research organisation, a data protection implementation description must be documented and it is recommended that it should be presented by the AS to the DC.

AS-2: The analysis work stations must be physically protected.
Analysis work stations used for either remote virtual access to the DC or for handling downloaded data must be protected in such a way that unauthorized access is prohibited. Work stations must be placed in either locked rooms, or by other means placed so that contents on screens can be seen only by the analyst or annotator.

AS-3: Analysts must have received relevant training in data protection and integrity issues.
Before data access can be granted, analysts must present proof that mandatory training, possibly prescribed by the initial project, i.e. US NIH education (http://phrp.nihtraining.com/users/login.php) has been followed.

AS-4: A confidentiality agreement for any involved AS personnel must be in place.
The research organisation must require signed confidentiality agreements with all analysts (researchers, research assistants, students), before data access can be granted. Agreements can be either specific for the project (for guest researchers, students, etc.) or implicit through means of employments contracts.

AS-5: The AS leader administers access requests and forwards to the DC authentication manager.
When the AS has presented its data protection implementation to the project management, access requests for personnel may be sent directly to the DC.

AS-6: Specified procedures for data extraction must be used.
Extraction of a portion of the data must be according to the participants consent and the data extraction procedures must be used. Video snippets and screen shots are also subject to this requirement. All extraction is administered through the DC.

AS-7: The analyst must not extract or re-distribute data.
As regulations for data extraction procedures are in place, the analyst must not circumvent these procedures nor disclose data outside of the AS in any other way

AS-8: The project data must not be used for research areas not covered by the consent forms in the project.
The data must not be used for any other purpose than those stated in the consent forms, except if given an Ethical Review Board approval. In the case national law has required approval from ethical review board (or similar) for the project, other usage of personal data is normally not permitted, and must be sought explicitly.

AS-9: Visitors/guests to the AS must sign a non-disclosure agreement.
If any portion of the analysed data is presented for a visitor, the visitor is required to

sign a non-disclosure agreement. By definition visitors do not have access to the data, and are always accompanied by an authorized person.

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

With the drivers consent forms, national ethics regulations on research together with project agreements set the conditions for post-project research. All research, but especially if not previously covered by the project, might need to be submitted to local ethics committee and/or competent national authority for approval or additional consent might be needed from the drivers. Project agreements including agreements with sensor providers, might restrict the use of the data.

Documents

The following specifically required documents within the context of the AS were identified:

- Confidentiality or Non-Disclosure Agreement (CDA or NDA), between analyst/visitors and AS organisation.
- Approved training certificate for analyst.
- Data protection implementation documentation signed by the AS leader.
- Approval from ethics committee for intended research (if applicable).
- Data extraction request.

3.4 Education on data protection related to personal data and IPR

The addition of video can add substantial value to a data set. The reason for a sudden brake or steering manoeuvre can easily be understood by simultaneously looking at the video. The inclusion of video in the data set brings at the same time another level of need for protection of the data. Especially for those data sets where video 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 the web course. At the same time, it is important to get information 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 content of such an education could involve

- Description of the data with special focus on personal data and Intellectual Property Rights (IPR)
- Requirements on the data handling from a legal point of view
- The content of the consent forms, especially the specific active consents related to data sharing
- Data ownership and access rights for partners/third parties
- The set-up of the physical workspace
- Local procedures on how to perform analysis

All training needs to be documented, most conveniently done by an analyst's information sheet, which the participant needs to sign. Though the analyst might have a non-disclosure agreement in his/her certificate of employment, the signing process of the document enhances the protective level of the data.

3.5 Support and research services

Huge data sets have been collected worldwide and in the future, even larger datasets will have to be handled. The data is rich and can be used for research in a number of different research disciplines, such as safety, mobility, eco driving, traffic planning, infrastructure etc. Most importantly for many research questions, multi-disciplinary joint research is needed to be able to answer the questions. .

The support and research services is one of the cores of data sharing, which was pointed out by both data providers and researchers during the discussions in FOT-Net 2. Depending on the knowledge of people re-using the data, either just support is given or both support and research services. As examples, SHRP2 and SAFER have both of these services. The support services will assist the researchers during the process, while the researcher is doing the actual work. The analysis tools are an integral part of the support services. The research services are more targeted to perform the research itself or extract usable datasets.

3.5.1 Support services

The support service targets the researcher and his/her possibilities to perform analysis. The support starts already at the application stage with discussions on the usability of the data to answer the specific research questions at hand. If the data application is approved, the researcher is given training in security and integrity matters, thus providing a deeper understanding of the sensitivity of the data. The analysis platform is described and training in using the tools is also included in the education. After having signed a document, stating that the analyst has understood the set-up, he/she is given access to the data. The support could give some additional support at this stage, but in general, here is where the research service takes over. After the analysis is done, the support services could offer a discussion on the result of the analysis, to enhance the result of the project and also see to that no misunderstandings have led to wrong conclusions.

The 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 SW, such as SPSS, but can 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 with graphical interfaces or other constraints other than the raw data formats and data descriptions. That is, as mentioned in 3.2.2 "Description of data and metadata" it is important that data from all 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). This is important since different analysts have different ways to analyse data. Support services should impose as few constraints as possible to what processes analysts can analyse the data (within the data protection framework). Examples of different ways to analyse data are given in 3.2.2 "Description of data and metadata". Data description formats and data formats will have to be able to deal with different analysis processes to be acceptable and used by as large community as possible. It is also important that the dependency on third party software for

access is kept to a minimum. The support function could also include basic maintenance of the analysis platform, also including further development of the tools.

3.5.2 Research Services

The research services are beyond the initial start-up provided by the support services. In this case, the data provider takes a larger part in the actual research to be performed, depending on the needs of the analyst. If the analyst comes from another discipline and/or is unfamiliar with the type of data and therefore would like to have it aggregated to a more suitable format, the research services, in this case sometimes called the data extractionist, can assist. From this level, the work performed by the research services could stretch as far as performing a complete package of analysis, answering specific research questions.

3.6 Financial models for post project funding

Many FOT/NDS datasets have been collected and the issue of post-project funding is a shared issue. Some projects are fortunate to have supporting funding, but it seems to be a key issue that the vast majority do not have the finances to keep the data available for further research. The waiting and search for new projects to come and finance the revival of the data is not fruitful, as it seldom happens - the start-up cost is too high. The data also need to be taken care of directly after the project, while the persons having worked with the data are still present. They need to do the final clean-up of the data, before they start with new project. In the case where no additional funding is available, the data might just be taken off the infrastructure and hopefully stored in a structured way. Another factor is the time period, where the data set is still interesting enough to attract a larger quantity of research projects. All these factors point in the direction of planning for data availability funding to be present directly when the project ends.

The next section elaborates on the items to take into account to have a successful funding of the data after the project. The items to be funded are identified, the interest from the financing bodies is investigated and finally, different funding models are discussed.

3.6.1 Items to be funded

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. The most urgent datasets to receive funding for re-use are the FOT/NDS that has been collecting continuous video. The amount of data could be more than tenfold that of the project collecting only signals, based on experience from euroFOT and SHRP2.

Table 6: 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

3.6.2 Financing bodies

There are different financing schemes for the projects applying to conduct an FOT/NDS. In Europe, it is common that the partners pay part of the cost of the project and own the data after the project, which results in that the data ownership often are un-evenly distributed across the project. This makes the ownership issue in the US, generally, a bit easier as the large FOT/NDS projects are fully paid and the data belongs to the authority. The projects are requested to hand over the data or may receive money to keep the data available.

Both on a national level and internationally in Europe and also in the US, the awareness of the value of the data is rising. Many countries, such as the US, Sweden and Finland, have written policies on making data open and available. This would imply that funding is redirected to hosting of and provision of data, which is the case in the above-mentioned countries. By doing so, the use of the data is facilitated, as the project does not have to pay for data, and thereby the re-use of data is enhanced. The former/current situation with the projects having to pay for accessing the data is not sustainable, as the projects are unaware of the funding issue for data and the data provider is often not participating in the project application. This is born out of the many international discussions and also expressed by people outside of Europe. People wanting to re-use data have no or too little money to pay for data, only man hours to use it.

3.6.3 Financial models

There are several different ways of funding the cost of maintaining and providing data for re-use. The following models are widely used among the data providers giving access to external users, for instance for accessing data from SHRP2, euroFOT and data at JARI.

Per project

The infrastructure gets funding by the projects utilising the data. In conjunction with the application, the cost is discussed. The cost is usually a generalized cost split per year, distributed over the estimated amount of projects, but 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 from 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. This model seems to be the most appreciated, based on all discussions during FOT-Net 2 with different organisations

hosting data for re-use and people wanting to re-use data. 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.

3.7 Application Procedure

The project should agree early on in the project on an application procedure for re-use of data, so that all project partners and possibly also third parties know the conditions for additional research using the specific dataset. 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, 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).

The suggested list of information to be provided by the applicant for a decision within the set response time is:

- Applicant details
- Short project description
- Requested data set
- Use and expected results

- Information on the intended publication of the data
- List of persons to get access and the related access time period
- Need of training in data protection and integrity issues
- Need of support and research services

4 Overview of procedures, documents, templates and standards related to data sharing

The framework of documents, providing assistance in preparing for data sharing, consists of a variety of different procedures, templates and “standards”. An overview is presented in the table below, where some of the content is provided in this document and others need to be developed.

Table 7: Procedures, documents, templates and standards

Data sharing area	Procedurs	Related project documents	Templates	”Standard requirements”
Project documents		DoW, CA, PA	Template text in CA and PA	
Data description		DoW, data description deliverable, CA	Data description and data format	Data and meta data description, data format
Data protection	Data extraction request, data download request	DoW, CA, PA	Data protection implementation documentation, data extraction request, NDA for analysts/visitors, Application to ethical review board	Level of protection at data providers/ analysis sites, data extraction format
Education		DoW, CA	Data security presentation, approved training certificate	Level of data security education
Support/research functions	X	CA	X	
Financial models			Form to describe the content to fund	
Application	X	DoW, CA	Application form, Data sharing agreements	X

5 Main Challenges

There are several large challenges in setting up a common data sharing platform. To make the platform 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 as in EU-funded projects, participating organisations pay between 50-75% of the cost and also own the data. This leads to different situations when it comes to the possibilities to gather and share the data after the project. Also 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.

The efforts to create and maintain such a platform 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 platform 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.

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.

Documentation of data and metadata, the most essential part of data sharing, 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.

Working Group discussions suggest that 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 what motivates a data owner to share the data.

6 Conclusions

The goal for the data sharing group was to create common data sharing rules for European projects. This report sets out the elements of a data sharing platform that would be required to facilitate re-use of the large amount of FOT/NDS datasets, stored in databases around the world. Such a platform would also facilitate data sharing within new projects, as the content of the platform is general and could be used whenever data sharing is performed.

The report constitutes the essence of the discussions made during the FOT-Net 2 time frame and there are many hands-on recommendations in the text. Through all the discussions, it is obvious that the text is universal, not only useful for European projects. At the end, it is always up to the specific project, national or international, to decide on their data sharing strategy and what parts in this data sharing platform that is applicable for their project.

The platform consists of the following seven items: pre-requisites that must be part of project documents such as the consortium agreement and the consent form, if the data should be able to be shared, descriptions of data and metadata, data protection, education on data security, support and research services, financial models for post project funding and the content of the application procedure. All parts need to be in place to efficiently form a data sharing platform.

The concept needs to be further developed and to be more in-depth adapted to the different national laws and research settings worldwide in order to be usable in as many FOT/NDS countries as possible. In particular, the data description of the list of possible FOT/NDS data types needs to be developed in close connection to the developments of the data protection concept. Still, if using this concept, with the suggestions and requirements involved, future FOT/NDs will be much better prepared for data sharing during and after the project than previous projects.

List of Tables