EUROPEAN COMMISSION

DG CONNECT

SEVENTH FRAMEWORK PROGRAMME

INFORMATION AND COMMUNICATION TECHNOLOGIES

COORDINATION AND SUPPORT ACTION

FOT-Net Data

FIELD OPERATIONAL TEST NETWORKING AND DATA SHARING SUPPORT

DATA SHARING FRAMEWORK

Deliverable no.	D3.1
Dissemination level	Public
Work Package no.	WP3
Main author(s)	Helena Gellerman, Erik Svanberg, Riku Kotiranta and Ines Heinig (SAFER), Clement Val (CEESAR), Sami Koskinen and Satu Innamaa (VTT), Adrian Zlocki (IKA) and Jörg Bakker (Daimler)
Status (F: final, D: draft)	F
Version number	1.0
Date	31 January 2017
Project Start Date and Duration	January 2014, 36 months

www.fot-net.eu

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 610453

Document Control Sheet

Editor: Helena Gellerman (SAFER)

Work area: WP3

Document title: Data Sharing Framework

Version history:

Version number	Date	Main author	Summary of changes
1.0	31.01.2017	Helena Gellerman (SAFER)	Final version

Review:

Name	**Date**
Yvonne Barnard (Leeds)	31.01.2017

Circulation:

Recipient	Date of submission
EC	31.01.2017

Table of Contents

1 Executive summary

In the past 15 years, we have seen great growth in the number of field operational tests (FOT) and naturalistic driving studies (NDS) performed worldwide. The data, mainly collected through naturalistic driving by volunteer drivers, have been used to answer the research questions in the original projects. As the number of different datasets has increased, so has the awareness of the substantial effort and funding needed to perform these FOT/NDS; as a result the interest in data sharing has grown across the globe.

The availability of a common data sharing framework, in which projects are set up in a similar manner (integrating data sharing pre-requisites into the project agreements from the start and using procedures and templates with the same content), will facilitate greater use of the collected FOT/NDS data. Researchers setting up new FOT/NDS would not need to decide on data-related issues for each specific project, but can focus instead on the project specifics, such as research questions and study design. Also, researchers wanting to re-use already collected datasets or several different datasets in the same research can utilise a more-or-less standard application procedure, rely on previous training that is widely accepted and plan for the costs to the project that using a specific dataset might incur.

The above-mentioned concept is elaborated in more detail in this data sharing framework. The framework consists of the following seven topics: (1) project agreements; (2) data and metadata descriptions; (3) data protection; (4) training; (5) support and research services; (6) financial models; and (7) application procedure. The topics have been discussed at the many international workshops, stakeholder meetings and topic-specific workshops arranged by FOT-Net, involving a variety of stakeholders from Europe, the US, Japan, Australia and China. The results of the discussions have been merged into the framework.

FOT-Net's Data Sharing Framework can be used by different stakeholders such as data providers, data re-users, consortia setting up a new data collection project and funding organisations. It is also applicable regardless of the current phase of a project, the category or amount of data collected or the size of the consortium.

The biggest constraint to sharing FOT/NDS data openly is the presence of video. Efforts targeting the development of methods for feature extraction from video would improve accessibility to key data for research.

2 Introduction

2.1 Background

Over the past 15 years, the methodological developments facilitating NDS and FOT have been primarily driven by two factors: the need to better understand the causal factors behind incidents and accidents and the continuous developments of inexpensive sensor and storage-capability technology."

The data has mainly been collected through volunteer drivers performing their day-to-day driving in their normal traffic environment. The data have been used to answer research questions in the original project, still many research questions usually remain unanswered due to lack of time and money which opens an opportunity for new projects re-using the data. The datasets vary in size, from less than 1 terabyte (TB) to several petabytes (PB), mainly depending on whether the data are collected continuously and whether they include video. The largest datasets so far were collected in the US (e.g., IVBSS, SHRP2 and Safety Pilot) and in Europe (e.g., euroFOT, DriveC2X and the on-going UDRIVE). In Japan, large datasets based on event recorders have been collected. Both Canada and Australia have several FOT and NDS datasets, such as CNDS and ANDS. It is noteworthy that data collection has also started in Korea and China, as the traffic environment and culture are so different from the above-mentioned countries.

As the awareness of the substantial effort and funding needed to carry out these FOT/NDS has increased, the interest in data sharing has grown. How to share data (also including concepts as Big Data and Open Data) is becoming even more important; it emerged as a key theme at the ITS Congress 2014 in Detroit. Numerous presentations addressed the problems associated with sharing all kinds of data, not only FOT/NDS. Regardless of the category of data, the key questions were the same; who should provide the data and how?

In FOT/NDS, the main focus is on evaluating the driver's behaviour in relation to the vehicle and the environment; the behaviour is recorded on both video and GPS. The drivers volunteer to be recorded in their daily lives and it is essential that the collected personal data is well protected. The presence of personal data in the dataset, together with the fact that this specific data is key to the research, makes FOT/NDS datasets one of the most challenging datasets to share.

Most of the earlier projects focused on learning the FOT/NDS methodology and answering the research questions set out by the individual project. These were major achievements in themselves. There was a lack of awareness of the implications of possible re-use of the collected data in the future. Therefore, many of these projects did not have the necessary pre-requisites in the consortium agreement and in the participant consent forms to share the data, at least not outside the project partners. Due to lack of time and funding, many datasets are not documented sufficiently, further hampering their re-use potential. Also, if tools were developed in the project, they were often tailor-made for the project and the tool requirement sheet did not include the view of a non-partner user. During these earlier projects, awareness about personal data increased, so data protection and security measures were developed. And finally, many projects did not discuss the nature of a data-sharing procedure, how to approve data applications and assist new projects in re-using the data or how to fund data maintenance after the project. The experiences from these earlier projects were first gathered into the Report from the FOT-Net's Data sharing Working Group (Gellerman, H., Bärgman, J., & Svanberg, E., 2014). The report forms the foundation for FOT-Net's Data

Sharing Framework and has since been revised and further elaborated, based on the discussions taking place during the course of the FOT-Net Data project.

2.2 Why share and re-use data?

There are different points of view on data sharing, depending on whether you are a data provider or a data user. The owner of the data has gone to great lengths (and usually used their own funding) to collect data and build up the data infrastructure and tools. Sharing the data after the project requires devoted persons to bring the data and tools to a level where they are easily understandable for someone who didn't participate in the project. It is therefore essential to understand how to compensate data providers for the efforts made to provide easily accessible data. Providing some benefit would also increase the number of data providers who are interested in opening up their datasets.

The data provider is usually, at least so far, also performing research, so the possibility of getting additional funding for further analysis through collaborations is probably the biggest motivation to provide data for data sharing. Opening up access to the dataset can stimulate a larger variety of research projects and increase the possibility of additional research funding.

The original project usually only performs a small part of the possible research that could be done on the collected dataset. From a funding organisation's point of view, utilising the already collected datasets for further analysis is an efficient return on investment. For project partners who already know the data, being able to further explore the data is good payback on invested efforts. During this additional 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 across FOTs and NDSs could provide a more quality-assured result than drawing conclusions from a single dataset. Further, using global datasets to research specific groups (e.g., older drivers) in different contexts and countries could provide insights into cultural differences in traffic behaviour for that specific group.

If funding for additional research is made conditional on international collaborations and data sharing, the global research community will be strengthened. Research collaborations create trust between organisations and thereby promote increased willingness to share data, enhancing the flow of ideas and knowledge.

Greater data availability is a great advantage for PhD candidates, making it possible for them to base their studies on a variety of real-world datasets. This variety also opens up more possibilities for the data to be used in other educational contexts.

These are some of the general advantages of sharing and re-using datasets. It is important, though, to identify the special circumstances that create a win-win situation between the data provider and the researcher in each specific case, and yet still provides adequate protection for the data collected from research participants.

2.3 Document usage

This document details a data sharing framework developed to facilitate sharing of data from FOTs and NDSs. It recommends procedures and document content on the following topics: (1) project agreements; (2) data and metadata descriptions; (3) data protection; (4) training; (5) support and research services; (6) financial models; and (7) application procedures.

The Data Sharing Framework facilitates data sharing regardless of the size or content of a dataset. The document as a whole is suited for large datasets, including both confidential/commercial data and personal data (including video and GPS tracks). Sharing large datasets imposes a greater effort in all the above-mentioned areas compared to a dataset with only a few signals and no video. The data-sharing of smaller datasets could, depending on the dataset, make parts of the respective chapters less relevant. Still, each chapter gives advice and recommendations that could apply to a variety of situations. Descriptive titles are given to the different sections, so that it is easy to find the content applicable to the data-sharing situation at hand.

The document describes seven topics of data sharing: five address primarily administrative issues, whereas two (the chapters on data and metadata and data protection) are more technically oriented.

2.4 FOT-Net Data project

FOT-Net is a networking platform open to all stakeholders interested in FOTs. It was established in 2008 as a European support action to let FOT experts benefit from each other's experiences as well as to give an international dimension to local activities. It organizes international workshops, publishes a series of newsletters and promotes FESTA – a European handbook on FOT methodology (FESTA, 2014).

FOT-Net Data is a Coordination and Support Action in the EU 7th Framework Programme for Research, submitted for the call FP7-ICT-2013-10. It stands for Field Operational Test Networking and Data Sharing Support. FOT-Net Data is a continuation of FOT-Net's activities. In external communication the activities will be referred to as FOT-Net in order to show continuity.

The main objectives of FOT-Net Data are to:

- Support efficient sharing and re-use of FOT datasets
- Develop and promote a framework for sharing data
- Build a detailed catalogue of available data and tools
- Operate an international networking platform for FOT activities.

The duration of the FOT-Net Data is 36 months, effective from 1 January 2014 until 31 December 2016. The project is funded by the European Commission (EC) under Grant Agreement number 610453. The EC Project Officer is Ms. Myriam Coulon-Cantuer from Directorate General for Communications Networks, Content & Technology (DG CONNECT).

The project partners are VTT Technical Research Centre of Finland Ltd., ERTICO – ITS Europe, SAFER Vehicle and Traffic Safety Centre at Chalmers University of Technology, Institut für Kraftfahrzeuge (ika) at RWTH Aachen University, Galician Automotive Technology Centre CTAG, University of Leeds, the European centre of studies on safety and risk analysis CEESAR and the automotive company Daimler. The project coordinator is Dr. Sami Koskinen, VTT.

3 Overview of the Data Sharing Framework

In the following chapters, the content of the Data Sharing Framework is described. Overall, the following seven areas, as shown in Figure 1, need to be addressed by the framework.

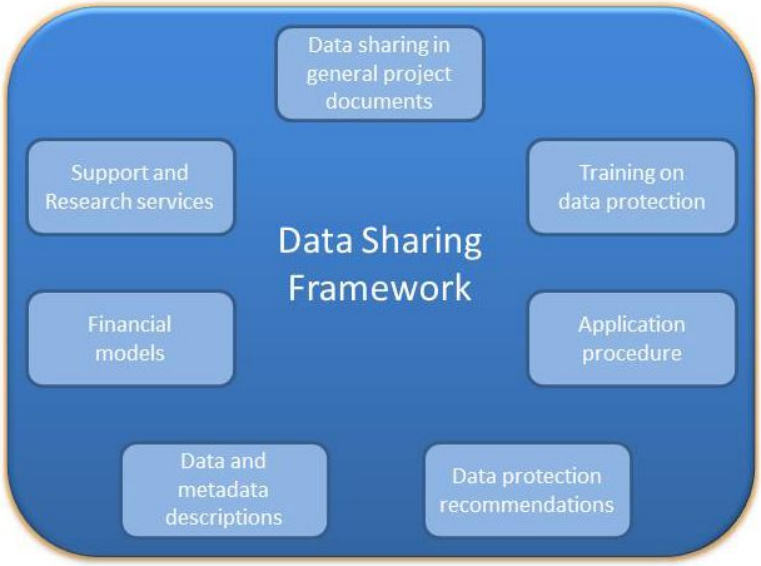

Figure 1: Data Sharing Framework

The Data Sharing Framework consists of:

- Project agreement content, including guidelines and checklists to incorporate the pre-requisites for data sharing in the agreements, which together with legal and ethical constraints form the conditions for data sharing. The project agreements include the grant agreement (together with the description of the work), the consortium agreement, the participant agreement and external data provider agreements.
- Data and metadata description recommendations to facilitate the understanding of the context in which the data was collected and the validity of the data. These include a suggested standard for the documentation of the data and metadata, divided into 5 categories: FOT/NDS study design and execution documentation, descriptive metadata (e.g., how the data is calculated), data (e.g., sampling frequency), structural metadata (e.g., how the data is organised) and administrative metadata (e.g., access procedures).
- Data protection recommendations, focusing on FOT/NDS personal and confidential data issues. It consists of security procedures and requirements at both the data provider and analysis sites, including detailed implementation guidelines.
- Security and human subject protection training for all involved personnel. The guidelines consists of 4 parts: who should be trained and when, what content should be part of the training (including detailed suggestions), how to do the training, and how to document it.
- Support and research services, proposing functions such as providing information/training to facilitate the start-up of projects, offering (for example)

processed data for researchers less familiar with FOT/NDS data, making analysis tools available or performing complete research tasks.

- Financial models to provide funding for the data to be maintained and available, and data access services. Eight financial models are discussed and a list of data management costs is provided.
- Application procedures which provide detailed content lists to address when developing application procedures and data application forms.

Another way of describing the common data sharing framework is by the contents of its documents, as in Table 1.

Table 1: Data sharing framework 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, non-disclosure agreement (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, training

Generally, the data can be managed either by the project itself or by an external data provider. An external data provider could also just provide test samples of the different datasets and guide the interested researchers to the organisation hosting the complete dataset. The current recommendation, however, is to let one or more project partner(s) from the original project maintain the data, possibly with test samples. Analysis of the datasets and research support services in most cases require a deep knowledge of the data and the way they were collected.

4 Data sharing in general project documents

The initial process of setting up a project is crucial in order to be able 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; the partners have entered the consortium 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 at the beginning of the project to agree on the conditions for data access and use (including data re-use after the project) is well invested.

The main documents to focus on are the funding agreement (including the description of the work), the consortium agreement among the project partners, the participant agreement and potential agreements with external data providers. This chapter discusses which topics to concentrate on, from a data-sharing perspective, for each specific project document.

4.1 Funding agreement including the description of the work

In the funding agreement and the description of the work, the result of the project and the funding are 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 on them during the project application and also during a possible negotiation phase. It is especially important to pay attention to the possibilities of providing open data after the project, based on the scope of the project and the data to be collected.

The requirements in a funding agreement are based on general requirements for projects collecting valuable datasets. As an example, in the European Commission's Open Research Data pilot, projects which are voluntarily part of the pilot are required to upload datasets to an archive of their selection before the project ends. It includes exceptions for projects collecting personal data.

The description of the work should include most of the topics listed in Table 2, at least on a high level. During the project application phase, it is especially important to address the possibility of post-project funding and other conditions which will keep the data available for sharing after the project—especially if there is such a requirement in the project funding conditions. For the following questions, it is beneficial for the project partners to develop common answers as early as the application phase:

- Who will own the data?
- May third parties access the data? To what extent? Under which conditions?
- Who owns the analysis tools and who will have access to them, if they are not generally available?
- Where will the data be stored during and after the project? Who is responsible for maintaining the data?
- How will the data be accessed? Who makes decisions on data provision?
- Are there legal and ethical or post-project funding constraints to be considered?

4.2 Consortium Agreement

The consortium agreement is an important document for setting the required conditions for data sharing and re-use of the data. Numerous topics need to be discussed and resolved in order to establish a legal platform for the handling of the data during and after the project. In Table 2, the topics to be included in the consortium agreement are listed. The questions should be seen as providing guidance in identifying the issues that would need to be solved for a specific project: they are all related to the possibility of sharing and re-using the data after the project. The text of the consortium agreement is fairly general; any details are developed during the project. Before the project ends, it is important to have a comprehensive written agreement for how the data should be handled after the project.

Table 2: Data-sharing topics within the consortium agreement

Topic	Comments
Ownership and access to data and data tools	• Who owns the data? • Is it necessary to add a specific ownership clause for the collected data? • How could the data be used and on which conditions? • Will all partners have access to all/part of the data? • May the data be licensed to third parties? • May third parties have access to the data and on what conditions? • Are there constraints related to personal data, especially video? • Are there future agreements with data providers to take into account? • Who will own the analysis tools and on what conditions are they licensed during and after the project? • Has a partner included previous work as background in the tools? • Who own the IPR to this work and how does it effect the project? • How can data be re-used if the data is owned by one partner and this partner ceases operation or leaves the project?
Storage and download of data	• How will the data be stored—centrally or distributed? • What are the general requirements for data protection and how are they assured? • Shall all/part of the data be downloadable for all partners and if so, under which conditions? • Shall all/part of the data be downloadable for third parties and if so, under which conditions? • Is there a time limit for requesting data for download? • Is there a time limit for keeping the data?

Access methods	• Can the data be downloaded, remotely accessed, or only accessed at the premises of any partner? • Shall a specific access procedure be used, and if so by whom? • Who should manage the access procedure? • How will the data be accessed? • What are the requirements for data protection for partners/third parties analysing the data?
Areas of use	• Shall it be possible to use the data for education, research and commercial purposes? • Are there special conditions for 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 provider of the project data after the project? • Which application procedure shall be used? • Who will grant access to the 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? • Are there time limits after which the data need to be deleted?
Post-project financing	• How will the storage and support services for data re-use be financed after the project? • Known or to be decided? • How will this funding be distributed?

4.3 Participant agreements including consent forms

The participant agreement explains the project to the participant and outlines the commitments required of both the project and the participant. It includes informed consent on several topics (e.g., the participants release their data for research). As the participants allow the project to follow them in their private lives for a period from a few weeks up to several years, it is important to be very clear on the use of the data during and after the project.

From a data-sharing standpoint, it is especially important to describe:

- what data are collected;
- where the data will be stored and who is responsible for the data;
- who (project partners/third parties) will have access to what data and on what conditions, during and potentially after the project;

- an overview of the access procedures;
- how anonymity will be ensured;
- the responses to the three YES/NO options below, directly related to data sharing.

It is recommended that the participant actively consent to these vital aspects of data sharing. Example text, which needs to be adapted to adhere to specific national regulations, is provided here for European conditions:

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

□ Yes□ No

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

□ Yes□ No

I also consent to the collected data (including video recordings and pictures) being re-used in other research projects by research partners/third parties, focusing on factors regarding:

- *the driver (e.g., drowsiness, distraction, driving style);and/or*
- *the vehicle (e.g., fuel consumption, system activation);and/or*
- *the traffic environment(e.g., road geometry, weather conditions); and/or*
- *… (to be completed by the specific project)*

□ Yes □ No

4.4 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 dataset. Contracts and NDAs should be signed. It is important to be aware of topics that can affect future research due to possible restrictions in data use. Attention from a data-sharing perspective should be given to answering the following questions:

- What is regarded as confidential information and what can be shared?
- Can confidential data be anonymised/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 restrictions on what the data can 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?

5 Data and metadata descriptions

FOT/NDS studies collect a large amount of raw data, especially when continuous data-logging is favoured over event-based data collection. Moreover, these studies also generate considerable amounts of derived data. Derived data can take different forms to address different needs. They can, for instance, be very similar to the raw data, simply representing the same information in a different format (e.g., in-vehicle signal values decoded from raw CAN frames). They can also be cleaned-up, filtered and/or discretized versions of raw measures. They can be a derived measure, where several pieces of information have been combined together to compute a new, more directly interpretable measure (e.g., time headway is the distance to the forward vehicle divided by speed; traffic density is calculated from traffic volume and speed). Lastly, they can be aggregated data, obtained using a data-reduction process, in which the most important aspects of the dataset have been summarised. The summarised data generally consist of a list of relevant events or driving situations and their associated attributes, the result of a mix of algorithm and annotation-based processes.

Depending on the aim of data re-use, simply re-using data in their most transformed/aggregated form may be sufficient. Occasionally, and when not prevented by intellectual property agreements (e.g., in the case of CAN data provided by vehicle manufacturers), it might be necessary to go back to the original, raw form. In most cases, however, cleaned-up, derived, annotated data will be the most useful. Whichever form of data is used, the core of data sharing is that the data provided are valid, or at least documented to a level where an assessment of the level of validity can be performed. This is potentially problematic if the data re-user was not part of the project and does not know in detail how the tests were performed, which sensor/version was used or how the data were processed from the raw data. The main problem is usually that the data are insufficiently described.

Data re-use requires precise knowledge about the data. Therefore, it is vital to have extensive and high-quality metadata, providing the following information:

- the conditions in which they have been collected,
- the purpose;
- how they have been stored, cleaned-up, processed and aggregated; and
- how they can be accessed.

A well-documented dataset inspires trust when being used, and also reduces the risk of less confident conclusions—something that all stakeholders benefit from.

In addition, before researchers/analysts/business developers even start to use a dataset, it has to be identified as potentially interesting and then selected as relevant for their purpose. These first steps only require a subset of the aforementioned documentation, which gives an overview sufficient to compare several datasets but is compact enough to ensure efficiency both in terms of creation and consultation. This results in the choice of items to be documented in the data catalogue. The content and structure of the FOT-Net Data Catalogue are described in FOT-Net Data deliverable D4.1, to be published spring 2017.

The aim of this chapter is to address these issues and provide methods for efficiently describing a dataset and the associated metadata. It suggests good practices for documenting a data collection and datasets in a structured way.

5.1 Definitions

5.1.1 Data

This document defines **data** as '**any information whose value might be used during analysis and impact its result**'.

This means that information which may be considered 'contextual', such as, for instance, participants' characteristics or weather, traffic and driving conditions, is considered data, and part of the dataset.

A categorization of data is proposed in the following chapter.

5.1.2 Metadata

This document defines **metadata** as '**any information that is necessary in order to use or properly interpret data'.**

This document presents four different categories of metadata, each providing a different kind of information about the data. The categories are described below and in Figure 2:

1. *FOT/NDS study design and execution* documentation, which corresponds to a high-level description of the data collection—its initial objectives and how they were met, description of the test site, etc.;
2. *descriptive* metadata, which precisely describe each individual category of data, including information about its origin and quality;
3. *structural* metadata, which describe how the data are organized; and
4. *administrative* metadata, which set the conditions for accessing the data and how access is to be implemented.

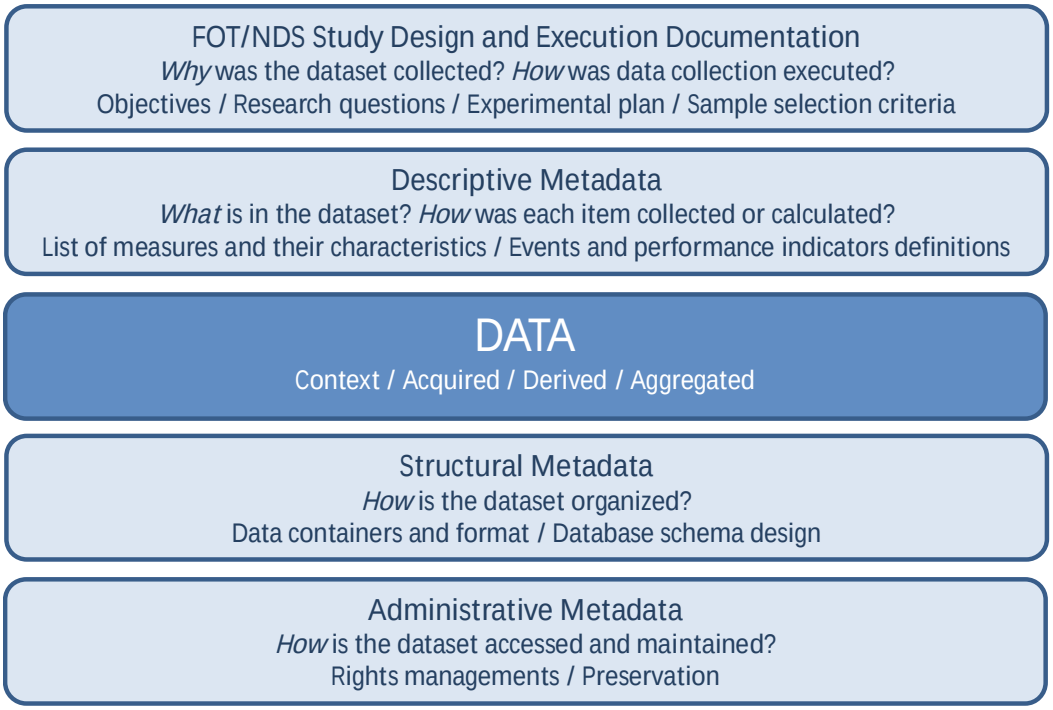

Figure 2: Types of metadata

In the following sections, FOT/NDS data are introduced and classified, including examples and recommendations for which data should be systematically collected (see 5.2). The four metadata classes are more precisely defined, and what should be documented (according to good practice) for each of them is described (see 5.3).

The recommendations are based on best practices from NDS and FOT projects—but additional information will probably be needed for each specific study.

5.2 Data categories

Data can take many successive forms, from raw collected data to very high-level aggregated data, with many steps in between. A dataset is not only the result of data collection, but also of an iterative process, comprising pre-processing, integration of different data sources, calculation of derived measures and manual and/or automatic data reduction. Aggregated data are usually the easiest to use, but may only be suitable for analysing research questions similar to the initial study. In contrast, raw data can meet a larger variety of needs, but usually requires a deep technical understanding of the data collection process and sufficient data storage and operational capacity in order to be used in a relevant and efficient way. A trade-off, using intermediary states of the data, generally has to be found.

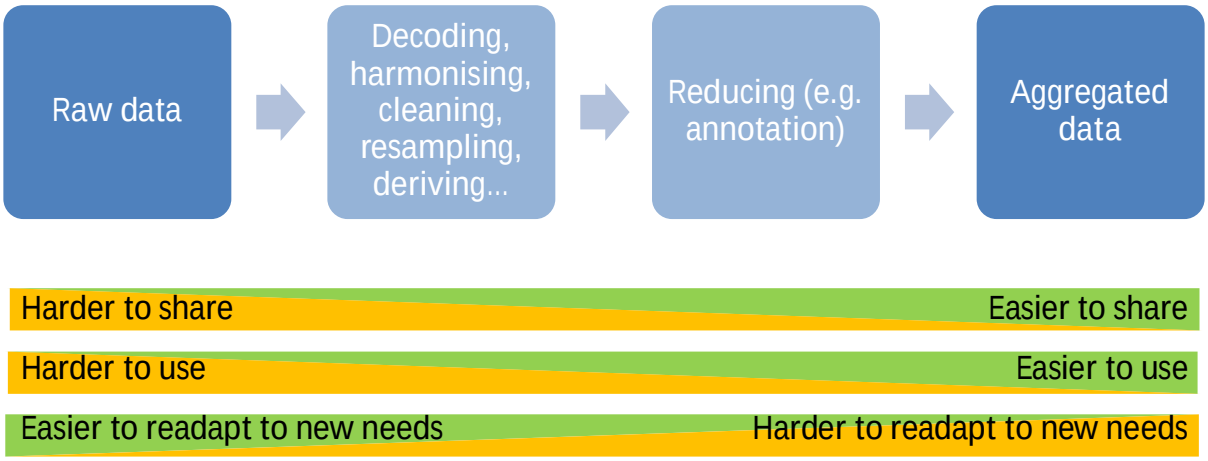

Figure 3: The trade-off between usability, usefulness, and availability

As a result, a data re-use case will typically require a combination of very different forms of data. This document proposes a way to classify them, based on two characteristics: the relations between the different entities (vehicles, users, infrastructure, etc.) addressed during FOT/NDS data collection, and the information which typically captures the entities' different aspects (measures). This classification system is as close as possible, but also complementary, to the definitions in FESTA, which essentially relate to data collection and analysis. This system emphasizes the typical structure of an FOT/NDS dataset and contains the following main categories: context data, acquired and derived data and aggregated data (see Figure 4). The sub-categories are described further in the following sections. Each sub-category may contain either objective data (which is normally quantitative data), subjective data (which can be either qualitative or quantitative data), or a mix of both.

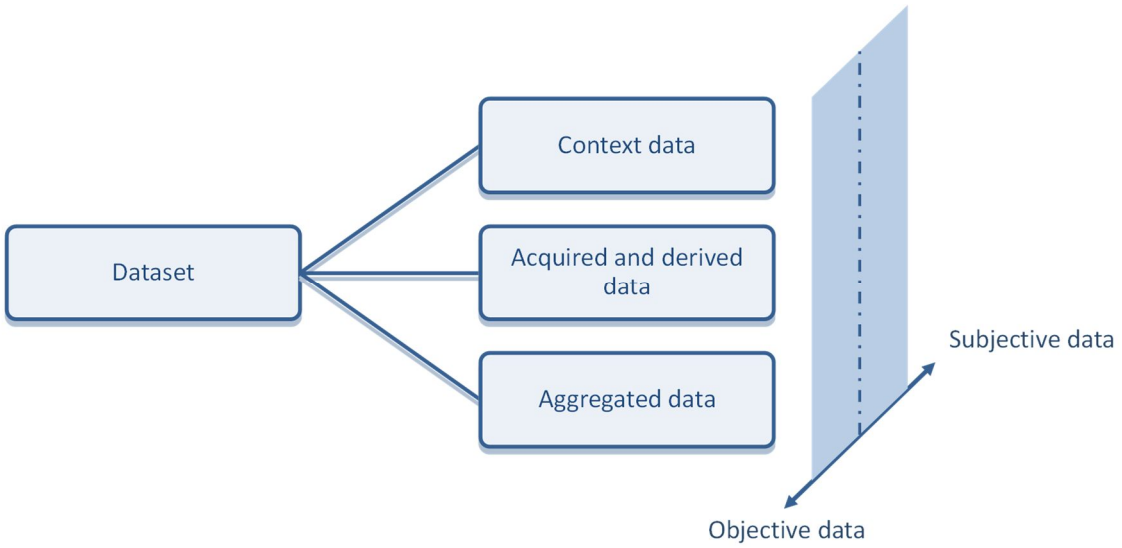

Figure 4: Categories of a FOT/NDS dataset

Objective data are collected through direct physical measurement, without any influence from the experimenter or the participant's subjective impression. They are collected using sensors, which can be pre-existing or installed on purpose, and data acquisition systems, which can be installed inside vehicles or on the roadside.

Subjective data are provided by the participants or observers, based on their impressions, feelings, memories or opinions—collected (for example) by questionnaires, travel diaries (usually quantitative data) or interviews and focus groups (qualitative data).

This categorization will be used as a basis for recommendations regarding what should be recorded in an FOT/NDS study, and how the corresponding metadata should be created.

5.2.1 Context data

Context data correspond to all information which doesn't change during the study, but helps explain the observations or document their values. They may be either directly collected, generated for the purpose of the experiment or previously existing and retrieved from external data sources.

They contain, for instance, background information—such as infrastructure characteristics (e.g., *map data*) and vehicle and driver characteristics, including questionnaire results.

Questionnaires collect qualitative and quantitative data reported by each individual participant. They typically cover basic data, such as age, gender and general attitudes about driving. They can also cover more specific aspects, such as personality traits (e.g., sensation-seeking, introverted). Quantitative data is obtained by means of closed questions (e.g., multiple choice, scales) whereas qualitative data is obtained when specific questions are open for rich text information, often of a more interpretive nature.

5.2.2 Acquired or derived data

Acquired data are all data collected during the course of the study for the sole purpose of the analysis.

Derived data are obtained by transforming raw data into more directly usable data through, for instance: data fusion, filtering, classification and reduction. They typically contain derived measures (such as, for instance, time-headway, which derives from both longitudinal speed and headway), and performance indicators (PI), referring to time- and location-based segments such as particular events.

In most cases, transforming acquired measures into derived measures during pre-processing or processing doesn't change their nature. For instance, an acceleration low-pass-filtered to remove noise doesn't cease to be a vehicle-dynamics measure; the depression of a pedal converted to a discrete *pressed/not-pressed* state doesn't cease to be a driver-action measure. As a result, in most cases, the subclasses presented below apply to both acquired and derived data.

However, in some cases, several kinds of measures are combined together to form new, more interpretable measures, which can't be categorized simply. For instance, speed and acceleration from several vehicles can be combined together to form a time-to-collision variable.

This category includes both objective data, in the form of measures from sensors (referred to as sensor data in FESTA), and subjective data, collected from either the participants (referred to as self-reported measures in FESTA) or analysts. Subjective data can be as varied as time-history data, subjective classification of time segments, or rich-text information from travel diaries, interviews, and focus group discussions. Questionnaires can also be seen as acquired data when collected periodically during the project (compared to the static questionnaire data described in 5.2.1).

Time-history data

Time-history data describe the history of a measurement over time. Time-history data can be collected with a specific measurement frequency, or when triggered by an event, typically a value change.

Time-history data may consist of the variation over time of single physical values (e.g., speed), a collection of physical values (e.g., 3-axis acceleration) or more complex media, such as sound or video.

Time-history data can either be collected from the vehicle perspective, by means of (for instance) an instrumented vehicle, smart device application or travel diary, or from the infrastructure perspective, by means of roadside measurements. They can be historical or real-time observations.

Time-history data consist of both direct measures, i.e. raw data measured over time, or derived measures, after any kind of transformation (such as resampling, offset correction, filtering and removal of incorrect values) has been performed.

In-vehicle measures

Instrumenting vehicles enables the collection of vast amounts of data, using either original sensors (tapping their communication networks, such as CAN) or additional sensors. Applications on smart devices (i.e. smartphones) can also collect important information in the following categories, and the data they collect can basically be treated the same way as the data from instrumented vehicles.

Vehicle dynamics

Vehicle-dynamics measurements describe the motion of the vehicle. Typical measurements are longitudinal speed, longitudinal and lateral acceleration, yaw rate and slip angle.

Driver state and actions

In addition to variables describing driver actions which command the vehicle, like steering wheel angle, pedal activation or HMI button press, variables characterizing the physical and emotional state of the driver can also be measured. For instance, cameras and computer vision can measure driver position, detect engagement in a secondary task or detect eyelid closure (which highly correlates with alertness).

In-vehicle systems state

The state of in-vehicle systems can be accessed by connecting to the embedded controllers. The data category comprises continuous measures, like engine RPM, or categorical values, like ADAS and active safety systems activation.

Environment detection

A precise understanding of the environment can be obtained by advanced sensors like radars, LIDARs, cameras and computer vision, or by simpler sensors (e.g., optical or temperature). For instance, luminosity (indicating the presence of rain), characteristics and dynamics of the infrastructure (e.g., lane width, road curvature) and surrounding objects (e.g., type, relative distance and speed) can all be measured from within a vehicle.

Vehicle positioning

The geographical location of a vehicle is most frequently determined with satellite navigation systems (e.g., GNSS or simple GPS) and the aforementioned advanced sensors. It can also be determined by information from the cell phone network, surrounding Wi-Fi networks, or a combination of these and GNSS.

Media

Media data are usually video, but in some data-collection projects audio is recorded. Media data also include the index files used to synchronize these data with other data categories.

Human behaviour measures

Complementary to sensors and instrumentation, some continuous measures can also be built through the perception of analysts or annotators using video data. Eye glance and driver state (e.g., drowsy, impaired, angry) can be evaluated manually by analysing video from driver-face-oriented cameras.

Roadside measures

Roadside measures comprise vehicle counting, speed measurement and positioning—using radar, LIDAR or simpler rangefinders, video-based counting, inductive loops or pressure hose. In the case of ITS systems, they may also contain more complex information remotely transferred from vehicles to roadside units. Media data (typically video—for instance in traffic conflict observations) are also often

collected from beside or above the roadside. Roadside measures are evolving rapidly, with data being collected by drones or open-data services, for example.

Experimental conditions

Experimental conditions are the external factors which may have an impact on participants' behaviour. They may be directly collected during the experiment, or integrated from external sources. Typical examples are traffic density and weather conditions. Controlled factors, such as the ability to use a system, also need to be included in the dataset, depending on which phase of the experimental plan a participant is currently participating in.

Time and location segments

For the purpose of the analysis it can be relevant to analyse the data aggregated for a delimited period in time or space (such as journeys, certain events as defined in FESTA or e.g. road segments). These data segments are defined by a combination of specific conditions and characterised by specific attributes, some which are automatically computed, and some which are manually annotated from video. The attributes mostly consist of situational variables and/or PI, depending on the studied phenomena and its expected contributing factors; they can also consist of links to other segments or contextual data. For instance, each trip might link to a specific driver and vehicle, each of which have their own characteristics. Finally, the segments might serve as a container for time-history data: a trip can contain the history of the vehicle speed and an event may contain successive eye-glance values, manually coded by an annotator. As a result, the segments contain a large amount of initial data, which is structured, reduced and summarized into more manageable tables, suitable for data analysis.

The creation of the segments can either be automated (i.e., they are created in response to a specific value or threshold of one or more variables), manual (when a specific event is observed on a video), or a combination of both (e.g., automatic detection of candidate events, accepted or rejected in video annotation). In the same way, attribute values can either be automatically computed (i.e., the mean or maximum value of a measure during a time segment) or manually annotated, typically from video. In the latter case, standardized annotation schemas are used to enrich data with information available from video recording. Annotation variables are thus a subjective assessment of the situation by an analyst or annotator. They can be quantitative, using single or multiple choices (i.e., present/not present or level of rain); they can consist of specific time stamps, for example when the driver is first aware of a hazard; or they can be qualitative narratives, which describe a specific event or situation.

Finally, subjective, participant-reported data can be collected as certain kinds of segments, such as self-declared events, or they can populate some segment attributes (e.g., travel diaries that contribute some characteristics to trips).

Time segments

Time segments are the most common type of segments, collected and/or generated during data reduction. They correspond to a time period when some specific conditions are met. Depending on the kind of conditions which define them, their typical duration, and the researcher's own vocabulary, they are identified as trips (a vehicle is started, driven for a period of time by a driver, then stopped), events (typically a short period of time with very particular characteristics), situations or

chunks (division of the complete dataset into segments of comparable size according to a combination of situational variables, characterized with PI).

Locations

While time segments take the perspective of a driver in a vehicle during a trip, locations take the perspective of a place, where multiple trips might pass through. Roadside observations will typically generate locations, and typical location attributes are vehicle counts or speed measurements, which can also be associated with the infrastructure attributes. Furthermore, using geographical information systems (GIS), data from in-vehicle collection can also be projected over a geographical reference system to characterize, for instance, one or several participating drivers' behaviour at a specific location such as an intersection.

5.2.3 Aggregated data

Using relations between segments, reduced data (e.g., segment attributes) are typically aggregated into smaller, more usable tables, suitable for data analysis or data interpretation. For instance, driver characteristics can be grouped together with attributes from one type of situation, to evaluate the impact of drivers' characteristics on their behaviour in that situation. The data resulting from aggregating different kinds of reduced data together are called aggregated data. Although they are generally linked to a specific research question, the aggregated data may be re-used with different statistical methods, or re-aggregated with other data, to quickly answer new questions without the need to go back to harder-to-use, raw data.

As they don't contain instantaneous values, they don't allow potentially problematic re-use, such as pinpointing illegal behaviour from one specific driver or benchmarking a driving assistance system without authorization from its supplier. As a result, aggregated data are generally easier to share than other categories of data.

5.3 Metadata

Metadata basically provide information about data and can be divided into different types, describing different traits of a dataset. Descriptive metadata, describing the content of a dataset, is perhaps the most useful type for data analysis. In contrast, structural metadata are the prerequisite that helps the analyst understand the structure of the dataset, by describing 'data about the containers of data' (Roebuck K., 2012). Administrative metadata are collected for the effective operation and management of data storage. Finally, the FOT/NDS study documentation provides an overall description of how the study was performed.

5.3.1 Descriptive metadata

Descriptive metadata shall include detailed information needed to understand each part of a dataset. The purpose is to describe the dataset and build trust in it—by providing not only the characteristics of each measure or component, but also information about how the data were generated and collected.

Descriptive metadata shall preferably be available close to the actual data to facilitate analysis. The descriptive metadata need to define the dataset and include detailed descriptions of measures, PI, time and location segments and their associated values. In addition, external data sources, subjective data from self-reported measures and situational

data from video coding must be described in detail. Not only must the output of the data be described, but how the data were generated and processed is equally important; this is where one can build trust in the dataset. The more thoroughly the origin of a measure is described, the greater the trust. The proposed structure of descriptive metadata follows the data categories in 5.2.

Context data description

The level of detail when describing contextual data can vary. Information about drivers and vehicles is often obvious from the name of the variable (e.g., gender and age for participants, and model, brand and year for vehicles). Other information, such as questionnaire data acquired from participants, might need a more in-depth description (e.g., a definition of the self-assessed sensation-seeking measure).

As FOT/NDS databases often consist of a variety of different external data sources, it is very important to document them all to get a full picture of the data. The external data sources can include static contextual data from map databases or dynamic data from weather services and traffic management services. In these cases, a more in-depth description is needed where it is important to describe the origin of the data, the methods used to match the different datasets (e.g., a description of the map-matching algorithm), and each output variable.

Some additional data might be merged with the acquired data (e.g., map attributes or weather codes). These data are described in their respective sections below.

Acquired or derived data description

A description of every measure in a dataset is mandatory, making the data re-usable for future analysis. The origin of the data and the processing steps performed are equally important for drawing correct conclusions in the analysis.

It is important to include definitions of time and location segments in descriptive metadata, as the definitions vary between different datasets depending on the purpose of selected segments. The segments need to be defined (i.e. how the segment start and stop times are calculated), and so do the associated attributes (e.g., summaries, situational variables and PI). The different types of time and location segments are often important products of the dataset, providing easy-to-use references to the actual data.

This section also includes a suggestion for describing PI and summaries—data which are often attached to time or location segments, but may also be used independently of them.

Direct or derived measures in time-history data description

The description of direct measures is often beyond the project's control and needs to be requested from the supplier of the equipment generating the data. If the data are acquired from the CAN bus of a vehicle, the OEM can supply information which describes the data. Understanding the origin and full history of direct-measure data is important, but often overlooked. To get access to this information, the use and restrictions of direct-measure metadata should be included in the contracts and NDAs with the suppliers. The origin of the measure should at a minimum include where the data were generated (e.g., CAN or other equipment), the frequency, the units, whether they were derived from other data and error codes.

When direct measures are being processed into derived measures, it is important to document all the data processing steps. Derived measures are often processed

several times, and the final product might consist of more than one measure. The need for a detailed description is crucial for creating trust for data re-use.

The output of the data processing must be documented and include information on data precision, unit and sample rate. This metadata must also include information about how the data were processed (e.g., synchronization policies, re-sampling filters, harmonization rules). In an ideal scenario, an analyst performing an analysis can quickly understand not only the meaning of the measure, but also its origin and history, and use this information to interpret the results.

Proper naming conventions for all data containers can go a long way towards helping interpret data's origin and understanding how it can be used. Tags describing the data type and origin can, for instance, be used. However, naming conventions are always a trade-off between comprehensiveness and legibility, and although necessary, are not sufficient for the proper documentation of a dataset.

Preferably all information in Table 3 should be included for each major data-processing step. As an example, interpolation filters must be documented in detail, so that the analyst can understand whether the measure can be used for a specific research question. Additionally, the tolerance for missing data (e.g., the number of frames or seconds) and how these values are stored should also be described in the metadata, because the values are often managed differently in different data formats (e.g., NaN in MATLAB, but NULL in Java and relational databases). Describing the measure in detail avoids misinterpretation.

Table 3: Metadata attributes of time-history data measures

Data description item	Instruction/example
Precision	What is the accuracy of the measure?
Unit	What is the unit of the measure (e.g., m/s, RPM or if an enumeration)?
Sample rate	What is the current frequency of the measure (e.g., speed resampled at 10 Hz or 1 Hz)?
Filter	Which filters were applied (e.g., low-pass, interpolation or outlier filters)?
Origin	How was the measure generated and from what data source? For instance, it is important to know if the speed measures originated from CAN at 20 Hz or GPS at 1 Hz. This could also refer to another described measure.
Type	Is the measure an integer, float, string or picture file?
Range	What is the expected range (minimum and maximum values) of the measure?
Error codes	Which values trigger error codes? What is a null value? It is

	also important to describe how the errors are managed.
Quality	Are there any quality measures related to this measure and how are they defined? The quality could be set on a per-trip, per-measure or even per-sample level (e.g., for GNSS data: HDOP, number of satellites).
Enumeration specification	Can enumerations be translated into readable values (e.g., 1 means left and 2 means right for the turn indicator)?
Availability	Can the measure be shared? What are the conditions to access it?

Time segment data description

Calculated time segments or triggered events represent singularities over time, which may be as short as a single time instance, or longer based on a specific set of criteria. The definitions of time segments differ among datasets; the more common ones are trips, legs and events. This variation makes it even more important to describe the purpose and how the segments were designed, including their origins. It is also important to understand the conditions that define the start and stop of a time segment.

Events are often described by type, which explains why an event was triggered or threshold met. To understand the event properly, event type descriptions must include references to the measures and method used to calculate the event, as well as threshold values.

Different segments can have different associated PI, summaries or attributes, and these should also be described: for example, a trip record might include the duration, distance travelled, average speed, number of times passing intersections, or just the number of samples. Time segments should include the attributes in Table 4.

Table 4: Metadata attributes of time segments

Data description item	Instruction/example
Type	What is the purpose of the trigger (e.g., a hard braking event, swerving at high speeds, overtaking or entering an intersection)?
Definition	What is the definition of the time interval? How are the time series grouped? The output could be a single point, fixed or variable in time.
Origin	Which measures were used to create the entity? What was the overall principle of the data computation that generated the entity?

Unit	What is the unit of any output value (defined by type)?
Enumeration specification	Description of enumeration values.
Attribute, PI or summary specification	Time segments might have associated data that need description. It could be attributes, such as driver ID or duration. It could also be computed data, such as PI or summaries (e.g., distance travelled, number of intersections passed, average speed or the number of times a button was pressed). The definition of all PI and summaries associated with the object are described later in this chapter.
Availability	Can the segment be shared? What are the conditions for accessing it?

Location data description

In many studies the vehicle is not the main entity; rather it simply provides values for locations. Locations must be defined, usually by position or a set of positions. This could be an intersection, a sharp bend, the specific position of a roadside unit or a stretch of road (anything from a city street to a European highway). The definition is of great importance because of this great variance. As with time segments, the value of the locations is not only the encapsulation of time or position, but also the determination of associated attributes and the output of computations. The metadata attributes of location segments are presented in Table 5.

Table 5: Metadata attributes of locations

Data description item	Instruction/example
Type	What is the purpose of the location segment?
Definition	What is the definition of the location, in terms of position, scenario or equipment? Can locations be grouped or arranged in a hierarchy?
Attribute, PI or summary specification	Location segments might have associated data that need description. It could be attributes, such as number of exits at a roundabout. It could also be computed data, such as PI or summaries (e.g., number of vehicles passing or average speed). The definition of all PI and summaries associated with the object are described later in this chapter.

PI and summaries definitions

PIs are used to measure the performance of one or more measures, and are often associated with a specific analysis project, although some might be re-used for other

purposes. Each implementation of a PI should therefore be described precisely; see metadata attributes in Table 6.

PIs as summary tables are pre-computed data, used to make the analysis more efficient. The summaries are stored as attributes, often with time or location segments as a base; the summaries could, for example, describe the mean speed of a trip or the number of passes through an intersection. Summaries are convenient in data reduction. They are especially useful in a larger dataset for excluding data not needed for the analysis.

Table 6: Metadata attributes of PI or summaries

Data description item	Instruction/example
Purpose	What is the purpose of the PI or summary?
Definition	Details about how the PI or summary was calculated.
Origin	Which measures were used to create the entity? What was the overall principle of the data computation generating the entity?
Unit	What is the unit of the output value?
Precision	What is the accuracy of the PI or summary?
Availability	Can the attribute be shared? What are the conditions for accessing it?

Description of video annotation codebook

Documenting the video annotation codebook is important for helping the person coding the data to understand the instructions, but also for defining enumerations (for incident severity there are the conditions that define a crash, near-crash, increased risk or normal driving). It is also important to document the process of coding the data, whether inter-rater reliability testing was conducted, and other important aspects of the persons coding the data; typically this information is part of FOT/NDS study design. For each measure (as part of the video annotation codebook) the recommendation is documented in Table 7.

Often the reduced data are coupled to time or location segments. Because it is important to know why those segments where selected for video coding, the reference must be documented.

Table 7: Metadata attributes of video annotation code book measures

Data description item	Instruction/example
Description	What is the purpose of the measure?
Instructions	In what way was this measure described to the person coding the data?
Type	What type of input is expected (single or multiple choice: e.g., present/not present or level of rain, continuous, free text or voice)?
Options	What are the possible alternatives (often coded as enumerations)? How reliable are the data expected to be?

Description of self-reported measures

Other subjective data include travel diaries, interviews, and documentation from focus groups. These data are often in rich text format and the data description should cover why, when and how the data were collected.. Questionnaires acquired during the data collection period should also be described in this section. These data are very similar to video annotations and could be described by answering the questions in Table 8.

Table 8: Metadata attributes of self-reported data

Data description item	Instruction/example
Description	What is the purpose of the self-reported measure?
Instructions	In what way has this measure been described to the participants?
Type	What type of data is expected (single or multiple values, continuous, free text or voice)?
Options	Descriptions of possible alternatives (often coded as enumerations) and how non-answers should be handled.

Aggregated data description

The shape of aggregated data can vary to such a degree that it is difficult to propose a structured format. Depending on the level of aggregation, the data could be described as time history measures or time segments. Also, in many cases the aggregated data are shared with the promise that the underlying data will not be revealed; the algorithms are not described in depth (to eliminate the risk of making raw data information available by means of reverse engineering), and only a high-level description is allowed. The trust in this data will be reduced and it is up to the recipient to judge if it is good enough for re-use. An appropriate set of metadata questions is proposed in Table 9.

Table 9: Metadata attributes of aggregated data

Data description item	Instruction/example
Description	What is the purpose of the aggregated data?
Definition	Which algorithms were applied to the underlying measures?
Origin	Which underlying measures were used to calculate the aggregated data?
Unit	What is the unit of the output value?
Precision	What is the accuracy of the output value?

5.3.2 Structural metadata

In a typical FOT/NDS study, different parts of the dataset will use different storage technology, such as file systems, SQL and Not-Only SQL databases.

Structural metadata are used to describe how the data are structured in relation to other data. Data are organized into a system (e.g., a database and/or file system), a structure or database schema and a data content format. The aim of structural metadata is to facilitate the initial phase of data re-use by providing the necessary documentation about how the data is organized. The description should include the file system, the file structure and how to interpret the contents of a data container. All components of the dataset need to be described.

Since data may be stored for a very long time, it also becomes important to describe and preserve tools that can read the data. This issue is highlighted when it comes to data archives. Even only five years after a project has ended, the knowledge about specific tools might have been lost and the cost of building up the competence again might exceed the data's value. It is therefore recommended that the tools, platform and prerequisites be described—in even more depth if using a non-standard data container, file format or file structure.

File system/Database

At the lowest level the file system format, or encapsulation, must be known. This information gets especially important as the years go by, as tools and formats slowly depreciating and are replaced by newer technologies.

Popular formats include NTFS (for Windows), EXT4 and XFS for Linux, or FAT32 (supported on many platforms). However, the particular demands and scale of NDS/FOT studies might require less common file systems. Examples are ZFS (Unix) and ReFS (Windows), which offer superior reliability for large volumes. Some file systems also contain metadata for each file, such as the 'forks' in HFS. For large projects requiring scalability and distribution of calculations over many servers, data may also be stored on a distributed file system such as HDFS.

If data are stored or archived in a relational database (e.g., Oracle, MySQL) or a Not-Only SQL database (e.g. Cassandra), it is important to know the type and version, to facilitate data import to an identical system or conversion to a different product.

Files themselves can also be encapsulated in archives (with or without compression and/or encryption) or in binary objects in databases.

File structure/database design

The file structure should be described. As an example, it could be described as Vehicle/Year/Month/Trip.

Files might not always be accessed with a traditional file system; if not, it is also important to describe how to access them. Examples include Content-Addressable Storage (CAS). The analyst accesses the content, without knowing its location, using a key.

It is recommended that the schema be documented graphically to indicate the relations between the different tables, a task usually easily accomplished using data management software. This principle should be applied whether data are stored in a relational database data or an alternative (i.e., in a file system or Not-Only SQL environment)

Data container

The data container describes the format of a file. This could be avi for a media file, csv for a text file or mat-file for data used by MATLAB. With a non-standard format it is important to describe it in detail, including file content structure, header length, data type and indices. It is also good practice to include information about tools that can interpret the data format of the container.

Content

The content description should include how the data are organized in a file or object. Thus codec and indices could be provided for an avi file, the description of a row for a csv file and the object design for a mat-file. It is recommended that the data descriptions be kept in a readable format; XML is recommended, since most tools/programming languages have built-in methods for reading xml files. A description of the file contents gets even more important if a non-standard format is used. Similarly, when different data types are mixed in the same file (e.g., video and CAN data) it is vital to have a precise description of the content. The content description of a database includes detailed information about the tables, such as columns and their respective data types, indexes, triggers, sequences and views.

5.3.3 Administrative metadata

Administrative metadata are collected for the effective operation and management of data storage and catalogues. This administrative information, covering various topics, is stored along with the datasets. From a FOT data re-use perspective, the key role of administrative metadata is to cover access conditions, rights, ownership and constraints. Generally, administrative metadata can include (Puglia, S., Reed, J., and Rhodes E., 2004):

- version number
- archiving date
- information about rights, reproduction and other access requirements
- archiving policy

- digital asset management logs
- documentation of processes
- billing information
- contractual agreements
- end of life of the data.

The method for storing administrative metadata depends on the specific repository/catalogue. Many of the items above need to be stored at least as supplementary documentation, according to repository/catalogue guidelines, if not directly as attributes of a dataset. The administrative metadata also have a role in data protection: defining processes, personal data management, access rights and keeping track of (for example) periodic backups.

For online FOT catalogues, information about a contact person/organisation and licensing options or required agreements must be included, so potential analysts know how to gain access to a dataset. Another required administrative feature of a catalogue is usage logs of information queries and retrieved data, in order to be able to summarize the level of interest for different datasets.

Assigning persistent identifiers for datasets is necessary for references and citations. Some persistent identifiers like the Digital Object Identifier (DOI) also support dataset version management. Each time there's a change in the data, a new DOI is assigned and a log of changes collected.

5.3.4 FOT/NDS study design and execution documentation

The study design and experimental procedures must be documented well enough so that persons and partners who did not take part in executing the test can perform analyses. The main purpose of this documentation is to describe in free form the purpose of the data collection, the experimental procedures and the important details of the actual execution—including a description of the test site, which must be known before the data are interpreted. As a result, this documentation should contain not only initial plans, but also the final details of the study. More information is available in the FESTA Handbook. The document should give an overview of the following (at least):

- purpose of the field tests or data collection;
- research questions;
- sample selection criteria and overall description of recruitment;
- possible grouping of participants (e.g., test groups 1 and 2 and a reference group); description of the groups;
- overall description of equipment used, functions, HMI, additional driver support in the vehicle (navigators, etc.) and vehicle fleet—preferably with links to videos demonstrating usage;
- description of the test site (if it was within a specified perimeter), including maps and photos;
- date and timing of different phases of the study;
- description of scenarios/test runs/study phases, if relevant—with photos of key locations and views from participants' perspectives;

- test plan and execution, describing (for example) what the participants were asked to do, how and when the briefing was given, what questionnaires were administered or what interviews were given;
- in the case of an FOT, how the participants were introduced to the system
- how contact was maintained during the study;
- special events and changes that may affect data analysis (e.g., roadwork, strikes, economical changes, special weather);
- summary information of the project and cooperation partners, duration, budget etc.

6 Data-protection recommendations

Data protection is the key to creating trust between a data provider, data owner(s) and the researcher. The data provider is responsible towards the data owner(s) to ensure that data are being handled according to agreements or contracts as well as the legal context in the country where the data is managed. Subsequently, if the data provider knows that the researchers have good, proven procedures in place to keep control of who is using the data, and that the researchers have knowledge of the legislation surrounding the handling of personal and IPR data, they will be more willing to share.

This chapter applies whenever the data are shared between two organisations. There are many different scenarios where data can be shared and the organisations must discuss the following questions beforehand:

- How are the data going to be accessed between the organisations?
- Should each organisation have a dataset?
- How can the data be transferred?
- What physical security requirements must be in place?

When data are collected and used within the same organisation there might be greater control of how the data is handled, but this chapter could still be applicable.

This chapter discusses the different demands imposed on data protection by different kinds of data. The scope of data protection includes unauthorized access, data theft, data loss and the proper documentation of the implementation. The chapter also includes a suggestion for data-protection requirements to facilitate the setup of the necessary data-protection framework, for a data provider in the role of a data centre (DC) and a data user in the role of an analysis site (AS).

6.1 Stakeholders

There can be many stakeholders involved when two or more organisations decide to share data. The researcher, later referred to as the *data user*, is the person who will use the data for analysis. The *data provider* shares the data with another organisation. The *data owner* is the organisation that owns the data according to contracts with the data provider(s). In many cases the data provider and data owner are the very same organisation.

In addition to these three, this document defines a *data centre* as the organisation that makes FOT/NDS data available to more than one data user. The distinction between the data centre and the data provider is that the latter shares data with another organisation, but does not interface directly with the data user. This document also defines *analysis site* as the organisation establishing data access for a group of data users.

It is important to state that a single organisation can act in one, many, or even all, of the roles.

A *study participant* is defined as the person who generates the data being collected. This is an important definition, as this person is protected by legal rights concerning the usage of the data.

Data centre

The data centre must implement appropriate data-protection means to ensure responsibility and liability, as stated in agreements with data providers, data owners and the study participants. A person downloading datasets to a computer is therefore not considered a data centre unless the data is made available to others.

When transferring the data to another organisation, the data centre organisation itself becomes a data provider. This can lead to chains where many organisations host data as data centres, while also sharing data as data providers. It is important for all parties in the chain to have a clear picture of the data flow, comply with the data-protection requirements and thoroughly understand the data ownership and privacy laws in the country where the data are being managed. In fact, a dataset (or parts thereof) can be owned by one or more organisations. The data ownership and usage are regulated in the agreements between the organisations (see Chapter 4).

Data user / Analysis site

A data user might be allowed to download data from a data centre or operate within an analysis site. An organisation can establish an analysis site where the requirements stated by the data centre are implemented. The data users within an analysis site must accept and follow the data-protection principles. In many cases an organisation establishing a data centre also acts as an (internal) analysis site, although it might be practical to keep the distinction between the two, especially in large organisations when managing personal and/or confidential data.

6.2 Data classification

The level of data protection required depends on the harm the data could do if revealed and the legal requirements. If the dataset consists of personal or confidential commercial data, it is mandated by law that action is taken to ensure data protection, regardless of the size of the dataset. Confidential commercial data is usually accompanied by agreements stating the conditions for access and use, whereas the use of personal data is regulated by law and the agreement with the participant (consent form). This document classifies data into personal, sensitive personal, confidential, and non-sensitive data. This classification will be used frequently in this document and it is therefore important to understand it; the different categories are defined below.

Personal data that need protection

The term *personal data* is defined in European Directive 95/46/EC Art. 2:

> *´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.*

The directive also specifically defines *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.

Personal data are therefore classified as either sensitive personal data or more general personal data. The suggested data-protection requirements in this chapter aim to guide the data centres and analysis sites in setting up a data-protection concept that meets the regulations and respects the will of the participants as stated in the consent form.

The new European General Data Protection Regulation (GDPR, REGULATION (EU) 2016/679) will reform the usage of personal data; the regulation will enter into force on May 24th 2018, and apply from the day after. GDPR will strengthen the rights of the individuals and set a common legal framework for all European Union countries. The definitions of personal and sensitive personal information will still be applicable but the term sensitive will not be used explicitly (see Article 9, GDPR). It is absolutely vital for any organisation operating in the European Union and managing personal data to investigate and ensure that personal data are managed according to the new law.

Even if GDPR harmonizes the regulations in a European context there will still be differences in implementation between the US, Australia and Asian countries. For example in the US, 'personal data' are known as 'personal identifiable information' (PII) and 'sensitive personal data' are known as 'sensitive personal information' (SPI or SPII). The definitions are not exactly the same as in Europe, and it is therefore advised to take any necessary actions to ensure that data are managed according to the laws of the country/ies where the data are located.

Confidential commercial data that need protection

Confidential commercial data is information which an organisation has taken steps to protect from disclosure, because disclosure might help a competitor. The sensitivity of confidential commercial data usually dictates the data-protection requirements stated in the data-sharing agreements. When contracts for providing the data are being signed, it is advisable for both parties to discuss, and agree on, the level of protection level that will be suitable. Some data might be less sensitive whereas some might not be sharable at all.

Several considerations affect the protection level. Confidential commercial data could be categorised as described in Table 10.

Table 10: Categorisation of confidential 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
Licensed	Confidential commercial data shared with all/certain project partners during the project. Available on a per-project with approval by the owner.	Data provider (usually the data owner who holds the IP rights)
Proprietary	Confidential commercial data that are never shared, as the commercial value of potential loss or misuse of the data is too high to allow data sharing.	Data provider (usually the data owner who holds the IP rights)

Licensed data could be made more open by, for example, aggregating some signals in order to produce a non-sensitive level of information—thus avoiding commercially harmful misuse of the original data.

Non-sensitive data

The definition of non-sensitive data is data that are completely anonymised and do not include any confidential commercial elements (unless an agreement with a data owner allows public usage). This means that no personal identifiable data are available in the dataset (i.e. video, images or GPS traces). If video or image material is included in the dataset, any identifiable traffic participant and any other objects that can be used to identify a person (e.g., number plates) must be anonymised (e.g., by blurring) to ensure confidentiality. If GPS traces (including position and time) are included in the data it is important to use proper methods to protect the participant from being identified. For more information on anonymisation, see Section 6.3.

If the data are classified as non-sensitive, there are few if any mandatory requirements for data protection, but it is still recommended that all possible requirements be investigated—including those regarding reliable data storage.

6.3 Anonymisation and feature extraction

The term *personal data* relates not only to data used in the actual analysis, but also to any other pieces of information connected to the dataset that could somehow identify a person (e.g., any references to a person in a file on your local computer or a printed document stored in a safe with contact information to a participant). The HIPAA Privacy Rule (The US Code of Federal Regulations including the HIPAA Privacy Rule at 45CFR Parts 160 and 164, 2007) lists 18 elements as direct identifiers, including the following data types commonly used when performing an FOT: names, zip codes, all elements of dates (except year), telephone numbers, electronic mail addresses, social security numbers, account numbers, vehicle identifiers and serial numbers (including license plate numbers, full face photographic images and any comparable images).

De-identified data are obfuscated data, making a person's identity less obvious and minimizing the risk of unintended disclosure (Nelson G., 2015, GDPR), whereas *anonymised data* are data that cannot be traced back to an individual by any means (Nelson G., 2015). A FOT/NDS data provider must strike a balance between identifiable and anonymised data, using different approaches to obfuscate the participant's identity by implementing a variety of features or algorithms. Possible methods include: record suppression, randomization, pseudo-identification, masking and sub-sampling (Nelson G., 2015).

The European Union Agency for Network and Information Security (ENISA) propose different strategies for "Privacy by design in the era of big data" (D' Acquisto et. al. 2015), described in Table 11:

Table 11: Privacy by-design strategies

Privacy by-design strategy	Description
Minimize	The amount of personal data should be restricted to the minimal amount possible (data minimization).
Hide	Personal data and their interrelations should be hidden from plain view.
Separate	Personal data should be processed in a distributed fashion, in separate compartments whenever possible.
Aggregate	Personal data should be processed at the highest level of aggregation and with the least possible detail in which it is useful.
Inform	Data subjects should be adequately informed whenever processed (transparency).
Control	Participants should be provided agency over the processing of their personal data.
Enforce	A privacy policy compatible with legal requirements should be in place and enforced.
Demonstrate	Data controllers must be able to demonstrate compliance with privacy policy into force and any applicable legal requirements.

Having a completely anonymised dataset could mean that the usefulness and value for analysis is so low that keeping the dataset available is meaningless. In any case, legal and ethical restrictions on how long one is allowed to keep a personal dataset might force its deletion anyway. These restrictions apply to several current datasets, increasing the need for methods to extract essential, anonymous data before the data is discarded.

For rich media (such as video or images), feature extraction is the key to preserving privacy. Feature extraction could be used to translate media data into measures, thus removing the identifiable elements. Efficient feature extraction would solve two major issues for FOT/NDS: current datasets could be shared and features could be extracted from data before they are purged.

The first decision to make is which features should be extracted from the data; if the extraction is being performed prior to data deletion, data owners, providers and researchers must collaborate on this difficult task. The next step is to select an extraction method and evaluate its performance. Some interesting cases have been published regarding the SHRP2 dataset (Smith et. al., 2015 and Seshadri et. al., 2015). Promising efforts are ongoing to evaluate and improve extraction methods, and interesting results were presented at the two consecutive Anonymisation workshops in Gothenburg (2015 and 2016), giving an overview of European and American efforts. The presentations can be found on the FOT-Net website (http://fot-net.eu/library/?filter=workshops). Finally, the project must decide if it has the extensive computational resources required to extract features from a large dataset.

The main benefit of feature extracting is the possibility of enhancing existing datasets with new attributes or measures, previously only available from costly video coding processes.

GPS traces are also considered personal data, albeit indirect, as they can potentially reveal where people live and work and even their children's schools. Similarly, no detailed travel diaries covering long periods of time can be made public if they contain addresses, even though a person making a single trip in the diary could actually be anyone living or working at those addresses. There are many approaches being explored to ensure personal integrity, e.g., k-anonymity and differential privacy (D' Acquisto et. al. 2015). The trade-off here, between anonymisation and maintaining usefulness of the data for research, is difficult.

6.4 Data access methods

This chapter presents data accessed in one of four different ways: 1) downloaded via a public website, 2) transferred on hard drives to the research organisation, 3) remotely accessed at the data provider, or 4) accessed exclusively at the premises of the data provider. Each method has its own implications; usually, the data category has the greatest impact on method selection.

Public download

This means that the dataset is downloadable from a public space (e.g., a web or an ftp server). This option is suited for non-sensitive data, as it is not possible for the data centre to control the use of the data. The dataset could be under a license that sets conditions or restricts the usage of data. The license could also state that any papers or public material must include a reference to the data provider. The organisation downloading the data will, by definition, be considered a DC—and also an AS, if it performs analyses on the data.

Conditioned download

This means that the dataset is transferred between two (or more) parties that agree on the conditions. The data is transferred from the data provider to the requesting organisation using portable disks or by an agreed-on Internet protocol. There are no restrictions on data categories but it is mandatory for the parties to consider all related agreements. The dataset could be under a license, agreed on between the parties, that sets conditions or restricts the usage of the data. The requesting organisation downloading (and therefore managing) the data will by definition be considered a DC and, as noted, if it performs analyses on the data it will also be considered an AC.

Remote access

In this case, the data will not leave the data centre; all analysis is performed within the data provider's IT-infrastructure. There are no restrictions that depend on data classification but it

is mandatory for the parties to consider all related agreements. The dataset could be under a license agreement between the parties that conditions or restricts usage of the data. The requesting organisation is considered an AS.

On-site access

When remote access is not possible due to network bandwidth limitations, or legal, contractual or data-protection requirements, on-site access might be the only option. It is then up to the data provider to allow external partner(s) access to the data on the premises. In this case the data provider will be acting both as a DC and an AS.

6.5 Data protection at data centres and analysis sites

Two sets of requirements are suggested below, one for data centres and one for analysis sites. This document recommends eight requirements be considered by a DC, called DC1–DC8, and ten be considered by an AS, called AS1–AS10. Moreover, documents related to both the DC and the AS are listed. Depending on the classification of the data involved, the needed level of protection will vary, regardless of the data size.

It is important to state that these requirements should be seen as a starting point for the FOT/NDS project organisation to further investigate the issue together with their IT department. The requirements and implementation plans need to be adopted according to the categories of the dataset as well as the existing IT-infrastructure of the organisation.

Note that additional requirements may need to be considered for sensitive personal data.

6.5.1 Data centres (DC)

It is imperative that any organisation hosting FOT/NDS data document its data management processes. Depending on the level of sensitivity of the data, different levels of precautions have to be taken. If the data include personal identifiable data or confidential data, stronger requirements need to be formulated. The data handling needs to be documented; there are frameworks that must be considered (if not already established) to ensure that the necessary processes are documented and traceable. For example, ISO 9001:2008 for Quality management systems, ISO/IEC 27001:2013 for Information security management or ITIL (IT Infrastructure Library) could be used. Additionally, similar (although not formally acknowledged) quality assurance procedures might also be suitable; the most important consideration is that the organisation reflects on data security and access—and implements routines that ensure data protection.

Most organisations have established general routines, but it is important to check for specific updates if the data include personal or confidential information. It is also important that third-party organisations (e.g., a cloud-based data-hosting company or a third-party organisation managing parts of the IT infrastructure) comply with the requirements.

DC1: The DC must document its data-protection implementation.

The DC data-protection implementation must be documented. It is recommended that the documentation be accepted by the data providers and owners.

Implementation guidelines:

A data centre shall not be allowed to manage personal or confidential data before the data-protection implementation is documented. The data centre must fulfil legal requirements and document how to meet requirements DC2-DC8. It is recommended that there be a

transparent process with the data providers and legal instances. It could be valuable to have the external partner(s) get an independent review of the implementation. The following steps should be part of the review process:

- Appoint an individual as DC data supervisor. The data supervisor is responsible for mapping, implementing and following the requirements for data protection.
- If personal data are included, the organisation (in a European context) handling the data also assumes the responsibilities of being a data controller.
- Data hosting should not be allowed before the DC and the data provider(s) have agreed on the level of data protection.
- Compile documentation meeting the requirements stated for a DC.
- The DC documentation should be reviewed by the data provider or a third party organisation.
- Data-protection implementation actions to address:
- Present DC.
- Define start and end date (if applicable) for data hosting.
- Provide name of the appointed DC data supervisor and description of organisational structure.
- Provide overview of personnel who will have access to data.
- Briefly analyse responsibilities of DC in the context of data protection and privacy issues.
- Describe in detail the compliance by the DC with numbered requirements. In the documentation, known deliberate deviations from requirements should be listed, analysed, and motivated separately. Why is compliance not needed, and how will issues be addressed instead? Any changes (additions, modifications, deletions) to the implementation must be documented.
- Provide status of the described implementation; is it planned or already implemented? Provide time plan with technical details where applicable.
- Provide disaster recovery plan, with risk assessments.
- Provide documentation of incident response plan for data security breaches, with risk assessments.
- Provide documentation of relevant internal routines/guidelines, as well as training for personnel.
- Describe relevant contracts/agreements.
- Analyse national legal status; what legal issues must be handled specifically for the data centre, and how will this be done?

DC2: Data stored and processed at a DC must be protected from unauthorized access.

Servers, computing environments (physical as well as virtual) and network connections must be protected, using measures sufficient to prohibit access to unauthorized parties.

Implementation guidelines:

This requirement covers many aspects of operating a data centre, but the requirements should focus on the most important processes for protecting data from unauthorized access.

Physical protection

The servers and other equipment must be kept in a secure environment to stop physical intrusion; within the organisation only entitled personnel should have access to the server rooms. Logging of individuals' access/activities regarding the servers might also be required. This rigid requirement should be applied when called for by the type of data and the associated data protection requirements.

Logical protection

The personnel having access to the data need to be identified. It is important to consider not only the analysts, but also IT-administrative staff members who might have full privileges on the servers.

The FESTA handbook recommends using group-based privileges instead of giving individual users access to very specific parts of the data. This might cause overhead in the initial phase but in the end it will help control data access.

It is recommended that future use of the data be considered when designing the data access patterns, even in the main data collection project, as this can avoid costly updates later on. As a starting point it might be suitable to consider the main data-collection project equal to any other post-project analysis.

The use of Internet firewalls is recommended when managing personal or confidential data, to restrict traffic to the services in the data environment. One way to implement this is to allow access only from a limited number of analysis workstations to the services that host the data. It is recommended that only specific workstations have the special privileges required to perform data uploading. The network traffic between the firewalls should be encrypted if transferring any personal or confidential data. Be aware that merely using the IP address is rather insecure (due to 'spoofing'), and more efficient measures should be investigated.

Disk cabinets and USB sticks can be used for transporting data but should not be used as either main data storage or backup. If used for transferring personal or confidential data the disks should be entirely encrypted.

DC3: Data stored and handled at a DC must be protected from accidental deletion or corruption.

Secure backup and disaster recovery solutions must be in place.

Implementation guidelines:

Corruption or accidental deletion of data can result from user error (unintended deletion), malware (e.g., ransomware) or alteration of physical media (hardware failure or a disaster such as flooding, fire or theft).

The consequences vary depending on the type of data: loss of users' own data (such as their own processes, algorithms, results and derived measures) will affect only that user and those depending on the outputs—whereas loss of original experimental data will affect all users, and therefore needs to be considered more critical. Further, it is frequently a legal

requirement that the data be kept a certain number of years in order to re-do analyses (in case of doubts about scientific results, etc.).

Several good practices can limit the impact of accidental data deletion or corruption. First, to prevent alteration of storage media, only a minimum necessary amount of trained IT professional should have access to the actual storage places (server rooms in particular). Professional security measures must be taken in order to prevent unauthorized access. It could be a part of a quality management system. Proper disaster prevention and mitigation measures, such as fire detectors and extinguishers and staff training, also have to be taken in order to avoid storage destruction.

Data must also be replicated, preferably in different locations, in order to survive a disaster scenario. Given the size and confidential nature of the typical datasets involved, this may prove challenging for technical or data-protection-related reasons. For instance, synchronizing data in an uncontrolled cloud service would pose a bandwidth issue as well as a lot of legal problems. One possible solution which, although non-ideal, is easily implemented, is to keep the original data collection media (such as hard drives) in antistatic sleeves in a fire-resistant safe, in a building other than the data centre.

Appropriate IT measures should also be taken to ensure resilience to hardware and/or software failures: hard drives in RAID arrays with sufficient redundant hard-drives and hot-swap units would handle multiple simultaneous failures of different hard drives before any data would be lost. Regular backups of machines and virtual machines must also be scheduled.

In order to avoid unintended data deletion by users (whether the data are in files or databases), the users' access rights must prevent them from deleting any part of the original data without authorisation.

If parts of the original data have to be deleted on some occasions (data privacy laws specify, for instance, that participants in data collection can request the deletion of some of their data), the deletion has to be done in a controlled way, following strict procedures which ensure that only the data to be deleted are actually deleted.

Users' own data have to be backed up on a regular basis using standard methods (such as a backup server, associated with a magnetic tape archive). The impact of any action from the users´ side must be made clear to them, both through the user interface of the tools they are using and their initial training. The training should include information regarding which data are backed up and which are not. It must also be explained that data can be permanently lost if they are created/modified and then deleted before a backup point, or deleted and not retrieved after the archiving period. Additionally, users should be trained in procedures to recover unintentionally deleted data and receive a general introduction to structural metadata (see Chapter 5).

DC4: Confidentiality agreements for any involved personnel must be in place.

The DC must require signed confidentiality agreements from all involved personnel before they start handling the FOT/NDS data. Agreements can either be explicit, for the specific project (for guest researchers, students, etc.), or implicit, through employment contracts.

Implementation guidelines:

Employees usually have a confidentiality statement in their employment contracts. It is important to understand what is applicable for the specific organisation before starting to manage personal or confidential data. If consultants, students and other temporary personnel are involved but not covered by an agreement with a confidentiality statement, then they must sign a separate NDA.

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.

Implementation guidelines:

A set of policies needs to be in place for storing and protecting the data at the DC after the end of a project. These policies should permit effective data re-use and be proportionate to foreseen risks and damages from leaking or losing collected data.

It is also important to make decisions about what to do with the data; they could vary depending on data type. Agreements could include requirements regarding when data need to be erased. The responsible organisation must adhere to national or other applicable legislation when deciding if the data are to be:

- Kept online

  The organisation must guarantee the same data-protection level after the main data collection project.

- Anonymised

  Personal data could be anonymised or essential features could be extracted from the data and shared, leaving the original data secure. Commercial confidential data could be aggregated to a sufficient level that the owner might be willing to allow their less restrictive use. Combined, these two possibilities could result in data considered non-sensitive, lowering the data-protection requirements and possibly allowing public usage.

- Archived

  If there is no possibility of keeping the data online, all or parts of the data could be archived. It is important to evaluate the overall requirements for infrastructure and data management during the archiving period and also to decide when and how the data shall be erased. If personal or sensitive personal data are to be stored as such at all, they must be encrypted. The same applies for confidential data; separate agreements usually set rules for deletion and archiving.

- Erased

  Policies should describe how to erase data securely. It is important to consider all media where data have been stored (e.g., storage and backup systems, portable storage such as USB drives, and even paper copies). The policies have to be in accordance with the requirements for recreation of results (see DC2).

DC6: Data sent between a DC and an AS must be encrypted.

Data may be transferred between a DC and an AS by electronic means or, alternatively, transported on physical media. The DC must ensure that the data cannot be accessed during the transfer.

Implementation guidelines:

Encryption should be applied to all data transfers containing personal or confidential information. Files exchanged on portable media (e.g., USB drives) or sent over the Internet should be encrypted using software. For data streams between a DC and an AS, encryption practices are needed to ensure that the transfer cannot be listened to by outsiders.

Re-using personal data requires the laws and standards of all involved countries to be investigated. Several countries have standards in place for encryption and protection of personal data, e.g., British Standard 10012:2009 and FIPS 140-2 (Federal Information Processing Standards in the US).

Use of encryption software is usually straightforward, as it is often integrated into normal processes by design. However, a specialist could be consulted for the selection of encryption software if the data-protection requirements are high.

Encryption methods and standards evolve. The chosen encryption product should be checked to ensure it has been certified to meet current standards.

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/all of a project's data. The project agreement should regulate the possibilities of doing this. Also, the participants must have given their consent to the data being disseminated outside the project partners.

Implementation guidelines:

Any data downloaded for data sharing must comply with the terms of the project agreement(s), data ownership, and the participants' signed consent form.

The receiving partner should operate under the same requirements as the DC. It is recommended that the same approval process be used for both of them. Depending on the data categories shared, some requirements might not be applicable—but it is important to justify and document any deviations.

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

The difference between data extraction and data download is that when extracting data, the recipient will not become a DC; the data extracted is the output of analysis (e.g., plots, statistics or images), not data for re-use. Depending on what the participants have agreed to in the consent forms, different extraction policies can be used. Video and GPS extraction must be treated with special care. The recommendation is to anonymise the personal data content in the videos, especially faces and vehicle number plates. Each extraction must be in accordance with the project agreement(s).

Implementation guidelines:

To facilitate data extraction, information describing the level of consent from all participants (e.g., a participant did or did not permit third parties to access recorded video) should be accessible within the analysis environment. This will make it possible to keep track of which data can be extracted and under which circumstances. It is recommended that all parties agree on and implement a process for data extraction within the project. All extractions must be documented. A person (the data supervisor in Figure 5) shall be responsible for managing data extraction requests and forwarding the decision to the person performing the extraction (*data extractionist*). This schema considers three questions:

- Do the requested data relate to participant consent and is the extraction valid?
- Do the requested data include data that need approval from the data owner?
- Do the requested data meet the size requirements for extractions defined in the project?

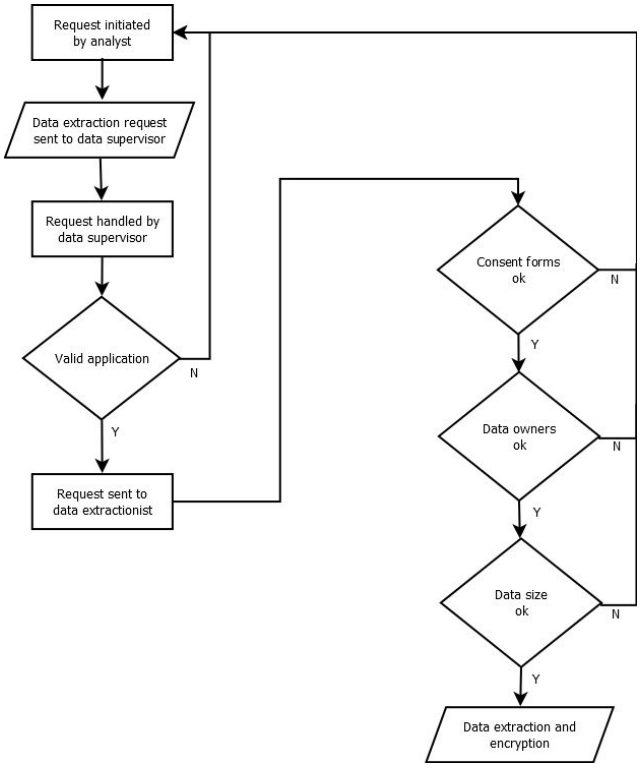

Figure 5: Data extraction process

Documents

The following specific documents within the context of the data centre are identified:

- an agreement and/or a data license with a data provider (if applicable);
- an agreement with an external IT infrastructure provider (if applicable);
- a confidentiality-disclosure agreement (CDA) or NDA for involved personnel;
- data-protection implementation documentation;
- data-extraction requests.

6.5.2 Analysis sites (AS)

An AS gets access to the data hosted by a DC by downloading or remotely accessing them. The AS must document the data-protection implementation plan, which should be agreed on with the DC. Depending on the sensitivity level of the data, different levels of precautions have to be taken. If the data include personal or confidential data, stronger requirements need to be fulfilled.

AS-1: The AS organisation must document its data-protection implementation.

In order for data access to be granted to the analysts from a research organisation, the data-protection implementation must be documented and it is recommended that it be agreed on by the DC.

Implementation guidelines:

An AS shall not be allowed to analyse any data before the data-protection implementation is documented and accepted. It could be valuable to have an external partner (or partners) provide an independent view of the implementation. The AS must fulfil the legal requirements and document how to meet the requirements AS2-AS10. The following process is recommended:

- Appoint an individual to be AS data supervisor. The data supervisor is responsible for mapping, implementing and following the requirements for data protection.
- Compile documentation that meets the requirements specified for an AS.
- The AS documentation should also be approved by the DC.

Data-protection implementation actions to address:

- Present AS and intended data usage.
- Define start and end dates for data usage.
- Provide name of appointed AS supervisor and description of organisational structure.
- Provide overview of personnel to be granted access to data.
- Briefly analyse responsibilities of the AS in the context of data protection and privacy issues.
- Describe in detail the compliance by the AS with numbered requirements. In the documentation, known deliberate deviations from requirements should be listed, analysed, and motivated separately. Why is compliance not needed, and how will issues be addressed instead? Any changes (additions, modifications, deletions) to the implementation must be documented.
- Provide status of the described implementation; is it planned or already implemented? Provide a detailed time plan with technical details where applicable.
- Provide documentation of incident response plan for data security breaches, with risk assessments.
- Provide documentation of relevant internal routines/guidelines, as well as training for personnel.
- Describe relevant contracts/agreements.

- Analyse national legal status; what legal issues must be handled specifically for the AS, and how will this be done?
- Determine whether the intended data usage requires approval from a national ethics committee.

AS-2: The analysis work stations must be physically and logically protected.

Analysis work stations (used either for remote virtual access to the DC or for handling downloaded data) must be protected in such a way that unauthorized access is prohibited.

Implementation guidelines:

For physical protection, work stations must be placed in locked rooms, or otherwise placed so that the screen content can only be seen by the data user. Some organisations also restrict users from bringing mobile phones or laptops into the analysis rooms, a stringent measure that requires security controls and usually limits the access to the computers. These guidelines have to be adapted to the level of confidentiality of the data in question.

As for logical protection, here are a few other ways to reduce the risk of intentional or unintentional data re-distribution:

- Restrict the computer's access to network services. The analysis computer should only be allowed to communicate with the necessary services. This could be configured in the Windows or Symantec firewall or, if using Linux, iptables. The ruling principle should be: deny all access, and then open up only the necessary services.
- Restrict the ports (USB, printer, SATA) of the computer. This could be done using Group Privilege Objects in Microsoft Windows. Or, even more drastically, by physically disabling the ports.
- Restrict the analyst from having administrator or root privileges in the operating system, since this capacity could eliminate the restrictions above.
- If unwarranted attempts to re-distribute data are detected, they should be reported automatically to the data supervisor.

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

Before data access can be granted, analysts must present proof that they received mandatory training, possibly prescribed by the initial project (e.g., US NIH education: http://phrp.nihtraining.com, accessed on December 27th 2016).

Implementation guidelines:

The analysts must have relevant knowledge in data protection and confidentiality issues if the dataset includes personal or confidential data. The web-based NIH training could be used, but the project could also choose to design and set up a training package. In either case, it is important to provide information regarding the local implementation of the security precautions, such as the data-protection procedures and the analysis environment capabilities and restrictions. The training should cover rules for the specific dataset at hand and provide analysts with general, basic information about the study. Further, the material should be designed according to the national regulations that concern personal data.

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

The research organisation must require signed confidentiality agreements from all personnel, before data access can be granted. Agreements can either be explicit, for the specific project (for guest researchers, students, etc.), or implicit, through employment contracts.

Implementation guidelines:

Employees usually have a confidentiality statement in their employment contracts. It is important to investigate what is applicable for the specific organisation before starting to manage personal or confidential data. If consultants, students and other temporary personnel are involved but not covered by an agreement that includes a confidentiality statement, a separate NDA must be signed.

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

It is recommended that the number of persons having accounts for the data be restricted. The fewer persons are involved, the easier it is to keep track of users, accounts and privileges.

Implementation guidelines:

The process is exemplified by the following schema:

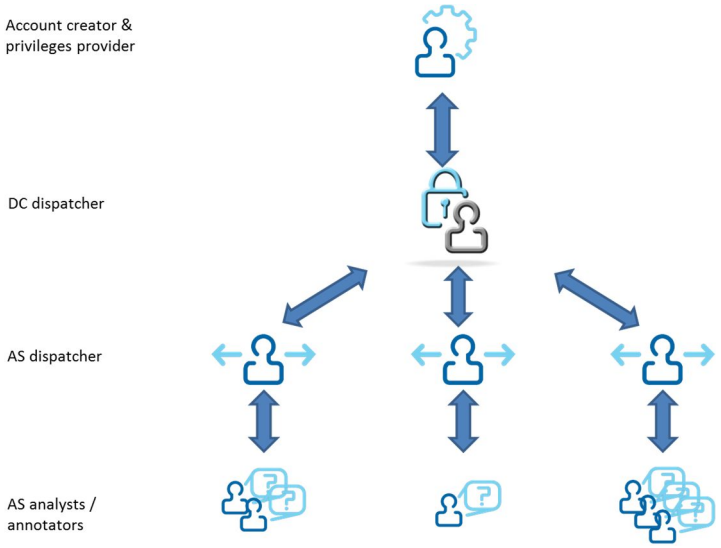

Figure 6: Data access process

The analyst asks for access from the AS dispatcher, the person in the AS organisation who is responsible for providing access to the DC. At the DC a person is appointed DC dispatcher. The DC dispatcher then forwards the information to the account creator, and the credentials or privileges are extended to the AS analyst. Depending on the size of the AS, the AS dispatcher and the AS analyst can be the same person. Similar simplifications could be suitable for the DC's involved personnel.

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

Extraction of a portion of the data, including video snippets and screen shots, must be performed according to the participant's consent form and the data-extraction procedures. All extraction is administered through the DC.

Implementation guidelines:

The DC and AS should agree on, implement and document the process for data extractions. The AS data user must send a data-extraction request, approved by the AS data supervisor, to the DC data supervisor for approval. The request could include:

- intended use of the extracted data;
- list of data types;
- text description of the data;
- total size of the data;
- list of files or folders to extract.

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

Regulations for data extraction procedures are in place; the analyst must not circumvent these procedures or disclose data beyond the AS in any way.

Implementation guidelines:

This information should be included in the training package, which should be given to the analysts before they get access to the data.

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 purposes other than those stated in the consent forms, except with ethical review board approval. Even if national law requires approval from an ethical review board (or similar) for the original project, other usage of personal data is normally not permitted, unless further approval is sought explicitly.

Implementation guidelines:

The need for ethical review board approval depends on the national legislation. The consent form signed by the participant defines what research can be done using the data. The participants give their consent to specified organisations to perform certain research during, and possibly after, the project. Special care should be given to the selection of research areas. The only ways to use the data for other purposes later are to ask the participant for another approval or get approval from an ethical review board using the regular national processes. Also, if ethical approval is mandatory and has already been given for a specific use of the data, additional ethical approval is still needed if new research areas are to be investigated. The specific procedures might differ between countries, but the overall principle must always be applied.

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

If confidential or personal data are to be presented to a visitor, the visitor is required to sign a confidentiality agreement. As a general rule, visitors do not have access to the data and are always accompanied by an authorized person. Generally, confidentiality agreements between organisations are preferred over personal agreements with visitors.

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

The drivers' consent forms and national ethics regulations together with project agreements set the conditions for post-project research. All research proposals, but especially those not previously covered by the project, might need to be submitted to a local ethics committee and/or national authority for approval. Additional consent might even be needed from the drivers. Project agreements, including agreements with sensor providers, might also restrict the use of the data.

Implementation guidelines:

Early in the proposal phase of the project that will be re-using the data, possible data providers should be contacted to figure out the boundary conditions for re-using the data. The conditions might affect research areas, data availability, data access restrictions and approvals and could therefore affect the proposal content. The data providers should be able to present an overview of the requirements for re-using the data, including the necessity for additional approvals. The responsibility remains, though, with the researchers performing the new research to investigate the need for approvals. Obtaining additional approvals can be time-consuming; it may involve former participants, an ethical review board and external data providers to the original project (e.g., providers of map data) and thus the additional time required should be considered in the time plan for the project.

Documents

The following documents are specifically required by the AS:

- CDA or NDA between analyst/visitors and AS organisation;
- agreement and/or data license with DC;
- approved training certificate / documentation for analyst;
- data-protection implementation documentation;
- approval from ethics committee for intended research (if applicable);
- data-extraction request.

6.6 References to accident databases

Accident data are a special type of data, related and connected to FOT/NDS data as a collection of special situations which usually form a very small (but highly interesting in the safety context) subset of FOT/NDS data, and are widely used globally. They are discussed here as a use case.

There are several projects world-wide that collect and protect accident data for scientific analysis. The context of these projects is differing and also very mixed. Partners range from governmental institutions to universities and companies. With this great variety of ASs, there is a need for effective data protection. Interestingly, the ASs and DCs even within one

accident data project are located in several countries and form different types of legal organisations.

Moreover, accident data projects are long-term, so the process of anonymisation is crucial for their survival. There is always a chance that persons involved in an accident may ask for data related to their case. Safety-related data, especially accident data (which imply legal aspects), need more care than non-safety related data (e.g., data for driver behaviour analysis) when collected in a scientific context and thus are a good test-bed for data protection. The level of anonymization is largely independent from the level of sharing the data, it can even be accident data collected and stored only by one OEM. It is a matter of the legal requirements that have to be applied at the DCs location.

When data are anonymised, the link between a dataset and a specific person, accident or geographic location is cut. Then, the data can be used for a scientific purpose but you cannot use it anymore in the context of legal affairs. Anonymisation is crucial, as those who are responsible for data protection would stall the project without it.

Technically, accident data is protected by the DC, who removes any details which can be directly connected to a single accident, or a person involved in that accident, before entering the data into the database. In particular, participants' identities, exact geographic locations and exact dates are removed. Usually, pictures are also included in the data, necessitating a more complex process of anonymisation: for instance, faces and company logos (e.g., printed on vehicles) have to be blurred to make them unreadable, which cannot be done fully automatically—manual intervention is required.

An interesting challenge arises when there is a need to link third party data to the already anonymised accident data supplied by the DC. It would be useful, for example, to know the equipped safety features of a car involved in an accident, to analyse their effectiveness. However, direct access to the equipment information for a single vehicle (other than standard equipment, which can be determined by make, model and year) requires the vehicle identification number (VIN) of the vehicle, which is not usually available to the DC, but is to some third party. On the other hand, any information to track a single vehicle, such as the VIN, is not available to the third party.

One solution is to provide a list of VINs, without any accident data information, to the third party. But as these VINs identify vehicles known to have been involved in accidents, this solution is not compliant with common data-protection requirements. In fact, to date this problem has only been solved in a closed environment (like an OEM), where the DC and the third party supplying the VIN are a single entity. However, hosting data in a closed environment also needs to honour the legal restrictions which are valid at the DCs location. This differs between legal systems and also the type of personal data stored, for example names and other details have to be removed from medical data and faces on pictures have to be blurred. In this example of information linkage, it has to be considered that the VIN only points to the owner of a vehicle, not directly to the persons involved in the accident. This example shows how important data protection is, and how seriously it is handled in current scientific accident databases. This situation is not necessarily restricted to accident data and should be considered in other domains, too.

In some legal and political constellations, an increased level of data protection has to be practiced. Such constellations can occur in mixed environments, when public and private institutions run a joint project. The DC has to meet certain additional requirements: for

example, it has to be a server at a university, and the anonymised data are transferred over secured lines to the AS.

There is some variance in data-protection requirements around the globe. For example, in the US, accident data collected by the government are made public and can be downloaded from websites. Access is regulated by the US Federal Research Public Access Act (FRPAA) and the Fair Access to Science and Technology Research Act (FASTR). The main reasoning behind public access is the social benefit from publicly funded research to all taxpayers which on the other hand is opposed to the protection of each individual's data in the case of accident data. There is no other country with similar regulations worldwide. When data is published, it is highly important to remove/hide personal information. It should be noted that the anonymisation level of US accident data is about the same as that of non-public databases in Europe, including blurred pictures and cut-off vehicle identification numbers.

In practice, data protection has proven for decades to be feasible when dealing with accident data in a scientific context.

7 Training on data protection related to personal data and IPR

All personnel handling data from FOT and NDS need to undergo training in data protection, if the data is personal or conditioned based on intellectual property rights (IPR). Personal data are any data that could reveal a person's identity—such as video, national identification number, address or GPS positioning, and any data that could be connected to identifying data. Persons or organisations collecting and managing personal information must protect it from misuse and respect certain rights of the data owners.

Protection of intellectual property rights is another important aspect when working with extensive datasets, including video, especially when research partnerships include industrial partners. The data could reveal algorithms of certain systems if re-engineered, and therefore need to be protected.

Training on personal privacy issues and IPR needs to accompany the general training on the data security measures put in place to protect the data. The level of training should be adjusted to the content of the specific dataset to be protected.

7.1 Set-up and content of the training

Who and when?

In order to ensure protection of personal data and IPR, training procedures must be in place and provided prior to any data access. Training material and procedures can be created by the organisation providing the training or possibly bought from the data provider's Support Services. Training must be given to analysts, video annotators, those responsible for the database, visiting researchers and all other staff handling, analysing or looking at personal or IPR data. Even persons to whom data are shown (during a demonstration, for example) must be informed about relevant data protection and IPR issues beforehand.

What?

The training needs to cover the following topics (the level of detail can be adapted to target audience's needs):

- description of the data with special focus on personal data and IPR:
  - What are personal data, in general and in this specific context?
  - What are intellectual property rights, in general and in this specific context?
  - What data are collected with (for example) video, questionnaires or GPS tracks?
  - Information about data ownership and access rights for partners/third parties.
- data-handling requirements originating from national and other applicable laws, regulations, and rules. Explain the purpose and implementation of each of the data-protection principles listed below; give practical examples and answer frequently asked questions. Personal data must be:
  - processed fairly and lawfully;
  - obtained for specified and lawful purposes;
  - adequate, relevant and not excessive;
  - accurate and up-to-date;

  - not kept any longer than necessary;
  - processed in accordance with the participants´ rights and acceptance;
  - securely kept;
  - not transferred to any other country without adequate protection in situ.
- explanation of the consent form content, especially the specific active consents related to data sharing (voluntariness, comprehension and disclosure):
  - How should the study participant be informed about data collection, purpose, handling, storage, and access—including re-use after the project ends?
  - What is included/excluded in the participant's consent? (For example, participants give their consent to collect videos for analysis purposes and to video of them being shown in conference presentations).
- explanation of data-handling procedures:
  - practical rules and procedures for data access (rooms and workspaces with limited access, personalized keys, password protection);
  - data structure;
  - how the data are anonymised, pseudo-identified and/or encrypted;
  - how the data are accessed, in order to (for example) perform analysis;
  - the contact persons for different procedures including the data protection responsible;
  - whom to inform in case of deviations.
- information about publication rights.

How?

It is recommended that a personal training session be organised in order to answer questions and make sure that all staff members know their responsibilities. Online courses might be helpful to provide additional valuable information, but they are not considered sufficient on their own as they cannot cover local implementation of the security precautions. For illustration purposes, videos on data protection (e.g., Data Protection Act training video (https://www.youtube.com/watch?v=wAe4358amJc&list=PLBEEA03BA780B128E&index=1, accessed on December 27th 2016) or case studies (e.g., https://dataprotection.ie/docs/CASE-STUDIES-2013/1441.htm#CS6, accessed on December 27th 2016) can be included in the training material.

The US NIH online training course (http://phrp.nihtraining.com/) can complement the training session by providing a basic understanding of the three principles essential to the ethical conduct of research with humans: respect, beneficence and justice.

7.2 How to document?

Documentation of all training is recommended, most conveniently recorded on the analyst's information sheet, which the participant needs to sign. Although analysts might have an NDA in their certificate of employment, the process of signing the document enhances the protective level of the data.

It is recommended that the following records be kept:

- persons who have undergone training;
- training procedures;
- process descriptions;
- contact persons for different procedures including the data protection responsible.

8 Support and research services

Support and research services are essential to data sharing. Depending on the knowledge and responsibilities of the persons re-using a dataset, either support services alone are provided, or research services are also required.

Support services comprise all activities in which support is being provided for successful data re-use. Support can be provided in various forms, starting with supplying information and ending with assistance with data analysis methods and procedures.

Research services comprise all activities where research work is carried out for the client, ranging from advice on specific research questions in different research stages to a more complete research endeavour providing a detailed analysis of specific research questions. The services are more targeted to the latter.

Analysis tools are an integral part of support and research services. The efforts and costs are to be included in the business model for the re-use of the data. These are discussed in Chapter 9.

8.1 Support services

Support starts as early as the application stage, with discussions on the suitability of the data to answer the specific research questions at hand. Support services target the researcher's ability to perform analysis and re-use existing data. The services are divided into different stages depending on the degree and impact of the support. These stages are:

- information and data provision;
- supporting tools;
- assistance with dedicated research needs;
- data-protection and analysis facilities.

Information and data provision

The first stage of support is to make researchers aware of available datasets and tools for data handling. This information is usually provided in online data catalogues. Furthermore, discussions may be necessary to answer questions about data usability (based on feedback from initial data analysis or from already performed data re-use) and which procedures have been established and proved to be successful. Metadata and other detailed background information on the data collection and initial study design can provide a better understanding of the dataset and improve data handling. Additional services, such as basic data aggregation and data extraction and transfer, could also be provided.

Supporting tools

Tools are an integral part of the support services. These tools consist of viewing and annotation tools, scripts to extract useful datasets from a database and licensed SW—and can also include entire frameworks for retrieving, processing and uploading data back into a database. However, it is important that the analysts are free to choose what tools to use without being constrained by factors other than the raw data formats and data descriptions (for example, by complex frameworks with graphical interfaces). It is, as mentioned in Chapter 5, important that raw data can be read in a clearly described format directly from the

data storage source (e.g., database or file storage), regardless of what analysis tools are used in the project. Note that appropriate access restrictions should always apply. Allowing analysts to choose their tools is important, since different analysts have different ways of analysing data. Support services should impose as few constraints as possible on what processes analysts can use to analyse the data (within the data-protection framework). Examples of different ways to analyse data are given in Chapter 5. Data description formats and data formats will have to be able to deal with different analysis processes, in order to be accepted and used by as large a community as possible. It is also important that the dependency on third-party software for access is kept to a minimum.

The FOT tools are available online on the FOT-Net website at the following link: http://wiki.fot-net.eu/index.php?title=Tools_for_FOTs (accessed on December 27th 2017) and the content is described in the deliverable D4.2 Tool Catalogue, to be published spring 2017. The tools are divided into three sections:

- Tools for Preparing: Operationalization of high-level FOT goals to specific study design and measures;
- Tools for Using: FOT operation and data acquisition;
- Tools for Analysing: Data handling and evaluation.

Support may consist of providing dedicated tools for specific tasks (if available) and setup and basic maintenance of the analysis tools. Due to the complexity of data analysis, the setup of these tools requires a profound understanding of the datasets. Further developments of the tools fall under the stage Assistance with dedicated research needs of the support services.

Assistance with dedicated research needs

Assistance, the most advanced stage of support services, can take the form of dedicated advice on analysis methods and the custom modification of tools. In a strict sense, analysis methods are not applied (this would be part of research services, see 8.2); instead, this service selects, provides and adjusts analysis methods.

Data protection and analysis facilities

The following support services can also be provided:

- analyst training;
- support relating to privacy issues;
- data-protection measures;
- secure facilities for analysis work.

The researcher could be given training in security and privacy matters, thus gaining a deeper understanding of the sensitivity of the data. Training in using analysis tools could also be included (see Chapter 7).

Support for new research projects on confidentiality and privacy issues is a common role for data warehouses.

Advice and support could be given on the need for data-protection measures.

Certain data warehouses offer secure sites/rooms for analysis. In these cases, the data may not be transferred, but must be analysed on-site to fulfil security requirements.

8.2 Research Services

Research services have a role beyond the initial start-up provided by the support services. In this case, the data provider takes part in the actual research to be performed, if required by 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 (sometimes called the data extractionist) can assist. A deep understanding of the research questions is necessary in order to aggregate the raw data in the best way without losing relevant information. The work performed by the research services can extend as far as performing the complete analysis, answering specific research questions.

The three levels of research services are:

- research advice on methodology;
- research involvement/research support;
- complete research performance.

The three levels are not necessarily distinct, but can overlap each other.

Research advice

On this level, advice is provided on data analysis. The advice, based on experience from data collection or previously performed analysis of the dataset, focuses on the best practice to answer the actual research questions and the related hypothesis. That is, the advice does not deal with how to solve a problem (using tools, data handling, data protection and/or data processing), but focuses on what methods should be used to get to the desired results.

Examples of research advice are:

- determine whether a dataset can be re-used;, review the scientific approach/method for re-using data;
- review whether a dataset is appropriate for the research questions, hypothesis and indicators.

Research involvement/research support

The second stage of research services is an active involvement in the research to be performed in terms of:

- support in the identification/selection of data for re-use;
- development of specific tools for:
    - data handling
    - data analysis and evaluation.
- performing parts of the analysis, such as:
    - formulating research questions based on research content,
    - formulating a hypothesis based on research questions,
    - deriving indicators from hypothesis,
    - applying data analysis based on indicators,
    - statistical analysis of data.

Complete research performance

Finally, the highest level of research services consists of the data provider, or a third party, performing all the research. In this case, complete work packages for research on the datasets are taken over by the research service provider. Work packages can consist of work in one or more of the following fields:

- selection and provision of data;
- selection and/or development of specific analysis tools;
- complete analysis;
- scientific reporting.

9 Financial models

Efficient management of FOT datasets is the key for successful re-use. If data sharing is not economically feasible for data owners and potential data re-users, re-use of data does not take place and the benefits of data sharing aren't achieved. Thus, in order for data sharing to gain popularity within ITS and FOT research, financial models are needed that cover data management costs.

Organisations supported by public funding are facing new requirements to plan long-term data preservation and management. Future work on financial models will have to take into account both the changing conditions of public funding to promote data sharing and the current trend opening ITS data for use in new services.

This chapter discusses options for organisations carrying out field trials to fund the sharing and upkeep of datasets after the project.

9.1 Data management costs

In terms of cost items, FOT data management has many things in common with open data efforts and large-scale user tests in various scientific disciplines. Documentation and user support have heightened roles, though, as FOT datasets are generally in non-standard form and have their origins in studies with specific goals. In addition, strict requirements to uphold user privacy and product IPR may require secure facilities and processes, raising the management costs of such datasets higher than those of fully open datasets.

Table 12 lists the items requiring funding in FOT data management. The items are related to data management after a project—or more generally, after the data collection has ended.

Clearly, storing a massive dataset and organising proper backups to avoid losing data incurs costs. Data may also have to be anonymised to enable wider sharing. When sharing a dataset, licences/agreements usually need to be completed, as well as financial arrangements. Further, to justify the benefits of data sharing to funding organisations, it is important to collect information on the use of the data. As a result of such requirements, the list of data management cost items can grow long. However, that does not necessarily mean that data sharing causes a huge burden on organisations. Effective processes, support and tools provided internally or externally by professionals, can reduce the stress on participants in single projects. Basic preparations to share data should become part of good scientific practise.

Considering the general costs of data management, it is unlikely that all test data can be stored for future science. A selection process is foreseen that would concentrate the efforts and funding on promising and valuable datasets. This selection could be carried out by those who fund the costs of data sharing and supported by the experts who collected the data for the original project. The selection could be based on the following criteria:

- potential for re-use, from both scientific and business perspectives;
- efforts needed to store the dataset;
- quality and amount of data.

Table 12 presents cost items and tasks involved in data sharing after data collection has ended. It is assumed that tasks enabling data sharing, such as concluding legal agreements, metadata documentation and data quality checking, have already been performed in the

original project that collected the data. Some of the cost items in Table 12 are optional, such as advertising datasets or participation in international harmonisation/standardisation efforts of data collection and data catalogues. However, such tasks are common in professional data management services and can also be foreseen in FOT data-sharing activities that have achieved an established status.

Table 12: Data-sharing costs

Cost item	Comments	Timing of cost
Data selection, enhancement of documentation (metadata), creation of entries in relevant data catalogues	Finalisation and structuring of data. As a pre-requisite for sharing, the datasets need to be comprehensively documented.	When project/data collection ends
Anonymizing data	The level of anonymisation and related efforts depend on how widely the data will be shared.	Before data is shared
Management & coordination personnel costs	Basic management of e-infrastructure, including user support, data catalogue operations and updates, data import to archives, backups, compilation of usage statistics, license management, agreements and finances	Continuous
IT operations	Database servers, storage, licenses and IT personnel costs	Continuous
Analysis or data handling facilities	Physically secure work space	Continuous
Analysis support services	Expert support at different levels	When data is shared and during analysis efforts
Promotion and advertisement	Informing potential data re-users and data-sharing funders Optional: Direct funding of further analysis projects, to ensure good use of valuable datasets Optional: Direct advertisement of datasets for potential research projects and those planning new projects, beyond common catalogues	When project ends/Continuous
Optional: Standardisation and collaboration regarding dataset formats	Taking part in national and international collaborations regarding dataset formats	Continuous

9.2 Financial models

This chapter suggests financial models for data sharing, starting mainly from the point of view of the organisation that has collected the dataset. As the main funding for transport-related research today comes from direct governmental grants, this is also likely to be the case for the re-use of FOT data. Future funding might be directed toward established data-sharing and e-infrastructure activities. In fact, the first two financial models in this chapter (A and B), are based on such activities.

Project-based funding is one of the current methods for financing data sharing. The models C–E consider the pros and cons of directly including data sharing and re-use in the project activities. In the models F–H, the costs fall mainly on the end user (e.g., through membership fees or licenses). Several funding sources might be required to keep data available and provide services for third parties. Therefore, the financial models can also be complementary.

A) Organisations' core activity

Digital preservation becomes a part of organisations' core activities. This model is motivated by conditions set by public grant agreements. A part of the grant for the original projects that collected the data will be directed toward central data preservation activity inside the organisation. This would cover data management and sharing for a certain period after the project is finished. The data availability for third parties should be based on reasonable conditions and costs.

A selection process may be required to decide which data will be stored, the way they will be stored and for how long. The operation of a repository can also be outsourced. However, when a dataset containing personal data is stored by a third party, it needs to be strongly encrypted to avoid misuse and liability problems (see Chapter 6).

Table 13: Model A (example: social sciences universities, possibly larger FOT/NDS datasets)

Pros	Cons
• Data would be considered IPR, valuable datasets would not be lost • Dedicated professionals would enhance the quality of the data provision procedures and analysis tools	• A burden for small organisations not prepared for such requirements • No existing selection process for funding

B) e-Infrastructures

Public funding is directed to data infrastructures, serving multiple organisations and disciplines. Centralised data management could offer professional data management services, general harmonisation and possibly greater cost-effectiveness when compared to distributed approaches. The roles of public infrastructure would cover certain tasks, but project-specific funding would still be needed when data-users or data owners request additional services.

Table 14: Model B (example: Supercomputing infrastructures and their services to universities)

Pros	Cons
• Professional data management services • Data and processing services are free (i.e., for academic re-use)	• The operators of the data infrastructure will have limited knowledge and means to provide detailed support for analysts, other than existing documentation • Dataset confidentiality sets limitations for storage by third party services • No selection process for funding exists • Valuable datasets from smaller projects might not be considered

C) Archiving included in project budget

Project budget allows for dataset finalisation and archiving in commercial services. In this model, the project budget allows for final cleaning, documentation and fees for archiving in selected data storages for a fixed period (e.g., 10 years). The project creates entries in relevant data catalogues.

Table 15: Model C (example: Research team storing its data—or making them open-source)

Pros	Cons
• The commercial service could get part of their funding from advertising, even enabling free storage	• Who answers questions regarding the dataset after a few years have passed? • Is the documentation of the required quality? • No existing selection process for funding

D) Project extension

The project is awarded a continuation to maintain its data. Model D is like model C, except the dataset is archived by the project partners. For notable projects, separate grants for operation (including data storage, promotion, calls for analysis proposals, etc.) would be awarded based on a review board decision, under specific conditions.

Table 16: Model D (example: Large research projects apply for extensions)

Pros	Cons
• Targeted promotion activities for datasets can also include funding for analysis activities and effective monitoring of results	• No selection process for funding exists • Valuable datasets from smaller projects might not be considered

E) New project funding

New projects finance maintenance or revival of a dataset. In a chain of projects, the benefit of using past data is obvious, encouraging efforts to be put into maintaining and exploiting the old dataset. Depending on the follow-up activity, the data owner might also be motivated to share data with third parties, to extend analyses for mutual or customer benefit (e.g., offering material for thesis work, benefiting the customer who originally funded the data collection).

If a data request from outside of the organisation meets the business interests of the data owner, it is welcome. Otherwise, it fails to motivate the efforts needed for data sharing.

Table 17: Model E (example: Various research projects analysing and benefiting from previous dataset)

Pros	Cons
• No changes to current funding methods (additions are needed in call texts to promote existing datasets) • When data is re-used by those who collected the data in a previous project, the re-use is very efficient	• Plain project-based funding may not be sufficient to keep datasets available and it should be seen instead as a complementary funding source • Project owners have difficulties estimating the costs required to access a dataset, when they are making an initial project plan/offer

F) Established network

A network of organisations with participation fees arranges data management jointly. Organisations within the same discipline form networks that share and promote data. Datasets are collected, documented and catalogued using agreed-on/standardised methods. The networks are likely to be formed for handling continuous operational data which meet their business interests. There could be various levels of memberships and fees.

Table 18: Model F (example: Accident data collection and sharing)

Pros	Cons
• Business aspects can be applied on fees, high-quality harmonised data • Could include freemium services, where the basic information is available for free but advanced services have a cost • Facilitates cooperation in research	• Only certain disciplines seem to reach this status

G) Analysis services

An organisation with several valuable datasets uses them to create business, offering both data and related services. This model can enable the original group that carried out a study to get further funding for their work. The model is for organisations with a prominent role in a discipline.

Table 19: Model G (example: Notable data owners/Data providers)

Pros	Cons
• Continuous research quite possibly results in high-quality results.	• Small organisations and partnership projects have difficulties setting up this sort of business and their data easily get lost. • Even valuable datasets become old and lose value for organisations purchasing analysis services.

H) Data integrators

Companies acquire and market FOT datasets along with transport and other related datasets. In this model, a data integrator markets particularly useful FOT datasets (among others, such as those containing real-time traffic data) licensed from original sources. Customers are offered a single access point for data so they don't have to go through negotiations with several parties, facilitating (for example) the development of mobile applications. In order for a dataset to be shared without fees for the re-users, the maintenance would have to be financed through the organisation that contributed the dataset—or supporting business operations.

Table 20: Model H (example: Road operators putting together information services)

Pros	Cons
• Easy licensing of various high-quality information resources	• Data integrators may have little interest in non-commercial academic work

9.3 Distribution of costs

Depending on the financial model and the activities set up for data sharing, the costs are shared differently among the project that collects the data, the organisation(s) owning the data and, finally, the re-users.

Table 21 considers the funders for data management and re-use in the different financial models presented previously. The costs are divided into three classes:

1) dataset finalisation, costs that often come at the end of a project;
2) continuous costs coming from management and upkeep of data;
3) costs when data are shared, e.g., selection of data and user support.

Additionally, the table considers the cost for the re-user in each financial model.

Those costs that are potentially funded by an external party or the organisation's non-project funding (i.e., part of the organisation that is not involved in sharing or re-using the data) are highlighted.

Table 21: Funding source and re-use costs in different financial models

Financial model	Funder			Costs for re-user
	Dataset finalisation	Continuous costs	Costs when data are shared	
A. Organisation's core activity	Project/ Organisation's selection process	Organisation (with public funding)	Organisation/ Re-user	Non-profit price
B. e-Infrastructures	Project	Publicly funded e-infra, where organisation may have a role	Publicly funded e-infra, additional services have a price	Free (basic services)
C. Archiving included in project budget	Project	Project or e.g., data storage service supported by advertising	Both project and re-user	Free or non-profit price
D. Project extension	Project	Project	Project	Free or non-profit price, even calls for analysis proposals
E. New project funding	Re-user	Re-user	Re-user	Commercial price
F. Established network	Project	Re-users via participation fees	Re-user	Different levels of memberships and fees
G. Analysis services	Project	Organisation	Re-user	Commercial price
H. Data integrators	Project	Integrator	Re-user	Commercial price

10 Application Procedure

The partners should agree on an application procedure for re-use of data early on in the project, so that all project partners (and any possible third parties) know the conditions for additional research using the specific dataset. This will provide the necessary information so that new research applications to utilize the data can already take the data application time and potential costs for re-using the data into consideration during the proposal phase, before the application is sent to the targeted call. The feasibility of disseminating this information in the proposal phase should be investigated. It should be noted that these procedures are often much more time- and resource-consuming than expected.

10.1 Contents of the application procedure

The application procedure shall address the following items (at least):

- where to apply:
    - information regarding where and how to send in the application;
    - contact person for questions regarding the application.
- information needed in order to be able to evaluate the application (see suggested content below);
- person/organisation approving an application;
- response times, and conditions to be taken into account in the approval decision;
- requirements for mandatory training in data protection and privacy issues;
- requirement for signing an NDA;
- information on the data-access procedure;
- requirements for data protection, including possible certification of data-protection implementation;
- conditions for access and use of the data;
- potential costs for data storage, access, support and research services;
- requirements for acknowledgements on publications, reports and presentations;
- documentation of data applications and the related approval decision(s).

10.2 Contents of the application form

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

- applicant details:
    - organisation(s) applying;
    - contact person(s) for each organisation;
    - project partners applying (when applicable)—list of partners that want data access for project analysis.
- short project description;

- requested dataset:
    - which dataset (if many available);
    - specific data requested (which time-series data, video, GPS, questionnaires, etc.)
- use and expected results:
    - What research questions are the data expected to answer?
    - How is the data to be accessed?
    - What are the expected results?
- information on the intended publication of the data:
    - How will the results be disseminated?
    - What data, graphs, etc. is intended to be disseminated?
- list of persons to get access, and the related access time period:
- need for training in data protection and privacy issues:
    - Have the researchers had previous training? If so, what kind?
    - Is training related to data protection required, or only in data analysis setup at the data provider?
- need for support and research services:
    - Level of knowledge of the concerned analysis tools? Using self-supplied tools or needing training on provided tools?
    - Other support needs for the analysis, such as extracting datasets, etc.?
    - Should the research facilities do part or all of the analyses/research?

11 Conclusions

This report details the elements of a data sharing framework that would be required in order to facilitate re-use of the many FOT/NDS datasets hosted at different locations globally. Such a framework would also facilitate data sharing within new projects, as the content of the framework is general and could be used whenever data sharing is performed.

FOT-Net's Data Sharing Framework consists of the following seven items: pre-requisites that must be part of the project documents (such as the consortium agreement and the consent form), descriptions of data and metadata, data protection, training on data protection, 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 form an efficient data sharing framework.

The report constitutes the essence of the discussions held during the FOT-Net 2 and FOT-Net Data time frames; there are many hands-on recommendations in the text. Through the discussions, it has become obvious that the recommendations apply to a wide variety of cases, including different national contexts. At the end, though, it is always up to the partners of the specific project, national or international, to select the appropriate data-sharing strategy and decide what parts of the data sharing framework are applicable to their project.

The Data Sharing Framework needs to be continuously discussed and applied by different stakeholders with good knowledge of their national requirements, who collect experiences and make the framework applicable to as many FOT/NDS countries as possible. The framework also needs to be updated as new technology and methods provide new possibilities, especially regarding anonymisation and feature extraction. Reliable tools for automated feature extraction are key to be able to provide large quantities of essential features from video. Still, if the suggestions and requirements presented here are taken into consideration, future FOT/NDS projects will be much better prepared for data sharing during and after the project than previous projects.

List of abbreviations

Abbreviation	Full text
ADAS	Advanced Driver Assistance Systems
AS	Analysis site
CA	Consortium Agreement
CAN	Controller Area Network
CAS	Content-Addressable Storage
CDA	Confidentiality-Disclosure Agreement
DC	Data Centre
DG CONNECT	Directorate General for Communications Networks, Content & Technology
DOW	Description Of Work
FASTR	Fair Access to Science and Technology Research Act
FESTA	European handbook on FOT methodology
FOT	Field Operational Test
FP7	EU 7th Framework Programme
FRPAA	US Federal Research Public Access Act
GIS	Geographical Information Systems
GPS	Global Positioning System
GNSS	Global Navigation Satellite System
HDOP	Horizontal Dilution of Precision
HMI	Human Machine Interface
IPR	Intellectual Property Rights
ITIL	IT Infrastructure Library
NDS	Naturalistic Driving Study
NDA	Non-Disclosure Agreement

OEM	*Original Equipment Manufacturer*
PB	Petabyte
PI	Performance Indicators
RAID	Redundant Array of Inexpensive Disks
RPM	*Rounds Per Minute*
TB	*Terabyte*
USB	*Universal Serial Bus*
VPN	Virtual Private Network
VIN	Vehicle Identification Number

List of tables

List of figures

List of references

D' Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y-A., & Bourka, A., 2015. Privacy by design in big data - An overview of privacy enhancing technologies in the era of big data analytics 1.0. Available at: https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport

European Directive 95/46/EC Art. 2. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=en, accessed on December 27 2016.

GDPR, REGULATION (EU) 2016/679. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN, accessed on December 27 2016

Gellerman, H., Bärgman, J., & Svanberg, E., 2014: Report from the FOT-Net Data Sharing working group, Available at: http://fot-net.eu/wp-content/uploads/sites/7/2014/06/WG-Data-sharing-report-140528.pdf

FESTA, 2014. FESTA handbook v5. Available at: http://fot-net.eu/Documents/festa-handbook-version-5-2014/

Nelson, G. s., 2015. Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. In: Proceedings of the SAS Global Forum 2015, Austin. Available at: http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf

Puglia, S., Reed, J., & Rhodes E., 2004. Technical Guidelines for Digitizing Archival Materials for Electronic Access. Report for U.S. National Archives and Records Administration (NARA). Available at: http://www.archives.gov/research_room/arc/arc_info/techguide_raster_june2004.pdf.

Roebuck K., 2012. Metadata Repositories: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Publishing.

Seshadri, K., Juefei-Xu, F., Pal D. k., Savvides M., & Thor, C. p., 2015. Driver Cell Phone Usage Detection on Strategic Highway Research Program (SHRP2) Face View Videos. Available at: http://www.cv-foundation.org/openaccess/content_cvpr_workshops_2015/W11/papers/Seshadri_Driver_Cell_Phone_2015_CVPR_paper.pdf

Smith, Wang, Hu, Dyer, Chitturi, & Lee, 2015. Automatic Driver Face State Estimation in Challenging Naturalistic Driving Videos. Available at: ftp://ftp.cs.wisc.edu/computer-vision/repository/PDF/smith.2016.trb.pdf

The US Code of Federal Regulations including the HIPAA Privacy Rule at 45CFR Parts 160 and 164, 2007.