Data specification and quality: TeleFOT deliverable D2.3.1

PUBLISHER

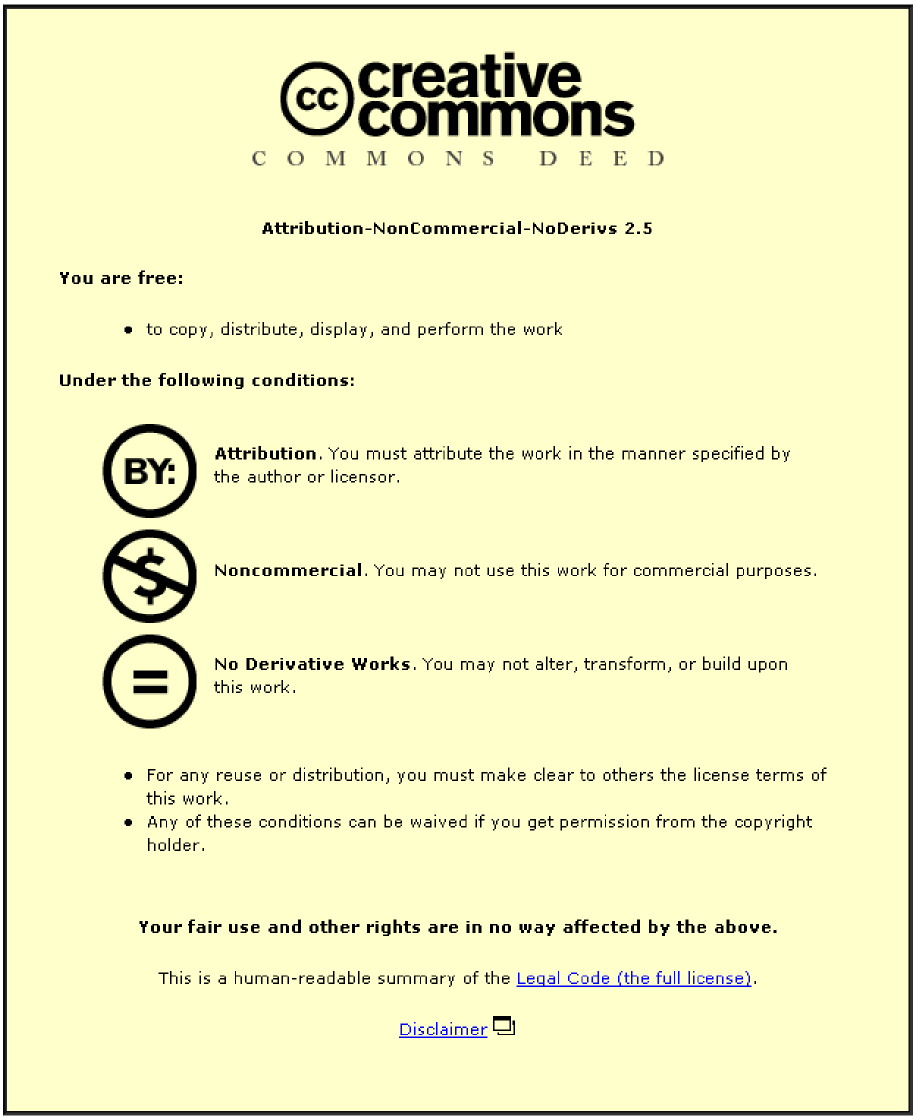
© TeleFOT

VERSION

VoR (Version of Record)

REPOSITORY RECORD

Welsh, Ruth, Andrew Morris, Jussi Vasama, Mark Fowkes, and Oskari Heikkenen. 2019. “Data Specification and Quality: Telefot Deliverable D2.3.1”. figshare. https://hdl.handle.net/2134/9333.



TeleFOT

Field Operational Tests of Aftermarket and Nomadic Devices in Vehicles

Large Scale Collaborative Project

7th Framework Programme

INFSO-ICT 224067

D2.3.1 Data specification and quality

Deliverable n.	2.3.1	Data specification and quality	
Sub Project	SP 2	FOT Framework	
Work package	WP 2.3	Data Specification	
Task n.	T 2.3.1, 2.3.2, 2.3.3	Data acquisition, Quality of data, Database structure	
Author(s)	Ruth Welsh Andrew Morris Jussi Vasama Mark Fowkes Oskari Heikkinen	File name	TeleFOT_D2.3.1_Data specification and quality_v3_1.doc
Status	Draft		
Distribution	Restricted Partners (RP)		
Issue date	2010-01-18	Creation date	2010-01-21
Project start and duration	1st of June, 2008 – 48 months		
European Commission Information Society and Media	Project co-funded by the European Commission DG-Information Society and Media in the 7th Framework Programme		SEVENTH FRAMEWORK PROGRAMME

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	DESCRIPTION
ASCII	American Standard Code for Information Interchange
CAA	Cockpit Activity Assessment Module
CAN	Controller-Area Network
COTS	Commercial Off The Shelf
DAS	Data Acquisition System
DB/db	Data Base

D-FOT	Detailed Field Operational Test
DoW	Description of Work
DVD	Digital Versatile Disc
DWG	Data Working Group
ECA	Environmental Conditions Assessment
FESTA	Field Operational Test Support Action
FOT	Field Operational Test
FTP	Field Transfer Protocol
GPS	Global Positioning System
L-FOT	Large Scale Field Operational Test
SQL	Structured Query Language
TMC	Traffic Message Channel
WP	Work Package
XML	eXtensible Markup Languge

REVISION CHART AND HISTORY LOG

REV	DATE	REASON
1	29/07/2009	Response to reviewers comments
2	13/01/2010	Addition of revised Chapter 1 Data Specification
3	18/01/2010	Editorial changes and summary file content

EXECUTIVE SUMMARY

This deliverable reports on the activities undertaken in WP 2.3 Data Specification. The WP originally comprised of three tasks; Task 2.3.1 Data acquisition, Task 2.3.2 Quality of data and Task 2.3.3 Database structure. The tasks were designed to meet the objectives of the WP namely, to prepare and check the whole data collection, transfer and management in order to ensure that the process can be made as automatic as possible.

The main purpose of this deliverable is to provide guidance firstly to those involved in the test site set up and secondly those designing the database. For the test sites, the specific data to be logged and the format that this should take is identified together with any responsibility for processing of the data. The deliverable also provides the information required for the test sites to develop a data quality procedure covering all aspects of the FOT. For the database developers, this deliverable aims to provide a high level description of the database requirements.

In addition, a Data Working Group has been established within TeleFOT due to a need to oversee and co-ordinate various data activities across SP2, SP3 and SP4. Since the Data Working Group (DWG) was not a part of the original description of work, no provision had been made for reporting on the activities undertaken by the group. It was therefore decided, subject to approval in changes to the DoW, that the DWG would be reported upon within D2.3.1 and a new task established within WP2.3. The activities thus far are included in Annex 2 of this report.

The report is structured in a manner that follows the ordering of the tasks in the DoW. A chapter is allocated to each of the three tasks.

Work related to Chapter 1, Data Specification, is still in progress. This task was delegated for consideration within the DWG discussions and activities and has continued throughout 2009. Chapter 2 therefore summarises the results of these discussions to that point. Issues considered include the minimum data requirements for both the L-FOTs and the D-FOTs, the flow of the data from acquisition to storage in the central TeleFOT database, pre and post processing requirements and definitions for derived variables. Specific issues related to the data requirements for the Large Scale Field Operational Tests (L-FOTs) and the supporting Detailed Field Operational Tests (D-FOTs) are identified.

Chapter 2, Quality of data, lays out the issues that need to be considered in order to assure, as far as possible, the quality of the data from acquisition through transfer and in to storage and analysis. These cover the following stages of the FOT development;

- FOT experimental design
- Data collection
- Data transfer
- Data storage
- Database quality control
- Data analysis
- Sampling requirements to ensure the quality of subsequent data analysis
- Subjective data collection
- Contingencies in the event of problems with the data

Each of these topics is considered in sections 2.1 to 2.9 and a summary checklist is provided in Annex 1. The advice given in this chapter should be seen as guidance and each test site should formulate their own quality control procedure that includes consideration of each issue. This is necessary since the guidelines provided are generic to all FOTs whereas each test site within TeleFOT will have their specific requirements.

Chapter 3, Database structure, specifies the requirements for Data Specification and provides a high level technical description of the database structure and layout as well as the input data structure. The following are considered;

- Performance requirements
- Availability requirements
- Security requirements
- Common framework for centralised long term storage of events
- High level data storage structure and layout
- Common schema for each data integration
- Input data structures

These points are considered in turn in sections 3.1 to 3.7.

Two scenarios for data collection are identified; in the first the data are sent from the data acquisition system (DAS) to the data management centre, in the second the data are sent to a local data centre and then later on fetched by the data management centre.

INTRODUCTION

TeleFOT is a Large Scale Collaborative Project under the Seventh Framework Programme, co-funded by the European Commission DG Information Society and Media within the strategic objective "ICT for Cooperative Systems".

Officially started on June 1st 2008, TeleFOT aims to test the impacts of driver support functions on the driving task with large fleets of test drivers in real-life driving conditions.

In particular, TeleFOT assesses via Field operational Tests the impacts of functions provided by aftermarket and nomadic devices, including future interactive traffic services that will become part of driving environment systems within the next five years.

Field Operational Tests developed in TeleFOT aim at a comprehensive assessment of the efficiency, quality, robustness and user friendliness of in-vehicle systems, such as ICT, for smarter, safer and cleaner driving.

This deliverable reports on the activities undertaken in WP 2.3 Data Specification. The WP originally comprised of three tasks; Task 2.3.1 Data acquisition, Task 2.3.2 Quality of data and Task 2.3.3 Database structure. The tasks were designed to meet the objectives of the WP namely, to prepare and check the whole data collection, transfer and management in order to ensure that the process can be made as automatic as possible. This is all started with a theoretical approach to see how the parameters selected represent the phenomena that are to be measured and what is required from the data and the devices to make a functioning system.

Following the FOT chain, this deliverable also builds upon the work carried out in WP2.2 where the research questions and hypotheses have been identified together with the performance indicators required to test the hypotheses. Data specification makes the next step and in turn guides the test site logging requirements and provides input in to the questionnaire specification.

In further detail this report consists of 3 chapters:

Chapter 1, **Data Specification** defines in detail the type and nature of data to be collected in the L-FOTs and the D-FOTs. It also considers the flow of the data from acquisition to storage including pre and post processing requirements. A core data specification for data capture that is common to both L-FOTs and D-FOTs is proposed and input data format for the TELEFOT central database identified. Comments are made on the integration of logged objective data with subjective reported information and other database augmentation as part of this data process. This is currently work in progress as piloting of physical data capture and transfer is performed, and development of necessary procedures and analysis scripting is developed.

Chapter 2, **Quality of data** gives guidelines in order to best assure the quality of the data that is uploaded to the database and the subsequent storage, retrieval and analysis of the data. This chapter will provide a reference protocol to be followed by the test sites, the database managers and the analysts at each stage of the FOT from the set up, through the implementation and into the analysis phase. The chapter also deals with sample sizes for the data to be collected. Guidance is given that relates the amount of data to the analysis requirements, recommending that which would be considered sufficient to answer the research hypothesis within an acceptable confidence level.

Chapter 3, **Database Structure** deals with the actual high level structure that the database will take. It

- Defines performance, availability and security requirements for the database environment used in this type of partly distributed (test sites and vehicles) and centralised (Data and User Management) system
- Defines a common framework for centralised long term storage of events
- Defines common schema (e.g. xml) for each data integration (including definition of the unified data format for vehicle data and for data collected manually)
- Defines high level data storage structure and layout for TeleFOT
- Defines (get/collect information about) input data structures from vehicles and test sites

In addition, and still depending upon the outcome of a proposed change to the DOW and suitable resources being made available, a further task has been included in WP2.3, that of the **Data Working Group**. At the start of the TeleFOT project partners perceived a need to increase the emphasis on the co-ordination of the work within the project to ensure that the different timescales for individual FOT could be accommodated effectively into the planning phases of the project. It was acknowledged that while individual tasks and roles had been described to cover data issues within the original project structure, there was not a specific defined overall data handling co-ordination task. It was therefore agreed at the first TeleFOT plenary meeting (Helsinki June 2008) that a co-ordinating body should be established having appropriate representation from SP2, SP3 and SP4 members and led by the project co-ordinators VTT. The resulting TeleFOT Data Working

Group (DWG) was first convened on the 14th October 2008 at a project meeting held in Brussels.

The main tasks of the data working group were agreed to be:

- to agree on the architecture of the data collection system
- to agree on the specifications for the transfer of data to the collection system
- to agree on the specifications for accessing the data

A report on the work carried out by the DWG thus far is included as Annex 2 to this report. The Annex discusses the Data Working Group process and its impact on the project. Since the working group will continue to function through the remainder of the project, this Annex will be updated in due course unless it is decided that the remaining activities are to be reported in an alternative deliverable.

1. DATA SPECIFICATION

1.1. Overall Requirements for TELEFOT Data

The main objective of the TELEFOT project is to assess the impact of nomadic device use on users. This will be assessed within the project by the completion of both Large Scale Field Operational Tests (LFOTs) and Detailed Field Operational Tests (DFOTs) across many member states. In these FOTs nomadic devices will be deployed to a large group of real-world users and data collected on their behaviour to assess the impacts of the use of these devices.

The particular focus of assessment within TELEFOT is under several specific identified impact areas. These are described as :

- Safety (SP4/WP4.3)
- Mobility (SP4/WP4.4)
- Efficiency (SP4/WP4.5)
- Environmental (SP4/WP4.6)
- Business and User Uptake (SP4/WP4.6)

Each of these impact areas have lead organisations who have established the specific research questions and hypotheses that are required to be answered in the subsequent analysis of the collected data from the use and operation of nomadic devices within the FOTs.

It is therefore necessary to ensure that the actual data collected is sufficient in quantity and quality to enable the appropriate analysis process.

This represents a top-down approach to defining data specifications for the project. In addition there are numerous specific nomadic device functionalities to be evaluated within these TELEFOT FOTs. These include

- Traffic Information
- Speed Limit Information
- Speed Alert
- Navigation Support (both static and dynamic)

- Green Driving Support
- Nomadic device based eCall

Clearly it is important that the data to be collected across the FOTs is sufficient to answer the various impact area research questions and hypotheses and address the particular functionalities of the nomadic device based applications. As the implementation of various FOTs differ in their data collection capabilities then these potential constraints on data collection possibilities form a **bottom up** approach to defining data specifications to for the project.

Finally it should be noted that the LFOTs and DFOTs planned have a range of different goals and possibilities in the ability to capture data. LFOTs are characterised as large numbers of vehicles/users real world experiments with minimal additional data capture possibilities beyond that available from the nomadic device functionality itself. As a result the data collected maybe restricted by practical device constraints and experimental methodology. DFOTs may have a smaller number of vehicles/users but with much greater data collection possibilities (e.g. additional sensors and data loggers) and can therefore provide more detailed data sets for analysis.

The Data Working Group described above was formed within the project to assess the needs of the project (**top down**) and the practical constraints that exist (**bottom up**) and define a core fixed data specification that constitutes data that should be collected as a minimum for ALL FOTs within the project. It is acknowledged that this will only form a subset of all data to be collected in the project, however definition of such a minimum core data set is required to ensure comparability between test sites and nomadic application settings.

Figure 1.1 below illustrates the data capture process and the relative influence of the logged Core Data Specification.

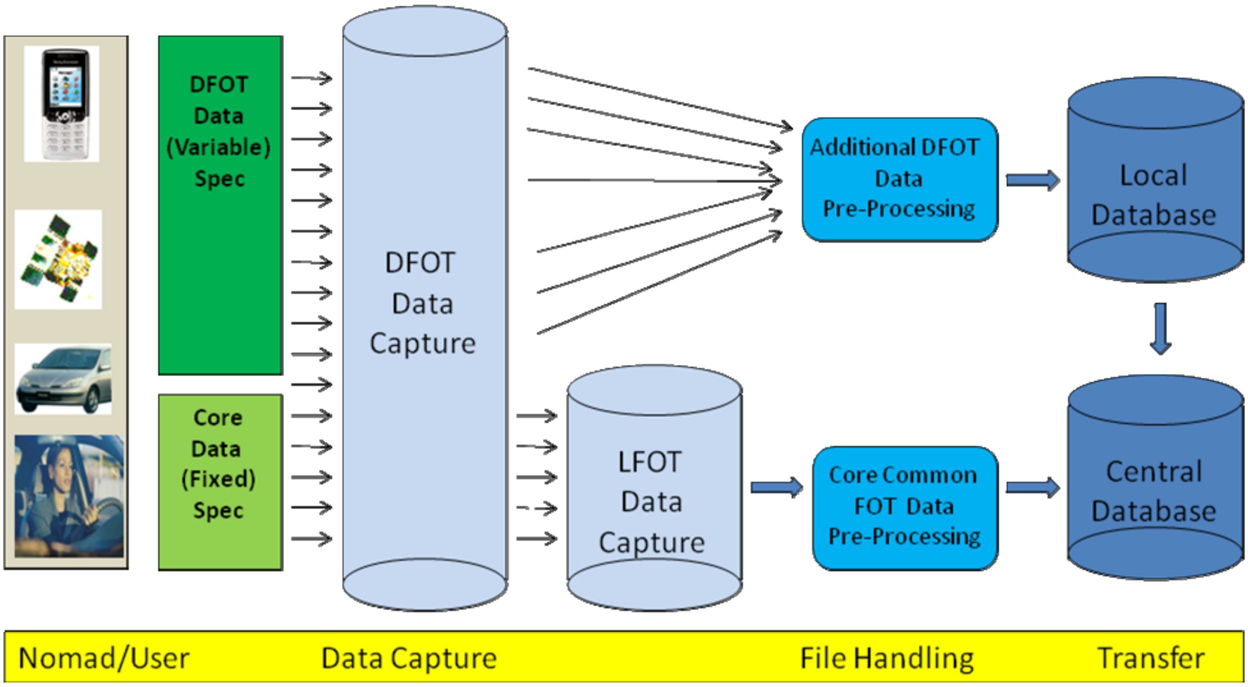


Figure 1.1 The Generic Logged Data Capture process TELEFOT DFOT & LFOT

This figure also illustrates the process for both LFOT and DFOT of capturing objective data from the nomad device, and associated datalogging systems, on the activities of the system and users. This is followed by transfer of data from data logging functionality and any required pre-processing before transfer to a local and central database.

It is important to emphasise that the focus of the Core Data Specification is on objective logged data that is common to all FOTs. It is acknowledged that several relevant research interests into the impacts of nomadic devices cannot be examined by logged data alone. It has always been accepted that the overall approach to be employed by TELEFOT would be to incorporate logged data in a complex data set with subjective user reported information (e.g. travel diaries and questionnaires) and other context augmentation information that is gained from post trial processing and other data sources (e.g. traffic flow data, local weather databases, digital map databases).

These additional information sources are supplementary to the Core Data Specification of concern here and will be described in later TELEFOT deliverables. Figure 1.2 below illustrates the relative processes for logged data

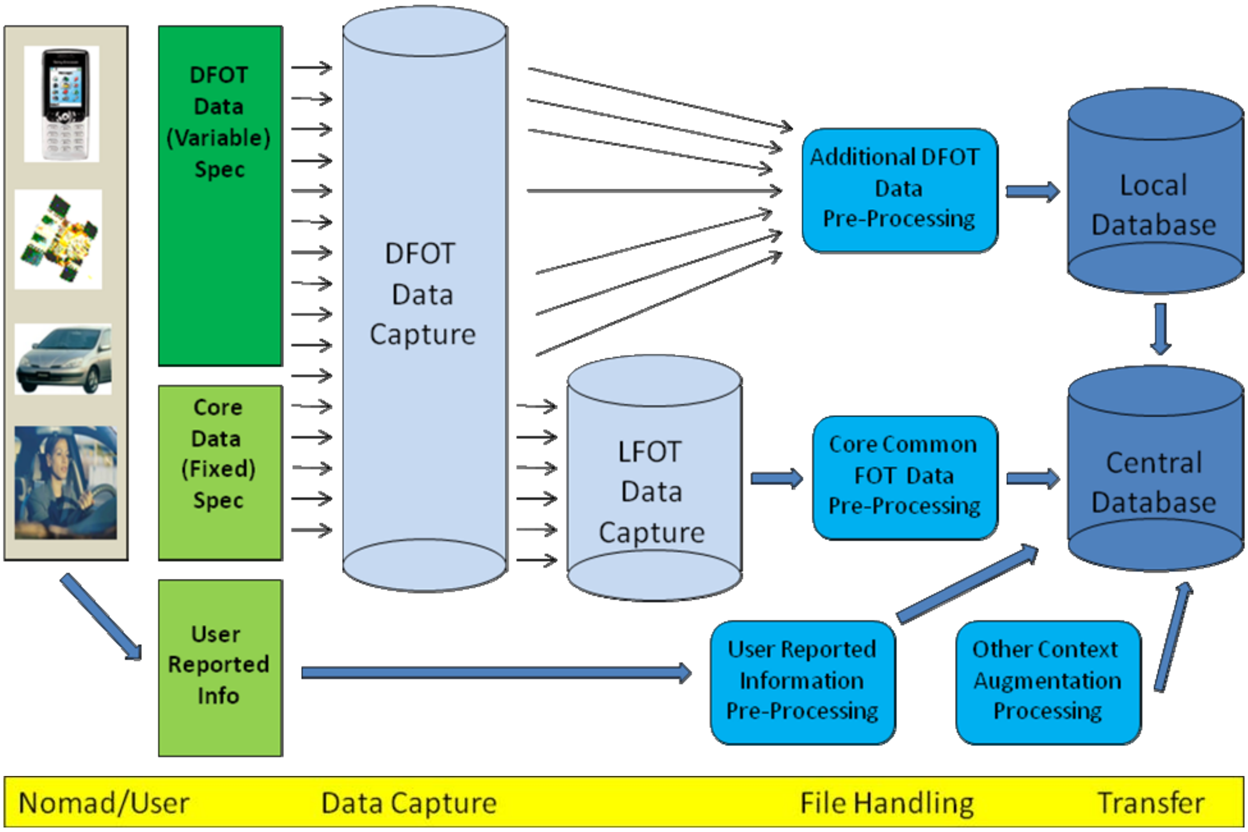

DFOT Data (Variable) Spec
Core Data (Fixed) Spec
User Reported Info
DFOT Data Capture
LFOT Data Capture
Additional DFOT Data Pre-Processing
Core Common FOT Data Pre-Processing
User Reported Information Pre-Processing
Other Context Augmentation Processing
Local Database
Central Database
Nomad/User
Data Capture
File Handling
Transfer

Figure 1.2 Overall Logged Data and Other TELEFOT context information

It may be noted that the largest influences on the content of this Core Data set represented by the data specification described below is equally that from the requirements of the top down analysis and constraints from the evaluation of the bottom up approach noted above.

1.2. Top Down Requirements

1.2.1. Expansion of Research Questions

TELEFOT WP 2.2 Methods and Tools has carried out the analysis of how the many identified high level research questions can be turned into definitive research hypotheses, which can then in turn be translated into data requirements, i.e. data variables that are required to be captured. To aid understanding of the how this process has influenced the core data specifications, a brief overview is given here with some selected illustrative examples given.

For each of the impact areas noted above a list of **primary** research questions was compiled. This was used to extrapolate a set of **second** level research questions and then a more detailed third level set was developed. This **third** level set was then used to establish specific hypotheses against which data variable requirements could established.

This is perhaps more easily illustrated by using a single example which is drawn from the safety impact area.

An overlying point of concern is whether the use of a nomadic device while driving has an impact on safety, or more specifically that use of such a device may increase the risk of the user encountering hazardous situations.

In this case the primary research question may be :

- *"Will there be a change in the number of hazardous events when a nomadic device is being used?"*

The secondary level research questions that were developed from this included :

- *"Exposure – Is there a change in the exposure of the user/vehicle to the road associated with nomadic device use?"*
- *"Focus of Attention – Is the user/driver distracted from the driving task when using a nomadic device?"*
- *"Driver Behaviour – Is there a change in driving style associated with nomadic device use that affects the user/vehicle and those around them?"*

If we take the last of these as an example, we can establish a number of third level research questions that are specific to driver behaviour assessment. Examples of these are :

- *"Is speed affected?"*
- *"Is proximity to other vehicles affected?"*
- *"Is risk-taking affected?"*
- *"Is lane position affected?"*
- *"Is braking behaviour affected?"*

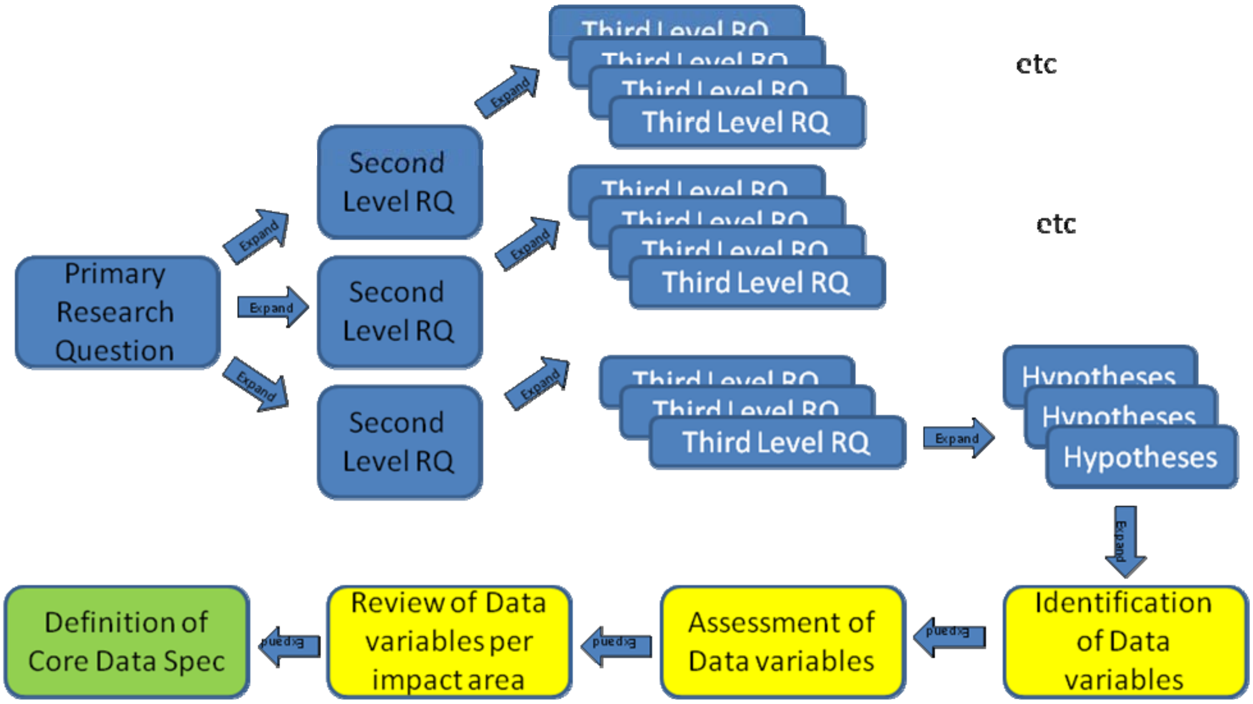
Once again taking the last of these as an example we can then establish at a fourth level what specific possible hypotheses exist which can then in turn identify specific data requirements. In this latter case a possible hypotheses to be tested is that :

- *"Incidences/prevalence of harsh sudden braking changes"*

From these hypotheses the method can then identify at a fifth level that it would be necessary in recorded to determine when harsh braking events occur and their duration. In this particular example this could be detected by data from the vehicle braking system, or acceleration data being recorded or being synthesized from recorded vehicle road speed. There is therefore a need to examine the trade off in resource cost and equipment capability to determine what is the most likely core data specification variable to address this hypotheses.

The expansion methodology to develop analysis of required data variables and their use in subsequent analysis is shown in Figure 1.2 below.

Figure 1.3 Expansion process of Research Questions per Impact Area


Primary Research Question
Expand
Second Level RQ
Second Level RQ
Second Level RQ
Third Level RQ
Third Level RQ
Third Level RQ
etc
etc
Hypotheses
Identification of Data variables
Assessment of Data variables
Review of Data variables per impact area
Definition of Core Data Spec

This identification of variables for data specification was then used to assess which variables would need to be selected for analysis from data records, which variables to analyse, and which variables would be needed to interpret results.

Results from this analysis of data and analysis requirements from each of the impact areas were then used to compare where data variables were identified that were similar to resolve any ambiguities and contrary requirements in the process of determining a core data set for all FOTs.

This methodology was adopted to ensure that there was a true top-down expansion of all identified research questions through successive expansion to full justify the inclusion of a specific variable, and the way it was to be utilised in the core data specification.

This methodology also identified where additional non-logged data was required to enable logged data to be interpreted adequately. This included the subjective and context augmentation additional information identified earlier in this chapter.

Taking the example of “harsh braking” given above this could include user reported incidences of experience of high levels of stop-go traffic on a particular journey (therefore causing more sudden braking events) which may in turn be supplemented by traffic flow information for that particular journey at a defined date, time and road location.

These expansions of high level research questions are now lodged in draft with the appropriate impact area analysis plans in SP4. These are currently only partially complete, but those that were available for Safety, Mobility and Efficiency were used to assess the content of the core logged data specification.

1.2.2. Other High Level Research Question Factors

There are several areas of the issues raised in the expansion of the research questions raised above that have common implications to how data is collected and processed and finally analysed. Further comments on these areas is given here.

Identity of sources of data – Clearly it is of vital importance that a data record related to the behaviour of a specific user/nomad device functionality in a phase of a national FOT can be traced back to that particular FOT, and related to other similar devices data records from that same FOT. This dictates that some unique identification of both the data logging unit and the FOT type and location is required. Ideally this identification should be available at point of data recording, e.g. from within the data logging unit itself, to avoid potential data misinterpretation. Ideally this should also be capable of being directly related to a specific FOT participant, i.e. person participating in the FOT. However it is acknowledged that this user and FOT context data may in some cases have to be associated with the logged data record by post-recording processing adding user/FOT information as a secondary process.

Data packages – As a common consideration in regard to the research questions is the impact of nomadic device use on journey behaviour, a means of determining how data recorded relates to "journeys" is required. This is particularly relevant when questions are posed that have an underlying hypotheses that nomadic devices that are intended to support driving, e.g. navigation support, green driving, traffic information etcetera, may be implicated in changing drivers underlying behaviour with regard to the timing, type, number and duration of "journeys" undertaken, as compared to when that nomadic device support is not available.

The DWG has considered the interpretation of what constitutes a "journey" in relation to travel behaviour and therefore TELEFOT concerns. Alternative terms that can be used are "trips" or "legs" in relation to "journeys". Essentially a "journey" or "trip" are the same. They both constitute an English meaning relating to the "act or process of travelling from one place to another". However this does have a relation as to how a "journey" can be interpreted in the case of data recorded in vehicle use such as in logged data in TELEFOT and real "journeys". In this case the real journey for an individual may constitute many partial legs, potentially using different transport modes. An example of a complete journey may be from an individual's home to a place of work some distance away. This may be broken down into a sequence of "legs", e.g. **walking** from home to car, **driving** in car from home to railway station, train from railway station to destination station, **bus** from railway station to a specific stop, **walking** from specific stop to the place of work. TELEFOT will seek to understand behaviour at a multi-modal level with supporting user reported information from questionnaires, travel diaries etcetera. TELEFOT will also examine in detail how behaviour in driving is impacted by use of logged data.

In this latter context the indication of the start and end of a "journey" maybe most appropriately triggered by determining when the vehicle is initially in "ignition on" state and when it reaches an "ignition off" state again. This however will require connection to vehicle systems which may not be possible in all FOTs. If this vehicle "ignition" status is not available then it will be necessary to derive start of journey from the change in location data available from GPS, i.e. unit in motion, to when that changing location stops.

In both cases this may therefore deliver a data record that can be identified with a "journey". Rules can then be applied to data to add further interpretation to "journey" records that constitute actual “legs”. In this example an “ignition off” termination of data may be considered as only an interruption if the time difference between the last “ignition off” and the next “ignition on” is less than a defined time period. This would account for temporary stops such as for refuelling. A similar interpretative rule would apply to GPS motion/location information. However this could still lead to inconsistencies and it is possible that later analysis definition WPs may select analysis by “journeys” as described above although noting that some of these “journeys” will be “legs”.

It is also noted that when data collection within the FOTs is related to logging within the nomadic device functionality, such as that for Navigation systems FOTs in Spain and Greece, that a journey start is defined and data collected when a user selects a new destination within the navigation system. In that case a single journey can consist of consecutive “legs” if the user changes the destination while travelling.

Human readable data sets – As the data recorded in the FOTs may require local processing and inspection prior to upload to the central database, it is important that the data streams captured and produced are able to be easily decoded in human terms. Therefore complex coding should be avoided.

1.3. Bottom Up Constraints

As has been previously noted the requirements for objective logged data in all FOTs potentially require some additional data logging equipment and/or functionality. While some of the nomadic devices selected have varying degrees of data logging, either within the nomadic device itself or within supporting infrastructure systems that the device communicates with, access to this data in a commercial system is problematic.

Therefore for most LFOTs an additional data logger was thought to be required. A number of Commercial Off the Shelf (COTS) data loggers will be employed within the FOTs that will offer additional logging facilities for TELEFOT purposes but also offer some constraints in terms of number and type of data logging channels, storage capacities etcetera.

The capabilities of these varied systems and their means of data transfer has been described in TELEFOT Deliverable D3.2.2a Test Tools. This has concluded that :-

For the large scale FOTs, GPS and acceleration data will be collected using in-vehicle data loggers, which usually have no connection to the vehicle electronics. The main challenge in the large scale FOTs is to get reliable data of the interaction between the driver and the nomadic device. A wide range of commercial devices will be tested in the TeleFOT sites, some of which are "black box" devices, for which additional logging software is difficult to add.

In the detailed FOTs, additional information will be gathered from vehicle sensors via CAN or OBD-II interfaces. Two additional project tools are also available for tests: CAA and ECA. The CAA module allows analysing the eye movement and activity of the driver using an in-vehicle camera, and the ECA module allows assessing the environmental risk (e.g. collisions). Both these tools allow processing collected sensor and camera information and extracting relevant information for further analysis.

Reference D3.2.2a Test Tools v30

These conclusions support the need for an agreed **minimum common core data** set across many devices and functions within the FOTs that is consistent across the project. This is described in the following section.

However it also highlights that there will be a wide range of data logged beyond the core data set in each specific FOT but this will not be consistent between test sites and FOT scenarios. It will therefore be necessary for each of the test sites/FOTs to define what the additional data specifications are for each scenario to enable formation of a holistic set of specifications. This is relevant for any additional data collected that will be subsequently intended for upload to the central database.

It is also possible that some specific additional logged data will only be used locally within a test site to assist data interpretation. This will probably include any video data recorded in the context of D-FOTs where this will be recorded at local level and summary data from the analysis only will be uploaded to the central database.

In this Deliverable 3.2.2a, a table (Table 13) has been compiled that summarises the data collection approaches and their associated advantages and drawbacks. This is included as Table 1.1 below within this deliverable.

Table 1.1 : Data collection methods for FOTs. (Ref Table 13 – D3.2.2a)

Logging method	**Parameters possible**	**Assets**	**Drawbacks**
Nomadic devices (data stored by function)	Use of devices and services, GPS, acceleration	Depends on the study. Mobile phones are personal loggers. Possibility to collect data on non-vehicle related trips	Software modifications may be required. Battery consumption when logging. Transmission of the data. Data is only collected when the device is switched on. Driver identification?
Nomadic device (FCD)	GPS, acceleration, use of service	Depends on the study. Mobile phones are personal loggers. Possibility to collect data on non-vehicle related trips	Agreement with FCD server provider. Data is only collected when device is switched on. Battery consumption.
Data logger (GPS + acceleration)	Mostly only coordinates with 1 Hz or less. Prototypes include detection of acceleration-based events	Cheap price, still enables business concepts	Installation costs: getting power and ignition signal (if required) is vehicle specific. No driver identification.
Access to vehicle bus via FMS (for trucks and buses) or OBD-II (for personal vehicles)	Subset of CAN, interest e.g. in fuel consumption	Variety of loggers. OBD-II for potentially easy end user installation and power. FMS also for driver identification (e.g. through digital tachograph).	Installation costs for FMS loggers, information content from OBD-II
Full access to CAN / vehicle bus	Numerous	Detailed logging	OEM specific

Server-side logs, traffic and weather data	Use of services, broadcasted traffic data	Valuable metadata without vehicle logging	Only weather data available internationally. Accessing logs requires contracts with service providers.
Questionnaires and travel diaries	Numerous	Collection of subjective data	Can be laborious to collect and analyse
CAA	eye direction	automated collection of video data	
ECA	headway, manoeuvers	collection of manoeuvre related data	
Video (/pre-processed)	Driver activity, headway, LDW…	Numerous	The amount of data and required analysis work, unless data is pre-processed. Permissions to use video.
Storing on hard drive / memory card in the vehicle	-	Nearly everything can be logged	Difficulties in changing and collecting the media, ensuring functionality
Sending data wirelessly to a data centre	-	Continuous data collection and monitoring. Driver feedback.	Maximally ~1 Hz continuous logging of a few values or events only. Big brother?

1.4. Core Logged Data Specification

The Data Specification described below therefore defines the data variables that are required to be logged in all TELEFOT FOTs **as a minimum**. It assumes that data logging is available either within the specific nomadic device under investigation and/or supported by an associated data logging unit or functionality.

There should necessarily be a correspondence between what is required as a core logged data and the required input data format for the central database. The specification defined below acknowledges this and the input data format is described in the following section.

Table 1.2 Core Common Logged Data Specification

Field	Data Name	Unit/Format	String Length	Notes
F1	Unique Identifier Logging Unit	DataLogger	15	Max of 15 characters for ID of logger (where available)

F2	Unique Identifier
	Function
		Nomad Device Function	3	NVS = Navigation Static
				NVD = Navigation Dynamic
				TRI = Traffic Info
				SPI = Speed Information
				SPA = Speed Alert
				GRE = Green Driving
				ECA = eCall

F3	Unique Identifier Device		4	Code for specific device name (FOT defined)
F4	Unique Identifier FOT Location	FOT Location Nation	2	UK, DE, SW etc

F5	Unique Identifier
	FOT Data Type
		FOT Scenario	1	T=Test
				D=Detailed
				L=Large

F6	Date	Year Month Date Yyyymmdd	4+2+2	E.g. 20100118 = 2010-01-08
F7	Time	Hour Minute Second Hhmmss	6	24 hour clock UTC E.g. 093005 = 09:30:05
F8	Data Record	East or West	1	E or W

	origin (x)	E or W		
F9	Data record origin location (x)	GPS x coordinates	8	3.5 format
F10	Data Record origin (y)	North or South N or S	1	N or S
F11	Data record origin location (y)	GPS y coordinates	8	3.5 format
F12	Altitude	Derived from GPS	5	Metres above mean sea level e.g. 265.8 m
F13	Speed	Derived from GPS	5	Kilometres per hour e.g. 125.5 km/h
F14	Number of satellites being tracked	Derived from GPS	2	e.g. 07 = 7 satellites being tracked

This indicates that GPS (GNSS) derived data with identifying data that associates a data stream with a specific FOT and nomadic device test forms the key common data log record for the TELEFOT trial.

This however is only the beginning of a process of adding further logged data, bespoke to each specific FOT, and additional subjective reported information and context augmentation information as indicated in earlier parts of this section. Each of these additional data and information sources will require agreed formats and input definitions and processes to be added to the central databases. These will also be needed to be subject to data quality issues described in later sections to this deliverable.

It has also been agreed that it is desirable for core data streams related to a specific journey to be led by a **header field** consisting as a minimum of fields 1-5 above to allow identification of that journey. The specific journey collected data then comprises the remaining fields 6-14 at the sampling rate possible for that FOT. It is noted that the target sampling rate for data collection is 1 Hertz for these fields however it is acknowledged that some L-FOTs may have restricted sampling rates due to the logging technology employed and may have longer sampling. In addition it is also noted that some higher specification data loggers in some D-FOTs may allow higher sampling rates.

In either case of variance from the target sample rate then it is the responsibility of the Test Site to indicate that to the central database manager as part of the FOT specific data specification.

It has also been agreed that post recording of data, each journey data file should have a summary file constructed that has appropriate metrics related for that journey calculated. This may include the following as examples :

- Distance travelled on journey
- Mean speed on journey
- Standard Deviation of Speed
- Total Time of journey
- Duration (Time) vehicle was stationary on journey

However the requirements for the content of this summary information should be defined by the impact area WPs of SP4 as these summary files will form the main way of accessing data for analysis. Specifications of these summary requirements are therefore requested from SP4 impact area leaders. This response for this specified summary items and methods of calculation of them can then be considered subsequently by the DWG, and the point in the data collection process where they are added to the base logged data can be identified by each FOT. This could be completed locally or centrally and the decision on which of these is more appropriate and/or efficient has yet to be concluded.

Pending this decision it is necessary to consider the current identification of the file and fields of the input data structures for the core logged data structures.

1.5. Input Data Structures

This section defines the default input data structure, i.e. the message from Data Loggers or Local Data Collection Server to the central Data Management Centre. The content on this section is based upon the latest version of the Input Data Structures definition provided by EMTELE (v12)

The defined message format is human readable where all the data in message are encoded in ASCII format. Each message must contain at least message header and message end. In addition it may contain arbitrary amount of GPS data minimum, acceleration data minimum and event data. Each record contains one snapshot and might include related processed data.

Header and each record is prefixed with identifier and the sequence CR LF is used as separator i.e. end of element marker for all elements except for end element. Each element consists of fields separated with comma (“,”). The format of data is fixed-length.

1.5.1. Message Header

The first row of the message is the header which tells the Data Logger ID and date of the records. This means that one message may contain records only from one day. As the date changes, a new message must be created.

Format: HEAD,xxxxxxxxxxxxxxx,dd,mm,yyyy

Example: HEAD,355632002225796,07,06,2009

Table 1.3 : Message header

Field	Description	Length	Example	Interpretation
1	Id of the header.	4	HEAD	
2	Unique ID of the data logger.	15	355632002225796	IMEI of the data logger
3	Day	2	07	7th day
4	Month	2	06	June
5	Year	4	2009	Year 2009

1.5.2. GPS Data Minimum Record

GPS Data Minimum Record contains the minimum set of the GPS data to be transferred from the data logger. Its format is as follows:

Format : GDM,hhmmss,llll.lll,a,yyyyy.yyy,a,M,78,008,78.34,08

Example : GDM,123337,4807.038,N,01131.000,E,545.4,78,008,78.34,08

Table 1.4 : GPS Data Minimum record

Field	Description	Length	Example	Interpretation
1	GPS Data Minimum, Record ID for GPS data.	3	GDM	
2	UTC time: Hours, minutes and seconds	6	123337	12:33:37
3	Latitude coordinates in the degrees/minutes format.	8	4807.038	48°07.038'
4	North (N) or south (S) side of the equator	1	N	North
5	Longitude coordinates in the degrees/minutes format.	9	01131.000	11°31.000'
6	East (E) or west (W) side of the Prime Meridian.	1	E	East
7	Altitude meters above mean sea level	5	545.4	545.4 m above sea level
8	Elevation in degrees, 90 maximum	2	78	78°
9	Azimuth, degrees from true north, 000 to 359	3	008	8°
10	Speed, km/h (what is the measured direction of the speed)	5	108.3	108.3 km/h
11	Number of satellites being tracked.	2	08	8 satellites being tracked

In case a certain GPS parameter is not recorded, the following values should be sent to represent a null value. Parameters in fields 1-7 are mandatory parameters which means that if a record does not include all of these parameters, it is rejected.

Field	Description	null value
8	Elevation	99
9	Azimuth	999
10	Speed	-9999
11	Number of satellites	-9

1.5.3. Acceleration Data Minimum Record

Acceleration Data Minimum Record contains the minimum set of the acceleration data to be transferred from the data logger. Its format is as follows:

Format : ADM,hhmmss.sss,x.xx,x.xx,x.xx

Example : ADM,190214.930,+0.03,+0.04,-1.00

Table 1.5 : Acceleration Data Minimum Record

Field	Description	Length	Example	Interpretation
1	ADM, Accelaration Data Minimum record ID	3	ADM	
2	UTC time: hours, minutes, seconds and milliseconds	10	190214.930	19:02:14.930
3	X acceleration in m/s^2	5	+0.03	+0.03 m/s^2
4	Y acceleration in m/s^2	5	+0.04	+0.04 m/s^2
5	Z acceleration in m/s^2	5	-1.00	-1.00 m/s^2

1.5.4. Event Data Record

Event Data Record contains the minimum set of event data to be transferred either from data logger or from nomadic device. Its format is as follows in Table 1.5 and 1.6 :

Format : EDR, hhmmss,xxxxxxxxx,xx,xxx

Example : EDR,123337,000000001,04,120

Table 1.6 : Event Data Record

Field	Description	Length	Example	Interpretation
1	EDR, Event Data Record, Record ID for EDR data.	3	EDR	
2	UTC time: Hours, minutes and seconds	6	123337	12:33:37
3	Trip ID	9	000000001	First trip ever
4	Event type	2	04	Speed limit change
5	Event value	3	120	New speed limit 120 km/h

Table 1.7 : List of event types

Event type	Description	Possible Event values
01	Navigation	000: OFF 001: ON
02	Speed limit information	000: OFF 001: ON
03	Speed alert	000: OFF 001: ON
04	Speed limit of the current road	020: 20 km/h

		120: 120 km/h etc.
05	Type of road	000: Motorway 001: Residential 002: Urban 003: Extra urban
06	Major / minor road	000: Minor 001: Major
07	Traffic volume	000: Low traffic 001: Medium traffic 002: Heavy traffic
08	Traffic requests	002: two traffic information requests
09	Trip to a new location	000: Route unknown 001: Route known
10	Re-calculations	000: no re-calculations 001: first recalculation 002: second recalculation etc.
11	Route selection	000: shortest 001: fastest 002: greenest
12	Speed cameras alert	000: OFF 001: ON
13	Green driving support	000: OFF 001: ON
14	Navigation mode	000: 2D 001: Real image
15	Sudden manoeuvre	000: No 001: Yes
16	Type of Journey	000: Commuting 001: Home related 002: Work related
17	Type of warning / message	000: none 000: visual 001: audial 002: visual & audial 003: turn left 004: turn right 005: speed limit

Data Structure End - It should be noted that the last row of the message includes the trailer mark which is “&” sign.

Example Messages – Below is an example message containing two seconds of GDM, ADM and EDR records as described above :

```
HEAD,355632002225796,07,06,2009
GDM,123337,4807.038,N,01131.000,E,545.4,78,008,78.34,08
EDR,123337,000000001,13,000
ADM,123337.000,+0.03,+0.04,-1.00
ADM,123337.200,+0.04,+0.05,-1.00
ADM,123337.400,+0.05,+0.06,-1.01
ADM,123337.600,+0.06,+0.07,-1.02
ADM,123337.800,+0.07,+0.08,-1.03
EDR,123338,000000001,04,120
GDM,123338,4808.009,N,01131.900,E,544.4,78,007,79.34,08
ADM,123338.000,+0.07,+0.08,-1.04
ADM,123338.200,+0.06,+0.07,-1.05
ADM,123338.400,+0.05,+0.06,-1.04
ADM,123338.600,+0.05,+0.06,-1.00
ADM,123338.800,+0.04,+0.05,-0.99
&
```

It should be noted that it is also possible to upload files containing only one or two record types. This means that the GDM, ADM and EDR records can also be uploaded each in a separate file in case that is more feasible solution.

1.5.5. File naming convention

Each file should be named after the following naming convention as shown in Table 1.8 below.

Table 1.8 : File naming conventions

Field	Description	Length	Example	Interpretation
1	Country code	2	ES	Spain
2	Name of the test site	variable	Vallaloid	The Vallaloid test site
3	Date	8	20091120	The data was logged in November the 20th in 2009
4	File number	variable	1	The first file today
5	Filename extension	2	tf	TeleFOT data file

The delimiter between each field described above is an underscore (_) character except for the filename extension of course where the delimiter is a dot (.).

Example filename : ES_Vallaloid_20091120_1.tf

1.6. Conclusions

The sections noted above have indicated how the core data specification was formed and stated the data fields that are required to be recorded from all FOTs within the TELEFOT project. This took into account the top down analysis of the research questions posed for each of the TELEFOT impact areas and the proposed data process for collection, handling and processing of data. The bottom up approach of identifying the data logging possibilities for both L-FOTs and D-FOTs and the technologies and loggers involved also formed an important backdrop to this specification.

A core common logged data specification based upon GPS has been defined which would enable the construction of a journey related data record, consisting of an identifying header, a main data file and a post recording summary file.

Issues remaining to be resolved include :-

- Specification of summary file components – SP4 impact area WPs
- Identification of mechanism of calculating summary file components
- Allocation of responsibilities for definition of additional logged data (local FOTs)
- Allocation of responsibilities for definition of subjective reported information formats
- Allocation of responsibilities for definition of additional contextual information formats
- Extended process definition of quality related issues for all three areas noted above

2. QUALITY OF DATA

In FOTs, assuring data quality during the experimental design, data collection, data management and data analysis activities is very important. The procedures for data quality assurance should therefore be well defined and guidelines for these procedures should be used.

As outlined in FESTA Deliverable D2.4 (Data Analysis and Modelling), data quality assurance is aimed at ensuring that data are consistent and appropriate for addressing hypotheses and research questions of interest. D2.4 states that data quality assurance starts from the FOT database and determines whether specific data analyses will be suitable for addressing specific research questions. However, data quality assurance should be ensured at all stages in the FOT – from the moment that the experimental design is conceptualised until the stage at which analytical outputs are derived and the answers to the research questions are produced.

Essentially the responsibility for data quality changes as the status of the FOT proceeds from the experimental design stage through to the data outputs stage. For example, whilst data is being collected during the FOT itself, responsibility for data quality should rest with the test-site managers. However, during the data outputs phase, responsibility for data quality rests essentially with the data analysts.

Therefore, there are a number of key stages at which data quality should be considered. The chain of events during most FOTs can be summarised in figure 2.1 (with the equivalent TeleFOT activities shown) and it is important to consider data quality at each point in the chain. As can be seen from the figure, data quality in respect of the TeleFOT project should be considered during the following stages;

1. FOT experimental design
2. Data collection
3. Data transfer
4. Data storage
5. Database quality control
6. Data analysis

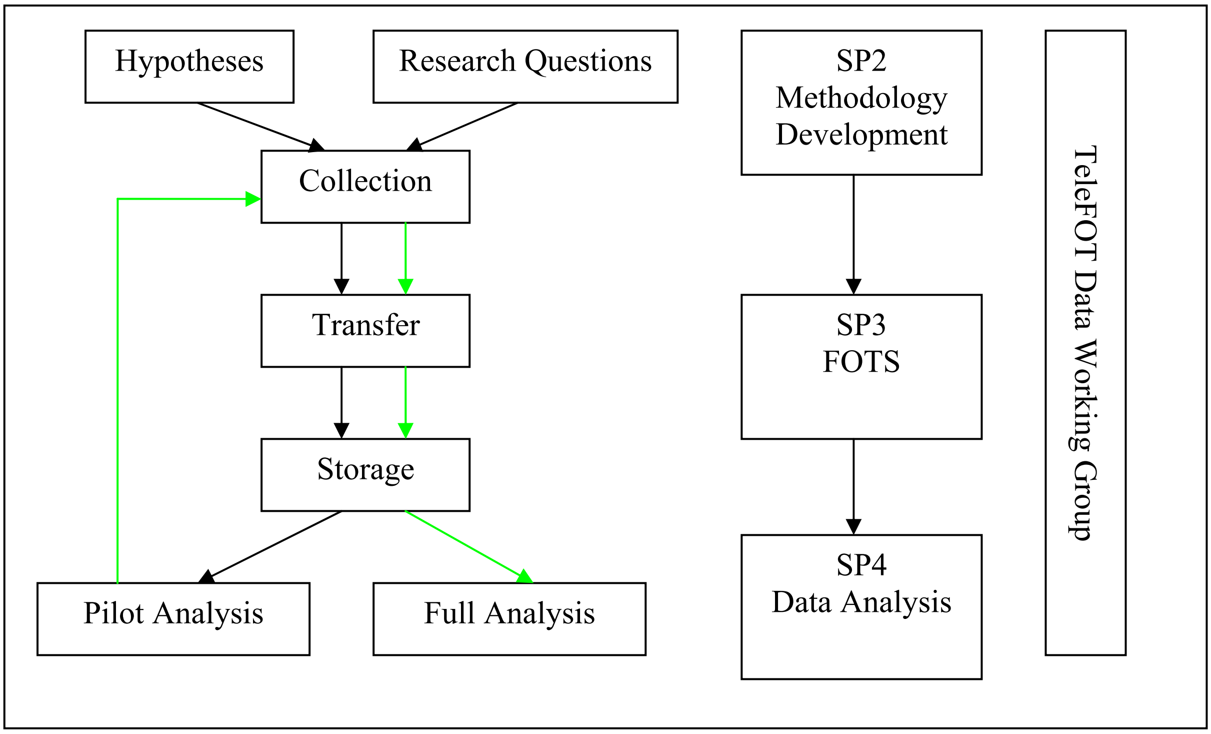


Figure 2.1 The FOT chain

In addition to the FOT chain shown above, data quality considerations can extend beyond the tasks defined. Therefore other activities and stages where data quality should be considered include;

7. Sampling requirements to ensure quality of subsequent data analysis
8. Subjective data collection
9. Contingencies in the event of problems with the data

Each of these stages is considered in turn in sections 2.1 to 2.9. A summary checklist covering the main points from sections 2.1 to 2.8 is provided in Annex 1.

2.1. FOT Experimental Design

The experimental design of an FOT can have an overwhelming effect on the quality of data and hence the results that are eventually derived. In particular, careful attention is required in terms of the functions to be tested, the participating subjects, the vehicles used in the FOT and the geographical location, time and date of the FOT. Many factors can influence the overall results and therefore, whilst a wholly rigorous scientific method is difficult to implement, steps should be taken to ensure that the parameters of the FOT are well-defined and caveats in relation to interpretation of the results are introduced where necessary. This is particularly true where the sample sizes are small and the consistency between an experimental and control subject group cannot be well maintained. Where inadequate experimental design is implemented, the quality of the outputs will be compromised and confidence in the results will at least be questionable.

The following considerations particularly relate to the experimental design of the TeleFOT project and how they may impact on the data quality

- Participant selection
  - Should be representative of the population that will ultimately be using the device;
  - Demographics (age, gender, experience etc) can be verified by questionnaire;
  - Personality aspects and their effect on the interaction with the device also need to be considered;
- Sample size
  - The study should be able to assess the functionality of the system and its impact on driver behaviour, traffic safety, environment etc;
  - If the sample size is too small statistical confidence can be difficult to prove;
  - Power of analysis increases with the sample size but is associated with additional expense. Furthermore, small effects may show a significance that may not actually be relevant;
- Study Design
  - The variables collected within the study should be comprehensive enough to allow the researcher to accept/reject the hypotheses;
  - Independent and dependent variables should be well defined at the start of TeleFOT.

2.2. Data Acquisition

Data Acquisition refers to the point in the FOT where the research questions and hypotheses have been determined and the FOT has now reached the ‘operational’ phase and data are being collected from individual journeys and trips. It is the responsibility of the test-site to ensure that the following data quality issues are considered.

1. Test subjects must remember to bring the nomadic device to the vehicle every time he/she uses it whether the device is itself being used as a DAS (Data Acquisition System) or not.
2. For all FOTs, the minimum requirement for robustness is that the entire system should operate under the normal driving conditions for the specific FOT including the harsher situations of normal driving.
3. When controlling the power supply to the DAS, the start-up and shutdown speeds must be optimised to reduce the loss of data. Loss of data can occur both during hardware initiation when no software is started and during hardware termination when no software is able to trigger on a vehicle restart. As much as 80% of the DAS hardware problems can be deduced to physical connector issues. Too high and too low temperatures (both static and transient) do affect the DAS. Components with moving parts need special attention.
4. Attaching any equipment to the in-vehicle CAN-bus systems has to be done very carefully. Transmitting data on vehicle CAN-buses should in most cases not be needed or not done at all in FOT implementation. Failure to adhere to this may be dangerous and result in vehicle operational malfunction that may result in significant cost, injury or death, or produce other very unwanted results.
5. When acquiring sensor data from a vehicle CAN-bus the information is passed through several stages before it can be read from the CAN-bus. These stages are likely to affect the signal value both in terms of amplitude and need to be carefully observed.
6. The nomadic device or system under evaluation (e.g. SatNav, SmartPhone, Green Driving Indicator etc) needs to be continuously monitored to ensure that it is operating properly. The system status signal should form a measure to be recorded in the data acquisition.

7. To ensure data validity and quality, a calibration and verification scheme is recommended. For data quality aspects it is important that all installed systems of the same category are calibrated and verified using the same procedures. During the verification process, a full dataset should be recorded for the analysts and quality management team in order for them to verify that the installation adheres to the analysis requirements.

Where video footage is being collected, the following should also be observed.

8. Direct real time observations should be carried out with great care and as unobtrusively as possible to minimise the risk of the driver modifying his/her driving behaviour.
9. The number and resolution and views captured by the cameras should be sufficient to address the hypotheses.
10. Pre evaluation of video image quality should be undertaken.

2.3. Data Transfer

Data transfer is the physical transfer of data usually over a point-to-point or point-to-multi-point communication channel. When data have been collected by the vehicle or system DAS, there is an obvious need to transfer these data from the respective systems to a data storage facility or location. Data Quality issues are evident during this particular stage of the FOT.

In TELEFOT, the following data quality issues should be considered;

1. When the point is reached where data transfer is required, checking procedures should be implemented to ensure that all collected data is backed up and stored in a safe place in order to minimise data loss. The aim is to prevent data loss, verify data completeness and to prevent data storage waste.

2. Data back-up should be initiated to prevent data loss by having multiple sets of the data stored in different places.
3. Data verification is aimed to assure that no data is lost during data transfer and data back-up.
4. Once data transfer has taken place and suitable data back-ups have been created, the test-site should consider data deletion to ensure that storage space is newly available in the vehicle.
5. Experience from previous FOTs suggests that data loss at the retrieval/upload stage is common even if it could be almost totally avoided with a robust and well-tested procedure. To prevent data loss during the data upload/retrieval procedure it is important to use a verification process to check that the data are consistent before deleting it from the vehicle. In the case that the verification process reveals that the data are not consistent, the vehicle data logger should be checked as soon as possible.
6. A process should be considered whereby synchronisation with subjective data can be ensured. Previous experience suggests that the accuracy needed in most cases is less than 5 seconds. For post-hoc structured comments or questionnaires on video or events, it is important to define a process of linking these events to the time.

The following figure depicts the data transfer scenarios relevant to TeleFOT. In the first scenario, the data is sent from the data acquisition system to the Data Management Centre. In the second scenario, the data is sent to a Local Data Centre and then later on fetched by the Data Management Centre.

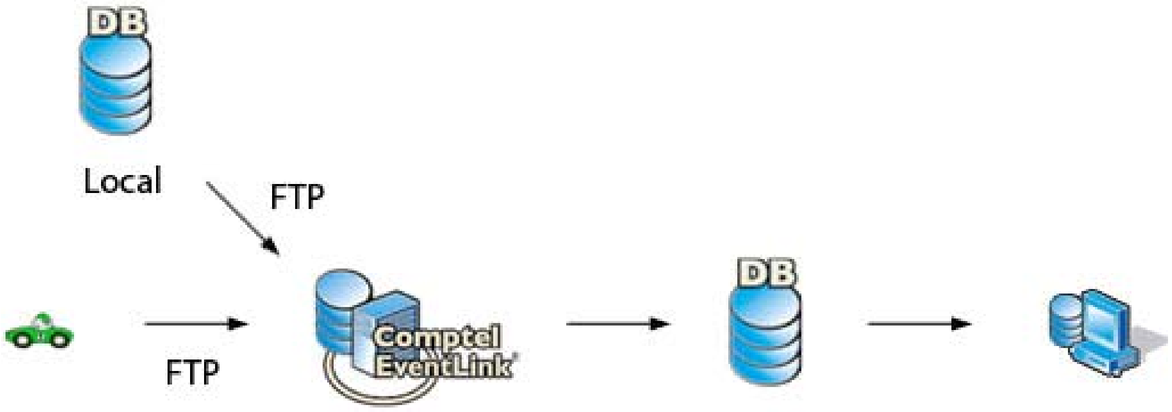


Figure 2.2: Different data collection scenarios

Input data structure concerns especially the Data Logger manufacturers and Local Database Management as they are expected to provide the input data using the format described in the Input Data Structures section. The protocol used for data transfer is FTP. Comptel EventLink is used as a tool by Emtele to collect and validate the data before storing it into the Data Management Centre's database.

2.4. Data Storage

It is expected that TELEFOT will generate thousands of hours of raw data during the collection phase. Therefore the quantity of data needs to be handled by an appropriate data management process to ensure data quality, to avoid data loss and to provide ease of access to the data analysts. Data storage has implications for data quality and the following points should be considered;

1. The main aim of the storage capacity estimation is to guarantee the availability of free space for recording the vehicle data. Ideally the sample rate for each signal should be the lowest possible able to guarantee no information (relevant for answering the research questions and hypotheses) is lost in the sampling process.
2. If there is no space available on the storage device this would inevitably result in data loss. Therefore, a 20% to 50% on storage size tolerance is recommended.
3. A safe data deletion procedure implies that no data should be deleted in the vehicle until a copy of the data has been backed-up, verified and stored in a safe place.
4. Storage of all data but in particular video data should be in a relational database. Implementation should consider what to do in the event of a data loss from a sensor (for example, a null value could be inserted).

2.5. Database quality control

As trials are completed, data will be entered and stored on a central database. The quality of this database depends on good management of the case materials and a well-designed user interface for data input and data downloads. Oracle, SQL Server and MS-

Access are examples of database applications available to serve as the central database. A logical hierarchy of relationships between the component data tables, if applicable, is an important foundation.

The trials may generate digital case materials that cannot be incorporated into the central database. This could include images, video, sensor output, and time-location data-streams. It is important that these files are systematically named according to rigorous protocols so that they can be identified by computer logic. This also applies to case directories and folders.

A user-friendly data input system with the capability for validation checks is necessary to create a good quality database.

The management of data records - creation (especially), modification and deletion - is the first general requirement. The system should also respond interactively to data input by only showing relevant sections of the forms, hiding those that are irrelevant or not applicable to the case at hand.

A very simple but critical aspect of the design of the database is to ensure that a value has been entered for each field (even if is ‘Not Known’ or ‘Not Applicable’) and the data that are entered are valid. A warning should be issued at data entry for values that are valid but extreme, rare or otherwise improbable.

At certain stages of data input it is recommended to have the user "sign-off". This signals the completion of part or all of the data input stage, including compilation of digital case materials. The data input system should then make cross-checks to ensure that the whole database is internally consistent. Such checks include;

- Calculation of derived values;
- Checking that all necessary parts of the database have been filled in;
- Reading the case folders to ensure that all core, required materials are present;
- Listing of all case materials.

This relies on the folder structure and file-naming protocols mentioned above. If problems are detected, the program should block the case from being marked as completed.

Figure 2.2 below illustrates the database quality control.

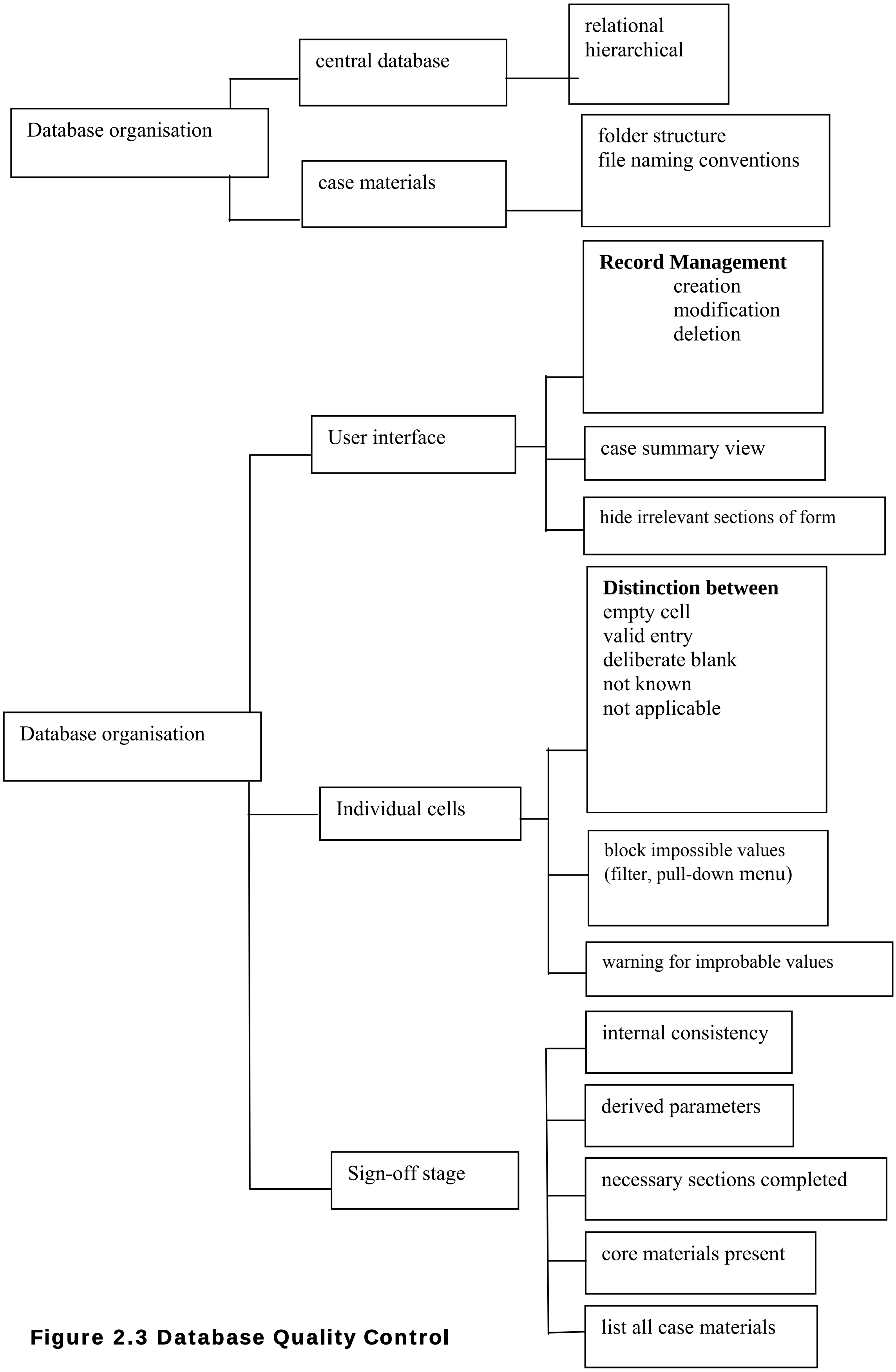


Figure 2.3 Database Quality Control

2.6. Data Analysis

The Data Analysis stage is the point where hypotheses and research questions are confronted and outputs are derived which answer the research questions and provide support for, or evidence for rejection of the hypotheses.

In the FESTA Manual on data analysis, three main difficulties are identified in the data analysis process. These are as follows;

- That there are huge and complex amounts of data coming form different sensors and data acquisition systems including subjective and video data;
- That there is the potential for bias in the data analyses due to the sampling including selection of drivers, location of FOT and the systems and functions that are being tested; and
- That extrapolation of the results to level of the whole transport system is not a straightforward process.

Pilot Analysis

Data quality analysis should be undertaken to ensure that the data are consistent and appropriate for addressing the research questions and hypotheses. A pilot data analysis should be conducted in TELEFOT as soon as data are available form the FOT to consider and ensure;

1. That the data storage system works;
2. That data can easily be downloaded form the data storage system;
3. The completeness of the data collected; and
4. The validity of the data – i.e. that no erroneous or extraneous values in the data are found

If failures are found in any of the stages listed, then the database management team should be alerted as soon as possible so that the error can be located and corrected at the first opportunity.

The pilot analysis should also be used for the following purposes;

1. To ensure that data values are reasonable and units of measure are correct (e.g. a mean speed value of 6 may be unreasonable unless speed was actually recorded in m/s instead of km/h);
2. To check that the dynamic data over time is appropriate for each kind of measure (e.g. if the minimum speed and the maximum speed of a journey are the same, then the data may not have been correctly sampled);
3. To guarantee that measurements satisfy the requirements for the specific data analysis; and
4. To check that participants can be asked how often they use a function and also that actual function activation and the different settings chosen by the driver can also be logged from the system.

2.7. Sampling requirements for quality of data analysis

Within the FOT study design the number of participants and the duration of each trail will have been determined. It is important to check that the data that will be collected, given these elements of the study design, will be sufficient in order to meet the analytical requirements.

It is likely that a number of analytical techniques will be employed in order to measure the effect that the nomadic devices have in the various impact areas. These will also vary according to the nature of the data, for example subjective versus objective or categorical versus continuous.

The power of the analysis, that is the chance that the hypotheses is accepted when it is in fact false, is largely dependent upon the frequency upon which the event under consideration occurs under control conditions, the effectiveness that is anticipated upon the introduction of the test condition and the sample size. Hence, the sample size required for a predefined power of analysis is related to the other two factors.

As a general rule, the more power required (i.e. the less chance of drawing a wrong conclusion) the larger the sample size. If small effects are anticipated then large sample are required, and if events occur infrequently then large sample sizes are required. If many confounding factors need to be analysed with the data using multivariate techniques to, for example explain why an effect has been noted, then this also increases the amount of data that is needed.

It is also recognised that budget allowances among the test site will play a part in determining the amount of data collected as will the constraints on the overall timescale of the project.

The following is taken from TeleFOT D2.2.1 which should be considered;

FOTs are described as studies in which a large number of individuals participate. In TeleFOT a difference has been made between D-FOTs and L-FOTs with implications for the number of participants to be in volved. A test involving e.g. 10-12 participants will not be regarded as a L-FOT whereas a test involving, e.g. 100 participants may. On the other hand, 10-12 participants may suffice in a D-FOT.

The required number of participants in an FOT will always depend upon a number of factors, e.g. the number of functions and/or systems to be tested, the hypotheses formulated, the choice of a between or a within subjects design, etc. If the number of participants is small, it is difficult to statistically prove any effects of the function/system that are actually there whereas a large number of participants increases the chance of finding an effect. On the other hand, a large sample implies a higher investment in terms of equipment, resources, etc.

In order to ensure that the chosen sample size is representative for the behaviour of a group of users and that it is possible to statistically prove any effects that are there, a power analysis is needed to calculate the desirable sample size. A statistical power analysis exploits the relationships between the four variables involved in statistical inference: sample size, significance criterion, population effect size, and statistical power (e.g. Cohen, 1992). For any statistical model, these relationships are such that each of them is a function of the other three. In line with the previous reference, the following formula can be adopted as a basic approach to determine the sample size to be used in the trials:

$$n = \frac{Z^2 pq}{E^2}$$

where **n** is the size of the sample, **Z** is the confidence level (determining the confidence of the generalisation of the data from the sample to the population), **p** is the estimated effect, **q** is equal to 1-p (where pq represents the level of variability in the computations to verify the hypothesis) and **E** error (the percentage of error to be accepted in the generalization).

If considering e.g. a confidence level of 95% (Z=1,96), an estimated effect of 5% (p=0,05 and q=0,95) and an allowed error of approx. 2,5% (E=0,025), the required number of participants would be n=292. Nevertheless, this final number would correspond to the total number of observations needed. Thus, depending on the registered data:

- if only a single observation per subject is stored, 292 subjects would be needed.
- if more than one observation per subject is registered (and this will be the case for most of the variables in L-FOTs as well as D-FOTs, e.g. will a large amount of speed values be stored): less subjects would be needed since for each of them, a large set of will be available.

This should be considered as a first estimation of the appropriate number of subjects/observations to be considered in a L-FOT since effect sizes, acceptable errors, etc. cannot be assured in advance.

2.8. Special Considerations for Subjective data collection

A number of factors contribute to the quality of the subjective data collected by means of interviews and/or questionnaires. To reduce errors, automatic transcription of subjective data is always preferable. Subjective data should be stored and handled logically and preferably stored electronically. When this is the case, the guidelines listed above apply.

Issues to consider when collecting subjective data are now discussed;

Questionnaires

If a questionnaire is distributed, the factors include

- The respondent (ability and willingness to respond as well as experience and knowledge)
- The content and design of the questionnaire (e.g., complexity; number of questions, formulation of questions, etc.).

In order to address the validity of the data, the formulation of the questions (and possible answers) is a key issue - even more so when designing a questionnaire to be distributed to respondents. Questions must evidently be formulated in a way so that they measure what is intended to be measured. However, questions must also be

- Specific;
- Not too complicated;
- Formulated in terms that can be understood by the interviewee.

Hypothetical questions are the most difficult questions and should be avoided. Data quality can also be improved by designing the interview/questionnaire so that interview/questionnaire itself checks for consistency – for example, the same question posed in different ways. Pilot tests must *always* be carried out in order to ensure the clarity and completeness of the questions.

Interviews

Some problems associated with missing data can be avoided by choosing interviews instead of questionnaires. For example, the interviewer can explain the question if it is poorly formulated and therefore not understood correctly. The interviewer can also probe for answers with open-ended questions, ensuring not only collection of data but also that more in-depth information is gathered. Also the questionnaire approach allows the respondents to answer the questions at a time of their own choice which may ensure a higher response rate. If data is missing, it is important to determine if there is a bias. For instance, one should check whether missing data results from a specific group or category of participants and how this may bias the analysis of the data. As a rule of thumb, if missing data is less than 10% and is randomly distributed, the analysis may not be significantly affected.

The interview situation is affected by more factors but at the same time allows for control which contributes to higher quality data. In an interview situation important factors to consider include;

- The interviewee (in terms of, e.g., social skills, ability and willingness to respond as well as experience and knowledge);
- The interviewer (in terms of social skills; training; motivation, etc.) and
- The content of the interview (in terms of, e.g., complexity; the sensitivity of the topics addressed) as well as the structure.

The interviewer plays an important role in collecting data in an interview situation. In order to ensure the quality of the data derived during this process, the interviewer should;

- Show the interviewee respect;
- Be able to listen as well as be able to communicate (”active listening”);
- Should not be afraid to wait for the answer;
- Should have good knowledge of the issues addressed and the specific conditions;
- Should never show dislike, irritation, or stress.

Interviewer bias (which is the influence of the interviewer on the participant's response) can be avoided by administering a questionnaire if necessary (see above). However the interviewer may be able to increase the quality of the data collected by being present to explain questions and by also using probing questions.

The Questions

Questions (in either a questionnaire or interview situation) can be open-ended or close-ended. Open-ended questions do not supply any answer categories while close-ended questions do.

If close-ended, the answer categories should be as few as possible in relation to the questions; be relevant in relation to the type of question; be mutually exclusive; be reasonable and make sense. They should allow the respondent or interviewee to be able to answer the question.

The answers to open-ended questions will take longer to analyse than close-ended. Missing data is more common for open questions than closed. Furthermore, most often these answers must be coded which in itself may result in errors. This can be avoided by the support of a clear and consistent code key. Furthermore, in an interview situation, the interviewer can summarise the answer or group of answers, and allow the interviewee to agree or disagree and/or to comment on the interpretation. Consistency in coding can be checked by comparing several independent analysts' coding of the whole or a subset of the collected data. Questions can also be direct or indirect. An indirect question directs the interviewee's attention to another person (or to other persons) other than the interviewer and can be a way to address more sensitive questions or areas where a "true" answer may not be anticipated.

Missing data is a threat to the quality of the data at all levels of operation, whether an entire interview or questionnaire is missing or the answers to individual questions are missing (or indeed, answers are not readable). In addition, data can be missing due to the respondent providing an answer, or providing a rating which is outside allowed categories. In the case of a missing questionnaire or interview, efforts must be made to ensure that data collection is as complete as possible and reminders must be administered. Furthermore, overall the number of questions should be considered carefully. Where possible, it is preferable to limit the number of questions. In addition, the number of open questions should be as few as possible in order to reduce the effort of the respondents.

2.9. Contingencies in the event of problems with the data

Contingencies may be required in the event that problems occur with missing, lost, erroneous and inconsistent data. Table 1 summarises the risks that are inherent in FOT data collection, how the risk can be managed and proposed solutions in the event that the identified risk becomes a reality;

Contingency Plan		
Risk including risk severity (e.g. low, medium, high)	**Reduction** **(how the risk can be managed)**	**Solution** **(if the risk happens)**
Missing data at point of collection (medium risk, medium severity)	Arrange check-list of required data-fields to ensure that collection is fully specified.	Missing data should be denoted as such in analyses and caveats will be applied to results.
Loss of data post-collection (low risk, high severity)	Ensure that data back-ups are provided (on a main server and DVD)	Back-up should be utilized. If data cannot be recovered, same caveats as above will be applied
Inconsistent data across test communities meaning comparisons cannot be made - in multi-centre studies (low risk, medium severity)	Data quality should be determined through pilot data analyses. Data consistency should be ensured through a review process.	In the event that this occurs, data analyses should not be conducted where data inconsistencies are found
Insufficient data to ensure scientific rigour/statistically valid outcomes (medium risk, high severity)	This should be established and addressed in a pilot study – any indications that the data will not give statistically robust results should result in revision of methods, tools and data specification	In event that this occurs, the data analyses should be modified accordingly and the validity of the outcomes described.

Late identification of needed analysis and analysis procedures cannot accommodate it (medium risk, medium severity)	Research questions & indicators should be established at the start of the project which should prevent late identification of required analyses. An early task in any experimental design should identify the database structure and thereby should incorporate flexibility to respond to unpredicted analysis requirements. A pilot study should test analysis procedures.	All efforts should be made to include the required analyses. Where this is not possible, the risk management strategies should have ensured that this analysis is not core to the needs of the impact assessment.
Privacy of participant data compromised (low risk, high severity)	All reasonable measures should be taken to ensure privacy. Protocols should be developed based on expert advice. Data should be stored in lockable filing systems – no personal data should be stored on databases. Participant identification details should be shredded shortly after use.	Participants should be informed of privacy compromise and appropriate remedial actions will be taken in consultation with participants.
Commercially confidential data compromised (low risk, high severity)	Protocols should be developed. Stakeholders should be informed before participation that all reasonable measures will be taken to ensure commercially sensitive information will be kept confidential. This should be documented.	Stakeholders should be informed of compromise and appropriate remedial actions will be taken in consultation with stakeholders.

Table 2.1.Data Contingency Plan

3. DATABASE STRUCTURE

This section contains the high level data specification. The contents have been divided in to requirements regarding the performance, security and availability. The High Level Data Structure and Layout is presented to give an idea of what type of data is stored into the Data Management Centre and the relations between different entities. The structure of input data is described to provide the Data Logger manufacturers and Local Database Managers a common data format for sending and storing the data into Data Management Centre.

Each requirement contains the following attributes: ID, name, description, use case, type priority and evaluation criteria. ‘ID’ is the unique identifier of a requirement. ‘Name’ is the name of the requirement. ‘Description’ lays out the requirement in more detail. ‘Use case’ identifies a type of situation where the requirement is relevant. ‘Type’ defines whether the requirement is functional or not. ‘Priority’ shows the priority of the requirement on a scale of interesting, relevant or critical. ‘Evaluation’ criterion defines the condition which can be used to determine whether the requirement is met.

The document does not describe the low level data structure and layout. The purpose of this deliverable is to use it as a backbone when designing the low level system. If required, the input data structure may also be slightly altered later on.

3.1. Performance Requirements

The performance of a system is a function of its operational workload, the underlying hardware and software infrastructure and the applications’ persistent data volume. The two most important concepts are response time and throughput. Response time defines how quickly the system responds to a request. Throughput means how much work can it do in some periodic of time. Response time is really important for user interfaces. Throughput is really important for a system that has to process a lot of data. It is also critical for systems which have many users.

The performance requirements guarantee the analysts access to the data in a way which enables them to operate as efficiently as possible. Meanwhile the system should also be able to handle new incoming data at a feasible rate.

The most common sources of performance limitation can be, for instance, peak hours regarding the number of active analysts as well as the amount of incoming data.

When calculating the cost forecast of possible performance issues, it should be noted that most of the costs are indirect and related to the delayed results of the project. On the other hand, analysts are unlikely to lose any extra working time since the amount of time each operation will take can be estimated by previous experiences.

The performance requirements are summarised in table 3.1

Req ID	Requirement
WP23_REQ_01	Business level performance requirements
WP23_REQ_02	Data logger related performance requirements
WP23_REQ_03	Monitoring performance requirements
WP23_REQ_04	Database performance requirements
WP23_REQ_05	Multiple transfers requirements
WP23_REQ_06	Maximum throughput time requirements
WP23_REQ_07	Maximum load and response time requirements
WP23_REQ_08	Minimal functional coverage
WP23_REQ_09	Data accessibility
WP23_REQ_10	Interface efficacy

Table 3.1 Summary of performance requirements

Details of the performance requirements are given in tables 3.1.1 to 3.1.10

ID	WP23_REQ_01
Name	Business level performance requirements
Description	Stakeholders may need to access the data at any given day. Data Management Centre must be able to process more data in a day than can be generated in a day. Data Management Centre must be able to provide also the “raw” input data it downloaded from data loggers or local data centres.
Use case	Data Management Centre
Type	Functional
Priority	Relevant
Evaluation Criteria	Availability of database

Table 3.1.1 Business level performance requirements

ID	WP23_REQ_02
Name	Data logger related performance requirements
Description	Data logger must be able to record data using sufficient sampling frequency. For GPS this stands for a sampling frequency of 1 Hz or higher and for acceleration data a sampling frequency of 50 Hz. Data Management Centre must be able to process the amount of data recorded by data loggers.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	Data logger specifications and average load on Data Management Centre

Table 3.1.2 Data logger related performance requirments

ID	WP23_REQ_03
Name	Monitoring performance requirements
Description	Data Management Centre must have a function for monitoring the data processing. In particular, the amount of records processed per unit of time.
Use case	Data Management Centre
Type	Functional
Priority	Relevant
Evaluation Criteria	Statistics related to Comptel EventLink performance

Table 3.1.3 Monitoring performance requirements

ID	WP23_REQ_04
Name	Database performance requirements
Description	In order to be useful when conducting analyses, the database has to be able to provide required data within a reasonable amount of time. The maximum time that an advanced and optimized database query can take up to is determined to be one hour.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	Duration of database queries

Table 3.1.4 Database performance requirements

ID	WP23_REQ _05
Name	Multiple transfers requirements
Description	Data Management Centre must be able to open multiple concurrent connections to receive and process all the data from the amount of around 3000 data loggers within the TeleFOT project. Data Management Centre must be able to handle the estimated amount of traffic generated by the data loggers.
Use case	Data Management Centre
Type	Functional
Priority	Relevant
Evaluation Criteria	

Table 3.1.5 Multiple transfer requirements

ID	WP23_REQ _06
Name	Maximum throughput time requirements
Description	The requirement for maximum throughput time depends on how quickly the data is needed for analysis. A reasonable time for the data to be ready for analysis purposes has been determined to be 24 hours after it has been collected. The same reasoning applies to throughput time for data processing and fusion purposes as well.
Use case	Data Management Centre
Type	Functional
Priority	Relevant
Evaluation Criteria	Comptel EventLink performance statistics

Table 3.1.6 Maximum throughput time requirements

ID	WP23_REQ _07
Name	Maximum load and response time requirements
Description	The number of different users and user interfaces will be rather limited which brings down the maximum load and response time requirements. However, the data collection, processing, storing into a database and database queries for analysis purposes will all happen concurrently at times. As a result, the data collecting, processing as well as the database have to be developed to be efficient as the amount of data will be large.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	The average server load should be less than half of its capacity.

Table 3.1.7 Maximum load and response time required

ID	WP23_REQ _08
Name	Minimal functional coverage
Description	Data Management Centre must be able to create all the data entries (and structures) reflecting the current business requests. The same is valid about the deletion requests, to be accompanied by the reference integrity constrains (no orphans, cascade etc.). The data migration routines if any must be provided.
Use case	Data Management Centre
Type	Funcional
Priority	Relevant
Evaluation Criteria	

Table 3.1.8 Minimal functional coverage

ID	WP23_REQ _09
Name	Data accessibility
Description	Data Management Centre should make accessible the business the contents: the must be readable and updatable, according to the rights of the user. Being a B2B service, real time access should be supported.
Use case	Data Management Centre
Type	Non functional
Priority	Relevant
Evaluation Criteria	

Table 3.1.9 Data Accessibility

ID	WP23_REQ _10
Name	Maintenance Interface efficacy
Description	Providing to the stakeholders a usable and effective user interface, making easy the db maintenance activities. It will improve the general performance of the database. Interface must support actions as: query lists, data sorting and visualization, data export in the main format available.
Use case	Data Management Centre
Type	Non functional
Priority	Relevant
Evaluation Criteria	

Table 3.1.10 Maintenance interface efficacy

3.2. Availability Requirements

The availability requirements are aimed to contribute to setting up the conditions that will guarantee the access to the database is as wide and constant as possible. The aim is to try to anticipate possible downtime and service interruptions, in order to predifine recovery strategies.

This family of requirements is strictly related to the conditions posed by the business scenarios the project looks at. The use of the device for e-Call, for instance, should resolve the context in order to provide remote assistance, using data about circumstances like: affected area, date, time, position, car's conditions, road's conditions, activities preceding the request (braking etc.). The availability of the data will depend on the goal and the actions/service that will kicked up as consequence.

The most common sources of availability limitation can be, for instance, electric power down, absence of connectivity, deny of service by hackers. Planned and unplanned downtime sources should be evaluated and ranked, in order to prepare mitigation strategies.

The cost forecast of possible availability problems and recovery strategies will be based on the quantity of additional peripheral storage needed (on nomadic aftermarket devices) to avoid data loss and the need of eventual additional synchronisation procedure between overloaded devices and the host.

The availability requirements are summarised in table 3.2.

Req ID	*Requirement*
WP23_REQ_AV_01	Redundancy
WP23_REQ_AV_02	Back-up frequency
WP23_REQ_AV_03	Mean Time of repair
WP23_REQ_AV_04	Mitigation strategies
WP23_REQ_AV_05	Recovery time objective

Table 3.2 Summary of availability requirements

Details of the availability requirements are given in tables 3.2.1 to 3.2.5

ID	WP23_REQ_AV_01
Name	Redundancy
Description	Data Management Centre must provide solutions to guarantee locally and/or remotely the back-up version of the DB requiring also some added storage capacity.
Use case	Data Management Centre
Type	
Priority	Relevant
Evaluation Criteria	

Table 3.2.1 Redundancy

ID	WP23_REQ_AV_02
Name	Back-up frequency
Description	Based on the amount of data and of users, the periodic back-up must be planned. The synchronization with the remote devices should be considered as well.
Use case	Data Management Centre
Type	
Priority	Relevant
Evaluation Criteria	

Table 3.2.2 Back-up frequency

ID	WP23_REQ_AV_03
Name	Mean time for repair
Description	Based on the determination of the criticity, the time of repair should be calculated and communicated to stakeholders.
Use case	Data Management Centre
Type	
Priority	Relevant
Evaluation Criteria	

Table 3.2.3 Mean time for repair

ID	WP23_REQ_AV_04
Name	Mitigation strategies
Description	In order to guarantee and make the access to the DB contents faster, possible strategies can be implemented as pre-calculated queries.
Use case	Data Management Centre
Type	
Priority	Relevant
Evaluation Criteria	

Table 3.2.4 Mitigation issues

ID	WP23_REQ_AV_05
Name	Recovery time
Description	If the local DB will prevent data loss, implementing redundancy and mitigation strategies, the recovery time in case of service interruptions would not be critical. The evaluation of the maximum time has to be defined according to the business scenarios. For instance the maximum time would be in the order of seconds, to meet eCall scenario requirements. A downtime of one week could be tolerable as insurance companies requirement.
Use case	Data Management Centre
Type	
Priority	Non-Relevant
Evaluation Criteria	

Table 3.2.5 Recovery time

3.3. Security Requirements

This section defines security requirements for the TeleFOT database environment which need to be fulfilled in order to guarantee the common goals of protecting confidentiality, integrity and availability of TeleFOT information.

For example, common security risks are incorrect handling or storing of usernames and passwords, malicious programs and intrusion attempts.

The security requirements are summarised in table 3.3.

Req ID	*Requirement*
WP23_REQ_SEC_01	Identification
WP23_REQ_SEC_02	Authentication
WP23_REQ_SEC_03	Authorisation
WP23_REQ_SEC_04	Immunity
WP23_REQ_SEC_05	Integrity
WP23_REQ_SEC_06	Intrusion detection
WP23_REQ_SEC_07	Non-repudiation
WP23_REQ_SEC_08	Privacy
WP23_REQ_SEC_09	Security auditing

WP23_REQ_SEC_10	Survivability
WP23_REQ_SEC_11	Physical protection
WP23_REQ_SEC_12	System maintenance

Table 3.3 Summary of security requirements

Details of the security requirements are given in tables 3.3.1 to 3.3.12

ID	WP23_REQ_SEC_01
Name	Identification
Description	Every entity, device or user, using database environment must be identified before granting access to database environment. Normally username is used to identify the entity.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Without identity it is impossible to access the database environment.

Table 3.3.1 Identification

ID	WP23_REQ_SEC_02
Name	Authentication
Description	Every entity using database environment must be able to prove its identity for database environment. The entity must have a credential issued by the database administrator for proving the identity. Password is the most common credential for accessing some system.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Without credentials it is impossible to access the database environment.

Table 3.3.2 Authentication

ID	WP23_REQ_SEC_03
Name	Authorisation
Description	Entities can only access data and resources for which they have been properly authorized. Access rights to database resources and data must be defined carefully and by using principle of least privilege. The entity should have access only to the resources and data that it is necessary and can be granted without breaking privacy policy of the database environment.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Entities cannot access resources or data that they are not properly authorized.

Table 3.3.3 Authorisation

ID	WP23_REQ_SEC_04
Name	Immunity
Description	Database environment must protect itself from infection by unauthorized undesirable programs, e.g. viruses and Trojan horses.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Unauthorized undesirable programs are immediately detected and their execution is denied.

Table 3.3.4 Immunity

ID	WP23_REQ_SEC_05
Name	Integrity
Description	Database environment must ensure that any data cannot be modified without authorization and all changes to data have to be legal, accountable and correct. The typical objectives of an integrity requirement are to ensure that communications and data can be trusted.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Unauthorized data modifications are denied and integrity violations in communications are detected.

Table 3.3.5 Integrity

ID	WP23_REQ_SEC_06
Name	Intrusion detection
Description	Database environment must detect and record attempted access or modification by unauthorized entities. Intrusion detection system should inform security personnel on intrusion attempts so that they can handle them.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	All unauthorized intrusion attempts are detected.

Table 3.3.6 Intrusion detection

ID	WP23_REQ_SEC_07
Name	Non-repudiation
Description	Database environment must prevent an entity to one of its interactions (e.g., message, transaction) from denying having participated in all or part of the interaction.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	There must not be any interaction with database environment that is not logged to the system.

Table 3.3.7 Non-repudiation

ID	WP23_REQ_SEC_08
Name	Privacy
Description	Database environment must secure the confidentiality of the sensitive information, i.e. ensure that unauthorized individuals and programs do not gain access to sensitive data and communications. Privacy is both political and technical issue. Policy defines what data is sensitive and technologies and processes implement the privacy.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	

Table 3.3.8 Privacy

ID	WP23_REQ_SEC_09
Name	Security auditing
Description	Database environment must enable security personnel to audit security mechanisms and their status.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Security mechanisms and their status must be possible to audit.

Table 3.3.9 Security auditing

ID	WP23_REQ_SEC_10
Name	Survivability
Description	Database environment must gracefully survive from unintentional loss or destruction of a component.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Every component of the database environment are run down and survivability of the system is monitored.

Table 3.3.10 Survivability

ID	WP23_REQ_SEC_11
Name	Physical protection
Description	Database environment must protect itself from physical assault. The typical objectives of physical protection requirements are to ensure that an application or centre are protected against the physical damage, destruction, theft, or replacement of hardware, software, or personnel components due to vandalism, sabotage, or terrorism.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	Every component of the database environment are run down and survivability of the system is monitored.

Table 3.3.11 Physical protection

ID	WP23_REQ_SEC_12
Name	System Maintenance
Description	Database environment should prevent authorized modifications (e.g., defect fixes, enhancements, updates) from accidentally defeating its security mechanisms.
Type	Non-Functional
Priority	Critical
Evaluation Criteria	

Table 3.3.12 System maintenance

3.4. Common Framework For Centralized Long Term Storage of Events

3.4.1. Fundamentals of data privacy

Regulations regarding data privacy are generally based upon human rights, e.g. Art.1 §1 and Art.2 §1 Grundgesetz (German constitution). This means that only data containing information about personal or factual relations of a defined or definable person are considered.

The European Union has issued the Directive 95/46/EG, which specifies the minimum standard of data privacy, and Directive 2002/58/EG, which is covering electronic communication. These Directives are in general not directly exercisable but have to be applied by the different countries in national laws.

In case of the German law the background of the company (private or public authority) determines the level of applicable terms of data privacy. For the Field Operational Tests the focus will be on the “Bundesdatenschutzgesetz” (BDSG), which has been adapted according to the European Directives. This law on data privacy is applicable to every private organisation and federal public body, except for federal public bodies within the federal state. Due to the fact that the data privacy act in Germany is very dense, the contents of this act should be able to be applied on most of the participating countries (whereas the specifications are only definitely valid for Germany). According to German law only companies collecting, process and using data by means of data processing equipment are affected by the BDSG. In this law any form of data acquisition and processing, that is not explicitly authorised by the test subject, is interdicted. In some countries, e.g. in Finland, it is mandatory to inform authorities about the collection and storing of sensitive and private data that is related to the test driver. In case this data will be transferred to a third country, an additional declaration can also be required by the authorities, like it is the case for Finland.

3.4.2. Data logging and storing

To assure that the data collection conforms to applicable law, the confirmation of the test participants for the logging of the data is needed. This confirmation has to be done in written form, if no other form, e.g. in case of web questionnaires, is necessary. This means that the participating subject has to give his explicit permission to the data collection, storing and the following evaluation of the data, preferably in the form of a contract. The driver has to be informed about the consequences of the data storage. Therefore it is required to inform the driver, which kind of data is logged and which conclusions can be derived from the data logged inside the vehicle. This information has to contain all possible combinations of internal and external data sources.

In addition to informing the driver the data has to be made anonymous. This means that all data sets have to be modified in the earliest possible state (Section 40 BDSG), so that the data cannot be traced back to the individual. To keep the test subject informed about what is done with their data, the process of making the data anonymous has to be explained to the driver, so that it is visible at which point of the data-handling procedure anonymity of the data takes effect. The term of data privacy in this context has to be considered as a relative term, because absolute anonymity cannot be guaranteed depending on the test setup, the contents of the data set and the number of test subjects involved in the tests. A simple and reversible method of making the data anonymous is the replacement of the subject’s name by a multi-digit code or a pseudonym. With this action the data privacy of the subject can be assured. However the protection of the privacy is dependent on the separation between key list, containing all sensitive and personal data, and data set, containing all data relevant for the evaluation. Therefore the key list can be stored at the single test sites/countries, while the measurement data will be forwarded to the central data base. In case this key is destroyed the data can be considered as anonymous. To ensure that the test subject stays in charge of their data up to the moment of delivering, all data can be stored on a storage medium, which can be deleted by the test subject, if this is intended by the driver.

Another issue with regard to data privacy is the purpose limitation, as mentioned in Article 6 EU Directive 95/46/EG and Section 14, 28 and 29 BDSG, in which it is said that it is permitted only to use data for the same purpose it has been collected for. For this regulation exceptions can be made if the test subject gives his consent to further use or the purpose of further use is criminal prosecution. Another exception can be made, if the purpose of the further use is for scientific research which overweighs the individual interest in data privacy and which cannot be achieved in another way (from an objective point of view, Section 14 and 28 BDSG). Because of the difficulties in estimating if the last exception is the case, it is advisable always to obtain the consent of the test subject for the further use of the data (Section 28, § 3 No.4 BDSG).

In the BDSG (Section 4, 13, 28, 29) and the EU Directive (Article 5-7) it is established, that data has to be collected openly from the test subjects. This means for the companies involved in FOTs, that the data to be acquired has to be specified in the letter of agreement (contract). In case of intimate and very private information (e.g. health and personality of the test person) data acquisition is restricted to the explicit consent of the test subject. Nevertheless the acquisition of these data will be permissible in most cases for research reasons. Generally the data acquisition should follow the principle of data economy, which means that only that amount of personal data should be collected that is necessary to answer the research questions.

To assure data privacy (defined in Article 16, 17 EU Directive and Section 9, 10 BDSG) different technical and organisational standards have to be fulfilled. These can be achieved by ensuring that third parties have no access to data access processing equipment (admission/ entry control), that only authorized personnel can operate the data processing equipment (access control) and extent of user’s access is granted only for the respective access rights and that no third party can read, copy, alter or delete data (access protection). Furthermore it has to safeguard that personal data transmitted, transferred or saved to a storage medium cannot be read, copied, altered or deleted by a third party (transmission control), that it can be determined which and by whom personal data have been entered, altered or deleted (input control) and that data acquired for different purposes can be processed separately.

3.4.3. Recording of third party video data

The acquisition of video data is highly demanding in terms of data privacy, because of the high quality video data, which allows identification of the driver. In Germany it is permitted to acquire video data in publicity accessible places (e.g. roads, Section 6b BDSG), which means that the surroundings of the vehicle can be recorded without any legal problems. It is only if video data is recorded in restricted areas (e.g. military locations) that special consent with authorised personnel is required. Regarding video data of the driver, different requirements have to be fulfilled. Video data are collected openly (camera is not hidden) and only with consent of the driver. For the logging of this data a legitimate interest has to be existing (according to Section 6 §1 No.3 BDSG) and it has to be necessary to log the data in order to achieve this legitimate interest. Nevertheless it has to be ensured that the video data is deleted as it is no longer necessary to reach the purpose of research. Should offensive sequences for the test subjects have been logged, it has to be ensured that the entire video data is deleted entirely and as soon as possible. During the logging of the data no other passengers should be recorded, if this can be avoided by technical means. Otherwise it must be ensured that the camera is well on view and that it is obvious that data is recorded. Furthermore the participating driver must inform his attending passengers that data is logged. The same applies for other drivers of the vehicle.

3.4.4. Influence of criminal law

In case of unexpected incidents, like accidents, the data might become of interest regarding criminal prosecution. For Germany this possibility is given in Section 94, Strafprozessordnung (StPO), so that the data can be barred for means of criminal prosecution. This can be avoided with regard to technical solutions (e.g. button for deletion of the data inside the vehicle). The action of deleting the data will not be considered as a criminal offense, because the suspected person has the right to avoid self-incrimination. In case the data has already been transferred to the central database the driver does not have access to the data, so that he is not able to delete his data from this point in time. Source [FestaD6.3], for further details have a look on [Roßnagel03], [Directive1995/46], [Directive2002/58], [BDSG90]

3.5. High Level Data Storage Structure and Layout

Objective data can be collected from in-vehicle data acquisition systems and subjective data from drivers. In addition, there is a need for background information regarding drivers, vehicles and data acquisition systems. The following tables describe the relevant sources of data as well as the background information required.

Data sources	Description
GPS	latitude, longitude, timestamp, speed, heading, number of satellites
Sensors	acceleration for Y axis, acceleration for X axis
Questionnaires	Questions and answers are to be defined later

Table 3.5.1 Data sources

Background information	Parameters
Driver	driver_id, logger_id, age, gender, country_id
Vehicle	driver_id, brand, type
Country	country_id, name
Logger Configuration	logger_id, configurations (such as sampling frequencies)
Logger Specifications	features and limitations

Table 3.5.2 Background information

Processed data	Parameters
Event	event_id, event_type_id, trip_id, start time, stop time
Event Type	event_type_id, description
Trip	trip_id, driver_id, start time and a stop time

Table 3.5.3 Processed data

The tables also illustrate the high level data storage structure and layout. The data will be managed in a relational database management system. Relations exist between the database tables: each vehicle has exactly one driver who has a unique driver_id. In the same way each event has exactly one event type. Each event is associated with exactly one trip. Each trip has exactly one driver. The data collected from GPS and sensors is divided in to multiple trips and identified using trip_id. Each data logger has a unique logger_id identifier which provides more information about the logger as well as managing the logger configurations. Each driver has a country_id to identify the country in which he/she is attending to the TeleFOT project.

In addition to the structure presented earlier, the following use cases are taken into account in designing the final database structure.

ID	WP23_REQ_01
Name	Business functions
Description	The general business functions and their database needs are analysed and based on results: ▪ the final data model, including entities and relationships (ER diagrams) is developed ▪ the final detailed data model, including all entities, relationships, attributes and business rules is developed
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	

Table 3.5.4 Business functions

ID	WP23_REQ_02
Name	External Data
Description	Database structure for external data (such as GIS data) is developed and implemented after which the data is inserted into the database tables.
Use case	Data Management Centre
Type	Functional
Priority	Relevant
Evaluation Criteria	The applicable database tables can be accessed and the data can be fetched.

Table 3.5.5 External data

ID	WP23_REQ_03
Name	Data Logger Configurations
Description	Data Logger configurations are saved into a database table for management purposes. Configurations include parameters such as address of the remote server and login information.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	Data Loggers are working as specified

Table 3.5.6 Data logger configurations

ID	WP23_REQ_04
Name	Data enrichment
Description	Data is enriched by using data from other sources (such as GIS). The enriched data should be saved into the database.
Use case	Data Management Centre
Type	Functional
Priority	Relevant
Evaluation Criteria	The applicable database tables can be accessed and enriched data can be fetched.

Table 3.5.7 Data enrichment

ID	WP23_REQ_05
Name	Personal information about drivers
Description	Drivers fill questionnaires which will be linked to their driver_id. The information included in the database about drivers must not contain any information that can reveal their identity.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	Questionnaires can be successfully taken.

Table 3.5.8 Personal information about drivers

ID	WP23_REQ_06
Name	Summaries
Description	Summaries of each trip, week or month containing information such as average speed or duration can be saved into the database.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	Summaries can be fetched from the database.

Table 3.5.9 Summaries

ID	WP23_REQ_07
Name	Events
Description	Events are automatically or manually classified and linked to a specific trip.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	

Table 3.5.10 Events

ID	WP23_REQ_08
Name	Results of analyses
Description	Results include the statistical models and their outcome which is based on the data saved in the database. Results can be stored as well.
Use case	Data Management Centre
Type	Functional
Priority	Critical
Evaluation Criteria	We hope to see some results.

Table 3.5.11 Results of analyses

The following figure summarizes the high level database storage structure including the relations between database tables. The database structure is presented by using UML class diagram [OMGUML].

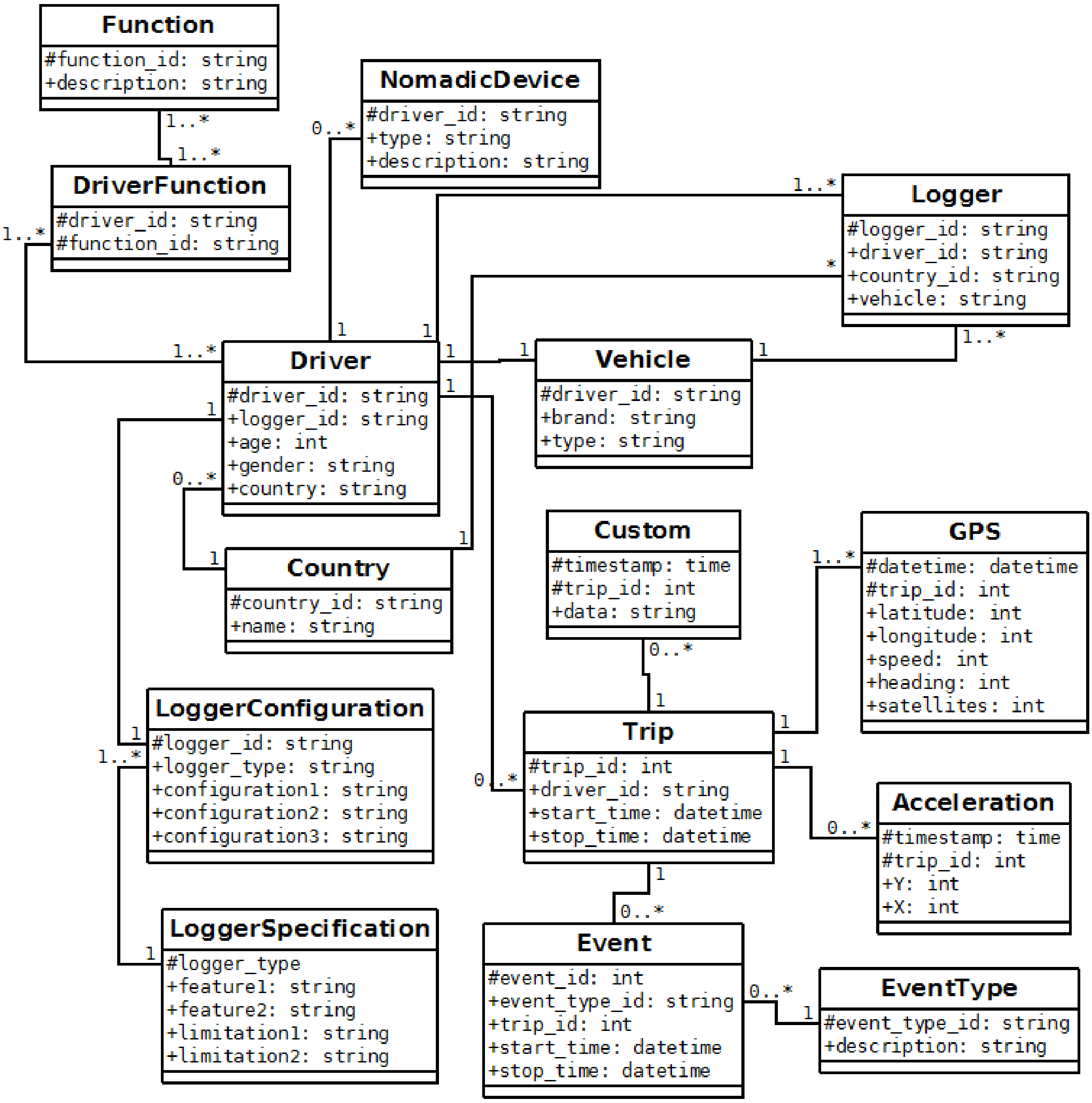


Figure 3.1: High Level Database Structure

3.6. Common Schema for Each Data Integration

The following XML schema [XMLSchema] defines the common XML schema where all the proprietary data structures should be converted.

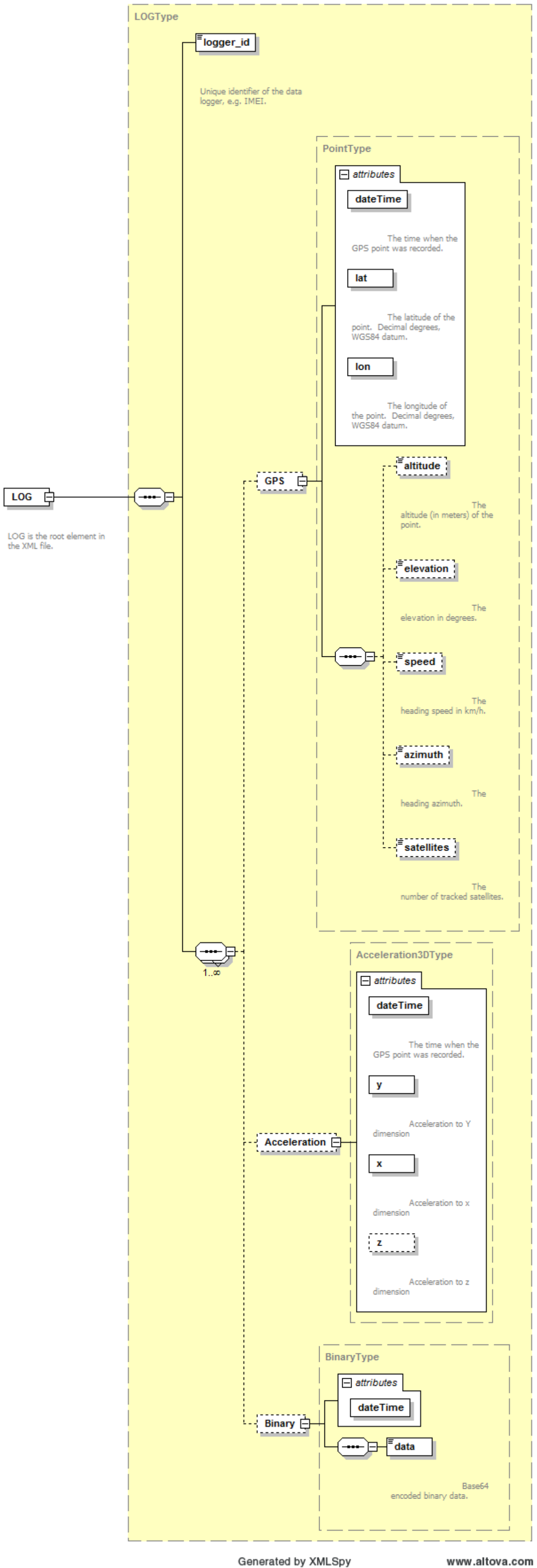

LOGType
logger_id
Unique identifier of the data logger, e.g. IMEI.
PointType
attributes
dateTime
The time when the GPS point was recorded.
lat
The latitude of the point. Decimal degrees, WGS84 datum.
lon
The longitude of the point. Decimal degrees, WGS84 datum.
altitude
The altitude (in meters) of the point.
elevation
The elevation in degrees.
speed
The heading speed in km/h.
azimuth
The heading azimuth.
satellites
The number of tracked satellites.
LOG
LOG is the root element in the XML file.
GPS
1..∞
Acceleration3DType
attributes
dateTime
The time when the GPS point was recorded.
y
Acceleration to Y dimension
x
Acceleration to x dimension
z
Acceleration to z dimension
Acceleration
BinaryType
attributes
dateTime
Binary
data
Base64 encoded binary data.
Generated by XMLSpy
www.altova.com

Figure 3.2: Common XML schema

The XML schema definition is as follows.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XMLSpy v2009 sp1 (http://www.altova.com) by Jussi Vasama (Emtele) -->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified"
attributeFormDefault="unqualified">
    <xs:element name="LOG" type="LOGType">
        <xs:annotation>
            <xs:documentation>
                LOG is the root element in the XML file.
            </xs:documentation>
        </xs:annotation>
    </xs:element>
    <xs:simpleType name="latitudeType">
        <xs:annotation>
            <xs:documentation>
                The latitude of the point.  Decimal degrees, WGS84 datum.
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:decimal">
            <xs:minInclusive value="-90.0"/>
            <xs:maxInclusive value="90.0"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="longitudeType">
        <xs:annotation>
            <xs:documentation>
                The longitude of the point.  Decimal degrees, WGS84 datum.
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:decimal">
            <xs:minInclusive value="-180.0"/>
            <xs:maxExclusive value="180.0"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="degreesType">
        <xs:annotation>
            <xs:documentation>
                Used for bearing, heading, course.  Units are decimal degrees, true (not magnetic).
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:decimal">
            <xs:minInclusive value="0.0"/>
            <xs:maxExclusive value="360.0"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="speedType">
        <xs:annotation>
            <xs:documentation>
                Speed of the object.
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:decimal">
            <xs:minInclusive value="0.0"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="accelarationType">
        <xs:annotation>
            <xs:documentation>
                Acceleration of the object to one dimension.
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:decimal">
            <xs:minInclusive value="0.0"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="PointType">
        <xs:sequence>


            <xs:element name="speed" type="speedType" minOccurs="0">
```

```xml
                <xs:annotation>
                    <xs:documentation>
                        The heading speed in km/h.
                    </xs:documentation>
                </xs:annotation>
            </xs:element>
            <xs:element name="azimuth" type="degreesType" minOccurs="0">
                <xs:annotation>
                    <xs:documentation>
                        The heading azimuth.
                    </xs:documentation>
                </xs:annotation>
            </xs:element>
            <xs:element name="satellites" type="xs:int" minOccurs="0">
                <xs:annotation>
                    <xs:documentation>
                        The number of tracked satellites.
                    </xs:documentation>
                </xs:annotation>
            </xs:element>
        </xs:sequence>
        <xs:attribute name="dateTime" type="xs:dateTime" use="required">
            <xs:annotation>
                <xs:documentation>
                    The time when the GPS point was recorded.
                </xs:documentation>
            </xs:annotation>
        </xs:attribute>
        <xs:attribute name="lat" type="latitudeType" use="required">
            <xs:annotation>
                <xs:documentation>
                    The latitude of the point.  Decimal degrees, WGS84 datum.
                </xs:documentation>
            </xs:annotation>
        </xs:attribute>
        <xs:attribute name="lon" type="longitudeType" use="required">
            <xs:annotation>
                <xs:documentation>
                    The longitude of the point.  Decimal degrees, WGS84 datum.
                </xs:documentation>
            </xs:annotation>
        </xs:attribute>
    </xs:complexType>
    <xs:complexType name="Acceleration3DType">
        <xs:attribute name="dateTime" type="xs:dateTime" use="required">
            <xs:annotation>
                <xs:documentation>
                    The time when the GPS point was recorded.
                </xs:documentation>
            </xs:annotation>
        </xs:attribute>
        <xs:attribute name="y" type="accelarationType" use="required">
            <xs:annotation>
                <xs:documentation>
                    Acceleration to Y dimension
                </xs:documentation>
            </xs:annotation>
        </xs:attribute>
        <xs:attribute name="x" type="accelarationType" use="required">
            <xs:annotation>
                <xs:documentation>
                    Acceleration to x dimension
                </xs:documentation>
            </xs:annotation>
        </xs:attribute>
        <xs:attribute name="z" type="accelarationType" use="optional">
            <xs:annotation>
                <xs:documentation>
                    Acceleration to z dimension
                </xs:documentation>
            </xs:annotation>
        </xs:attribute>
    </xs:complexType>
    <xs:complexType name="BinaryType">
        <xs:sequence>
            <xs:element name="data" type="xs:string">
                <xs:annotation>
                    <xs:documentation>
                        Base64 encoded binary data.
                    </xs:documentation>
                </xs:annotation>
            </xs:element>
        </xs:sequence>
        <xs:attribute name="dateTime" type="xs:dateTime" use="required"/>
    </xs:complexType>
    <xs:complexType name="LOGType">
        <xs:sequence>
            <xs:element name="logger_id" type="xs:string">
                <xs:annotation>
                    <xs:documentation>
                            Unique identifier of the data logger, e.g. IMEI.
                        </xs:documentation>
                </xs:annotation>
            </xs:element>
            <xs:sequence maxOccurs="unbounded">
                <xs:element name="GPS" type="PointType" minOccurs="0"/>
                <xs:element name="Acceleration" type="Acceleration3DType" minOccurs="0"/>
                <xs:element name="Binary" type="BinaryType" minOccurs="0"/>
            </xs:sequence>
        </xs:sequence>
    </xs:complexType>
</xs:schema>
```

CONCLUSIONS

Chapter 1, data specification, has been assembled to guide the core logged data specification that will be carried out by all FOTs within TELEFOT. This is only a small segment of all data to be recorded in the FOTs but constitutes the common denominator between all tests and is focussed on GPS related data records together with data to identify which FOT/nomad device functionality is related to the data. Further additional logged data which is specific to the individual FOTs, and subjective reported data formats and procedures, and context augmentation information processes remain to be defined.

Chapter 2 of this document has successfully laid out the issues that need to be considered in order to assure, as far as possible, the quality of the data from acquisition through transfer and in to storage and analysis. A check list is provided as guidance to the test sites, the database managers and the analysts. Data quality should be considered at all stages of the data flow and those of specific relevance to each user group considered appropriately. Each test site should formulate their own quality control procedure that includes consideration of each issue. This is necessary since the guidelines provided are generic to all FOTs whereas each test site within TeleFOT will have their specific requirements.

In Chapter 3 the high level process of data collection and handling is specified. The chapter describes the relevant requirements to support the creation and implementation of that process. In addition, the risks related to the data collection and handling were evaluated and solutions suggested in order to avoid the identified risks.

Ensuring the quality of the data was discussed and the use of pilot analyses was presented as a solution to detect many of the possible data handling errors throughout the data chain. The pilot analyses will also prove that the system as a whole works according to the specifications.

The fact that there is still some relevant information missing that is blocking the possibility of defining the low level database structure became evident during the process. One of the objectives of Data Working Group is to conquer this challenge. As described earlier, the data specification and architecture definition process has followed a path somewhat divergent from the essentially linear process identified in the original DoW.

The issues that still might have a significant impact on operational decisions include topics such as: pre-processing of data, analyst queries, DAS functions, external data sources, data enrichment, sampling frequencies and inferences, data formats.

Finally, the requirements for top level Data Specification were identified and a high level technical description of the database structure and layout as well as of the input data structure was introduced. Although slight modifications are possible, this deliverable described the previously mentioned matters on a level that enables the stakeholders concerned to start designing and implementing the relevant interfaces related to the Data Management Centre.

The Data Working Group activities so far (Annex 2) have identified the overall architecture for the data process from collection to central server storage and access. Analysis of this proposed architecture has identified where more detailed investigation is required of what specific data handling procedures are required to be developed that can be deployed across the multi-site TELEFOT operations in subsequent years of the project. In this respect the DWG is acting as an essential co-ordinator and manager of all data process issues throughout the project and is a vital link between SP2, 3 and 4.

REFERENCES

[BDSG90] Bundesdatenschutzgesetz. December 1990.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, pp. 155-159.

[Directive1995/46] European Union Directive on Data Protection (Directive 1995/46/EC)

[Directive2002/58] European Union Directive on Privacy and Electronic Communications (Directive 2002/58/EC)

Reports from the FESTA project, all of which are available at http://www.its.leeds.ac.uk/festa/downloads.php

- [FestaHandbook] Festa Consortium. FESTA Handbook, Version 2. August 2008. http://www.its.leeds.ac.uk/festa/, checked 27.7.2009.
- FESTA D2.4 - Data analysis and Modeling.pdf
- [FestaD6.3] Festa Consortium. D6.3: FOT requirements, legal aspects planning and development. May 2008. http://www.its.leeds.ac.uk/festa/, checked 27.7.2009.

[GermanConstitution] The Basic Law for the Federal Republic of Germany

[OMGUML] OMG Unified Modeling LanguageTM (OMG UML), Superstructure. Version 2.2. February 2009. http://www.omg.org/spec/UML/2.2/Superstructure, checked 27.7.2009.

[Roßnagel03] Roßnagel, Alexander. Handbuch Datenschutzrecht. March 2003.

[XMLSchema] W3C XML Schema Specification. http://www.w3.org/XML/Schema, checked 28.7.2009.

[XML] W3C Extensible Markup Language. http://www.w3.org/XML/, checked 28.7.2009.

ANNEX 1 DATA QUALITY CHECKLIST

The table below summarises the main checks that should be made to ensure the quality of the data. These checks apply throughout the FOT chain from the experimental design through to the data analysis and all stages in between. An indication is given as to the required frequency at which the checks should be carried out. The table should be used in conjunction with the more detailed text in section 2 to ensure that each test site builds a data quality procedure that is specific to each individual FOT.

Table A1 – Data Quality Checklist

		One off check	Each trip	Daily	Weekly	Monthly	As required
Experimental Design	Participants represent user population.	√					
	Representative in age, gender, driving experience.	√					
	Personality aspects that may effect interaction with device.	√					
	Large enough to assess functions and their impacts in each area.	√					
	Large enough for statistical confidence and sufficient power of analysis	√					
	Variables comprehensive enough to allow hypotheses to be tested.	√					
	Variables well defined.	√					
Data Acquisition	Reminder for subjects to take device to vehicle		√				
	Check that all systems operate under all normal driving conditions		√			√	
	Check that device is functioning correctly at all times					√	
	Ensure consistent calibration and verification of all systems installed					√	
	Optimise DAS start up and shut down speeds to minimise data loss	√					

	Ensure any equipment fitted does not interfere with any vehicle operations	√					
	Check amplitude of any CAN-bus data not affected during acquisition process					√	
	Ensure video data is acquired as unobtrusively as possible	√					
	Carry out pre evaluation of video image quality	√					
	Ensure number and resolution of views captured by video is sufficient to address the hypotheses	√					
Data Transfer	Arrangements made for suitable back up of data in a number of safe places	√					
	Ensure back up procedure is fully implemented prior to deleting DAS data						√
	Perform data verification (data are consistent) prior to deleting data from DAS. If inconsistent check vehicle data logger						√
	Ensure that logged data can be synchronised with subjective data	√					
	Raw data should be transferred to local server						√
Data Storage	Check the storage capacity is sufficient for all transferred data				√		
	Ensure back up data storage provision has been made.	√					
	Ensure that data is backed up and stored in more than one location before deleting vehicle data.						√
	If data is pre-processed at the storage stage, ensure the raw data is saved and backed up else where						√

Database	Ensure a user friendly data input system with the capability for validation checks	√					
	Ensure a logical hierarchy of relationships between component tables	√					
	Ensure files containing digital material (not incorporated into main database) are named systematically and can be identified by computer logic.	√					
	Ensure sufficient data management protocols in place with clear responsibilities and permissions for creating, modifying and deleting data records	√					
	Check values are entered for each filed (even if not known, or not applicable) and that data are valid (within reasonable bounds)						√
	Ensure consistent coding of variables such as missing values, not known values, not applicable values.	√					
	Issue warning if data are valid but extreme, rare or otherwise improbable					√	
	Data input system should make cross checks ensuring that database is internally consistent.					√	
Data Analysis	Conduct pilot analysis	√					
	Check that data storage system works	√					
	Check data can be easily downloaded from data storage system	√					
	Check the completeness of the data					√	
	Check the validity of the					√	

	data						
	Check that data values are reasonable and units of measure are correct					√	
	Check that dynamic data over time is appropriate for each kind of measure					√	
	Ensure that measures satisfy the requirements for specific data analysis	√					
	Check that participant use of function and function settings chosen can be logged	√					
Subjective data	Uncomplicated, clear, specific questions	√					
	Keep overall number of questions to the minimum sufficient to meet the analysis needs	√					
	Avoid hypothetical questions	√					
	Consistency of response checks built in (i.e. same question formulated in 2 different ways)	√					
	Check close ended questions are clear and unambiguous with a minimum number of possible responses whilst still making sense	√					
	Conduct pilot test	√					
	Consider interviews to minimise missing data	√					
	Choose suitable interviewers with appropriate interview skills and background knowledge of the issues	√					
	Ensure careful retrospective coding of open ended questions in order to minimise errors. This can be achieved by using a clear and consistent coding key	√					
	Check consistency of coding by comparing independent					√	

	coders work						
	Send reminders to complete questionnaires to participants						√

ANNEX 2 DATA WORKING GROUP - PROCESS AND IMPACT

Project Goals

The TELEFOT project was established to assess the impacts of functions provided by aftermarket and nomadic devices in vehicles and raise wide awareness of their traffic safety potential. It proposed to fulfil this goal by carrying out Field Operational Tests (FOTs) of these devices and applications in a number of member states. These individual national tests would be grouped into three European Test Communities : Northern, Central and Southern. To ensure that these results from these FOTs were comparable TELEFOT proposed a project structure that provided suitable co-ordination to ensure that FOT planning, data collection and analysis techniques were defined that were applicable to all national Test settings.

Data Working Group - Instigation

At the start of the project it was identified that the complex set of proposed national FOTs initially defined some 9 months before the award of contract were subsequently at a range of different stages of preparation. It was noted that while appropriate attention had been given to the overall process of FOT planning and data handling within the Description of Work (DoW) for TELEFOT in specific sub-projects, the practical timescale issues raised by the individual test sites required a more flexible and pro-active approach. This required co-ordination on, data collection procedures and specifications (SP2), the practical implementation of data recording procedures during the FOTs (SP3) and the subsequent analysis of impacts from analysed data (SP4).

Specifically the TELEFOT partners perceived a need to increase the emphasis on the co-ordination of the work within the project to ensure that the different timescales for individual FOTs could be accommodated effectively into the planning phases of the project. It was acknowledged that while individual tasks and roles had been described to cover data issues within the original project structure, there was not a specific defined overall data handling co-ordination task.

It was therefore agreed at the first TELEFOT plenary meeting (Helsinki June 2008) that a co-ordinating body should be established having appropriate representation from SP2, SP3 and SP4 members and led by the project co-ordinators VTT.

The resulting TELEFOT Data Working Group (DWG) was first convened on the 14th October 2009 at a project meeting held in Brussels. The purpose of the initial meeting was to establish the principles and scope for this ad hoc activity within the project.

The main tasks of the data working group were agreed to be :

- to agree on the architecture of the data collection system
- to agree on the specifications for the transfer of data to the collection system
- to agree on the specifications for accessing the data

It was recognised that the inputs/outputs to/from this DWG would come from a wide number of tasks within the TELEFOT structure. These included the following work-packages and sub-tasks :

WP 2.3.1	Data acquisition requirements
WP 2.3.2	Data quality
WP 2.3.3	Database structure
WP 3.2	Test tools development
WP 3.5	Large Scale FOT Execution
WP 3.6	Detailed FOT Execution
WP 3.7.2	Requirements for data processing, reporting & device management
WP 4.1	Tools for database handling and analysis
WP 4.3	Safety impact assessment
WP 4.4	Mobility impact assessment
WP 4.5	Efficiency impact assessment
WP 4.6	Environmental impact assessment
WP 4.7	Business models & User uptake assessment

It was also noted that until a more formal examination of the roles and responsibilities of partners versus tasks in an amended Description of Work was available, this essential co-ordination activity would remain in an ad hoc status, and would therefore have to rely upon a pragmatic and flexible approach by project partners.

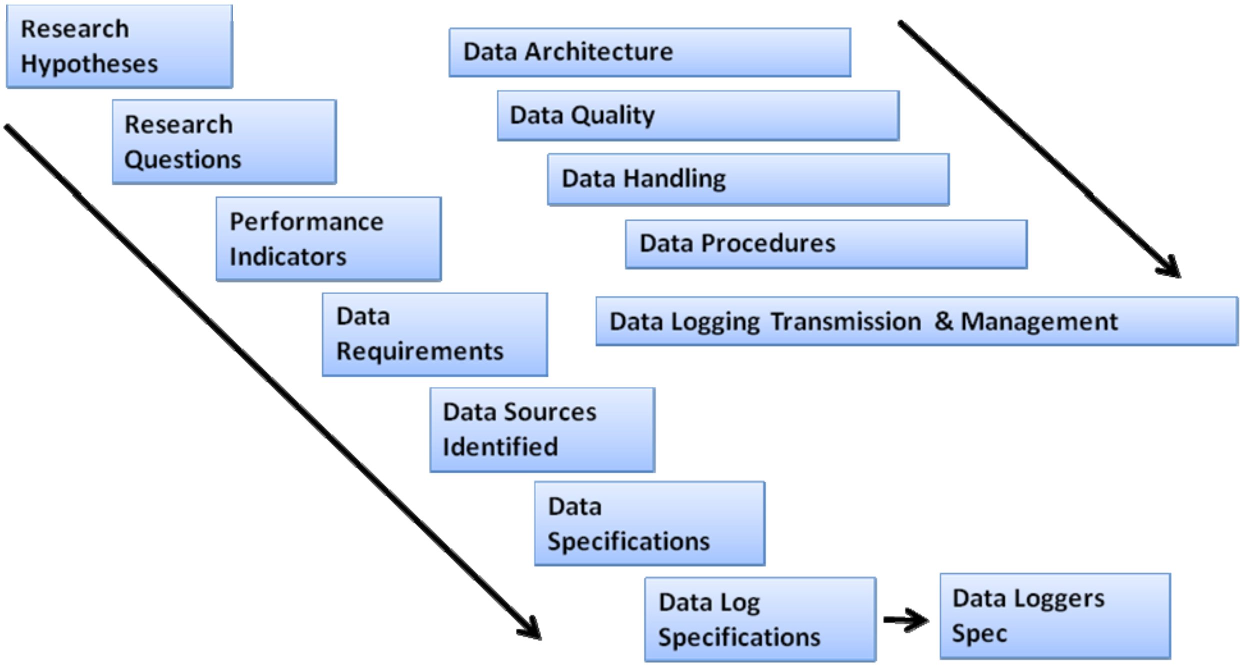
Data Working Group - Process

The original DoW for TELEFOT indicated that the project would follow an essentially linear process in identifying the necessary research, data definition and handling processes. This would initially identify the **research questions** of relevance to TELEFOT goals, and from that identify specific **research hypotheses** to be addressed.

Subsequently the relevant **performance indicators** required to assess the research questions would be derived and then **data requirements** could be specified, **data sources** identified and TELEFOT **data specifications** could be defined. This would then allow evaluation of **data logging options** and specification of **data logging equipment**.

In parallel with this a related investigation into the overall **data architecture** to be used, specification of **data quality** requirements would feed into definition of **data handling** methodologies and specific **data procedures** and **data logging, transmission and management** guidelines. This idealised model of research questions definition and data specification and handling process is illustrated in Figure 1.1 below.

Fig A2.1 : Idealised data specification and architecture definition process

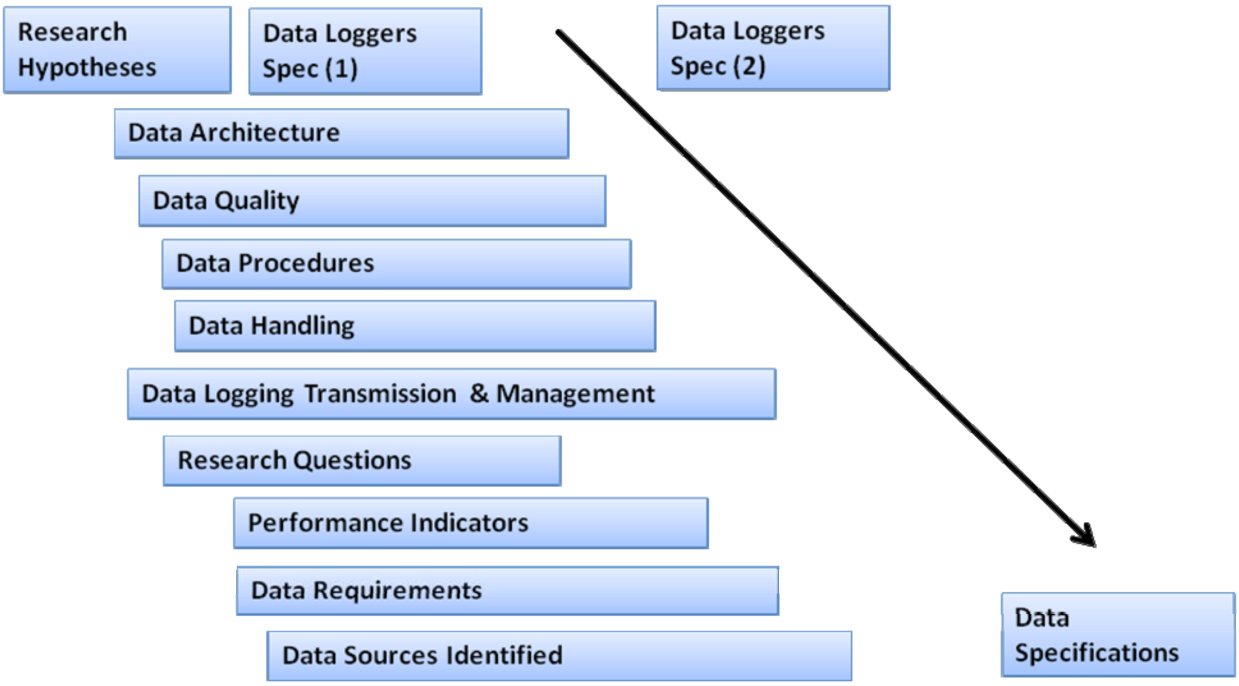


However, while logical, this idealised process in reality has to take note of potential constraints in specifying, collecting and processing data. This also relates to many practical issues that have to be considered in planning an FOT. This can include issues such as, required sample size to answer a specific research question versus logistical and financial costs of supplying that sample size that exceed available resources. Another practical constraint may be accuracy and reliability of available data logging equipment versus required data to provide a suitable research question related analysis. Factors such as these may cause an iterative loop within this process to define a suitable and achievable overall research-data-analysis chain for the FOT.

The identification of such constraints, and their impacts on the definition of research and data handling procedures, would be applicable to any individual real world FOT. However within TELEFOT this is confounded by the large scale multi-site nature of the FOTs proposed. It has already been noted that at the start of the project work, and that of the Data Working Group, not all national test sites were starting from the same point in the process chain. Some had relatively advanced positions with regard to data sources and data logger specifications. Other national test sites were at a much earlier formative stage in FOT planning and specification, and were not able to identify potential constraints until later stages in negotiation and planning with partners within the project, and other external groups such as equipment suppliers and test site hosts.

These practical considerations with regard to test site readiness and level of available information meant that the DWG had to adopt a more pragmatic approach that overlay several stages in the idealised process defined above in a hybrid approach. As the issues raised by this are complex and not easily detailed in terms of formal timescales a diagrammatic representation has been formed that perhaps reflects the overall sequence followed in discussions in the first year of the project. This is shown in Figure 1.2 below.

Fig A2.2 : TELEFOT practical data specification and architecture definition process



It can be seen from this diagram that while initial overall research hypotheses remained valid from project start initial discussions were focussed on overall strategy for data handling rather than identifying and agreeing research questions, performance indicators etc. It is acknowledged that this research question activity has been carried out in parallel to the DWG discussions. Figure 1.2 also indicates that as some test sites had a more advanced stage in planning that they had already made decisions about the data logging equipment to be used, where as other more formative FOTs had not yet reached that decision. The consequence may be that there would be two phases of data logger specification, one for the early adopters and one for later stage FOTs.

Data Working Group – Identified Topics

The initial priority items identified by the DWG are outlined in the list below.

- Data Collection Architecture
    - Overall Architecture for FOTs
    - Data Quality and Pre-processing
    - Assumptions as to analyst query of database
    - Translation of GPS data into road network data

The relevant points from the DWG discussions on these specific points that have an impact on forming operational decisions for the FOTs are outlined below. These discussion points are drawn from the initial DWG meeting (14/10/08) and subsequent telephone conferences (10/11/08, 17/12/08, 10/02/09, 04/06/09).

Data Collection Architecture - An initial task was to examine the assumptions and potential direction for the overall architecture for collecting data from individual field trials and how this data would be eventually delivered to the central TELEFOT database. It was also noted that in general terms the FOTs within TELEFOT fell into two separate categories, Detailed FOTs (D-FOT) and Large scale FOTs (L-FOT).

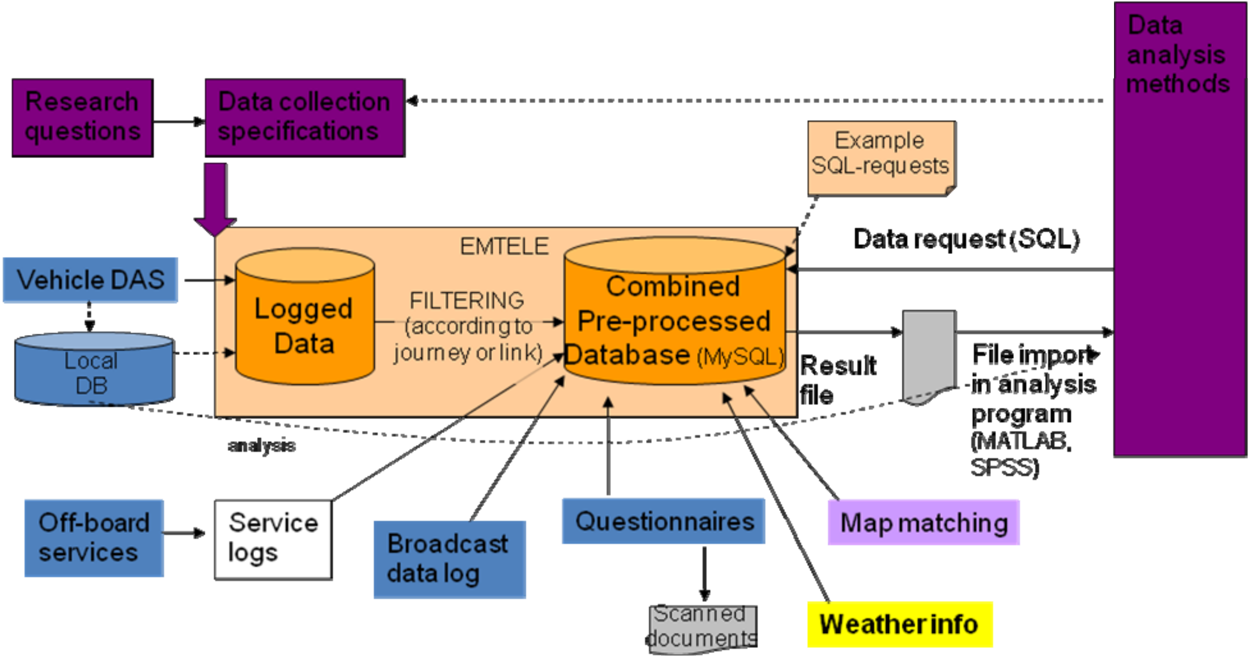
The main distinction between these two categories of FOT was their intended purpose and the level, or detail, of data being collected. The D-FOT trials were more likely focussed on the collection of a large amount of data per vehicle/nomadic device under a controlled experimental trial to support assessments of detailed impacts on the use of the device by drivers while driving. This could include a wide range of experimental parameters being recorded and include additional means of categorising the trial environment and driver responses.

This could include video recording of driver and road scene, or more complex equipment to assess driver visual behaviour, for example. These D-FOT trials may be of more relevance to detailed investigations of impacts on safety related questions and would have relatively small numbers of vehicles fitted with additional sensors and associated data capture systems. In addition these vehicles would be used to assess a relatively small number of subjects per FOT site, and trials would be carried out under more experimental control for a limited time.

In contrast the L-FOTs would generically collect less intense data from a larger number of vehicles/users under less controlled conditions over a longer period. Data logging would be more restricted in comparison to the D-FOT and require less intensive equipment installation and potentially data collection would be more automated.

It was assumed that the data collected from D-FOT and L-FOT trials would have some commonality, e.g. some common data field, but that this data set would be extended in the D-FOT scenarios. In the initial discussions on project level system architecture a simplified representation of the major elements was developed for both L-FOT and D-FOT. The schematic for the L-FOT is shown in Figure 1.3 below.

Figure A2.3 : Data Architecture for Large Scale FOT (L-FOT)

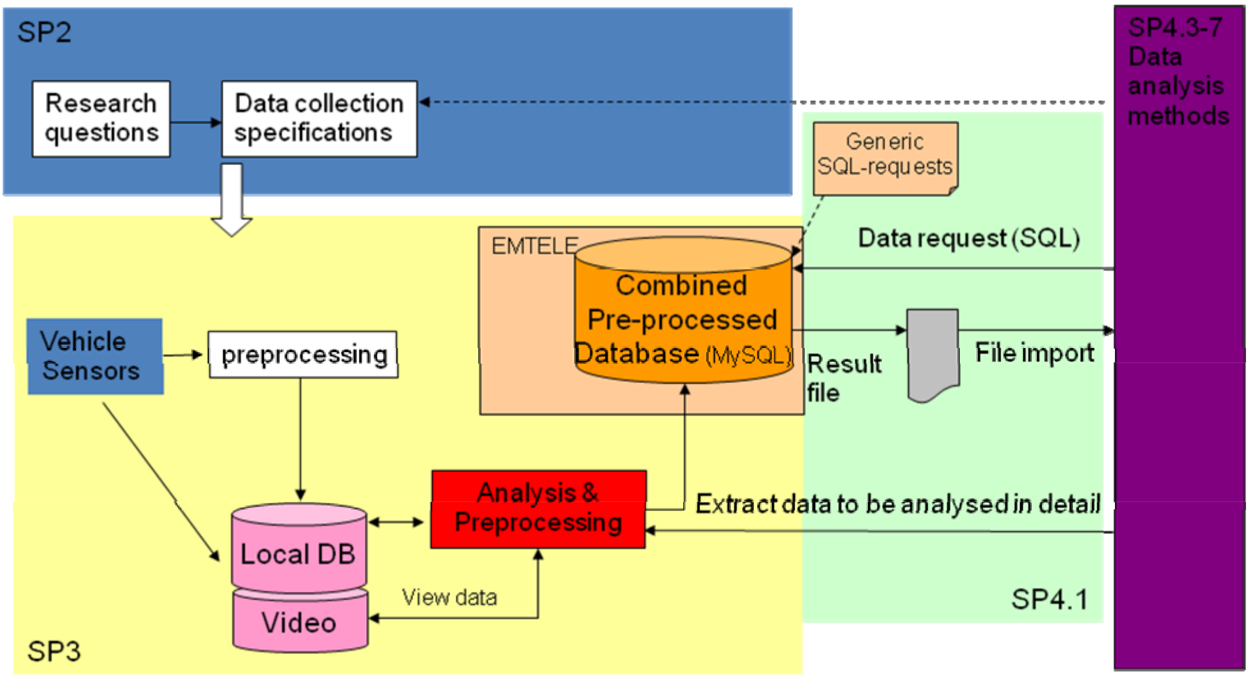


This indicates the concept that source data is collected in the L-FOT by some vehicle data acquisition system (DAS) and that post trial is transferred to a local database. This test-site based database is at some stage uploaded/transferred to a central TELEFOT database (EMTELE) as a central project-wide logged database. This source data would then be filtered (say by journey type or other pre-determined factor and combined with associated data from other sources relevant to; system performance (such as service or broadcast logs), location (via map matching), environmental conditions (via weather information sources) and subject related data (including questionnaires).

The resulting combined pre-processed database would then be available for analysts within the project to make data requests (SQL) of that database and receive an appropriate results file which could then be analysed.

For the D-FOT work it was assumed that there would be a need for more local data processing and abstraction before any data was uploaded/transferred to the central database. This is shown in Figure 1.4 below.

Figure A2.4 : Data Architecture for Detailed FOT (D-FOT) in addition to Fig A2.3



Data Pre-processing - This indicates that the assumption was that the D-FOT trials would utilise a range of additional vehicle and/or user related sensors to collect useful data in relation to the research questions to be defined. This may also include video data and would require perhaps pre-processing before being transferred to a local database where further analysis and pre-processing would be carried out before any upload to the central pre-processed database. It was noted that there may be a potential for some D-FOTs to collect large data files, particularly in relation to video capture, and these potentially may only be held on local databases. There may then be an operational consideration concerning how and why there may be a need to share such data beyond the local database/test-site management.

Subsequent discussions focussed on the D-FOT related overall project access to locally held data. It was agreed that analysis of D-FOT trial data required an understanding and knowledge of the context in which the data was collected. This included knowledge of the nomadic device, specific data related equipment, trial set-up, trial format and local operational conditions. As a result remote analysis by an analyst abstracted from this knowledge is problematic.

It was also noted that any recorded video data would require need local pre-processing to reduce raw data into a useable data set. Consequently only generated metadata would be potentially uploaded to the central TELEFOT server. A question was raised as to whether “critical situation” derived data could be held on a central ftp-server, however this would require definition as to what these “critical situations” were and the purpose for centrally held video data would need to be fully understood and its relevance to specific research questions identified

Analyst Query - Subsequent SQL data request enquiries by analysts could be made as identified in the L-FOT example and also directly from the local database. By definition this local database may not therefore have the additional data integration of enriched sources (e.g. weather or map matching) unless there was a two way link between the central server pre-processed database and the local databases.

Data Logging Functions and Devices – It was clear at the outset that there were some initial assumptions amongst some partners that the data logging devices to be employed within the project were going to be co-ordinated and focussed on the BroadBit and MetaSystem devices to be adopted by some test national sites. However it became clear as the detailed plans for the individual test sites that other data logging mechanisms would be utilised as well. Some of these would employ other commercial off the shelf (COTS) data loggers and some test sites would acquire data that is generated from the specific functions to be evaluated. In this latter case some nomadic devices act as a portal to some commercial back-office system that also can log data on GPS location, vehicle speed etcetera. It therefore became clear that a better understanding of the capabilities of all data collection methods was required and any constraints or restrictions identified that would have an impact on eventual centrally held datasets.

This could also have an impact on how the data could be transferred to the central dataset. For example, the BroadBit captured data could potentially be transferred directly to the EMTELE data server. Many other data collection mechanisms would require some translation and processing before any upload. Although it was agreed that a uniform description of the data to be made available for upload (as an SQL or XML description) was required.

Additional Data/Analysis sources – It was also noted that the project had identified the potential application and use of additional tools for assessment. The use of such additional tools would be particularly relevant to the D-FOT trials to be performed within the project. In particular this related to the ECA (TERA) and CAA tools supported by individual partners. During the course of the first year DWG discussions it was clear that the national test sites had not all formed a definitive view as to how one or both of these tools could be utilised. By June 2009 it became clear that all test sites were considering the use of CAA pending further information being supplied, but the ECA tool, as an on-line or off-line function may only be used in one test site. Towards the end of the first year of the project a proposed timescale for demonstrating the CAA and implementing ECN has been agreed to progress these related matters.

Central Database data enrichment - In discussions it appeared that there were a range of possibilities to enrich uploaded data held on the central server with other related data sources. These included the previously mentioned service logs and map matching to translate GPS co-ordinate location into road attribute data. Service log data could be available from some test sites such as TMC service logs for the Finnish national test site data which could be provided five times daily, however the position with regard to other such service data for other sites remains unclear. In the context of road attribute enrichment project partner NavTeq could provide a module to the central server that could perform the necessary map matching function and that in a wider context this should be a co-ordinated process in conjunction with the parallel project EUROFOT which would require similar inputs.

Sampling Rates & Inferences – There has been some discussions within the DWG on the data sampling rates and their impacts on ability to generate performance indicators. A general assumption appears to be that a 1Hz sampling rates for data channels recorded on vehicle is a default frequency. However it has been acknowledged that some applications may vary from this. It has also been noted that there may be a negative impact on the ability to derive some indicators if slower data rates are employed. For example if a test site is intending to collect vehicle road speed as a basic, and probably universal variable, then if slower data rates are used then the inference of jerk (second derivative of speed) would likely to be compromised unless supplementary direct measurement of acceleration is available as a supplementary variable. In this particular instance the means of calculating a derived value such as jerk would have to be standardised and used in an identical manner at each test site.

Conflict Analysis and specific data collection - The DWG also discussed the possible collection of more specific data related to conflicts although a formal definition of what constitutes a “conflict” has yet to be achieved. In this specific case some suggestions have been made of collecting data at a higher data rate (say 50Hz) if a “conflict” situation is detected triggered. This would require a better definition of what would trigger such an event and an evaluation made of the capabilities required to enact a higher sampling rate and its feasibility.

Data Formats – Overlying most of the discussion is the question as to where data should be “standardised” and appropriate agreed formats for initial data storage (local), post-processing (local and central) and final data storage (central). While discussions have attempted to begin to form a consensus on this issue this has still not been achieved.

This is also true of data acquired via other sources and means that add context to the basic data. This may also include the addition of trial subject subjective data gathered by questionnaire. The use and applicability of questionnaires, both manual and web-based, has also been discussed. If manual questionnaires are employed then they will require transcription into a format that is capable of being held and interrogated at local or central data store “levels”. This aspect relies on an agreement on what are the research questions that require such additional information to be collected, which questionnaires should be adopted across the project, and how the results can be individually or collectively analysed.

Interaction & Reliance on other TELEFOT activities - It was acknowledged at the outset that the DWG would rely on inputs from other TELEFOT SPs and WPs to focus on its co-ordination and review role. However, in the first year of the project several crucial areas of activity remain undefined and in progress. The DWG has therefore had to form discussions based upon premature, incomplete or unavailable information. This has necessarily made progress slow. In addition direct participation in the DWG discussions has relied upon flexibility from a sub-set of the partners to set aside the resource available in specific WP activities allocated to related tasks to enable the DWG to proceed. While the DWG has had positive inputs from participants it has not always had the benefit of involvement with all test-site representatives, aside from where there is an overlap of responsibilities from individuals and organisations within the collaborative partnership.

Emerging applications – It has also been clear that those national FOTs planned for later stages in the project are still in the process of formation at the end of the first year of the project. While these later FOTs may benefit from the knowledge generated from discussions and conclusions formed by the DWG it is also possible that they will also bring new considerations and constraints on how data is collected, processed, uploaded and enriched prior to final analysis. A key area is perhaps to maintain a focus on the practice and procedures to be adopted to ensure data compliance and quality with these trials in collaboration with that used in the early phase of FOTs.

Overall Commentary on DWG Year 1

From the comments made in the previous section it is apparent that the need for such a co-ordinating activity as the DWG was entirely appropriate in the context of the overall project and its goals. However it is also clear that as this was an “additional” task on the partners within an original defined DoW and responsibilities, it has only been able to proceed with the flexibility and enthusiasm of those partners who have contributed. These contributions have been made mainly by those tasked with realising a viable data capture, local processing, transmission and central server processing methodology.

A key set of inputs into the DWG at the early stages for TELEFOT come from SP2. This is particularly true regarding scene setting for the FOTs in terms of the research goals, questions and performance indicators for which the data capture and analysis process is supposed to serve and directly address. However outputs from SP2 are still arriving at the time of writing and so the DWG has still to receive representative consensus from these activities. Some DWG participants are active within SP2 but resources to support both activities in parallel timescales has proven difficult.

Another area of reliance relates to specific roles within the TELEFOT structure. The DWG has not had the involvement of a more representative group of the national test site managers directly during the first year. Ultimately those who have responsibilities for management of national test sites have a significant role in identifying how the actual FOTs and data will be handled and procedures implemented. While their interaction may be more focussed on assigned responsibilities and interaction within SP3, some mechanism should be found for ensuring their greater involvement. This will ultimately require the attention of the co-ordinators and the core group.

Given the lack of formal responsibilities and resources to support the additional task of the DWG it would seem appropriate to analyse whether a subsequent revised DoW should be proposed that **adequately resources** the essential activities and identifies those who are required to be involved as an ongoing task throughout the lifetime of SP2, SP3 and eventually SP4. Careful thought should also be given to the strategic nature of the DWG within the context of the project and whether the method of telephone conference should be the sole means of partner interaction. The outline of discussions held in Year 1 indicate the somewhat circular nature of discussions related to elements within the data process chain that have trade-off analysis, technical issues, logistical considerations, resources issues and project management consequences at national and project level which are all relevant but confounding factors. These issues would benefit from a more formal set of face-to-face meetings between partners on the DWG matters to progress discussions and decision making in a **more effective manner**.