Police Enforcement Policy and Programmes on European Roads

Contract No: 019744

Deliverable 8

Database structure for an enforcement data collection system on speeding, drink driving and restraint systems

Due date of deliverable: 30/04/2008
Actual submission date: 21/05/2008

Start date of the project: 1/3/2006
Duration: 30 months

Organisation name of lead contractor for this deliverable:

Authors: Jensen, C. (DTU)
Masson, D. (DTU)
Siren, A. (DTU)

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Dissemination level:

PU Public
PP Restricted to other programme participants (including the Commission Services)
RE Restricted to a group specified by the consortium (including the Commission Services)
CO Confidential, only for members of the consortium (including the Commission Services)
**PEPPER Project Consortium**

<table>
<thead>
<tr>
<th>No.</th>
<th>Logo</th>
<th>Institution / Description</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>![VTT]</td>
<td>VTT Technical Research Centre of Finland</td>
<td>FI</td>
</tr>
<tr>
<td>2</td>
<td>![4Sight]</td>
<td>4Sight, Ergonomics &amp; Safety</td>
<td>ISR</td>
</tr>
<tr>
<td>3</td>
<td>![IBSR-BIVV]</td>
<td>IBSR-BIVV Institut Belge Pour La Sécurité Routière</td>
<td>BE</td>
</tr>
<tr>
<td>4</td>
<td>![POLITIE]</td>
<td>KLPD/TISPOL Korps Landelijke Politiediensten /</td>
<td>NL</td>
</tr>
<tr>
<td>5</td>
<td>![bfu]</td>
<td>bfu Schweizerische Beratungsstelle für Unfallverhütung</td>
<td>CH</td>
</tr>
<tr>
<td>6</td>
<td>![CERTH]</td>
<td>CERTH, Hellenic Institute of Transport</td>
<td>GR</td>
</tr>
<tr>
<td>7</td>
<td>![bastian]</td>
<td>BASt Bundesanstalt für Strassenwesen</td>
<td>D</td>
</tr>
<tr>
<td>8</td>
<td>![CDV]</td>
<td>CDV Transport Research Centre</td>
<td>CZ</td>
</tr>
<tr>
<td>9</td>
<td>![DTU]</td>
<td>DTU Department of Transport, Technical University of Denmark</td>
<td>DK</td>
</tr>
<tr>
<td>10</td>
<td>![IBDIM]</td>
<td>IBDIM Road and Bridge Research Institute</td>
<td>PL</td>
</tr>
<tr>
<td>11</td>
<td>![INRETS]</td>
<td>INRETS Institut National de Recherche sur les Transports et leur Sécurité</td>
<td>FR</td>
</tr>
<tr>
<td>12</td>
<td>![KfV]</td>
<td>KfV Kuratorium für Verkehrssicherheit (Co-ordinator)</td>
<td>AT</td>
</tr>
<tr>
<td>13</td>
<td>![SWOV]</td>
<td>SWOV Institute for Road Safety Research</td>
<td>NL</td>
</tr>
<tr>
<td>14</td>
<td>![TØI]</td>
<td>TØI Institute of Transport Economics</td>
<td>NO</td>
</tr>
<tr>
<td>15</td>
<td>![TRL]</td>
<td>TRL Limited</td>
<td>UK</td>
</tr>
<tr>
<td>16</td>
<td>![UPM]</td>
<td>UPM Universidad Politécnica de Madrid</td>
<td>ES</td>
</tr>
<tr>
<td>17</td>
<td>![VTI]</td>
<td>VTI Swedish National Road and Transport Research Institute</td>
<td>SE</td>
</tr>
<tr>
<td>18</td>
<td>![ETSC]</td>
<td>ETSC European Transport Safety Council</td>
<td>INT</td>
</tr>
</tbody>
</table>
Abstract

Deliverable 8 is part of the work carried out in PEPPER Task 2.3. Deliverable 8 presents a conceptual model for a TLE data collection system, a model that is build upon the previous work in the WP 2. The present Deliverable is structured as follows: First, the overall approach as well as the challenges in designing a TLE data collection system are being described. Second, a preliminary data base model, used as a conceptual tool for designing the data collection system, is presented. Third, the structure for TLE data collection system itself is presented; and finally, conclusions are drawn. The present Deliverable contributes to the task of creating a common data base by providing a structure for the design and will hopefully provide the tools needed for well-informed decisions about the future TLE data base and data collection. After the PEPPER project, political will as well as consensus about the importance of establishing such a system is needed in order to finally establish a data base system for Europe wide TLE data.
Contents

ABSTRACT ..................................................................................................................................................3
EXECUTIVE SUMMARY ..........................................................................................................................7
LIST OF ABBREVIATIONS .......................................................................................................................14
1 INTRODUCTION .................................................................................................................................15
  1.1 Background ....................................................................................................................................15
2 PERSPECTIVES AND CHALLENGES .............................................................................................17
  2.1 General challenges .......................................................................................................................17
  2.2 The data chain ................................................................................................................................18
  2.3 Demands on the output side .........................................................................................................19
  2.4 Database modelling ......................................................................................................................20
  2.5 Data collection – demands and strategies ....................................................................................21
3 A PRELIMINARY DATABASE MODEL ............................................................................................24
  3.1 Introduction ....................................................................................................................................24
  3.2 Overview .......................................................................................................................................24
  3.3 Temporal dimension (Calendar) ....................................................................................................26
  3.4 Spatial dimension (Geography) ......................................................................................................27
  3.5 Speeding ........................................................................................................................................28
  3.6 Restraint .........................................................................................................................................29
  3.7 Driving Under Influence (DUI) ......................................................................................................30
  3.8 Driver ............................................................................................................................................30
  3.9 Vehicle ..........................................................................................................................................31
  3.10 Casualty .......................................................................................................................................31
  3.11 Campaign .....................................................................................................................................32
  3.12 Equipment ....................................................................................................................................33
  3.13 Legislation .....................................................................................................................................34
  3.14 Sanction .........................................................................................................................................35
  3.15 Statistics .......................................................................................................................................35
  3.16 Security .........................................................................................................................................36
  3.17 Combined facts ............................................................................................................................36
4 DATA COLLECTION SYSTEM ............................................................................................................38
4.1 Input data – availability in the short/long run ........................................................... 39
4.2 User interface ............................................................................................................. 41
4.3 System administration ............................................................................................ 46
4.4 Data security/data confidentiality ............................................................................ 49

5 CONCLUSIONS ......................................................................................................... 52

6 REFERENCES ............................................................................................................. 54
EXECUTIVE SUMMARY

1. Introduction and background

Deliverable 8 is part of the work carried out in Task 2.3. The aim of this task is to analyse and evaluate the pilot data collected in selected countries and to make a suggestion for a TLE data collection system. The pilot data was previously evaluated, and this has been reported in Working paper 40. Deliverable 8 presents a suggestion for a TLE data collection system, which is build upon the previous work.

2. Perspectives and challenges

2.1 General challenges

There are certain key challenges in creating a model for a TLE data collection system. At present, a major challenge or obstacle is the collection of TLE data. First, the data is largely unavailable and second, the usefulness of the data is often not optimal because of a number of compatibility issues. Thus great effort is needed in order to satisfy the expectations of the European Commission recommendations of 6. April 2004 as well as the demands from the scientific world. Another challenge lies in the structure of the PEPPER WP 2. From a technical standpoint, the ideal approach would be to have the total data flow from data collection, through database modelling to the definition of needs for output regarded as integrated parts of one common task. For various reasons, however, it was decided to divide this into separate tasks in WP 2 attended to by different partners, which has posed some general challenges regarding the actual task of defining the demands for a data collection system (DCS).

As there is no specific description of the demands of the output side it is necessary to base the work in this deliverable on a number of assumptions regarding these demands. This way it is possible to be precise when describing how the data is going to be handled throughout the whole system from data collection to output.

In order to develop a data collection system it is necessary to have exact knowledge of the database model and the structure in which input data is going to be stored. However, the deliverable describing the database model has its deadline after the completion of this deliverable. Therefore, it was necessary to draft a preliminary database model in this deliverable as well.

It was aimed to structure the deliverable in such a way that it can be used as important input to the demands specifications for a future real-world PEPPER Data Collection System. It identifies and describes important issues that need to be taken into consideration.

2.2 The data chain

There are several reasons why it is important to focus on the entire data chain, even though the task only is to present design considerations on the input side:
It should be possible to get output data, which combine data from different sources – this has to be taken into account when designing the data collection system.

Data on the input side is much more disaggregated than on the output side. This has implications on how data is to be collected and stored.

The database is the middle layer where conflicts between demands on input and output have to be resolved. Therefore it can not be regarded as an isolated entity.

As already mentioned, the logical way to approach this task is backwards by first defining the needs of the output side, then designing a proper database model and finally defining demands on the input side. However, as this was not possible in WP 2, the whole data chain will have to be considered to some extend here (even though this is beyond the scope of the present deliverable).

2.3 Demands on the output side

Different assumptions about the output demands have been made.

First, the demands for the output data is expected to change heavily over the lifetime of the system. There is expected to be an increased need for more and more detailed and complex data.

Second, the European Commission recommendation of 6 April 2004 on enforcement in the field of road safety (2004/345/EC) outlines which types of information the Commission expects to receive in the fields of speeding, drink-driving, and non-use of seat belts. So this is to be regarded as the main objective if this deliverable. These expectations are likely to reflect the expectations of national and regional authorities in the field of traffic safety. Also, for scientific and some administrative purposes data on a disaggregated level will be essential. At the time being the data available is probably insufficient to meet this demand but in the future these challenges are expected to be overcome.

Third, it is very important that the authorities who deliver the data can see that their data are being used and their efforts therefore worthwhile. In addition, they themselves should be able to use the output data and thereby stay motivated to deliver the data.

Fourth, some of the more aggregated data should be made available to the public in order to raise awareness of the benefits of most TLE-activities.

Finally, the way the output data is presented is important. It should be through standard tools such as tables and graphs etc. Also, it is very important that the system can handle geography. Many presentations of data from the system will be significantly easier to understand if presented as maps from a Geographic Information System (GIS). An example is shown in the text.

2.4 Database modelling

Some important aspects need to be taken into account when designing the database model. The database must be able to handle data at very different levels of aggregation and it should be
able to check the quality and completeness of the data in order to ensure sensible and valid output. In addition, it must from the outset be able to handle time and geography on a very detailed level so it is prepared for future needs for very detailed information. Finally, the database should be able to calculate a great number of supplementary variables including external statistics.

2.5 Data collection – demands and strategies

The data collection system (DCS) has to meet a set of fairly high demands, which are elaborated in more detail in chapter 2.5. Among these are e.g., flexibility, data availability, data aggregation levels, user friendliness, security and benefits for the providers of the data.

Even though all the requirements are important, and must be taken into account when developing the DCS, this can be done at different levels of ambition. In this deliverable a fairly complex system will be described.

The reason for deciding on describing a relatively complex system is that it is always easier to make a system description based on a high level of ambition, and then afterwards, during the demand specification phase, downsize the solution according to a prioritization of needs against a set of budget and timetable constraints, than it is to have to develop a more ambitious solution, based on a basic design that is too limited in its approach to the problem that has to be solved.

3. Preliminary database model

In order to describe the DCS it was necessary to define a preliminary data model on how the TLE database could be organized, solely for the purpose of being afterwards able to describe the major issues in the DCS. The preliminary model is not supposed to be a complete description of the model in all details (which will be the topic of task 2.4.) but it is believed to have taken all major issues and connections between different types of data into consideration.

The result is a model with two primary dimensions: temporal (“calendar”) and spatial (“geography”), referring to time and place which ties the different parts of the model together. In addition, there are eleven areas that are analysed. This number can be expanded to include future relevant areas. The present eleven areas are: Driver, vehicle, speeding, equipment, campaigns, legislation, restraint, sanction, statistics, DUI (driving under influence), and casualty. Finally, there are two areas called “combined_facts” which is used to combine information across areas, and “security” which obviously deals with security issues.

4. Data collection system

In this deliverable work is done at a high level of abstraction which means that only very little will be written about technical issues and choices between different technical solutions. As mentioned earlier, the work should be regarded as input to the demand specification that must be produced at a possible later stage if the TLE data system is going from the planning phase to the development phase.

This chapter will go through the main issues associated with the establishment of a DCS.
4.1 Input data – availability in the short/long run

It has been shown earlier on in the PEPPER project that data availability in the fields of speeding, driving under influence and restraint use is very limited at present. It must, on the other hand, be a common goal for all involved countries that these problems with data availability will be remedied. This is going to take time and will also pose an economic challenge as these improvements are likely to compete for resources. Thus, it is necessary to put in substantial political effort to keep a common focus on the goals of this work.

It is assumed that the data availability situation will improve over time. As this process goes on the DCS will face different challenges from having to deal with small amounts of data of questionable quality and different levels of detail to having to deal with large amounts of complex data. It also has to be able to cope with a transition from manual input to more and more automatic input from for example other databases.

4.2 User interface

A good, user-friendly design of the DCS user interface is a fundamental prerequisite to ensuring the success of the system. In order to reach this goal several issues must be taken into consideration.

The terms questionnaire and question are used in the sense that the DCS is a technical substitute to the questionnaires with a large number of questions that have been used during the PEPPER pilot data collection process. The DCS is not intended just to be a collection of predefined questionnaires, but a much more flexible data input environment.

Web-based interface: In the later browser based tools have been developed to such an extent that the benefits clearly outweighs the few drawbacks. Therefore it is the obvious choice for this type of system. Some of the advantages are that it can be controlled and maintained centrally, it is standardized so most people/institutions/organizations will have easy access regardless of type of computer and provides a possibility to take advantage of and implement new technology to make data collection more efficient.

Personalized questionnaire: As the DCS on the long run will be able to input huge amounts of data it is important that each user is only presented with the relevant questions. Thus, the system has to be able to present a selection of relevant questions in a logical and easy understandable way. Otherwise the user will be overwhelmed and unlikely to provide the information. This is primarily handled by assigning specific user rights and roles to the user.

User guidance/clearly defined questions: A large number of users will input data to the system. This poses a big risk that many interpretations of the questions will exist and the data provided thus no longer comparable. The easiest way to overcome this is to build up a framework behind the questionnaires, so that clicking on hyperlinks in the text gives the user access to extensive text explaining concepts, describing variables etc.

Data input: The system has to support the user in such a way as to minimise errors and optimise efficiency. It must also deal with having both manual and automatic input. This can
be achieved by for example allowing for copy paste actions, presentation of input from latest similar period for comparison etc.

Different types of data: The TLE data to be entered into the system comes in many forms. Some are simply numbers others are text maybe even large amounts of text. This can potentially lead to a very complex questionnaire likely to put off most users. It is therefore important to design the DCS in such a way that it deals with this diversity of data in a way that does not compromise user friendliness unnecessary. Some possible options are to present the user with multiple choice like options where appropriate such as ticking boxes stating e.g. “yes” or “no” or chose from predefined categories. Also, some of the large texts can be kept out of the general database.

Delivery management: In a system where sufficient data delivery is dependent on a great number of suppliers, it will be a very complicated task to manually check and administer the data delivery status for every supplier. Thus it would an advantage to design the system in such a way that it automatically notifies the system administrators if suppliers have not provided data at the agreed time.

Data validity/quality checks: When loading the data into the database, a great number of data checks must be performed in order to secure data consistency, data quality and data completeness. But data checking in the database can be supported and made less complex by performing a great deal of the validity checking already at the input stage by e.g. requiring numbers to be inside a given range, not differ too much from previous periods etc.

4.3 System administration

Setting up a DCS capable of collecting data on a European wide basis, where data delivery is very fragmented and the structure varies from country to country requires a well organized system administration. This is the case with respect to the administration procedures available in the DCS, and just as well the practical organizational set-up and the people who are given the task of handling the system administration. If this is not handled on a high level of quality, it is very likely that the credibility of the system will be severely reduced

Organization: One central authority on the EU-level should be responsible of running and developing the PEPPER DCS (as well as the databases and output delivery tools). This authority should also be responsible for handling user roles and rights as well as system security issues. However, it is not likely, that one authority on the EU-level will be able to keep a full overview of how data collection must be handled inside each country and which authorities are responsible for the single parts of data delivery. Therefore system administration must be handled in 2 or more “layers” where administrative rights are delegated from the top level to a more locally based authority.

It is important that the delegation of administrative rights is followed up by a system of checks and balances, where the authorities on different levels keep each other informed on the status of administrative matters.

User roles: The system must have a fairly advanced set of rules defining which questions the different authorities have to answer. These rules shall be able to handle differences in
organizational structures etc. from country to country on a single-question level. This must be
done by defining different user roles, which in combination with identification of the authority
in charge of delivering each single piece of data and identification of the single user, makes it
possible to identify which entity shall deliver a certain information and who has actually done
it (for follow up purposes and quality assurance).

The hierarchy of user roles can be defined in many ways, but a preliminary suggestion could be
the following overall rules, each given a set of user rights: Top level administrator, national
administrator, local administrator, data provider, data validator, and data viewer.

4.4 Data security/data confidentiality

General data security: Data security is a vital issue to take into consideration when designing
the DCS. Data security is generally handled through the use of the different user roles.

Data confidentiality: Even though data on a certain subject is actually available, the authority
responsible for the data is very unwilling to report these data to a common database or other
system due to confidentiality issues. It is therefore important to have a set of built in
mechanisms in the DCS which can make sure that confidential data are treated in the right way.
One solution could be to apply the following rules:

- That a certain amount of data or a certain result is permanently made unavailable for one or
  more groups of users - only public users/public users + certain groups of PEPPER-
  users/users outside specific countries/users outside specific authorities etc.

- Data is entered into the system, but is in a certain period of time for restricted use only

- Data is entered into the system, but must only be available to certain groups on specific
  levels of aggregation.

- Data is entered into the system, but shall always be aggregated before use. This could be
  the case with very detailed data from national databases, automatic enforcement equipment,
  statistical databases with data on an individual level etc.

If PEPPER DCS is to have the needed credibility among its core users, it is important that
these rules can be applied and will be handled properly.

5. Conclusions

There are several challenges related to establishing and designing a Europe wide TLE data
system including data collection, data storage, and data output. If it is regarded important by
the authorities and the scientific community to establish such a Europe wide TLE data system
which can help supporting work in the field of TLE throughout Europe, the system must be
able to handle these challenges.

If we are going to succeed in the task of building up a coherent TLE data collection system,
capable of doing the job at a fairly high level of ambition, the process of planning and
developing the system must be considered carefully. In order to secure the best possible
systems development it is important that the work on demand specifications is conducted in a
way where the three parts of the system – DCS, database and output systems – are defined in an integrated process.

The present Deliverable presents the key challenges related to designing a Europe wide TLE data collection system (and a database), and gives conceptual guidelines for establishing such a TLE data system. The present Deliverable contributes to the task of creating a common database by providing a structure for the design. When used together with the upcoming Deliverable 12 “Conceptual model for the European traffic law enforcement monitoring database”, it will hopefully provide the tools needed for well-informed decisions about the future TLE database and data collection. After the PEPPER project, political will as well as consensus about the importance of establishing such a system is needed in order to finally establish a system for Europe wide TLE data.
List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCS</td>
<td>Data Collection System</td>
</tr>
<tr>
<td>DUI</td>
<td>Driving under influence (of alcohol or drugs)</td>
</tr>
<tr>
<td>EPI</td>
<td>Enforcement Performance Indicator</td>
</tr>
<tr>
<td>TLE</td>
<td>Traffic Law Enforcement</td>
</tr>
</tbody>
</table>
1 INTRODUCTION

The present Deliverable 8 (D8) “Database structure for an enforcement data collection system on speeding, drink driving and restraint systems” is prepared by Work Package 2 (WP2) of the PEPPER project. The general objectives of the WP2 are to

W Design pilot studies on data collection to describe enforcement efficiency in EU member states
W Carry out pilot studies in six EU member states
W Set up a model database to be able to monitor the enforcement chain in the future
W Creation of a model database to monitor EU-recommendations on enforcement

Deliverable 8 is part of the work carried out in Task 2.3. The aim of this task is to analyse and evaluate the pilot data collected in selected countries and to make a suggestion for a TLE data collection system. The pilot data was previously evaluated, and this has been reported in Working paper 40. Deliverable 8 presents a conceptual model for a TLE data collection system, a model that is build upon the previous work in the WP 2.

1.1 Background

As part of Work Package 2 of the PEPPER project, various information on Traffic law enforcement (TLE) data has been gathered. In the first phase, information was gathered on a very broad range of variables based on the European Commission recommendations on enforcement in the field of road safety (EC, 2004). This information was collected in four EU Member States (Belgium, Denmark, Greece and Spain) and a detailed representation of the availability information can be found in Working Paper 5 (Akkermans, Vavakos & Orozova-Bekkevold, 2006). In short, it was found that although each country provides a very unique picture on which information is available, in general only little or no TLE information is available from the fields of speeding, drink driving, and restraint use.

The fact that data is very scarcely available posed several problems for the next tasks at hand in Work Package 2 of the PEPPER project. In particular, it made it difficult to identify possible TLE data variables that could serve as Enforcement Performance Indicators (EPI’s) based on what is known at the moment. In the process of defining the EPI’s an analysis was made taking into account the contents of the European Commission recommendations (EC, 2004), the findings on the actual availability of information along the traffic law enforcement chain in different EU member states (Akkermans, Vavakos & Orozova-Bekkevold, 2006) and information related to EPI’s that could be found in Working Paper 24 on good practices identified by Task 4.3a of the PEPPER project (Schagen et al., 2006). The resulting Working Paper 26 (Orozova-Bekkevold, Akkermans & Meng, 2006) presents the results of this analysis with the identification of a group of EPI’s that would in theory be capable of providing preliminary, yet valid, information on the efficiency of traffic law enforcement practices. The set of EPI’s that is presented in Working Paper 26 (Orozova-Bekkevold, Akkermans & Meng,
2006) was then used in the second phase where information was gathered. This time actual data were requested for the very specific set of variables that would be used as EPI’s in six member states: Austria, Belgium, Czech Republic, Denmark, Sweden and Greece. In addition, limited information from Switzerland was also received. The data collection process is described in Working paper 29 (Siren, Meng & Orozova-Bekkevold, 2007), and the collected data were presented streamlined in Working paper 33 (Vavakos & Papaioannou, 2007). In Working paper 40 (Siren, Akkermans, Jensen & Meng, 2008) the collected pilot data and the data collection process were evaluated. Also the the challenges, problems, and barriers encountered while collecting the data were analysed. Working paper 40 also presented a preliminary suggestion for a data collection system.

The present publication, Deliverable 8, continues the work from here. It builds upon the previous work mentioned above and presents a suggestion for a TLE data collection system that could serve as a valuable tool in the future if TLE data is to be collected at a European level.

The present Deliverable is structured as follows:

w First, the overall approach as well as the challenges in designing a TLE data collection system are described

w Second, a preliminary data base model, used as a conceptual tool for designing the data collection system, is presented

w Third, the structure for TLE data collection system itself is presented

w Finally, conclusions are drawn
2 PERSPECTIVES AND CHALLENGES

2.1 General challenges

From the outset, the PEPPER WP2 has been facing some central challenges. First, in the area of data availability, much of the necessary information regarding Traffic Law Enforcement (TLE) is at present unavailable, and the information at hand is not always easy to use due to a number of compatibility issues. As previous work in the PEPPER WP 2 has shown, a major effort is needed if the amount and quality of data shall meet the expectations regarding data output, as they are defined in the European Commission recommendation of 6. April 2004 on enforcement in the field of road safety (2004/345/EC), not to mention the demands from the scientific world. The problems related data availability are discussed in more detail in the previous working papers of WP 2.

Second, in the area of defining the demands for a system, capable of collection and dissemination of data, it would, from a technical standpoint, have been ideal to have the total data flow from data collection, through database modelling to the definition of needs for output had been regarded as integrated parts of one common task. For various reasons, however, it was decided to divide this into separate tasks attended to by different partners, which has posed some general challenges regarding the actual task of defining the demands for a data collection system (DCS).

The demands on the output side form also a challenge in the task. The demands are to some extent defined by the European Commission recommendation of 6. April 2004, but a more specific description of the demands would be helpful, in order to be very precise regarding how data is going to be handled throughout the whole system, from data collection to output. This is however not a part of the PEPPER project, so we have based our work on a number of overall assumptions regarding output demands.

The fact that the work on the data collection system (in the present Deliverable) is finished earlier than work on the design of the database model (Deliverable 12), has also posed certain challenges. A somewhat vital prerequisite for the development of a data collection system is the exact knowledge on the data model and the structure in which input data is going to be stored. In order to overcome this challenge, we have decided it was necessary to draft a preliminary database model in this deliverable. This conceptual model has served as a working tool to ensure that expected demands from the output side can be handled with the input data and the data handling procedures that will be part of the data collection system. It has not been our aim to fulfil the objectives of task 2.4 regarding the design of a database model, but some overlap with this task has been unavoidable.

In the present Deliverable it has been our goal to bring the statements and conclusions in chapter 3 of working paper 40 into a more well defined and precise form. By doing this we have tried to structure the deliverable in such a way that it can be used as an important input to the demand specifications for a future, real-world TLE Data Collection System (DCS).
However, it should be noted that is not part of PEPPER WP 2.3 to describe and design a fully functional DCS, including all relevant considerations regarding technical issues, the best possible design strategy, user interface design etc., but merely to identify and describe the considerations that have to be taken in order to present a future developer of the system with the best possible demand specifications.

2.2 The data chain

A central conception in our work is that it is important to focus on the entire ”data chain” from input to output of data, even when the specific task is to present design considerations on the input side only. We have the following reasons for this approach:

W The demands on the output side focus on presenting well-defined, structured data – typically in a context where data from a number of different sources is presented as a whole to give the user an overview on a given subject. This means that data must be joined together across data sources (e.g. from different countries) and across different variables, as for instance when comparing planned and actual data.

W On the input side, focus lies on how to support data delivery on a single question and single user level. In other words, the focus is much more disaggregated than on the output side – and this has clear implications on the way data must be collected and saved.

W The data model and the resulting physical database is the middle layer between input and output. It is here the conflicts between demands on the input side and the output side have to be resolved. This is done by designing a database structure that allows for the very detailed needs on the input side, while at the same time securing that variables and tables in the database are linked together so it is possible to present data in the right, flexible way on the output side.

W The bottom line of this is that development on the input side is inextricably linked to the demands for data on the output side.

W Among the issues that have to be taken into account on the input side to secure a reasonable quality of output, are: different aggregation levels, different units of measurement, incomplete data, flexibility towards changing definitions and changing needs etc.

Normally it is most reasonable to define demands on the input side by starting “backwards” – that is, first defining the output needs, then designing a proper database model, and finally defining demands on the input side. As this was not possible here (due to the scheduling of different tasks), we will start by giving a very brief description of our assumptions regarding output from the system in chapter 2.3. Then we will turn to the database modelling issues in chapter 2.4, later handled in more detail also in chapter 3. Finally, we will focus on the demands for a data collection system in chapter 2.5 and chapter 4.
2.3 Demands on the output side

Our assumptions regarding output data are as follows:

W The demand for output data from a TLE database must be assumed to change heavily during the lifetime of the system, moving towards a need for more and more detailed data and more complex types of analysis as the quality and amount of input data increases.

W European Commission recommendation of 6 April 2004 on enforcement in the field of road safety (2004/345/EC) outlines which types of information the Commission expects to receive in the fields of speeding, drink-driving, and non-use of seat belts. If the PEPPER project is able to ensure that this information can be delivered to the Commission we have met one of our main objectives. This document must therefore be regarded as a primary reference on how data is expected to be used on an aggregated level, when the system is ready. In the short run it must however be expected that the fulfilment of these demands is not possible; therefore a set of more limited demands, in the form of Enforcement Performance Indicators (EPI), have been discussed in previous working papers of WP 2.

W The information required from the Commission is also likely to fulfil most requirements that are to be expected from national and regional authorities in the field of traffic safety. But for some purposes there will be a demand for more detailed information on distribution over time, geography and so on. This means that the system design will have to be flexible in terms of handling more detail where data is available and there is a demand for the information.

W For scientific and some administrative purposes data deliveries on a disaggregated level will be essential. The amount of data available will probably not be sufficient for the delivery of disaggregated data until a later phase.

W The authorities responsible for delivering data are an important factor. It is vital that they feel their efforts are worthwhile, so they must be reassured that the system is actually used according to their expectations, and that the system can also provide knowledge in areas and with a level of detail that is useful for them.

W It must also be expected that many of the more aggregated results from PEPPER can be made available to the public, in order to support the growing public awareness of the benefits of most TLE-activities.

W Access to the output data will most likely be through a web-based user interface, where data can be consumed through the use of a set of standard tools, as tables, graphs, maps, download of detailed data sets etc.

W We would like to emphasize the importance of handling the spatial dimension (geography) in the system. Many presentations of data from the system will be significantly easier to understand if presented as maps from a Geographic Information System (GIS). A very simple example of this, based on Danish statistical data, is shown in Figure 1 below.
2.4 Database modelling

As mentioned earlier, it has been decided to sketch a preliminary database model, where we have tried to take the many problems and complications in the transformation of data, from input in the DCS to final output, into consideration.

Decisions on how to implement the final database model, based on the input from task 2.4., as well as from the work in task 2.3., will be taken at a later stage. However, these important aspects need to be taken into consideration:

W The model must be able to handle data on very different levels of aggregation. It must at the same time be possible to input aggregated data, where that is feasible, and very disaggregated data, e.g. data from speeding campaigns on a “single offence”-level

W This means that data on the same subject from different authorities might be delivered in very different forms and in different measurements. The database model must include procedures to check data quality of diverse data and transforming them to a common scale, in order to be able to make use of the data on the output side.

W There must be procedures securing that data output across countries, e.g. total statistics on the EU-level, is checked for data completeness before use. For instance it must be checked that data is available from each country in a common form, if there is a need for output as EU-totals.
Data model must from the outset be able to handle temporal and spatial dimensions on a very detailed level (as well as more aggregated measures), to be prepared for future needs for very detailed information.

Inside the database there must be procedures for calculating a great number of supplementary variables, for instance based on combinations of input data and external, statistical data.

2.5 Data collection – demands and strategies

During our work it has become increasingly clear, that a data collection system has to meet a set of fairly high demands.

We have identified a number of overall challenges, which such a system has to be able to take into consideration in a proper way. In short, they can be divided into the following bullets, but a more thorough description will follow in chapter 4:

- **Flexibility**: In line with the demands at the database level, the DCS must in the short term be able to handle very aggregated input from a very limited number of sources, in such a way that the limited amount of data can still be presented to the end user in a simple form (e.g. EPIS) that makes it possible to fulfil some of the important goals of the PEPPER project. At the same time, the DCS shall be prepared for handling much more complicated data from diverse sources in a later stage of the project. The DCS must also be able to handle very different organizational structures between the different countries. Development of the DCS is not done in one coherent process; it is more likely that the DCS will be under continuous development while at the same time being in full-scale operation.

- **Data availability**: The amount and quality of data will differ from country to country, between organizational units and between the three main areas of traffic law enforcement. The DCS must be able to handle data input on all these different levels, without making it unnecessarily complicated for the users to input data into the system.

- **Data aggregation levels**: The DCS must include procedures to handle data input on varying levels of aggregation – and be able to handle changes in aggregation levels over time. It is vital, that the system secures that output can only be produced on data that fulfils common criteria regarding aggregation levels etc. I.e. data from different countries must be enumerated into one common aggregation level on the output side, even though input data is delivered on different aggregation levels.

- **Planned data/actual data**: The system must be capable of handling the input of planned as well as actual data. Planned data covers areas as data in National Enforcement Plans, data on legislative structures etc.

- **User friendliness**: The DCS must be designed to confirm to high standards regarding user friendliness. The nature of the data covered by this project makes it unavoidable that a high degree of data complexity must be handled. In order to succeed in developing a DCS capable of handling all these complex types of data, without making it very difficult for the
users to input data into the system, user friendliness is probably the most vital of all the issues that we have to take into consideration.

W Security: Most of the TLE data are by nature sensitive data – at least when reported on a fairly disaggregated level. Handling of security issues must therefore be an integrated part of the system. Issues covered include: handling of user rights, general system security, certain types of data can be regarded confidential so output must be restricted to selected groups of users etc.

W Open data structure: The DCS structure must be open and flexible, making it possible to implement changes over time, when new types of data or new areas of Traffic Law Enforcement need to be integrated into the system. Openness must also be ensured in the form of data interfaces that makes it possible to integrate a TLE data system with other systems holding other types of data (general statistical data etc.)

W Compliance to standards: On the technical side it must be ensured, that the DCS is designed and implemented in a way that is compliant to relevant technical and usability standards. As an example, the use of the system should be open for as many users as possible through the use of most common browser types, possibly supplemented with a need for standard add-in tools (such as pdf-document readers)

W A platform for common definitions/terms: There is today a lack of common understanding regarding many of the important terms and definitions in the area of TLE-data. We find it important to build tools into the DCS, which makes it possible to reach common understanding of the central terminology and which delivers a common interface for the users, functioning as a “TLE-Wikipedia” (for further information on this, see “user interface” in chapter 4.2.).

W Incentives: Special emphasis must be given to the needs of the organizations and users responsible for delivering data to the system. If they are not convinced that they benefit from taking part in the work, data quality and data availability will be at risk.

Even though we believe that all the above-mentioned issues are important, and must be taken into account when developing the DCS, this can be done at different levels of ambition.

Different strategies can be applied when deciding for the level of ambition. First it must be taken into consideration, as has been proven during the project, that data on a sufficient level of detail and quality will most certainly not be broadly available from the day, where a data collection system is ready for use. Any reasonable strategy must therefore take into consideration, that the system can handle this situation, and that the system will have to be flexible towards future development as new demands and possibilities arise.

In the present deliverable we have decided to take a strategic position, where we aim to describe the issues that have to be taken into consideration when building an advanced data collection system capable of handling a large number of the relevant issues, which we have been able to identify during the work process. There might very well be problems that we have not taken into consideration, as they may have simply not been identified at this state of the development process – but we believe that we have covered most of the important areas.
The reason for deciding on describing a relatively complex system is actually fairly simple: It is always easier to make a system description based on a high level of ambition, and then afterwards, during the demand specification phase, downsize the solution according to a prioritization of needs against a set of budget and timetable constraints, than it is to have to develop a more ambitious solution, based on a basic design that is too limited in its approach to the problem that has to be solved.
3 A PRELIMINARY DATABASE MODEL

3.1 Introduction

As mentioned earlier, the PEPPER WP2 suffers to some extent from the division of the project into different parts conducted by different PEPPER partners. The project might have greatly benefited from a clearer definition of the WP2 goals – or in other words, what data/reports are we supposed to be able to extract from the system?

For future reference, when similar projects are planned, it might also be of great benefit to rethink the chronology of the different project phases. For example it is very difficult to describe the Data Collection System (DCS – deliverable 8) when the data model (deliverable 12) has not yet been described.

In order to describe the DCS we have therefore had to define a preliminary data model on how the TLE database could be organized, solely for the purpose of afterwards being able to describe the major issues in the DCS. The preliminary model is not supposed to be a complete description of the model in all details, including all tables, variables and relationships, but we believe to have taken all major issues and connections between different types of data into consideration.

Focus areas of the PEPPER project are defined as speeding, drink-driving, and restraint-use. However, we intend to have an open structure, where it is fairly easy to implement new elements into the data model, if necessary. We have decided that measurements regarding driving under influence of drugs will be part of the data structure from the beginning. This is in line with the recommendations from the European Commission.

3.2 Overview

Collecting TLE data is not an easy task. Among the primary challenges are:

- The substantial difference in the amount of data available and their level of detail in the different member states.
- The difference between data availability now and data availability in the future.
- The very different focus on the output side between the more aggregated reporting that authorities on the national level and on the EU-level wishes and the more detailed analysis that researchers wish to perform.

These issues have the following impact on how we conceive our data model:

- The model has to be flexible, and it must be easy to incorporate new “areas” of data.
- The model has to be able to handle issues concerning different levels of detail (data granularity).
Certain security issues have to be considered, especially when considering the future scenario where data are imported from databases on a “single offence”-level.

The above points have led us to the following overall model:

![Diagram of data model elements]

**Figure 2. Overview of data model elements**

The light grey area consists of the two primary dimensions: Temporal (Calendar) and spatial (Geography), which ties the different parts of the model together.

The green area consists of all the different areas that are analysed. The reason for splitting the model into these many different parts is a wish to facilitate the easiest possible input of data into the model. Had we instead chosen to make the tables “broader”, incorporating data from different areas into one table, this would have made the model less flexible concerning future changes. It would also make it much more difficult to input and use data, where a combination of information from different areas is necessary.

The dark grey area consists of the security and “combined_facts” areas. Whereas the security area is self-explanatory, the “combined_facts” area is used to combine information across areas. An example could be an analysis of how many speeding infringements also included a DUI (driving under influence) infringement. In this case the “combined_facts” area is used to extract data on cases that include information on both DUI and speeding.

The rest of this chapter will be used to describe the different parts of the data model in more detail.
3.3 Temporal dimension (Calendar)

<table>
<thead>
<tr>
<th>Calendar</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
</tr>
<tr>
<td>WEEK</td>
</tr>
</tbody>
</table>

A primary entity is the calendar, which defines how data is linked to specific periods of time. Data can be delivered on different aggregation levels regarding time periods, and in the fact tables on speeding, DUI and restraint use we always register data with a “valid from date” and a “valid until date”. We then need a tool to manage that data can be “translated” between different scales. For example data from one country might be delivered on a monthly basis, while another country delivers once a year – it is then the task of the Calendar table to provide the translation between these two types of information.

When very detailed data from enforcement campaigns etc. is delivered into the system, there might be a possibility of registering time of day (hh:mm:ss). This is not handled specifically in our preliminary data model, but could easily be implemented by adding a TIME-table.
3.4 Spatial dimension (Geography)

The smallest geographic entity in our model is a road piece, which is defined as a road section where attributes are unchanged. For example if speed limit on a road changes at a given point, this will lead to the creation of a new road piece.

The geographic part includes entities like:

- Each road piece is part of a road whose main information is the road name or road number.
- Each road piece has one general speed limit. Depending on national traffic legislation, different speed limits might exist depending on driver seniority, weather, or day type.
- Each road piece is part of a region. Each region is part of a country. A country can be part of one or more groups of countries grouped for different purposes, for example EU versus non-EU.
3.5 Speeding

<table>
<thead>
<tr>
<th>SPEEDING_FACT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PK ID</td>
<td></td>
</tr>
<tr>
<td>GIS_ID</td>
<td></td>
</tr>
<tr>
<td>ROAD_PIECE_ID</td>
<td></td>
</tr>
<tr>
<td>GEO_AGGR_LEVEL_ID</td>
<td></td>
</tr>
<tr>
<td>VALID_FROM_DATE</td>
<td></td>
</tr>
<tr>
<td>VALID_UNTIL_DATE</td>
<td></td>
</tr>
<tr>
<td>LEGISLATION_ID</td>
<td></td>
</tr>
<tr>
<td>EQUIPMENT_ID</td>
<td></td>
</tr>
<tr>
<td>SANCTION_ID</td>
<td></td>
</tr>
<tr>
<td>VIOLATION_ID</td>
<td></td>
</tr>
<tr>
<td>VEHICLE_ID</td>
<td></td>
</tr>
<tr>
<td>NUMBER_OF_VEHICLES</td>
<td></td>
</tr>
<tr>
<td>SPEED</td>
<td></td>
</tr>
<tr>
<td>COMBINED_FACT_ID</td>
<td></td>
</tr>
</tbody>
</table>

The speeding area consists of one table registering measurements regarding speeding:

- Number of vehicles in each registration
- Measured speed / average speed for vehicles registered
- Time/date of measurement (or the total period for which speeding measurements have been collected, e.g. a full calendar year)
- A number of id’s connecting the measurements to other relevant tables

  - Geographic id’s to define where measurements took place; the speed limits that apply at this place etc.
  - Connections to data regarding legislation, violation and sanction define how the measurement(s) shall be related to rules etc. in these areas
  - Vehicle_id and equipment_id define connections to information on the equipment used for speeding measurements and on the vehicle (type of vehicle, nationality etc.)

In this way we secure the necessary links between data on speeding and all the needed supporting information in other parts of the data model.
3.6 Restraint

As in speeding the restraint area is given analytical value through connections to the other areas.

It primarily counts the number of vehicles and the number of violations.

It also includes a classification on which type of restraint violation occurred.

3.7 Driving Under Influence (DUI)

The Driving Under Influence (DUI) area is used to register infringements concerning driving under the influence of alcohol or drugs.

The primary point of interest is the number of vehicles registered.
Different types of influences (one at a time or in combination) can be described.

But also the registered abuse levels are of interest.

As above, the connections to other areas is secured through the use of id’s.

### 3.8 Driver

The driver area consists of a number of different tables used to describe/categorize the driver with respect to:

- **Age**
- **Seniority**, number of years the driver has had her or his driving license.
- **Gender**
- **Nationality** (as there is a need for specific data on cross border infringements)
- **Type** – is the connecting link between information on infringements and the driver involved.
Category – a table used for categorizing into specific groups, such as private vs. business.

Information regarding the driver is connected to other areas of data through the combined_facts area.

### 3.9 Vehicle

The vehicle area consists of tables describing the vehicle:

- **Age**
- **Nationality** (the country in which the vehicle is registered)
- **Use** - e.g. private vs. business

Information regarding the vehicle is connected to other areas of data through the combined_facts area.

- **Vehicle specific speed limit**

### 3.10 Casualty
The casualty area can be considered as the area where we measure the effects of infringements on the three selected areas of TLE. At the same time the number and severity of casualties is an important benchmark for the effects of initiatives carried out in the different areas.

### 3.11 Campaign

The Campaign area is used to describe different types of campaigns planned and carried out as part of the TLE-process. The term “campaign” is used very broadly, e.g. covering efforts as national campaigns in newspapers, billboard campaigns in a given area, and point specific speed measurements along the roadside.

- One of the expressed wishes regarding traffic safety work has been to monitor the effects of campaigns; therefore we have tried to build a rather extensive description of the planned and actual effects of campaigns into the data model.
- Many enforcement activities and campaigns are conducted in relatively short periods and as “one-off” events. The data model must take this into consideration, to secure that this type of data can be utilized in the best possible way.
- Among the information registered is:
  - Campaign owner
  - Campaign participants
The equipment area is used to describe all equipment used in the TLE work. Points of interest are:

- Categorisation into DUI, speed, restraint equipment or perhaps a combination.
- Visibility
- What hours the equipment was used
- Number of hours used (planned/executed)
- Equipment model, including unreliability
3.13 Legislation

The legislation area includes overall information on the different geographical entities’ legal limits concerning DUI, speeding and restraint use. The data model is not complete in this area, but on a conceptual level the above illustrates the important issues on how the juridical system is organized.

One important thing is that changes in legislation and administrative rules over time will be properly handled by the system. It will be vital to secure that measurements from a given period of time are compared to the right (time-stamped) version of legislation, and that comparisons across time take these changes into consideration.
3.14 Sanction

The sanction area is an overall attempt of categorising the sanctions that a registered traffic infringement should lead to. The follow-up on infringements, how the different countries executes sanctions, is important because if there are no sanctions then speeding and other infringement will be taken lighter on. More work will have to be put into modelling this area.

3.15 Statistics
In order to make meaningful comparisons between countries, regions etc., the database will need an area that contains general statistics. This could be data concerning subjects as for example economy, vehicles or traffic. By this the TLE-data collected can be put into a general context, and data from different countries is comparable e.g. by calculating the number of casualties 10.000 inhabitants. Supplementing the database with statistical data can also give added value to the collected data. The model is fully flexible, and allows for other types of statistical data to be added.

Among the information stored in this part could be:

**Traffic**

- Average speed (km/h)
- Number of cars per hour/day/year

**Economy**

- GNP per inhabitant

**Vehicle**

- Number of cars per 1000 inhabitants
- Number of trucks per 1000 inhabitants

### 3.16 Security

This area of the data model has not been specifically modelled, but many of the issues that are relevant in this area are discussed in chapter 4.4.

### 3.17 Combined facts

<table>
<thead>
<tr>
<th>COMBINED_FACT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>ID</td>
</tr>
<tr>
<td>SPEED_ID</td>
<td>RESTRAINT_ID</td>
</tr>
<tr>
<td>DUI_ID</td>
<td>CASUALTY_ID</td>
</tr>
<tr>
<td>SANCTION_ID</td>
<td>DRIVER_ID</td>
</tr>
<tr>
<td>VEHICLE_ID</td>
<td>CAMPAIGN_ID</td>
</tr>
<tr>
<td>EQUIPMENT_ID</td>
<td>COUNTRY_NR</td>
</tr>
<tr>
<td>GEO_AGGR_LEVEL_ID</td>
<td></td>
</tr>
</tbody>
</table>
As mentioned earlier the combined_fact table is used to group measures across the different areas into specific measurements. This table is vital in order to link data together, as e.g.:

- Identifying cases where both DUI and speeding were involved
- Connecting information regarding the driver with information on the vehicle etc.
4 DATA COLLECTION SYSTEM

Our work regarding the design of a Data Collection System is mainly carried out on a conceptual or logical level. Therefore we have focused on a presentation of the structural issues that have to be taken into account in order to design a DCS, which is likely to be able to fulfil the ambitions presented in chapter 2.1.

Working on a fairly high level of abstraction means that only very little will be written about technical issues and choices between different technical solutions. This is especially so because of the number of phases that the project will go through, each demanding its own technical solution. We regard our work as an input to the demand specification that must be produced at a possible later stage, where the database system is going from the planning phase to the development phase. Afterwards it will be the job of the preferred bidder to outline which technical solutions in practice are best suited for handling the demands.

In addition, solving problems regarding data availability and data quality is beyond the scope of the data collection system task. Nevertheless it is of great importance to the development of the DCS to take these issues into consideration. In our work we have tried to incorporate these issues into the system framework, but we are quite sure that with our limited knowledge of all the different problems that can and will occur, there are problems that we haven’t been able to foresee.

Our work is based on a set of assumptions on how to best solve the task of developing a DCS, which will be accepted by the users and will be able to handle data in such a way that demands on the output side can be met. These assumptions might be subject to change later on in the development process, as demands regarding database structure and output are clarified.

At present we have tried to state the assumptions clearly in the text, so the reader is able to judge whether they are realistic or not. Our choice of assumptions lead to the conclusion, that only a rather ambitious DCS will be able to fulfil the demands, as stated in chapter 2.

This chapter will present the main issues associated with the establishment of a DCS:

- In chapter 4.1 “Input data – availability in short/long run” we will go through the major phases that the availability of data is going to go through. The data availability will have great impact on the DCS, which is why the DCS is not going to be “one system”, but more a system that is going to evolve greatly over time.

- In chapter 4.2 “User interface” we present the major issues concerning the user interface.

- In chapter 4.3 “System administration” we present the major issues concerning the organisation of the TLE database and the DCS and their user roles.

- In chapter 4.4 “Data security/data confidentiality” we bring up the questions concerning data security and confidentiality. An important issue if the TLE database is going to be the import of the data on a more detailed level in the future.
4.1 Input data – availability in the short/long run

It has been shown earlier on in the PEPPER WP 2, that data availability in the fields of speeding, driving under influence and restraint use is at present very limited. Even among the countries with the strongest tradition for collection of traffic safety data there are major problems with delivering necessary data for even a limited overview on the status. This is described in more detail in the previous working papers of WP 2.

In the short run, the available data will by no means be able to fulfil the demands of the European Commission recommendations (2004/345/EC), not to mention the demands for more detailed data from scientists in the traffic safety field. Therefore a set of more limited EPIs has been proposed in the previous working papers of WP 2.

It must, on the other hand, be a common goal for all involved countries that these problems with data availability will be remedied over time, but it will most likely take a number of years before each country is able to deliver data that to a reasonable degree satisfies the demands from the political/administrative systems as well as from the scientists.

There are also serious economic considerations to take into account, as the collection of data can be a heavy burden on the involved countries and organizations, and can be in direct competition with the use of resources for other purposes that might be regarded as more important for the single authority. As demands for detailed data rises, the traditional, manual collection of data must be replaced by more automatic methods, leading to a need for IT-development, which can be expensive as well as complicated.

Therefore the single country/organization might not always have the necessary incentive to focus on improved data collection, even though benefits from a more consistent knowledge of best practises in the field of TLE across the EU might be enormous. To ensure a proper development process regarding the issues that the PEPPER project is aiming at, it is necessary to combine the technical and conceptual development process with a serious political effort to keep a common focus of the goals of this work.

However, this must in our view not lead to the conclusion, that a very simple data collection system is sufficient to cover the needs in the short/medium perspective, so the development of a more advanced data collection system could be postponed to some time in the future.

It is our belief, that if the development in this field shall continue in the right direction, it is important that we can deliver useful output from the outset, and that the DCS has a structural design that allows for the implementation of new data and new data sources over time.

We have therefore concluded that the development of a reasonable data collection system must be based on a probable development scenario regarding the changes in data availability over time. We believe the following 5 phases to be a realistic assumption on the future development, even though this is a stylized description of the development, and in real life the different phases will undoubtedly overlap and some of them may never occur:
Phase 1: Limited amount of data from a subset of countries. Data delivery on country level and on a yearly basis. Aggregated data manually input into a questionnaire-like user interface.

Main issues:
- Delivery of useful output difficult due to limited amount of data
- Data delivery very diversified (different countries deliver data in different fields)
- Data quality checking complicated because of lack of historic data

Phase 2: Data from more countries on the basic level. The more advanced countries deliver data on more and more areas, but still in an aggregated form through the questionnaire interface.

Main issues:
- As in phase 1

Phase 3: General development towards more data on more areas. The most advanced countries begin to deliver data in slightly more detail – shorter periods of time, geographic information, results from actual campaigns, and detailed data will perhaps only be delivered from some areas of the country, depending on administrative setup etc. Data delivery still in questionnaire format. The gap between advanced countries and countries with only basic data delivery will be wider.

Main issues:
- Combining data on different levels of detail
- Different level of detail in different regions of the same country
- Handling of users gets more complicated as users on different levels are involved

Phase 4: Detailed data from a few local/national databases is delivered to the DCS through a well-defined database connection, but most of the data deliveries are still on a more aggregated level through the questionnaire interface.

Main issues:
- Handling detailed data from databases – securing that formats and contents are correct
- Combining data from questionnaires and databases
- Large amounts of data

Phase 5: TLE DCS is increasingly changing into a “data hub”, where data from a multitude of local/national databases are converted and quality checked to ensure compliance to agreed standards. The questionnaire interface will still be needed for input of figures regarding the goals that shall be achieved etc.

Main issues:
- Very large amount of data / high degree of complexity
- Administration of the system complicated (and vital)
Figure 3. TLE data system development scenario

The term “questionnaire interface” does not imply that the DCS will be based on the layout of the current, paper-based questionnaires, but rather that the questions from these questionnaires will be presented to the user in a form, where he/she can manually input the necessary data to answer the question.

It is assumed, that data deliveries with a high degree of detail – a typical example could be where single observations from an automated speed enforcement equipment is registered – are only input to the TLE DCS through a connection to another database, with the specific task of collecting and handling this type of data. It is not likely that this type of information will be manually input into the DCS, as this would be a very time consuming task with a great risk of problems with data quality. The exact borderline between manual and automated input must be defined more precisely later on.

4.2 User interface

A good, user-friendly design of the DCS user interface is a fundamental prerequisite to ensuring the success of such TLE data system. In the following we have listed the major issues that must be taken into consideration, when designing the user interface.

We are using the terms “questionnaire” and “question” in the sense that the DCS is a technical substitute to the questionnaires with a large number of questions that have been used during the PEPPER WP 2 pilot data collection process. As can be seen in the following, we do not intend
the DCS to just be a collection of predefined questionnaires, but a much more flexible data input environment.

**Web-based interface:** Just a few years ago it could be argued, that the use of an internet browser as tool for giving users access to a business critical system, was not necessarily the right decision, as the technology had a number of drawbacks when it comes to designing very user friendly, very efficient and powerful data handling systems. At present, however, this discussion is no longer relevant – browser based tools have been developed to such an extent that the benefits clearly outweights the few drawbacks.

- Easy implementation (updates of the system can be handled centrally, without installing new versions on each PC using the system)
- System can be used on any standard-PC with a reasonable Internet connection
- Security/user rights can be handled centrally
- System can be developed using market-standard tools from a number of suppliers, customized to our needs using standard programming tools and interfaces
- The user interface can be developed in different language versions if preferred. This can be done by using a number of “parallel” language templates for each part of the DCS
- The technical solution shall balance two contradicting considerations: Utilization of the possibilities of state of the art web-technology vs. making sure that users with older computers, slow internet connections and older software (older browser versions) can still use the DCS
- Access can be made independent of browser type (Explorer, Firefox, Opera etc.) and operating system (MAC, Windows, Linux etc.), and must not demand download of supporting software, unless this can be regarded as de facto standard software (Adobe Reader etc.)
- In the design of the system, the best possible speed and performance must be a goal. The performance of the Internet connections is difficult to control, but in the design of the user interface and the database system, as well as in the technical setup of server environment, high performance must be aimed at.
- The DCS-user interface must conform to industry standards regarding the set-up of menus, the use of keyboard shortcuts, general usability etc.
- It must be considered, if a lightweight version of the DCS shall be available for smaller, web-enabled devices such as mobile phones or PDA’s. In this way registration of certain types of data could be done “on the spot” during control activities

**Personalized questionnaire:** To make data input easy to handle, the system must be able to present exactly the relevant parts of the questionnaires to each single user. This is primarily handled by assigning specific user rights and roles to the user (see chapter 4.3. “System administration”).

- The system must at the same time be able to present the right questions to the user, and still secure that they are presented in an understandable context.
This means that questions are usually presented to the user in logical “blocks”, to make sure that the user understands the role of the question in a wider context. If responsibility for questions in a logical block is actually divided between different users/authorities, this might be solved by presenting the whole set of questions to the user, with some of the input-fields as “read-only” and others as “read/write”.

It will be important to ensure, that questions are presented in a logical sequence, both inside the logical blocks and between blocks.

**User guidance/clearly defined questions:** In the pilot questionnaires a great deal of effort was put into the phrasing of the questions, in order to secure common understanding of the questions between the users and input of data based on this common understanding of the question. This was clearly necessary, but had the disadvantage of making the questionnaires very complicated. In the DCS it will be possible to create an environment surrounding the basic questionnaire, which will be able to support a high level of common understanding and interpretation of the meaning of each question and each concept in the questionnaire. We have identified the following issues and solutions in this area:

- Questions must be short and at the same time precise. This can only be achieved if users have a common “language”/terminology, so misinterpretations become rare.

- We believe that the easiest way to handle this is to build up a framework behind the questionnaires, so that clicking on hyperlinks in the text gives the user access to extensive text explaining concepts, describing variables etc.

- This framework could be organized as a sort of “Wikipedia”-clone – an encyclopaedia which is maintained and kept up to date by the users themselves, with certain super users responsible for quality assurance etc.

  - In this “TLE-Wikipedia” all relevant concepts, legislative rules etc. are described and explained clearly, to facilitate a common understanding among users.

  - Description includes variable definitions, e.g. what are the “legal” interpretations of a given variable, when data is to be delivered – can data be delivered as average/day, sum/month, sum/year etc., and how does the DCS convert the input to a common denominator?

  - Description includes national rules etc. in order to avoid complicated questions in the questionnaire, where national rules must be described in detail (see later under “different types of data”).

  - “TLE-Wikipedia” shall also describe changes over time, as new legislation on a given subject is implemented.

  - Parts of the description can be links to external (reliable/official) sites, where a given subject is treated.

  - Building such a “TLE-Wikipedia” is not a simple task, but it will make it possible to use the TLE DCS not only as a platform for input of data, but also as the reference source on concepts and rules regarding TLE across Europe.
Data input: In the situation where the user inputs data manually in a questionnaire, he/she will be presented with one or more input fields, where text, numbers etc. can be entered, depending on the question

- If numbers are to be input, the system should present input from the latest similar period(s), to give the user the possibility to check if new results are in line with historic data. In this way one can avoid simple errors of reporting data in different format from year to year, as the data from last period is assumed to have been checked for consistency etc.

- It could be useful to give the user the possibility to copy data from a historic period to the actual period – this will be especially relevant when updating planned/budgeted figures etc. Checking routines shall give system administrators a warning message if a user copies data in an unacceptable manner.

- As a change towards automatic delivery of data from local/national databases is expected over time, the user shall have the possibility to override the manual input. This will transfer the user to another page, where details regarding the set up of a database connection can be defined. Setting of database connections can also be handled by an administrator (national or at the system level). Settings will include: Definition of the physical connection – database type, connection type, addresses, data aggregation level, update frequency etc.

- Once a database connection is defined, the user will not be prompted to deliver input in this field. It might be shown in the user interface that a certain question/block of questions is accounted for by delivery from a database, giving some overall information on the database name, purpose etc. The user has the possibility to change the database connection settings, and the system must track deliveries from the database to ensure that the expected information in the expected format is delivered.

Different types of data: The demand for data in the field of TLE is very complex, which also influences the ability of the DCS to handle different types of data in diverse combinations.

- Planned vs. actual data: In the same system we have to be able to handle planned data and actual data. Planned data includes subjects as defined in National Enforcement Plans, descriptions of legal and administrative procedures etc. Actual data covers a very wide range of different subjects in the three areas of TLE covered by PEPPER – measurements from actual counts/enforcement campaigns, registrations of sanctions etc. The different types of data are described in earlier working papers and in the data base model presented in chapter 3 in this deliverable.

- Much of the data is numeric, and is fairly easy to handle. Numeric data can be handled according to well-defined rules and diverse checking procedures can secure the quality of the input. This does not imply that it is a simple task to ensure data consistency over time and between inputs from different countries/organizations.

- In the pilot questionnaires many questions must be answered by filling in a lot of text. In our view there is a critical trade-off between the inevitable need for text descriptions, e.g. descriptions of complicated, national legislative rules, and the demand for
connecting data between countries etc. There is no simple solution to this, but the following possibilities should be considered

- If possible, free text input should be combined with a set of checkboxes/multiple choice boxes, where the text input is put into an overall context. A simple example could be that legislation on speeding on motorways in the different countries can be divided into 3-4 main categories with common characteristics. Inside each category there can be differences between countries, and this is described in the relevant text field. Comparisons between countries can then be handled via the categories, with supplementary details in the text input.

- A lot of text input might be kept out of the general database if the DCS user environment is designed as described above, with extended use of a “TLE-Wikipedia” where concepts etc. are defined in detail. This could incorporate descriptions of national legislation, administrative rules and other relevant issues (or hyperlinks to official sites, where these issues are described). There will still be a need for handling important changes in legislation etc., as this could influence the use of data on the output-side, but we see a possibility of reducing the questionnaire complexity in this way.

- Some questions will probably be “classic” questionnaire questions, where the user must choose between different pre-selected values on a scale by checking boxes – or simple choices between “yes” and “no”.

- In some cases answers must be picked from a predefined list of alternatives – countries, road types and many more. The user can then select the relevant value in a so-called “drop down-menu”. This presentation of data will be very helpful to the users, but demands that data can be extracted from tables in the database.

**Delivery management:** In a system where sufficient data delivery is dependent on a great number of suppliers, it will be a very complicated task to manually check and administer the data delivery status for every supplier. We suggest that delivery status is managed on two different levels:

- System administrators on system level and national level are presented with regular status reports from the system, stating which suppliers have not met their scheduled delivery (all or part of the required data), based on information from the suppliers on how often and when they expect to input data to the system. Status reports can be delivered based on different levels of seriousness, so follow up from the administrators can focus on the most serious problems – data from one country is missing before a certain report can be presented etc.

- Status messages are sent to the suppliers, by e-mail and/or as messages at logon to the DCS, requiring a reply.

**Data validity/quality checks:** When loading the data into the database, a great number of data checks must be performed in order to secure data consistency, data quality and data completeness. But data checking in the database can be supported and made less complex
by performing a great deal of the validity checking already at the input stage. Among the checking procedures that might be reasonable to implement in the DCS, could be:

- All relevant fields shall be filled out.
- Numbers shall be inside a given range.
- Checkboxes must be checked in a correct way, and corresponding text fields must be filled out.
- Numbers must not differ more than a certain percentage from input in the previous period.
- Values must be in line with the chosen input format, in cases where more formats (e.g. average/day vs. sum/year) are allowed.
- Checks must protect against conflicting input of data – as when two users erroneously have been granted rights to submit data to the same question.
- Checks can be “soft”, meaning that a warning is displayed, but the user can override this warning, or “hard”, meaning that the user will have to change the input before being able to continue.

4.3 System administration

Setting up a DCS capable of collecting European wide data, with very fragmented data delivery and variations in the structure from country to country, requires a well organized system administration. This is the case with respect to the administration procedures available in the DCS, and just as well the practical organizational set-up and the people who are given the task of handling the system administration. If this is not handled on a high level of quality, it is very likely that the credibility of the system will be severely reduced, which in the end will lead to a lack of confidence from the people responsible for data input and consequently to serious problems with data quality.

System administration procedures are therefore a very important task in the development and implementation of the DCS. To define these procedures in detail must be done at a later stage of the project, but as it is so important for the functionality of the DCS, we have made the following preliminary assumptions.

Organization:

W One central authority on the EU-level must be responsible of running and developing the TLE DCS (as well as the databases and output delivery tools). Whether the authority is part of the EU-administration or an independent authority appointed to take this role, can be decided at a later stage. Parts of the job, e.g. technical operation of the systems, can be subcontracted to external, professional partners.

W The central authority should also be responsible for the handling of user roles and user rights on the general level – this task can to some extent be delegated to a national authority (see below).
The same authority should be responsible of handling system security issues, including a set of confidentiality rules, securing that critical data will only be used by persons or organizations authorized to do so.

It is not likely, that one authority on the EU-level will be able to keep a full overview of how data collection must be handled inside each country and which authorities responsible for the single parts of data delivery. Therefore system administration must be handled in 2 or more “layers”, where administrative rights are delegated from the top level to a more locally based authority.

We find it reasonable to have a national, administrative authority which has the ability to monitor the situation regarding data delivery inside the country. This authority defines in detail, which users at the different institutions that are responsible for data delivery. It is also responsible for keeping track of the national data delivery status.

In more advanced phases of the project, with many data suppliers and large amounts of data from different sources, including external databases, it might even be necessary to delegate some administration to local administrators, where a “hands on”-knowledge of data delivery from external systems is needed.

It is important, that the delegation of administrative rights is followed up by a system of checks and balances, where the authorities on different levels keep each other informed on the status of administrative matters.

We assume that all data delivery is handled through the DCS user interface. In early phases of the project, there might be a few local suppliers of data who wish to deliver through paper-based questionnaires. In such a case, the national administrator will be responsible for the input of data into the DCS.
Figure 4. TLE system's administration/data delivery

User roles:

- The system must have a fairly advanced set of rules defining which questions the different authorities have to answer.

- These rules shall be able to handle differences in organizational structures etc. from country to country on a single-question level. The same set of rules will also define which questions are used in individual countries in a given period. There might even be situations, where a question must be answered by two authorities in cooperation, so the possibility of shared responsibility for a single question must be considered.

- This must be done by defining different user roles, which in combination with identification of the authority in charge of delivering each single piece of data and identification of the single user, makes it possible to identify which entity shall deliver a certain information and who has actually done it (for follow up purposes and quality assurance).

- Roles must not be set up around a single, personal user, as changes in staff will very likely lead to problems identifying the responsible persons. Instead they shall be connected to a combination of authority and one of a number of standard user roles.
This system can be designed so flexible that it can accommodate the different administrative and legislative structures from country to country. At the same time it must be handling the different types of expertise that is needed in the work process – systems developer, systems technician, database administrator, data expert, responsible for quality assurance etc.

The hierarchy of user roles can be defined in many ways, but a preliminary suggestion could be the following overall rules, each given a set of user rights:

- **Top level administrator:** Is situated at the central authority and has in principle full rights to all parts of the system, including the right to assign user roles to other users and define/change user roles. Might in practice be divided between roles for technical experts and roles for data experts.

- **National administrator:** Is situated at the authority with responsibility for the national part of the entire system. Is assigned full rights to administer any part of the national system, including user roles etc. (can be overruled by top level administrator). Might in practice be divided between roles for technical experts and roles for data experts.

- **Local administrator:** Is situated at an authority with responsibility for given parts of the national system. Has full rights inside the authority, including user roles etc. (can be overruled by higher level administrators). Could be divided between roles for technical experts and roles for data experts.

- **Data provider:** Situated at an authority with responsibility for given parts of the national system. Is responsible for data delivery of a defined set of questions – either through the questionnaire or by connection to local databases.

- **Data validator:** Can be situated at different levels of the hierarchy. Has no (or limited) rights to input or change data, but is responsible for quality assurance and reporting back in the system, when problems are identified.

- **Data viewer:** Can be situated at different levels of the hierarchy. Has no rights to input or change data, but can see data input from others (in total or for a limited number of questions). This role can be used as described earlier, where a DCS user sees a whole “block” of questions, but is only responsible for supplying data for some of the questions.

### 4.4 Data security/data confidentiality

Data security is a vital issue to take into consideration when designing the DCS. This applies to the handling of the data inside the DCS and the database, but equally important is the handling of data with respect to their subsequent use by different groups of users, as parts of the input data will most likely be regarded as confidential information.

**General data security:**

Data security is generally handled through the use of the different user roles – no user can see, enter or change data without the user rights connected to one of the user roles.
On a more technical level, the systems and the transfer of data across the Internet must be designed to protect against unauthorized access to the data in any form. There will be a need for data transfer using secure lines and data must be encrypted according to best practise.

**Data confidentiality:**

Even though data on a given subject is actually available, the authority responsible for the data is very unwilling to report these data to a common database or other system due to confidentiality issues. This might also very likely be the case regarding TLE-data, for instance if an authority is worried that the publication of detailed data could be hampering the results of their work. We therefore find it important to have a set of built in mechanisms in the DCS, which can make sure that confidential data are treated in the right way.

Confidentiality is assumed to be important on different levels:

- It is considered critical, that a certain amount of data or a certain result is permanently made unavailable for one or more groups of users - only public users/public users + certain groups of users/users outside specific countries/users outside specific authorities etc.
- Data is entered into the system, but is in a certain period of time for restricted use only
- Data is entered into the system, but must only be available to certain groups on specific levels of aggregation – e.g. detailed data on a subject is only available for scientific purposes at selected institutions
- Data is entered into the system, but shall always be aggregated before use. This could be the case with very detailed data from national databases, automatic enforcement equipment, statistical databases with data on an individual level etc.

If TLE DCS is to have the needed credibility among its core users, it is important that these rules can be applied and will be handled properly. Definition of the rules must be handled by:

- Top level administrator or national administrator can set these rules in the database on a permanent/semi-permanent basis
- Local data provider can activate a parameter in the DCS user interface, setting confidentiality rules for a specific input of data. This parameter must be set for each new question, and is only valid for exactly this input (and not for the same type of input at the next period of time)
- Setting of this parameter by the local data provider must be followed up by a checking procedure from the top level administrator, in order to minimise unnecessary uses of the confidentiality tool, as it should be a goal in itself to assure the highest possible data availability.

The access to such advanced settings of confidentiality could to some extent prevent situations, where an authority refuses to deliver data to the system due to national or local
considerations regarding sensitive information, and in this way assure that data input to the TLE DCS is as complete as possible.

But a technical solution in itself is probably not sufficient; this is also a question of trust in the system, national tradition etc – matters that have to be considered in the process of disseminating the use of the TLE DCS.
5 CONCLUSIONS

Developing a functioning TLE-data system is not an easy task. In this Deliverable 8 we have tried to describe the many challenges that have to be overcome just to design the data collection system, which is only a third of the complete system from data collection through database to data output.

Some of the key challenges are:

- Data delivery from many authorities in many countries, with different levels of detail, different scaling of data etc.

- Organization and legislation in the area of TLE differs from country to country, and sometimes even between regions inside a country

- The field of TLE is complicated, and there is today a lack of common understanding of terms and definitions regarding TLE-data – the DCS should be a supporting tool to improve the common knowledge

- The demands on the input side are very different from demands on the output side, and it is important to design a system that satisfies both sets of demands

- Demands on the output side are very widespread, from aggregated results for reporting purposes to very detailed datasets for scientific purposes

- The system will have to develop over time, as data availability improves and the demand for output grows – and at the same time the system will have to be in continuous operation

If it is seen important by the authorities and the scientific community to establish a Europe wide TLE data system which can help supporting work in the field of TLE throughout Europe, the system must be able to handle these and other challenges.

If we are going to succeed in the task of building up a coherent TLE data collection system, capable of doing the job at a fairly high level of ambition, the process of planning and developing the system must be considered carefully. In order to secure the best possible systems development it is important that the work on demand specifications is conducted in a way where the three parts of the system – DCS, database and output systems – are defined in an integrated process.

There are two important prerequisites in particular that have to be fulfilled, if the development of a TLE system shall be successful:

- The organization and the people responsible for running and developing the system are extremely important
  - Running of the system will demand many resources, centrally and in each country, and it is vital that they are competent and enthusiastic about the project
– The organizational structure is also of great importance, as the complexity of collecting and disseminating TLE data will put a great pressure on the ability to keep a good overview over administrative processes

w The political and administrative processes in the field of TLE are at least equally important

– If there is a constant political pressure for continued and enhanced focus on the ability to deliver data in the field of TLE, the development will continue to move in the right direction, and we will in the end have the needed data for delivering answers to many of the questions in the field of TLE

– If, on the other hand, the political process does not focus on continued improvement in this field, it is most likely that a less ambitious TLE system than described here would be sufficient to handle the available data.

The present Deliverable presents the key challenges related to designing a Europe wide TLE data collection system (and a data base), and gives conceptual guidelines for establishing such a TLE data system. The present Deliverable contributes to the task of creating a common data base by providing a structure for the design. When used together with the upcoming Deliverable 12 “Conceptual model for the European traffic law enforcement monitoring database”, it will hopefully provide the tools needed for well-informed decisions about the future TLE data base and data collection. After the PEPPER project, political will as well as consensus about the importance of establishing such a system is needed in order to finally establish a data base system for Europe wide TLE data.
6 REFERENCES


