THE KNOWLEDGE FOR TOMORROW'S ROAD SAFETY BASES ON HARMONISED DATA - THE GLOBAL SAFETY DATABASE DOES ITS CONTRIBUTION

Johann Ziegler
Institute for Traffic Accident Research at Dresden University of Technology (VUFO)
Germany

Michael Düring
Germany

Michael Wagner
Germany

Paper Number 23-0086

ABSTRACT

Road traffic accidents remain to be a leading cause of death worldwide with nearly 1.3 million fatalities each year (WHO Global status report on road safety 2018) [1]. To develop safety systems according to real-world challenges, harmonized information is needed. Therefore, vehicle and road traffic safety experts are constantly looking for real-world data to answer the open challenges and to ultimately reach the "Vision Zero".

Numerous data on road traffic accidents exist and can be split into national and in-depth databases. The latter are characterized by a significantly lower number of cases then the national databases but a substantially higher level of detail and enable a microscopic view on the accident scenario.

By using in-depth databases, new safety systems may be developed and validated. The results of analyses are extrapolated to assess the impact on road safety for a specific country, continent or even for the whole world. However, it is not always obvious which database is suitable for which type of development approach or extrapolation.

The Global Safety Database (GSD) [2] solves this issue by offering access to a one of its kind up-to-date worldwide collection of road traffic accident statistics and database on a meta-data level. In addition to the objective evaluation of databases by matching them to research questions, the GSD also provides knowledge on the representativeness of each database. In order to identify similarities and differences in road safety within the countries, the latest publication of the Global Status Report on Road Safety from 2018 [1] is used to develop a clustering methodology. The goal of this method is to point out the possibilities and limitations of transferring information from the initial countries to other areas of interest.

The core of the investigation is the clustering methodology, which generates derivatives on countries or regions with similar road safety standards. The objective matching algorithm within the GSD helps to find the necessary information for the qualitative assessment of representativeness. Once the representative database within a country is identified, the clustering results are used to determine which countries represent the chosen database.

As the clustering relies on the latest Global Status Report from 2018 (and even partly from 2016), more recent data on road safety is desirable to narrow the spread to a steadily growing GSD. For a more integrated road safety approach, the GSD is also prepared to cover more topics related to road safety e.g., infrastructure or medicine. Additionally, an extension of the qualitative assessment of representativeness to a quantitative is more robust.

The clustering may be used to find derivations to the initial country and to transfer the results from these to the target countries by similarities in road traffic safety. From a global perspective, the GSD is one essential tool to push forth the worldwide harmonisation of traffic accident statistics and databases. Knowing what really happens on the roads by putting together everything we know empowers the data-driven development of safety systems and thus brings us one step closer to reach a road system without casualties – fulfilling the Vision Zero.

RESEARCH OBJECTIVE

Within the research and development on road safety, numerous interdisciplinary aspects are considered, particularly in the field of vehicle and traffic safety. The relevant issues of assisted, connected, and automated driving and the further development of passive and integral safety systems require reliable data sources. The heterogeneous traffic and accident situations in different countries and continents require taking data sources from numerous countries/regions into account. However, it is not always obvious which data source is suitable for what kind of research question or development approach.

The overall goal of the GSD is to provide a unified meta database that contains necessary information required for research and development departments on a meta based level (explicitly no raw data) for several countries. The database is designed dynamically on a platform that allows changes in the data sources to be effective immediately.

One of the main tasks is a detailed investigation on international data sources in the field of traffic and vehicle safety. This includes, but is not limited to, national road accident statistics based on police accident data as well as highly detailed investigations in smaller regions (so-called in-depth data sources).

In addition to the development of the meta database, a questionnaire is established and used to check the applicability of the developed meta database for specific questions regarding road safety. The questions are reformulated using variables which are necessary to answer the questions. An objective assessment of the meta database and the questionnaire requires developing a matching process. The aim of this matching process is to calculate the percentage of necessary variables covered in the various data sources for each question. The results of the matching process are collected in a result matrix within the GSD.

The result matrix offers a possibility for an objective assessment of data sources and provides the opportunity for data providers to improve their data quantity and data quality. Furthermore, the meta database can act as a platform to bring several data providers from different countries together and to encourage the global harmonisation of traffic accident data sources. For this purpose, a usage and management system are set up to improve and control data quality and quantity. Furthermore, it puts forth the search on new data providers and road safety experts.

In addition to the development of the database structure and user administration, the representativeness of individual data sources is examined and compared with the country-specific accident figures and investigation methods. The developed clustering procedure supports the process of searching for representatives in regions or for individual countries. The data analysis of the meta database is intended to show specificities within the data sources.

The GSD platform act as a tool to promote the global harmonisation of accident data and statistics for support the acceleration future research and development processes.

METHODS AND DATA SOURCES

Conception of the meta database

The basis for the conception of the meta database within GSD is the database structure of the interdisciplinary German In-Depth Accident Study (GIDAS) [3], one of the world's leading accident research projects. The meta database contains high-level information about the data sources and a content-related part in which available information is inventoried. Each part comprises several tables that are linked by primary keys.

The database not only contains information on the parameters and contents available in a data source, but also meta data that directly determine the suitability of the data source for certain research questions or the reliability of statements derived from it. An example of this is the representativeness of a data source, which is an elementary aspect for the usability and evaluation of data sources.

Further parameters are dedicated to the investigation methods as well as access options and costs to provide GSD users an overview of potential access paths in addition to an assessment of the usability of the content. For the derivation of exposure variables or basic figures, country-specific key figures are stored, e.g. the number of traffic accidents per year, information on demographics, vehicle fleet, and infrastructure. Content-related aspects comprise tables for the accident itself, involved participants, persons, and their injuries.

Search for data sources

The investigation of data sources for road traffic accidents is specific for national databases on the one hand side and in-depth databases on the other. Regardless of the origin of the data (national or in-depth databases), the researched databases or statistics for the GSD are designated as data sources and the developed database as meta database (*Figure 1*).

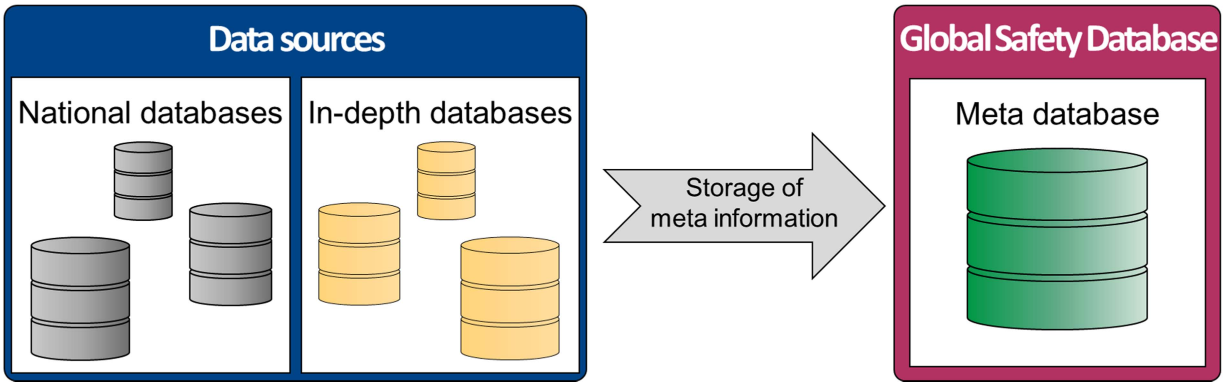

Figure 1. Scheme of the wording within the GSD

The so-called national data sources are based on data collected by the police, which consist of a large number of investigated accidents and give a macroscopic view of accident scenario. The use of a police reported data source on a national level allows a national coverage of the accident scenario and the acquisition of numerous cases.

The two main objectives of the police investigation and the national data sources is the collection of general accident-related data to gather evidence for prosecution and assessing the infrastructure safety. In-dept data sources take a different approach, in which the data providers primarily want to investigate how the accident happened without focusing on evidence subsequent prosecution.

In comparison to the national data sources, in-depth data sources are mainly characterised by a smaller number of cases, but usually by a higher level of detail in the investigation of road traffic accidents. This allows a microscopic view on the accident scenario. In contrast to the police reported accidents, the accidents of the in-depth data sources are usually investigated by accident researchers and medical experts.

The investigation for data source is based on a process that has be defined within the GSD. Figure 2 shows the process with all phases for contacting data providers and data immigration into the GSD.

Figure 2. Overview research process within GSD

Methodology of the objective assessment

A key idea of the project is being able to objectively assessment the quality of existing data sources in the GSD. A metric for evaluating the quality is defined with the suitability of the data sources for answering current and future research questions, also with a view to a practical application of the meta database.

For the object assessment of the meta database, the membering companies of the Research Association of Automotive Technology e.V. (FAT e.V.) were asked to collect questions relevant to the field of design, development, and evaluation of road safety measures. This collection of questions resulted in a questionnaire with more than 190 questions, of which 120 are research questions with content and the rest are questions about the characteristics and meta data of the researched data sources. The GSD user group has the possible to add new questions to the catalogue, which leads to a permanent growth and specification of the questionnaire.

The research questions are subsequently analyzed semantically in order to create a mapping of the content associated with the question to the variables of the meta database. The requirements for the data source contents are thus defined for each question, which are stored in binary codes at the parameter level.

Due to the basic binary coding structure of information, the linkage between questions and provided information in the meta database is based on the expertise of the user of the GSD. After translating, each “QUESTION_ID” with its coded variables is matched to each “SOURCE_ID“ and its variables (Table 1).

Table 1.
Example of the questionnaire and meta database content

QUESTIONNAIRE

QUESTION_ID	VARIABLE_1	VARIABLE_2	VARIABLE_3	VARIABLE_4
1	1	1	0	1
2	0	0	0	1
3	1	1	1	1

META DATABASE

SOURCE_ID	VARIABLE_1	VARIABLE_2	VARIABLE_3	VARIABLE_4
1	1	0	1	1
2	1	1	1	1
3	0	1	0	1

The content of the result matrix indicates the percentage of the “SOURCE_ID“ variables that match the necessary variables to answer the question. The result matrix (Table 2) is structured as follows:

- “QUESTION_ID” as row and
- the “SOURCE_ID“ as column

A complete match between one “SOURCE_ID“ and one question results in a value of 100 % (green box).

Table 2.
Example of the questionnaire and meta database content

RESULT MATRIX

QUESTION_ID	SOURCE_ID_1	SOURCE_ID_2	SOURCE_ID_3
1	2/3 = 66 %	**3/3 = 100 %**	2/3 = 66 %
2	2/2 = 100 %	2/2 = 100 %	1/2 = 50 %
3	3/4 = 75 %	4/4 = 100 %	2/4 = 50 %

In addition to the result matrix, the analysis of the data sources and questions as well as the matching process supports to identify important variables and labels, needed for current and future research questions. The analysing process described provides the opportunity for data providers to improve their data quantity and quality.

It should also be noted that an assessment completeness and plausibility of the actual (accident) data does not take place. This quality check and the decision whether a data source is used for an analysis even if the percentage value is low lies in the responsibility of the experienced users.

Assessment of representativeness
The representativeness is defined diversely, depending on the scientific mission and the goal to be aimed for. Within the GSD, the representativeness is defined by two key aspects:

- the existence of a “suitable” sample plan, and
- the unbiased chance of any accident to be part of data base.

A data source is representative if it is possible to draw conclusions from a sample to the totality, i.e. when certain elements of the totality have the same chance of being part of the sample. Therefore, the official road traffic accident statistics or national data sources collected by the police form the totality and represent the road accident scenario for a country or region. The sample of an in-depth data source corresponds to accident investigations in a certain region, which may also be investigated by the police, but are mostly carried out by in-depth collection units, which do not include all police-recorded accidents.

The assessment of representativeness can be determined with the statistics from official authorities as well as from the in-depth database via a quantitative methodology. The GSD with its “no raw data”-policy enables just a qualitative assessment of the in-depth data sources. For this purpose, a collection of variables is extracted from the GSD and compared to the national and the in-depth data sources of each country. The variable set can be chosen according to requirements of the analyses.

The methodology for the qualitative assessment is based on comparing the predefined variables between the national data source and the in-depth data source. In this process, it is queried whether a variable is present or not. If a variable is present in both types of data sources, it may be possible to use this variable for a representativeness assessment of the in-depth data source to the national statistics. The values within the variables are not considered in this qualitative method.

In addition to the qualitative assessment of representativeness, the data providers are also interviewed regarding the representativeness of their databases. These information as well as the result of the qualitative assessment form the variable representativeness within the GSD.

The variable "representativeness" in the GSD is defined by these parameters:

1 - representative for country (official statistics) without weighting

2 - representativeness possible (weighting necessary)

3 - not representative

999999 - unknown

Clustering methodology

The assessment methodology for the representativeness is used to compare national and in-depth data sources within a country. We investigate the feasibility of clustering by comparing and analyzing similarities and differences in road safety between two or more countries.

The main objective of the cluster analysis is to find out which countries have the closest similarity to a country selected for the analyses (target country) based on road safety standards. But it is not intended to carry out a full clustering calculation. The basis of the clustering method within the GSD is a simple comparison with rank assignments.

The clustering is carried out according to the following safety standards (description/category/unit) in Table 3.

Table 3.
Safety standards by category and unit

Safety standards	Category	Unit
Population density	A	[inhabitants / km²]
Gross Nation Income	B	[US dollars per capita]
Fatality rate I	C	[road deaths per 1 Mio. inhabitants]
Fatalities by road user	D	[percentage]
Vehicle rate	E	[vehicles per 1 Mio. Inhabitants]
Distribution of reg. vehicles	F	[percentage]
Fatality rate II	G	[road deaths per 1 Mio. vehicles]
Safety standards in traffic	H	[speed limitation, belt usage, helmet law]

The basis of the road safety standard is the latest WHO global status report on road safety for 2018 [1]. In total, 175 countries are mentioned in der WHO report, which are considered in the clustering process within this project. Not every country contains full information to all road safety standards. The boundary conditions for data collection by the WHO are mentioned in the report and can be found there.

In order to identify the countries with the closest similarity to regions or other countries, a ranking system is developed. A methodology checks within the safety standards which country is closest to the target region or country. The relative ratio between challenging country and target country/region identifies the deviations. The countries with the closest distance to the target earn the highest scores. The score depends on the number of countries to be compared. The average score of all safety standards results in the final rank for each challenged country. The highest average score gets the ranking position 1 and is closest to the target country.

The example of the Scandinavian countries shows which country is the closest to Norway (NOR) by using four of eight safety standards (Table 4). Depending on the three chosen safety standards, Sweden (SWE) is the closest to Norway. By using all categories of safety standards, Sweden is also the closest country to Norway.

Table 4.
Example of the clustering methodology for the Scandinavian countries

	Category: A		Score	Category: C		Score	Category: E		Score	Ø SCORE	**RANK**
NOR	14.1			26.31			730,307				
SWE	23.3	+65%	3	26.54	+1%	4	582,606	-20%	4	3,7	**1**
DNK	137.0	+871%	1	38.58	+47%	3	532,275	-27%	3	2,3	**3**
FIN	16.4	+16%	4	46.85	+78%	2	940,142	+29%	2	2,7	**2**
ISL	0.4	-97%	2	583.12	+2116%	1	7,673,373	+951%	1	1,3	**4**

The clustering process not only map which country is most similar to the target country. It also offers the possibility to find out which country within a region most closely represents the region. For this purpose, all countries of the chosen region are combined, and the average values of the respective safety standards is calculated based on the country-specific individual values. As an example, for Northern Europe and the Scandinavian countries, Sweden is the representative for this region after the clustering calculation.

RESULTS

Content of the meta database

In the two project phases commissioned by the FAT, 17 countries are researched in detail by national and in-depth data sources. In total, 52 data sources are found, where the half of the total is distributed between national and in-depth data sources (Table 5).

Table 5.
Number and type of data sources searched by country

Countries	Number of national data sources	Number of in-depth data sources
Australia	3	3
Brazil	1	1
China	0	2
Czechia	1	1
Denmark	2	-
France	1	1
Germany	3	9
Greece	1	1
India	1	1
Indonesia	2	0
Japan	1	1
Nigeria	2	0
Norway	1	1
Russia	1	0
South Africa	1	0
Sweden	1	2
USA	4	3
TOTAL	**26**	**26**

Additional information on other countries is recorded in the GSD fact sheet. In total the GSD contains 40 countries, where at least one national data source or one in-depth data source is mentioned. Currently, 103 individual data sources are inventoried.

Evaluation of the representativeness

For the qualitative assessment of the representativeness, a collection of variables is extracted from the GSD and compared for the national as well as for the in-depth data sources for each chosen country. The variable set can be chosen according to requirements of the analyses. Within this project, the following variables from the GSD were selected for the qualitative assessment (Table 6). These variables can be used for the general description of an accident.

Table 6.
Variables set from the GSD for the qualitative assessment of the representativeness

Variable description	Variable GSD	Classification
Accident month	ACCMONTH	
Weekday	WDAY	
Daytime	DAYTIME	
Accident location	LOCATION	Accident data
Maximum accident severity	ACCSEVERITY	
Accident type	ACCTYPE	
Number of participants	NR_PARTICIPANTS	

Due to a lack of information provided by the data provider or due non-collection of data some countries (e.g., Germany, Greece, France, Czech Republic, USA, Brazil, Australia) have better preconditions than other countries (e.g., Sweden, India, Norway, Japan). Consequently, the qualitative assessment is only carried out for the first seven mentioned countries (Table 7). For the methodological example, only one in-depth data source per country has been compared with the national database.

Table 7.
Example for the qualitative assessment of national data sources and In-Dept data sources

Country	Type of database	ACCMONTH	WDAY	DAYTIME	LOCATION	ACCSEVERITY	ACCTYPE	NR_PARTICIPANTS	**Total**	Share of Matches
GERMANY	National data source	1	1	1	1	1	1	1	7	
	In-depth data source	1	1	1	1	1	1	1	7	7/7 = **100%**
GREECE	National data source	1	1	1	1	1	1	1	7	
	In-depth data source	1	1	1	1	1	1	1	7	7/7 = **100%**
FRANCE	National data source	1	1	1	1	1	1	1	7	
	In-depth data source	0	0	0	1	1	1	1	4	4/7 = **57%**
CZECH REPUBLIC	National data source	1	1	1	1	1	1	0	6	
	In-depth data source	1	1	1	1	1	1	1	7	6/7 = **85%**
USA	National data source	1	1	1	1	1	1	1	7	
	In-depth data source	1	1	1	1	1	1	1	7	7/7 = **100%**
BRAZIL	National data source	1	1	1	1	1	1	0	6	
	In-depth data source	1	1	1	1	1	1	1	7	6/7 = **85%**
AUSTRALIA	National data source	1	1	1	1	1	1	0	6	
	In-depth data source	1	1	1	1	1	1	1	7	6/7 = **85%**

The qualitative assessment of Table 6 indicate some in-depth data sources with potential on representative statements to the official statistic. However, it depends on which set of variables is chosen for the comparison. For example, the in-depth data source in Brazil shows with 85% of the chosen variable potential for representative statements on the entire Brazilian accident scenario. Whereby a weighting process within the in-depth data source to the official statistic is recommended based on the selection of variables.

Clustering of the countries

In-depth databases on road traffic accidents are not provided in every country in the world. This leads to the idea of using the already known in-depth data sources to make statements about countries or regions with similar road safety standards as the origin country with in-depth data source.

According to the clustering method developed, the Table 8 shows which bordering countries and countries by continent are most similar to the selected countries.

Table 8.
Bordering countries and countries by continent with the closest similarity to the country selection

Country	Bordering country	Continent
GERMANY	Austria	Spain
GREECE	Bulgaria	Cyprus
FRANCE	Belgium	Slovenia
CZECH REPUBLIC	Poland	Slovenia
USA	Canada	Canada
BRAZIL	Suriname	Suriname
AUSTRALIA	New Zealand	New Zealand

The countries with the closest similarity to the country of origin give a possible indication of which countries could be served by the data sources. The developed clustering method is a rough comparison of static key factor, whereby mentalities and country-specific behaviours are not considered.

Objective evaluation of the data sources

The result of the objective evaluation of the meta database is the result matrix. This dynamic table is calculated automatically and continuously after every change or adjustment in the meta database or in the questionnaire. Finally, the content of this matrix indicates for each data source the proportion of variables that correspond to the variables necessary to answer the research question.

Figure 3 shows in a cross comparison of seven selected data sources how many of the 120 research questions can be answered completely or to what percentage. The seven data sources are based on the countries from Table 8. In order to preserve the data protection of the data provider and avoid distortion of competition, no countries or data sources are named in the Figure 3.

With the data source from "Country I", for example, 30 of the research questions included in the GSD questionnaire can be answered completely.

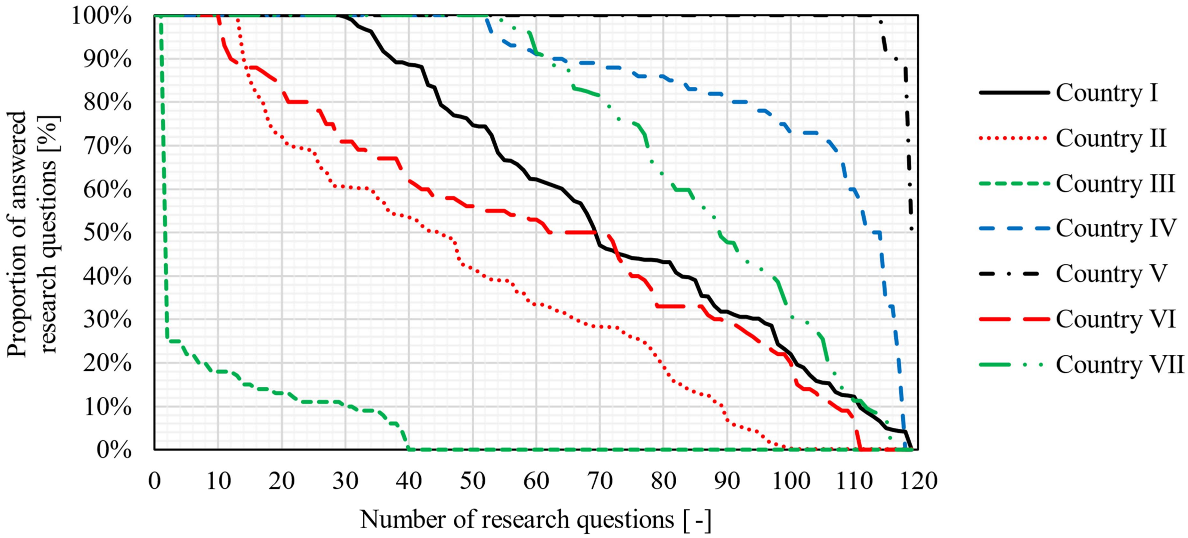

Figure 3. Visualization of the usability of selected data sources for answering the research questions stored in the GSD

DISCUSSION AND LIMITATIONS

The challenging task of storing meta data from several data sources around world from different data providers in a uniform database structure was successfully completed. The exchange of information between the GSD working group and the data providers is essential. Each data provider has different descriptions or meanings of variables and their collection. By this information the GSD working group have to decide which variables in the GSD match to the information of the data providers. The best scenario is when the data provider directly enters its meta data into the GSD.

The current questionnaire within the GSD and the evaluation of the suitability data sources is mainly based on research questions from R&D departments of automotive companies. For a more integrated road safety approach, the GSD is also prepared to cover more topics related to road safety e.g., infrastructure or medicine as well as the GSD questionnaire could be expanded to include questions from legislative bodies, authorities, associations, universities, or consumer protection organizations.

For a solid argumentation for future developments in road safety, up-to-date data is mandatory. The current basis of the clustering is the latest Global Status Report from 2018 [1] with data partly from 2016, whereby some of these data calculated and extrapolated by WHO. For a meaningful comparison, more recent data on road safety is desirable to narrow the spread to a steadily growing GSD. Current influences (e.g., covid pandemic, climate change, energy transition) can rarely be considered.

Additionally, an extension of the qualitative assessment of representativeness to a quantitative would be more robust. The GSD working group provides the methodology and framework for the quantitative assessment of representativeness, but on basis of the meta data stored in the GSD only a qualitative assessment is feasible.

In order to guarantee the data quality of the GSD, a steering committee (SC) is established. Every change that is made by authorised GSD users in the meta database is recorded in an integrated change log and will lead in a review. SC members are responsible for reviewing these changes. Each review is held according to the two-man rule, where disagreements can be escalated to the head of the SC. The SC is still in the process of being set up and the group of GSD experts will constantly looking for new SC members with expertise on accident statistics and databases worldwide.

CONCLUSION AND OUTLOOK

The Global Safety Database aims to investigate, inventory and objectively evaluate numerous accident data sources from different countries all over the world. Thus, it provides data for top relevant questions of todays and tomorrows road safety. For this purpose, a meta database is designed and filled with data sources of different kinds and origins. The subsequent objective evaluation is based on a matching process, which matches the content of current research questions of a questionnaire with the availability of variables in each data source. The result is stored in a so-called result matrix that shows the availability of necessary variables per data source for each and every research question.

An innovative methodology for clustering is used in order to identify countries or regions which data may be used to be transferred to other countries with similarities evaluating the road safety factors. Therefore, the representativeness of data sources especially of in-depth data sources is an essential aspect as knowledge of one country or region may accelerate the road safety improvement in other regions of the world as well.

As the GSD comprises only meta data, representative statements may also only be formulated on a qualitative level. The GSD is designed to only store meta data of data sources. This meta data is not sufficient for a qualitative assessment which requires raw data. Within the GSD, a qualitative assessment has been co-developed to show the possibilities how representativeness could be assessed with raw data if it is accessible.

In order to support the development process, the GSD is meant to be used by everyone, regardless of the position or organisation. [4] The GSD is freely accessible and the data quality as well as user management is ensured by the voluntarily voted Steering Committee.

From a global perspective, the GSD is one essential tool to push forth the worldwide harmonisation of traffic accident statistics and databases. The GSD is designed to be an ongoing meta database as well as questionnaire that engages its users by sharing information on global data sources for road traffic accidents and upcoming research questions for road safety. Thus, a community of experts may be formed and grow steadily. Furthermore, the GSD aims to support the increasingly data-driven vehicle development process by greatly reducing the effort finding required and suitable data sources to answer the top relevant research and development questions. This novel community is a key puzzle piece for enabling right developments and decision for future generations of safety measures, regulations, and systems. Thus, a Vision Zero with no road traffic casualties is getting closer and closer.

REFERENCES

[1] **World Health Organization**. “Global status report on road safety 2018” Geneva: World Health Organization; 2018; Licence: CC BY-NC-SA 3.0 IGO; ISBN 978-92-4-156568-4

[2] **Ziegler, J.; Liers, H.; Chanove, A.; Pohle, M**. “Objective assessment of database quality for use in the automotive research and development process”. FAT publication series 343, Berlin, 2021

[3] **The Research Association of Automotive Technology (FAT), Federal Highway Institute of Germany (BASt).** “German In-Depth Accident Study (GIDAS)“. [Online] [Cited: December 08, 2022.] www.gidas.org.

[4] **Global Safety Database** [Online] [Cited: December, 2022.] https://www.global-safety-database.com/