UTRECHT UNIVERSITY

Applying Semantic Integration to improve Data Quality

by

O.F. Brouwer

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Science
Department of Information and Computing Sciences

January 2016

Declaration of Authorship

I, AUTHOR NAME, declare that this thesis titled, 'THESIS TITLE' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"There's so much pollution in the air now that if it weren't for our lungs there'd be no place to put it all. "

Robert Orben (US comedy writer)

UTRECHT UNIVERSITY

Abstract

Faculty of Science
Department of Information and Computing Sciences

Master of Science (MSc)

by O.F. Brouwer

The research project is initiated to compose the mobility carbon footprint for the Netherlands, this footprint describes the impact on the environment of all travel movements in the Netherlands, expressed in emission of Carbon Dioxide (CO_2). There is no data source available that can provide an insight in both the number of movements and the location where these movements are made, this information has to be gathered from multiple data sources. The purpose of this research is to improve the data quality of both relevant datasets by applying semantic integration techniques. The resulting, integrated dataset can provide the information that neither of the data sources individually can.

A research method is proposed, consisting of several steps: the data quality of the two individual datasets is assessed, using a shared ontology semantic matches are identified. Using these semantic matches the individual datasets are integrated to a dataset with a higher data quality than the originating data sets. Lastly, the data quality of the integrated dataset is assessed to determine whether there is an increase in data quality.

The semantic integration results in an integrated dataset that is used to compose the mobility carbon footprint for areas in the Netherlands. The carbon footprint grants an insight in the mobility of all Mezuro areas in the Netherlands. Because of the integration of the Mezuro and OViN data it is now possible to get detailed information on CO_2 emission from mobility (for example highest emitting areas).

Acknowledgements

Firstly, I would like to thank my thesis supervisors, Wim Steenbakkers from Mezuro and Marco Spruit from Utrecht University, that have been supportive and provided guidance along the entirety of my research project. Their insights and motivation has helped me over the course of the research.

Secondly, I would like to express my gratitude to all my coworkers at Mezuro for showing me a great time during the months working with them. In particular I would like to thank Jasper Keij for his great help and patience while sharing a room with me. The time I've spent at Mezuro will stay with me as a great learning experience, including all the great times during lunch and the jokes in the office.

Contents

List of Figures

List of Tables

Abbreviations

BPR	**B**usiness **P**rocess **R**eengineering
CBS	**C**entraal **B**ureau voor de **S**tatistiek (Dutch Statistics Bureau)
CDR	**C**all **D**etail **R**ecords
CO_2	**C**arbon **D**ioxide
DQ	**D**ata **Q**uality
DQA	**D**ata **Q**uality **A**ssessment
DQM	**D**ata **Q**uality **M**ethodology
ERD	**E**ntity **R**elationship **D**iagram
IP-MAP	**I**nformation **P**roduction **MAP**
IPCC	**I**ntergovernmental **P**anel on **C**limate **C**hange
MON	**M**obiliteits**O**nderzoek **N**ederland (Mobility Research of the Netherlands)
NS	**N**ederlandse **S**poorwegen (Dutch Railways)
OVG	**O**nderzoek **V**erplaatsingsgedrag **N**ederland (Research Movement behavior in the Netherlands)
OViN	**O**nderzoek **V**erplaatsingen in **N**ederland (Research of Movements in the Netherlands)
PDD	**P**rocess **D**eliverable **D**iagram

Chapter 1

Introduction

Climate change and global warming are still a 'hot topic', possibly literally in the future. The Intergovernmental Panel on Climate Change (IPCC) released their 2014 report on Climate Change on November 1st, stating that greenhouse gas emissions have to be 0 (or even below 0) by the year 2100 to keep the temperature increase below 2 degrees Celsius (IPCC, 2014). The major component of greenhouse gases is Carbon Dioxide (CO_2), responsible for 76% of the greenhouse gas emission (IPCC, 2014). In 2013 in the Netherlands cars and other means of transport are responsible for over 20% of the total CO_2 emission (CBS, 2014). Therefore it is important to gain an accurate insight in the carbon emission caused by transport in the Netherlands, we call this the mobility carbon footprint.

The mobility carbon footprint is the amount of CO_2 emitted by vehicle transportation (measured in tons CO_2) for a certain area over a certain period of time. The emission of carbon dioxide is directly related to the amount of fuel used by a vehicle, therefore the fuel economy and the travelled distance determine the emitted CO_2. To draw an accurate picture of the mobility carbon footprint in the Netherlands it is crucial to have an overview of all travel movements and their distances. This is where the company Mezuro specializes in; they solve mobility problems using telecom operator location data. Mezuro has a unique partnership with a major Dutch telecom provider that allows them to have access to events (that are generated by Call Detail Records). These events include, among others, the hashed ID of a user's device (fully anonymous to protect the privacy), a timestamp, and the cellphone tower the device is connected to. The connected cellphone tower and its coverage area combined with the timestamp give a good insight in the location of a person on a specific time of day. Using these events over a longer period of time shows where a user has travelled, applying this techniques

FIGURE 1.1: An overview of the Mezuro area designation

to all subscribers of the partnered telecom provider results in an overview of the mobility in the Netherlands.

However, the data available to Mezuro contains some weaknesses that can be detrimental to the accuracy of the desired mobility carbon footprint. A known weakness is the absence of movements over short distances, this can occur when a movement remains within the coverage area of the cell tower (some cell towers have coverage up to 30 km). Another possibility is that the movement remains within a 'Mezuro area', the area designation used by Mezuro (see Figure 1.1).

For the mobility carbon footprint to be accurate the Mezuro data alone is not sufficient, more information (mainly regarding short movements) is required to create a useful carbon footprint. The Dutch bureau of statistics (CBS) also releases a dataset every year regarding the movements of the Dutch population, called OViN (a Dutch abbreviation of 'research of movement in the Netherlands). OViN is a survey based research conducted every year to provide insights on the day-to-day mobility of the Dutch population. The encountered problem is how both datasets can be put together in order to create a more complete integrated dataset. The purpose of this research is to integrate two datasets in

order to achieve a higher data quality, which is necessary to craft an accurate mobility carbon footprint. This leads to the main question:

"How can semantic integration of multiple data sources be used to improve data quality?"

The main research question is divided into more operational questions that can be answered more specifically. The combination of the three sub-questions illustrates to what extent semantic integration is applicable to improve data quality.

1. What approaches can be used to determine data quality of individual data sources?
2. What data quality (improvement) strategies and techniques can be applied to multiple data sources?
3. In what way can semantic integration techniques be used to impact data quality?
4. How does semantic integration improve the data quality for the Mezuro and OViN datasets?

1.0.1 Research approach

The research is designed using a qualitative approach, the first step is to carry out a literature study. Relevant methods and concepts from literature are used as a reference for building a method, steps from the method or techniques for building the method are addressed in the literature study. The subsequent step is the evaluation of the method with a case study, for this research a single case study is conducted at Mezuro. The research method will include a segment on the data quality of the involved datasets, this data quality assessment will be conducted using expert interviews. Results of the case study regarding the mobility carbon footprint will be critically discussed with the Mezuro company to ensure satisfactory results.

Answering the research questions will be done using a review of the literature in this domain, and by using a method that integrates the Mezuro and OViN datasets. The study is structured as follows: Chapter 2 discusses the domain of data quality, in Chapter 3 literature regarding semantic integration is addressed. The research method is introduced in Chapter 4, and the method evaluation is listed in Chapter 5. The results the integrated dataset provides are included in Chapter 6, Chapter 7 will be used for the conclusions, Chapter 8 includes the discussion and future work.

Chapter 2

Data Quality

In this chapter the theoretical background to data quality is discussed. Firstly the concept of data quality is explained; particularly the four data quality categories that define data quality. Thereafter the measuring of data quality is discussed; in specific how the data quality dimensions can be assessed. Lastly existing methodologies for data quality assessment and improvement are discussed, including techniques for assessing and improving data quality from these methodologies.

The concept of data quality (DQ) is defined by its "fitness for use" by data consumers; this definition emphasizes the importance of the data consumers' view on data quality (Wang & Strong, 1996; Strong, Lee, & Wang, 1997). Data quality can be described as a collection of data quality dimensions, each representing a different aspect of quality. Each of the DQ dimensions can be assessed using metrics. In general multiple metrics can be associated with each quality dimension, in some cases the metrics are unique and the metrics act as the operational definition of the theoretical dimension. Metrics ensure that data quality dimensions, and therefore data quality in general, is measurable (Pipino et al., 2002).

The data quality dimensions represent a single aspect of data quality; each data quality dimension can be described as a set of data quality attributes. A list of data quality dimensions is composed in a study by (Wang & Strong, 1996), using a data consumer focused approach. Three tasks for identifying quality attributes of a product are used (Churchill & Iacobucci, 2010) to identify the quality attributes of data:

1. Identifying consumer needs
2. Identifying the hierarchical structure of consumer needs
3. Measuring the importance of each consumer need

The first task of identifying customer needs consists of a survey in which data consumers have to come up with attributes they relate to data quality. This survey resulted in a list of data quality attributes. A second survey is used to assess the importance of the data quality attributes to the data consumer, the ratings of importance are used to yield an intermediate set of data quality dimensions that are important to the data consumer. Since the initial list of quality attributes resulted in a broad spectrum of data quality dimensions a follow-up study was conducted to group the data quality dimensions, grouping was done for three main reasons. Firstly, the high amount of quality dimensions is not critical for the evaluation process (Kriebel, 1979). Secondly, the highest rated data quality attribute by the data consumer may not cover the essential aspects of quality. And thirdly, the intermediate data quality dimensions can be split up into several families of quality factors.

There are several frameworks introduced for data quality, based on the research of Spruit (2013), the framework from Wang and Strong (1996) is selected for this research. *"It was found that the most influential framework is the one by Wang and Strong (1996)"* (Spruit, 2013).

Wang and Strong (1996) propose that a framework for data quality includes the following four aspects:

- The data must be *accessible* to the data consumer
- The consumer must be able to *interpret* the data
- The data must be *relevant* to the consumer
- The consumer must find the data *accurate*

Through further sorting and evaluation with data consumers, the list of data quality dimensions and the aspects of a data quality framework are combined to result a conceptual framework for data quality (2.1). The framework consists of four categories that represent the four aspects of data quality.

Data Quality Category	**Data Quality Dimensions**
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

TABLE 2.1: Data Quality Framework (Wang & Strong, 1996)

"Intrinsic DQ denotes that data have quality in their own right. Contextual DQ highlights the requirement that data quality must be considered within the context of the task at hand. Representational DQ and accessibility DQ emphasize the importance of the role of systems" (Wang & Strong, 1996)

Intrinsic data quality: The category of intrinsic data quality contains the data quality dimensions related to the data itself. Intrinsic data quality is viewed independent of the context in which the data is produced and used (Strong et al., 1997). Dimensions in the intrinsic data quality category are: accuracy, objectivity, believability, and reputation. Just like in product quality where the concept of quality includes the consumers (Deming, 1986), data quality has to include the people who use the data, the data consumers.

Accessibility data quality: accessibility implies that data should be available, obtainable or retrievably when needed (Wang & Strong, 1996). The dimension of access security is to ensure that data is not available to the wrong people. Accessibility data quality can be viewed as a category if overall data quality, or separated from other data quality categories, in both instances accessibility in an important concept to data quality.

Contextual data quality: "Data quality must be considered within the context of the task at hand" (Strong et al., 1997), this concept was derived from the field of graphical data representation, stating that the quality of a graphical representation must be assessed within the context of the data consumer's task (Tan & Benbasat, 1990). The contextual data quality dimensions are: relevancy, value-added, timeliness, completeness, and amount of data.

Representational data quality: representational data quality consists of two main aspects to make sure that data is well represented for the data consumers. The two aspects relate to the format of the data (concise and consistent representation) and the meaning of the data (interpretability and ease of understanding) (Wang & Strong, 1996)

2.1 Measuring data quality

Data quality is a multi-dimensional concept (Pipino et al., 2002), meaning that companies have to take both objective assessments and subjective perceptions into account. Objective measurements are based on the dataset, whereas the subjective data quality assessment is based on the perception of stakeholders. The subjective perceptions reflect the needs and experiences of the collectors, custodians, and consumers of data products (Pipino et al., 2002). The result of assessment of data quality with both subjective and objective measurements is displayed in Figure 2.1. The goal is to achieve high quality in both assessments (Quadrant IV in the figure), when a company falls in to quadrant

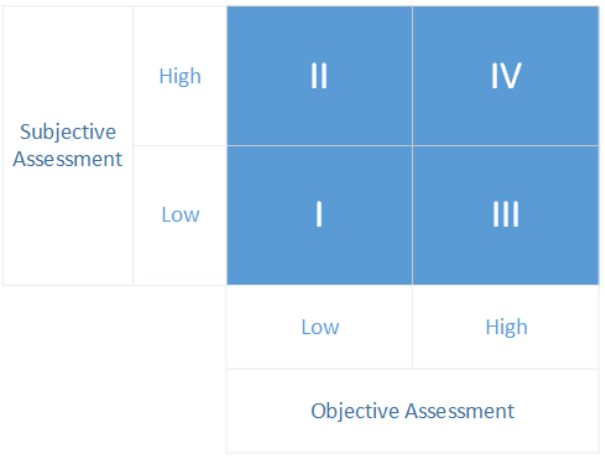

FIGURE 2.1: Quadrant of subjective and objective assessments (Pipino et al., 2002)

I, II, or III there is need for action (investigate the root cause of the quality issues and take action).

There can be discrepancies between the subjective and objective assessment (quadrants II and III), the cause of the discrepancy is generally caused by the subjective perception of stakeholders. For results in quadrant II the stakeholders perceive the quality as high, while the actual data quality is not. For quadrant III the perception of quality from the stakeholders is low, while the objective assessment reports high quality. It is important to find the cause of the discrepancy, an example for a quadrant III result by Pipino et al. (2002): *'the client's subjective assessment was based on the historical reputation of data quality'*.

Subjective assessment The subjective data quality assessment is conducted by measuring the stakeholders' perception of data quality dimensions, generally by using a questionnaire. When great differences are found in the results a follow-up investigation can provide a valuable insight what areas need improvement.

Objective assessment The objective measurement of data quality is performed by metrics, developed specifically to measure data quality dimensions. Based on the DQ dimension the metric is designed to track specific attributes that impact the data quality dimension, for example:

- Tracking the percentage of non-existing accounts, or accounts with missing values *(Completeness)*
- Tracking the amount of records violating referential integrity *(Consistency)*

For all the DQ dimensions that are being used in the assessment of quality metrics can be designed, providing an insight of the data quality level of the organization. Some

metrics use measurements where the observed data value is available (checking null-values; completeness), or compared to a reference value (accuracy). Other measures use constraints to determine the quality of a metric; the observed data value should be within a fixed range (Age needs to be within the range 0 to 120; consistency). Time-related metrics compare the change of data values to the change in a real world state (the measured delay; timeliness or the frequency of change of a value; volatility).

2.2 Data Quality Assessment and Improvement

We adopt the definition of a data quality methodology from Batini et al. (2009): *'A data quality methodology is defined as a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of data.'*

DQ methodologies can be analyzed and compared from several perspectives: phases and steps, strategies and techniques, dimensions and metrics, types of costs, types of data, types of information systems, organizations involved, processes involved and, services included. The first three perspectives on DQ methodologies include the most distinct differences and combined determine the capabilities of a methodology, therefore these perspectives (phases and steps, strategies and techniques and, dimensions and metrics) will be discussed in more detail.

2.2.1 Phases and steps of a data quality methodology (Batini et al., 2009)

In general, a data quality methodology starts with a state reconstruction. In this phase all contextual information is collected, this information relates to organizational processes and services, data collections and related management procedures, quality issues and related costs. The collection process of contextual information is done in preparation of the two main steps in data quality methodologies: data quality assessment, and data quality improvement.

Steps for data quality assessment (Batini et al., 2009)

Data quality assessment, or measurement, utilizes relevant data quality dimensions on the data collections to measure the quality of the data. The term measurement is used when the value of data quality dimensions is addressed; assessment is used to compare the measurements to reference values. The concept assessment is adopted in the majority of

data quality methodologies, since the reference values give an insight in the dimensions that are lacking in quality. The common steps to data quality assessment in a DQ methodology are (Batini et al., 2009):

- Data analysis, data schemas are examined and interviews are performed to come to a complete understanding of the data and related architecture and management rules.
- DQ requirements analysis, the opinion of data users and administrators is gathered to identify current quality issues and set new quality targets.
- Identification of critical areas, the most relevant databases and data flows are selected for quantitative assessment.
- Process modeling, the processes for producing or updating data are modeled.
- Measurement of quality, the quality dimensions affected by the quality issues (from the DQ requirements step) are selected, and corresponding metrics are defined. The measurement of the drafted metrics can be done both objectively (quantitative metrics) or subjectively (qualitative evaluations).

Steps for data quality improvement (Batini et al., 2009)

Data quality improvement aims to reach the new data quality targets; the goal is to select the right steps, strategies and techniques in order to fulfill the DQ targets. The improvement process start with preliminary steps to evaluate costs, assign responsibilities and identify the cause of the quality issues. After these steps suitable data improvement strategies and techniques are selected, these strategies can be either data-driven or process-driven. Data-driven strategies directly modify the values of the data to improve the data quality, process-driven strategies are focused to redesign the processes that create of modify data to improve the data quality.

All the steps for the data quality improvement phase (Batini et al., 2009):

- Evaluation of costs; during this step the costs (both direct and indirect) of data quality are estimated.
- Assignment of process responsibilities; the process owners are identified and their responsibilities on data production and management activities are defined.
- Assignment of data responsibilities; the data owners are identified and their data management activities are defined.

- Identification of the cause of errors; identifying the causes of quality problems.
- Selection of strategies and techniques; identification of all data improvement strategies and corresponding techniques, which comply with the contextual knowledge, quality objectives, and budget constraints.
- Design of data improvement solutions; the most efficient and effective strategy (and related set of techniques and tools) to improve data quality is selected.
- Process control; defines check points in the data production processes, these check points can be used to monitor quality during process execution.
- Process redesign; defines the process improvement actions that can deliver corresponding data quality improvements.
- Improvement management; defines new organizational rules for data quality.
- Improvement monitoring; periodic monitoring activities are set up in order to provide feedback on the results of the DQ improvement process. Additionally this enables dynamic tuning of the improvement process.

Strategies and techniques for data quality

In the data quality improvement process, methodologies adopt strategies and their corresponding techniques for DQ improvement. There are two types of strategies: data-driven and process-driven. "Data-driven strategies improve the quality of the data by directly modifying the value of data. Process-driven strategies improve quality by redesigning the processes that create or modify data" (Batini et al., 2009). An example of data-driven strategy can be the updating of obsolete values by refreshing the database with data from a more current database. A process-driven strategy example can be: a process is redesigned to include an activity that formats data before storing it in a database.

Strategies, both data-driven and process-driven, apply techniques (algorithms, heuristics, and knowledge-bases activities) to improve the data quality. A list of improvement techniques applied by data-driven strategies is (Batini et al., 2009):

- Acquisition of new data, improving data by acquiring higher quality data to replace the current values that cause quality problems.
- Standardization (or normalization), replacing or complementing non-standard data values to values that comply with the standard. An example for this technique is replacing abbreviated street names in a database to the full name: "Main Str." to "Main Street".

- Record linkage, identifies data representations in multiple tables that might refer to the same real-world object.
- Data and schema integration, defines a unified view of data provided by heterogeneous data sources. The main purpose of integration is to allow a user to access the data, stored in heterogeneous sources, through a unified data view. The heterogeneity of data sources can be classified in: technological heterogeneity, schema heterogeneity and instance-level heterogeneity.
- Source trustworthiness, selects data sources based on the quality of the data.
- Error localization and correction, data quality errors are identified and eliminated by detecting records that do not comply with the quality rules.
- Cost optimization, technique where quality improvement actions are defined along a set of dimensions by minimizing costs.

And there are two techniques that are used for process-driven strategies:

- Process control, inserts checks and control procedures on several points of the data production process: when new data is created, when data sets are updated, or when new data sets are accessed by the process. The purpose is to have a reactive strategy applied to data modification events, to avoid data degradation and error propagation.
- Process redesign, redesigns processes to take away the causes of poor quality and introduces activities to produce higher quality data. If the process redesign technique is applied radically, this technique is referred to as business process reengineering (Hammer & Champy, 2009).

Redman and Blanton (1997) discuss the improvement each technique can achieve along different quality dimensions, on both the short and the long term. The comparison of data-driven and process-driven techniques revealed that process-driven techniques outperform data-driven techniques on the long term, since the origin of quality problems is resolved. However, on the short-term, the technique of process redesign can be extremely expensive (Redman & Blanton, 1997), and data-driven strategies are generally cost efficient on the short term. Data-driven techniques are most suitable for one-time applications, and therefore best fit for static data (Batini et al., 2009).

Methodology Acronym	**Extended name**	**Main reference**
TDQM	Total Data Quality Management	Wang, 1998
DWQ	The Data Warehouse Quality Methodology	Jeusfield et al., 1998
TIQM	Total Information Quality Management	English, 1999
AIMQ	A methodology for Information Quality assessment	Lee et al., 2002
CIHI	Canadian Institute for Health Information methodology	Long and Seko, 2005
DQA	Data Quality Assessment	Pipino et al., 2002
IQM	Information Quality Measurement	Eppler and Münzenmaier, 2002
ISTAT	ISTAT methodology	Falorsi et al., 2003
AMEQ	Activity-based Measuring and Evaluating of product information Quality methodology	Su and Jin, 2004
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality)	Loshin, 2004
DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapeico et al., 2004
QAFD	Methodology for the Quality Assessment of Financial Data	De Amicis and Batini, 2004
CDQ	Comprehensive methodology for Data Quality management	Batini and Scannapieco, 2006

TABLE 2.2: Data quality methodologies and acronyms (from Batini et al. (2009))

Data Quality Methodologies

The data quality methodologies generally adhere to the assessment steps proposed by Batini and his colleagues in (2009), although some might use different names. Three of the five steps for data quality assessment are included in most of the methodologies (Table 2.3); Data Analysis, Identification of Critical Areas, and Measurement of Quality. The fact that an assessment step is addressed in multiple methodologies does not mean that step is similar in every methodology; there are different approaches to the same assessment step among methodologies. For Measurement of Quality step, for example, is performing with questionnaires in the AIMQ methodology, in DQA a combination of objective and subjective metrics is used, and QAFD uses statistical analyses. Different measurement approaches meet the needs of different organizational contexts, processes, users or services (Batini et al., 2009). Table 2.3 below shows what assessment steps are included in each of the data quality methodologies.

From the list of data quality methodologies, not all support both assessment and improvement steps. Some of the methodologies are focused solely on data quality assessment (AIMQ, CIHI, IQM, and QAFD), additionally DQA and AMEQ do not provide improvement activities. Data Quality Assessment (DQA) includes the identification of

	Data Analysis	**Data quality Requirement Analysis**	**Identification of Critical Areas**	**Process Modeling**	**Measurement of Quality**	**Extensible to other Dimensions and Metrics**
TDQM	+		+	+	+	Fixed
DWQ	+	+	+		+	Open
TIQM	+	+	+	+	+	Fixed
AIMQ	+		+		+	Fixed
CIHI	+		+			Fixed
DQA	+		+		+	Open
IQM	+				+	Open
ISTAT	+				+	Fixed
AMEQ	+		+	+	+	Open
COLDQ	+	+	+	+	+	Fixed
DaQuinCIS	+		+	+	+	Open
QAFD	+	+	+		+	Fixed
CDQ	+	+	+	+	+	Open

TABLE 2.3: Assessment steps for DQ Methodologies (from Batini et al. (2009))

the causes of errors step, but only emphasizes on its importance and do not discuss execution. AMEQ does not provide operating methods and tools, only general guidelines. The remaining methodologies that support data quality improvement activities are listed in Table 2.4. The data-driven and process-driven strategies and techniques adopted in these seven methodologies will be discussed in more detail.

Strategy/Methodology	**Data-driven**	**Process-driven**
TDQM		Process Redesign
DWQ	Data and schema integration	
TIQM	Data cleansing Normalization Error localization and correction	Process Redesign
ISTAT	Normalization Record Linkage	Process Redesign
COLDQ	Cost optimization	Process Control Process Redesign
DaQuinCIS	Source trustworthiness Record Linkage	
CDQ	Normalization Record Linkage Data and schema integration Error localization and correction	Process Control Process Redesign

TABLE 2.4: Methodologies and types of strategies (from Batini et al. (2009))

Data-driven techniques

Normalization techniques have been proposed in several domains, within the selected methodologies TIQM, ISTAT and CDQ provide normalization techniques. Normalization techniques improve data quality by comparing data with look-up tables and defining a common metaschema (Batini et al., 2009), in the ISTAT methodology the national street registry is used as a lookup table for territorial data. Record Linkage, a technique that has been investigated in database research since the 1950's and has been used in contexts that require efficient computer-assisted matching procedures to reduce personnel resources and minimize errors in matching (such as healthcare and administrative domains). The CDQ methodology mentions three types of record linkage techniques:

1. Probabilistic techniques; statistical and probability theory (for example: Bayesian networks, data mining)
2. Empirical techniques; algorithmic techniques (for example: sorting, tree analysis)
3. Knowledge-based techniques; extracting knowledge from files and apply reasoning strategies

DaQuinCIS performs record linkage in two phases: first the copies of same entities in different data sources are aligned. Secondly, identical instances are identified by the query results, returned by each data source.

'Data and schema integration' is a broad research area that overlaps with data quality. 'Data-driven improvement techniques applied in the methodologies are often based on the use of new data to improve the quality of a given data collection' (Lenzerini, 2002). Data quality improvement techniques focus primarily on instance-level heterogeneities, in order to identify similar records, detect conflicting values and select a single final instance (Batini et al., 2009).

Process-driven techniques

Methodologies that include process redesign as data quality improvement step tend to use techniques originating from the field of business process reengineering (BPR). BPR focuses to redesign value added processes in a company: "Reengineering is the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance such as cost, quality, service and speed" (Hammer & Champy, 1993).

TDQM is the only data quality methodology that proposes an original process redesign approach, in this methodology the IP-MAP model is introduced (Shankaranarayan, Wang, & Ziad, 2000). The IP-MAP (Information Production MAP) model is used to model the information products managed by the manufacturing processes; the IP MAP is a graphical solution that helps analysts visualize the information production process. The IP-MAP is a complex model and cannot always be applied due to high costs or unfeasibility (because of a thorough process modeling step), for this reason other methodologies adopt less formal solutions. The CDQ methodology is based on a set a matrices that describe relationships among data, information flows, processes, and organizational units.

Measuring data quality dimensions

Most of the data quality methodologies include the assessment of data quality. In this section the measurement of data quality dimensions in the methodologies is elaborated on. Different approaches are used in the methodologies, applying both objective and subjective techniques to measure data quality. The first data quality category that is discussed is the intrinsic data quality, followed by the contextual, accessibility, and representational data quality categories.

Intrinsic data quality consists of four dimensions that are included in several of the methodologies discussed by Batini et al. (2009), these dimensions are: accuracy, objectivity, believability, and reputation.

Accuracy, accuracy is the DQ dimension that is used for data quality assessment in all 13 data quality methodology included in the research from Batini et al. (2009). There are some differences in the measurement among these DQ methodologies, the majority of the methodologies (9 of the 13) choose to measure it by 'Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one'. An alternative measurement used in the Data warehouse Quality Methodology measures the number of delivered accurate tuples. The last option, used by AIMQ, CIHI, and IQM, is the use of user surveys to determine the accuracy.

Objectivity, this dimension is only used in one of the methodologies from Batini's (2009) comparison, namely AIMQ. AIMQ relies solely on user surveys (questionnaires) to determine quality, and focuses the output on benchmarking. DQ dimensions are classified according to their conformity to specifications and conformity to users' expectations, this classification is called the PSP/IQ model (Kahn et al., 2002). The measures for objectivity proposed by Lee, Strong, Kahn, and Wang (2002)) are: this information was

objectively collected, this information is based on facts, this information is objective, this information presents an impartial view.

Believability, this data quality dimension is only included in the AIMQ methodology as well. This dimension is measured using questionnaires. Lee et al. (2002) proposes the following four measures used in AIMQ: this information is believable, this information is of doubtful credibility, this information is trustworthy, this information is credible. The measures are rated with a score from 0 to 10 where 0 means 'not at all' and 10 means 'completely'.

Reputation, the reputation data quality dimensions is also a subjective dimension that can only be measured using user surveys (questionnaires). AIMQ includes this DQ dimension in the data quality assessment methodology. Reputation measures by Lee et al. (2002): this information has a poor reputation for quality, this information has a good reputation, this information has a reputation for quality, this information comes from good sources.

The data quality dimensions that describe intrinsic data quality are mostly measured by user surveys, which leads to a subjective quality assessment. Dimensions as objectivity, believability and reputation cannot be expressed in objective measurements therefore questionnaires have to be used. Information quality is classified using the PSP/IQ model using the classifications: sound, dependable, useful, and usable. Sound information is tangible and measurable against specifications, dependable information is current, secure and provided in a timely manner to support the task at hand. Useful information dimensions are task dependent characteristics, relevant to the consumer's task and sufficient to support decision making, usable information IQ dimensions are evaluated from a consumer point of view, often intangible and difficult to measure (Kahn et al., 2002).

The following five data quality dimensions relate to the contextual data quality of the dataset, these dimensions measure how the data fits the scope of the application (relevancy, value-added, timeliness).

Relevancy, a subjective measurement where the data consumer determines to what extent the data is relevant. This dimension is measured with user surveys, from the methodologies listed by Batini et al. (2009) there is only one methodology measuring relevancy; namely AIMQ. AIMQ includes four measures for relevancy: this information is useful to our work, this information is relevant to our work, this information is appropriate for our work, this information is applicable to our work. Just as in the AIMQ measures mentioned before, the relevancy measure are rated on a scale from 0 to 10 ('not at all' to 'completely').

	Conforms to specifications	**Meets or exceeds consumer expectations**
Product quality	Sound information - Free-of-error - Concise representation - Completeness -Consistent representation	Useful information - Appropriate amount - Relevancy - Understandability - Interpretability - Objectivity
Service quality	Dependable information - Timeliness - Security	Usable information - Believabiltiy - Accessibility - Ease of manipulation - Reputation - Value-added

TABLE 2.5: Information quality dimensions in the PSP/IQ model by Kahn et al. (2002)

Completeness. Completeness is a DQ dimension that can be either measured objectively using metrics, or be assessed with user surveys. 12 from the 13 methodologies from Batini's research (2009) use the dimension for completeness. The majority of these methodologies use an objective metric where completeness is measured as: *"Number of not null values/total number of values"*. An alternative metric for completeness is: *"Number of tuples delivered/expected number"*. The measures proposed by Lee et al. (2002) in the AIMQ methodology question the completeness more subjectively: this information includes all necessary values, this information is incomplete, this information is complete, this information is sufficiently complete for our needs, this information covers the needs of our tasks, this information has sufficient breadth and depth for our task.

Value-added, is described by Kahn et al. (2002) as: *"the extent to which information is beneficial and provides advantages from its use"*. This dimension can only be measured using subjective measures; user surveys. The opposite question can also be asked: to what extent the information provides a disadvantage.

Timeliness, a dimension that can be measured using objective metrics or with user surveys (subjective). There are two metrics used to measure timelines, the first relies on other quality dimensions: *"Timeliness = (max (0; 1-Currency/Volatility)S"*. The other measure is listed as: *"Percentage of process executions able to be performed within the required time frame"*. The alternative to using metrics is using a questionnaire, as is applied in AIMQ and CIHI methodologies.

Amount of data, the appropriate amount of data is measureable with a metric or with user surveys (questionnaire). The metric that is proposed in the Data Quality Assessment methodology by Pipino et al. (2002), is defined as follows: *"Appropriate amount of data = Min ((Number of data units provided/Number of data units needed); (Number of data units needed/Number of data units provided))"*. The measures included in the questionnaire from AIMQ are (Lee et al., 2002): this information is of sufficient volume for our needs, the amount of information does not match our needs, the amount of information is not sufficient for our needs, the amount of information is neither too much nor too little.

The accessibility category for data quality is divided into two dimensions: accessibility and access security. Both dimensions can be measured with either an objective or subjective approach, which is elaborated on below.

Accessibility. The data quality dimension accessibility is included in four of the thirteen data quality methodologies, of which two use an objective measurement and two methodologies use subjective surveys. The Information Quality Measurement (IQM) methodology applies an objective measurement where accessibility is measured by the time it takes for requests to deliver (on a scale from 0 to 1). This is expressed as: *Accessibility = max (0; 1-(Delivery time - Request time) / (Deadline time – Request time))* The ISTAT methodology measures the difference between broken links and broken anchors instead. The remaining two methodologies (TIQM and AIMQ) apply user surveys to compose a quality score for accessibility. The AIMQ methodology proposed four measures that used in the survey to compose a data quality score ranging from 0 to 10, these four questions are: *"This information is easily retrievable, this information is easily accessible, this information is easily obtainable and, this information is quickly accessible when needed."*

Access security is included in only two data quality methodologies, of which one measures this dimensions objectively (IQM) and the other method subjectively (AIMQ). The Information Quality Measurement (IQM) methodology measures the number of weak log-ins to gain an insight into the access security aspect of data quality. The AIMQ methodology poses four questions in the user survey to create a data quality score (on a scale from 0 to 10): *"This information is protected against unauthorized access, this information is not protected with adequate security, access to this information is sufficiently restricted and, this information can only be accessed by people who should see it."*

The representational data quality category consists of four dimensions, each cover a specific part how data in the data set is represented: interpretability, ease of understanding, concise representation and, consistent representation.

Interpretability is one of the dimensions used in the representational data quality category. This dimensions is included in two methods from the research of Batini et al. (2009). In one of the methods, DWQ, intepretability is expressed objectively by the 'number of tuples with interpretable data, and documentation for key values'. In the second methodology, 'A methodology for information quality assessment' (AIMQ), a subjective measurement is applied, five measure are used that combines form a data quality score for interpretability on a scale from 0 to 10. The five measure are as follows: *"It is easy to interpret what this information means, this information is difficult to interpret, it is difficult to interpret the coded information, this information is easily interpretable and, the measurement units for this information are clear."*

Ease of understanding is only included in one of the thirteen methodologies addressed by Batini et al. (2009). The only methodology that provides the measurements for ease of understanding is 'A methodology for information quality assessment' (AIMQ), where five user survey questions are proposed that deliver the data quality score. Within the AIMQ methodology measures are rated on a scale from 0 to 10 here 0 means 'not at all' and 10 means 'completely', the five measures from AIMQ are: *"This information is easy to manipulate to meet our needs, this information is easy to aggregate, this information is difficult to manipulate to meet our needs, this information is difficult to aggregate and, this information is easy to combine with other information."*

Concise representation refers to the clarity and briefness of the representation of the data. This concise representation is measured with a user survey in the AIMQ and IQM methodology. Lee et al. (2002) has composed four measures that used to score the data quality of the conciseness of the data representation: *"This information is formatted compactly, this information is presented concisely, this information is presented in a compact form, and the representation of this information is compact and concise."* The measures in the AIMQ methodology are rated on a scale from 0 to 10, ranging respectively from 'not at all' to 'completely'.

Consistent representation is a data quality dimension that is measured in a number of different way, the four methodologies that apply an objective measure for this dimensions measure consistency of the representation in three different ways. The TDQM and DQA methodology measure the ratio of consistent values relative to the total values: *"Consistency = Number of consistent values/number of total values"*. The Data Warehouse Quality Methodology (DWQ) measures quality by the number of tuples violating constraints and the number of coding differences. IQM chooses to measure 'consistent representation by the number of pages with style guide deviation. Lastly the data quality of consistent representation can also be measured using, subjective, user surveys. This approach is used by the CIHI and AIMQ methodologies, where Lee et al. (2002)

proposes the follow measures in his AIMQ methodology: *"This information is consistently presented in the same format, this information is not presented consistently, this information is presented consistently, and this information is represented in a consistent format."*

Concluding

Data quality can divided into four main categories, each of these categories (intrinsic, contextual, representational, and accessibility data quality) describes a different aspect of the data quality. The four data quality categories can be expressed using data quality dimensions, for each of the four categories there are some matching data quality dimensions that describe a more specific part of the data quality (such as accuracy, or believability). The data quality dimensions are measurable using data quality measures appropriate to the dimension, the quality scores of all (relevant) data quality dimensions combined determine the level of data quality.

Chapter 3

Semantic Integration

This chapter is used as a theoretical basis for semantic integration, with the purpose of giving an insight how semantic integration can be executed. Starting with the origin of semantics in literature, this is followed by a segment on the theory of ontologies. Lastly methods to build ontologies are introduced into detail.

Semantics is defined as 'the branch of linguistics and logic concerned with meaning', semantic integration is the process of integrating information based on its meaning. In the field of database research semantic integration has been a challenge to the database community since the early 1980's. The use of structured representations in database applications allowed more than one representation of the data, which introduced heterogeneity between schemas and their data (Doan & Halevy, 2005). Resolving these heterogeneities was necessary to enable manipulation of schema and data, and to translate data across schemas. In the field of database research the earliest application involved merging a set of given schemas into a single global schema, this is called schema integration (Pottinger & Bernstein, 2003). The integration process required establishing semantic correspondences, matches, between the given schemas and using these matches to merge the schema elements.

"Combining two models requires first determining correspondences between the two models and then merging the models based on those correspondences. Finding correspondences is called schema matching" (Pottinger & Bernstein, 2003). According to Rahm and Bernstein (2001) the process of schema matching is the first step towards schema integration. Before the semantic matches can be identified the involved schemas have to be able to 'communicate', since schemas are independently developed they often have a different structure and terminology (Rahm & Bernstein, 2001). Apart from that the schemas can also originate from different domains (for example a real-estate schema and

a schema for property taxes), in order to successfully integrate these schemas there has to be interoperability.

3.1 Semantic interoperability

The current information society shows a demand for access to all available information, this information is often heterogeneous and distributed. In order to access this information the relevant sources of information have to work together. The problem with heterogeneous and distributed information systems is that their differences prohibit information sharing; this is the so called interoperability problem (Wache et al., 2001). Interoperability is described using a technical and an informational level; meaning that interoperability requires full accessibility to the data (informational level), and it requires the data to be processed and interpreted by the system (technical level). Two types of problems may occur due to the heterogeneity of the data:

- Structural heterogeneity: different information systems store data using different structures.
- Semantic heterogeneity: heterogeneity in content of an information item and its intended meaning.

To achieve semantic interoperability, the meaning of the information that is interchanged between systems has to be understood. When two contexts do not have the same interpretation of the information, a semantic conflict occurs. Three main causes of semantic interoperability are identified by Goh (1996):

1. Confounding conflicts: information items have the same meaning, but differ in reality. This can happen due to different temporal contexts for example.
2. Scaling conflicts: issue where different reference systems are used to measure a value, an example is measuring a price using different currencies.
3. Naming conflicts: occurs when the naming schemes of information significantly differ, this happens when synonyms or homonyms are used between systems.

To make sure interoperability is possible between systems knowledge has to be explicit. *"The use of ontologies for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity"* (Wache et al., 2001).

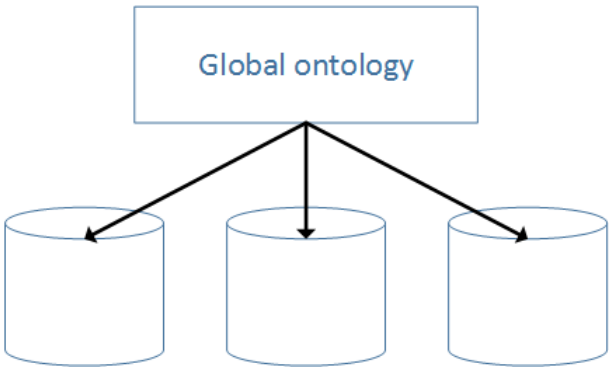

FIGURE 3.1: Single ontology approach

3.2 Ontologies

Gruber introduced ontologies as an "explicit specification of a conceptualization" in (1993). This means ontologies can be used to describe semantics of an information source to make the content explicit. Within the integration of data sources ontologies are used to identify and associate semantically corresponding concepts (semantic matches). Nearly all ontology-based integration approaches use ontologies to explicitly describe the semantics of information sources, however the way ontologies are used can be different. Three different approaches that employ ontologies to make content explicit are: single ontology approaches, multiple ontology approaches, and hybrid approaches.

Single ontology approaches

A single ontology approach uses a single global ontology that provides a shared vocabulary which specifies the semantics of all information sources (Figure 3.1). The global ontology can be the combination of several more specialized ontologies. The single ontology approach is best suited for integration problems where the individual information sources provide a similar view on a domain. The major disadvantage of this approach occurs when one of the information sources has a view on the domain that is very different, by providing another level of granularity for example. Another downside is that the single ontology approach is susceptible for changes in the information sources; this can affect the conceptualization of the domain.

Multiple ontology approach

In the multiple ontologies approach there is an ontology to describe each information source, every ontology contains the semantics of only one information source. The development of the multiple ontologies approach is because of the disadvantages the

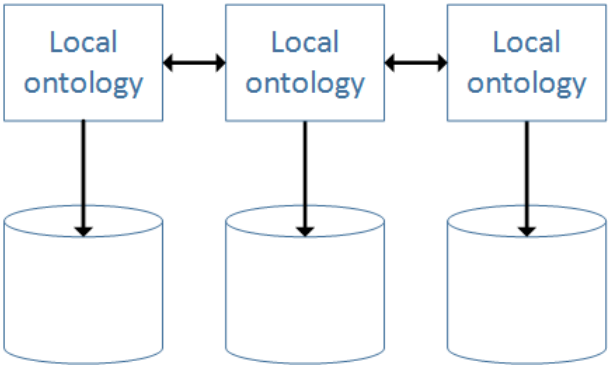

FIGURE 3.2: Multiple ontology approach

single ontology approach delivered. The advantage of using multiple ontologies is no common and minimal ontology commitment about a global ontology is needed (Gruber, 1995). The downside to this approach however, is the lack of a common vocabulary which makes it difficult to compare different source ontologies. This problem can be overcome by inter-ontology mapping, which identifies semantically corresponding terms between the local ontologies. This inter-ontology mapping also has to consider different views on the domain the different ontologies may have, this makes the inter-ontology mapping very hard to define (Wache et al., 2001).

Hybrid ontology approaches

The hybrid ontology approach was developed to overcome drawbacks from both single and multiple ontology approaches. The hybrid approaches describe the semantics of each information source by a local ontology (just as in the multiple ontology approach). But a shared vocabulary is constructed to make the local ontologies comparable to another. The shared vocabulary contains basic terms (primitives) of a domain; these primitives are combined in the local ontologies to describe more complex semantics (Goh, 1996). The advantage of the hybrid approach is that new information sources can easily be added without needing modification. The hybrid approach also supports evolution of ontologies. The drawback of this approach is the reusability of existing ontologies; they have to be redeveloped completely.

There are numerous ways the structure of ontologies can be described, different approaches incorporate different languages and the general structure of the ontology may vary. Many ontology representation languages use a form of description logics (for example CLASSIC or OIL). The different languages express the ontologies differently; the typical language constructs used on the descriptive logic languages vary. An alternative is to use frame-based representation languages for information integration systems that use ontologies. Frame-based languages use different elements to express ontologies

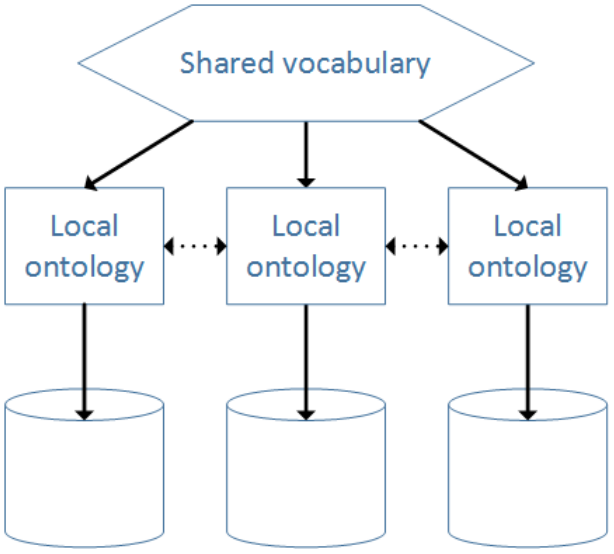

FIGURE 3.3: Hybrid ontology approach

for information integration. Descriptive logic languages provide logical operators to allow concept definition; frame-based languages provide elements to define functions and relations.

3.3 Building ontologies

Ontologies fulfill an important role in information integration activities; therefore it is crucial the ontology development process is supported. In literature there are several approaches to ontology development, for example METHONTOLOGY by (Gómez-Pérez, 1998) or TOVE (Fox, Barbuceanu, & Gruninger, 1996). Some literature proposed methods and phases to the development process that are independent of the domain of the ontology (Uschold & Gruninger, 1996), and (Gómez-Pérez, Fernández, & Vicente, 1996). The four main phases defined by Uschold and Gruninger are as follows:

1. Identifying a purpose and scope
2. Building the ontology
   (a) Ontology capture
   (b) Ontology coding
   (c) Integrating existing ontologies
3. Evaluation
4. Guidelines for each phase

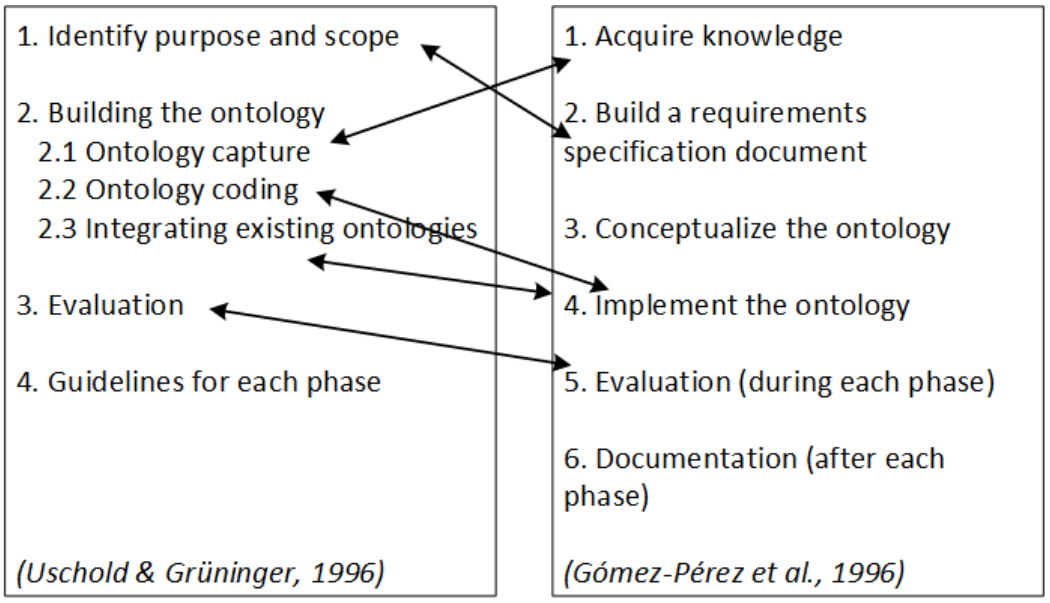

FIGURE 3.4: Relationship between ontology building methods by Fernández-López et al. (1997)

The approach Uschold and Gruninger (1996) apply in their method resembles that from Gómez-Pérez and his colleagues (1996). Figure 4 shows a comparison of both methods and their resemblance by linking the corresponding phases. The figure shows that the method from Uschold and Gruninger does not include a conceptualization phase, and their method provides guidelines whereas Gómez-Pérez et al. only provide documentation.

The phases from the ontology development method from Uschold and Gruninger (1996) are elaborated on below. For this research the most impactful phase is ontology capture, resulting in an ontology expressed in a natural language, the ontology capture step is discussed into detail later on.

Purpose and scope: The initial phase to the ontology development process is meant to provide clarity on why the ontology is being built. In this phase the intended uses of the ontology are identified, it can also be useful to identify and characterize the users of the ontology. The identification of purpose and scope is used as a target for building the ontology (Uschold & Gruninger, 1996).

The phase to *building the ontology* is divided in multiple steps, each step focuses on a different aspect of the building process: capture, coding, and integration. Ontology capture includes three processes that capture the conceptualization of a vocabulary. The first process is identification of key concepts and relationships in the domain of interest, followed by the production of precise unambiguous textual definitions for those concepts

and relations. The last process is the identification of terms to refer to the concepts and relationships, and agreeing upon all previous steps.

Ontology coding: the purpose of the coding step is to make the representation of the conceptualization explicit using some formal language. Processes within the coding are: committing to basic terms that will be used to specify the ontology (this is also called a meta-ontology), choosing a representation language, writing the code. Coding and capture processes are sometimes combined in a single step, developing the conceptualization while building the ontology.

Integrating existing ontologies is used to overcome the problem of whether and how to use (parts of) ontologies that already exist. Identifying synonyms and extending an ontology is rather easy, but when there are similar concepts defined in the existing ontology is often unclear if these concepts can be adapted and reused, and how.

Evaluation: "Evaluation means to carry out a technical judgement of the ontologies, their software environment and documentation with respect to a frame of reference" (Fernández-López et al., 1997). During the evaluation step the two main activities are verification and validation. Verification refers to the technical process that guarantees the correctness of an ontology, validation is the assessment whether the ontologies, the software environment and documentation correspond to the system they are supposed to represent. The eventual result is an evaluation document that states how well the designed ontology fits the specified requirements.

Documentation: one of the main barriers to effective knowledge sharing is the inadequate documentation of existing knowledge bases and ontologies (Skuce, 1995). In order to prevent this barrier it is important to document all assumptions, for both the ontology as for the primitives used in the ontology (the meta-ontology).

Guidelines for ontology design: a comprehensive methodology for building ontologies should include a set of techniques, methods, principles for the four phases of ontology design, and the relations between those phases (Uschold & Gruninger, 1996). Gruber (1995) describes principles to developing ontologies, these principles are summarized into several design criteria (clarity, coherence, extensibility, minimal ontological commitment, minimal encoding bias).

- *Clarity*: an ontology is supposed to effectively communicate the intended distinctions to the user. Ambiguity should be minimized, distinctions should be motivated, and exampled should be provided to help the user understand definitions that lack a specification using formal axioms. Definitions should always be documented using natural language and examples to clarify the intent.

- *Coherence*: it is important for an ontology to be internally consistent. Definitions should use the same logic (the axioms should have be logically consistent). Besides the consistency in the used axioms, coherence also applies to the natural language used in documentation and examples.
- *Extensibility*: an ontology should be designed to support shared vocabulary uses; extending and specializing the existing ontology. One should be able to define new terms (for special uses) based on the existing vocabulary, without the need to revise existing definitions. Extensibility is supported by the criteria 'minimal ontological commitment' and 'minimal encoding bias'.
  - *Minimal ontological commitment*: an ontology is supposed to have the minimal ontological commitment required to support the intended knowledge sharing activities. An ontology serves a different purpose than a knowledge base, with a different notion of representation and completeness. An ontology should make as few claims as possible about the world being modelled, allowing parties committed to the ontology the freedom to specialize the ontology as needed (Uschold & Gruninger, 1996).
  - *Minimal encoding bias*: the conceptualization should be specified at a knowledge-level without depending on a particular symbol-level encoding. The purpose of minimizing the encoding bias is to enable knowledge sharing across agents with varying representation systems or styles of representation.

3.3.1 Ontology capture

Ontology capture consists of identifying and defining the important concepts and terms. The output of ontology capture depends on the techniques used; informal techniques can result in a 'semi-formal' ontology expressed using natural language or a formal approach that is expressed using formal axioms. Using informal techniques, the procedure for capturing an ontology can be described using four phases: scoping, producing definitions, review, and development of a meta-ontology.

The first phase, scoping, uses the techniques brainstorming and grouping to respectively find (potentially) relevant terms, and structure the terms loosely in work areas. Part of the grouping is to identify semantic cross-references between the work areas, for example concepts that are likely to refer to (or be referred to) by concepts in other work areas (Uschold & Gruninger, 1996).

After the important concepts and terms have been identified, the main work of building an ontology is producing definitions. Uschold and Gruninger (1996) propose to start

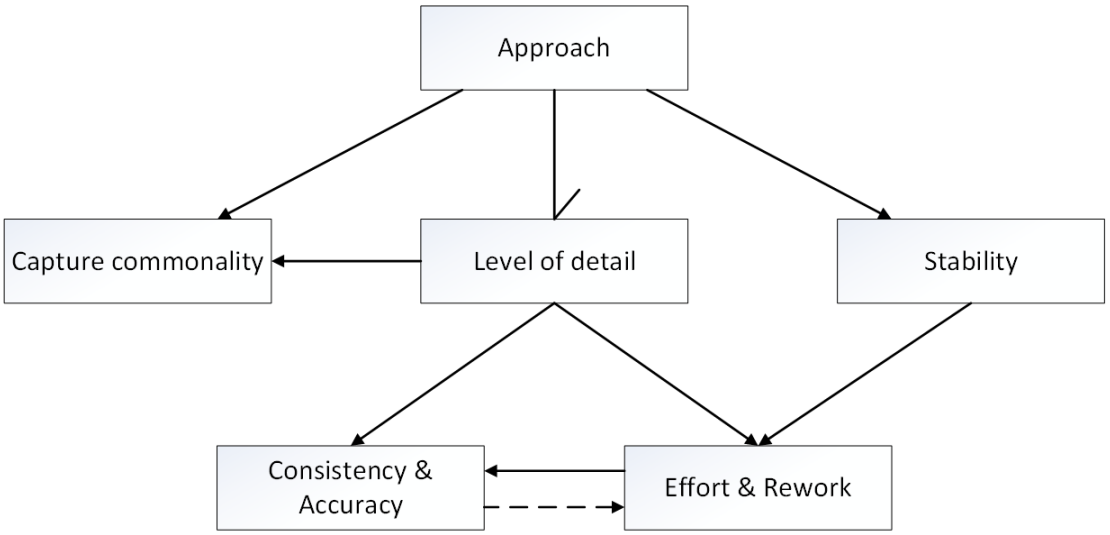

FIGURE 3.5: Middle-out approach, taken from Uschold and Gruninger (1996)

in the work areas with the most overlap, to define the terms with the dependencies. Mistakes in these areas lead to more re-work. For the definition of terms a middle-out approach is suggested, meaning the most fundamental terms in each work area are defined first, before continuing with more abstract and specific terms. For example: 'dog' is a fundamental term, 'mammal' is a generalization (abstraction), and 'cocker spaniel' is a specialization (Uschold & Gruninger, 1996). Using a middle-out approach has a number of advantages compared to bottom-up or top-down alternatives. A bottom-up approach does result in a very high level of detail, the downside is that this takes far more effort, makes it difficult to spot commonalities between related concepts, and increases risk of inconsistencies. A top-down approach has a better control of the level of detail, but starting from the top may result in arbitrary high-level categories. The emphasis on dividing up high level concepts results in missing the commonality of inter-connected concepts.

A middle-out approach offers a balance in the level of detail; fundamental concepts (or basic concepts) are only specialized when necessary. By starting with the most important concepts, and defining higher level concepts in terms of the basic concept, the high level categories naturally arise and are more likely to be stable (Uschold & Gruninger, 1996), the middle-out approach is shown in figure 5.

Executing the middle-out approach, to define the concepts into terms, results in a list of commonalities between concepts and their definitions. The following step is to reach an agreement on the definitions and terms for underlying concepts. It can occur that several terms, with different usage, correspond to one concept definition. Another option is that ambiguous terms are used to correspond to several, but different concepts. Uschold and Gruninger (1996) state several options to handle ambiguous terms, for example: suspend

use of the term, determine which concept is important enough to be in the ontology (usually one), or choose a different term for the concept ('thing' rather than object or entity). The review phase will critically review definitions, revise definitions and keep track of the changes in a set of historical notes.

The last phase in ontology capture is the development of a meta-ontology. The meta-ontology consists of the main terms and concepts used to define the ontology; in the Enterprise Ontology (Uschold, King, Moralee, & Zorgios, 1998) the meta-ontology contains entities, relationships and state of affairs. Where an entity is a fundamental thing in the domain being modelled (such as a person or a plan), a relationship is the association two or more entities have with each other (a relationship can be a sale between two legal entities that exchange a product for a sale price). A state of affairs is a situation characterized by any combination of entities being in any number of relationships with one another (Uschold et al., 1998). The relationship between the entities can be modelled using an Entity-Relationship diagram (ERD).

Concluding

The key element to semantic integration is to identify semantic interoperability between two data sources. To make sure that the two data sources are interoperable there is a need for a common, explicit, vocabulary; the ontology. A shared ontology ensures that values from different datasets have the same meaning, and therefore can be successfully integrated. The approach to building an ontology is taken from Uschold and Gruninger (1996), this method consists of four main steps: identify purpose and scope, building the ontology, evaluation, and setting up guidelines for each phase.

Chapter 4

Research Method

For this research a method is proposed where the integration of data is performed while adhering to data quality improvement. The designed method applies techniques and approaches from the field of data quality assessment and semantic data integration and combines best practices of both research fields. First step of the method is to assess data quality of the relevant datasets that are selected to integrate; the individual datasets should have the same domain to improve the success of integration. The second step consists of creating a shared ontology for all included datasets, during the third step the shared ontology is used to determine semantic matches between the datasets. This is followed by the fourth step, integrating the data using semantic integration techniques. The last step consists of another data quality assessment; this will be used to determine if the data quality has improved.

1. Assess data quality of individual datasets
2. Create a shared ontology
3. Perform Semantic matching
4. Perform Semantic integration
5. Evaluate data quality improvement

The method is visualized using a Process-Deliverable Diagram (PDD) in Figure 4.1, this meta-modelling technique is described by van de Weerd and Brinkkemper (2008). A PDD consists of two integrated diagrams; on the left side a process view, and on the right a view with corresponding deliverables. The overview PDD of the research method is listed on the next page (Figure 4.1).

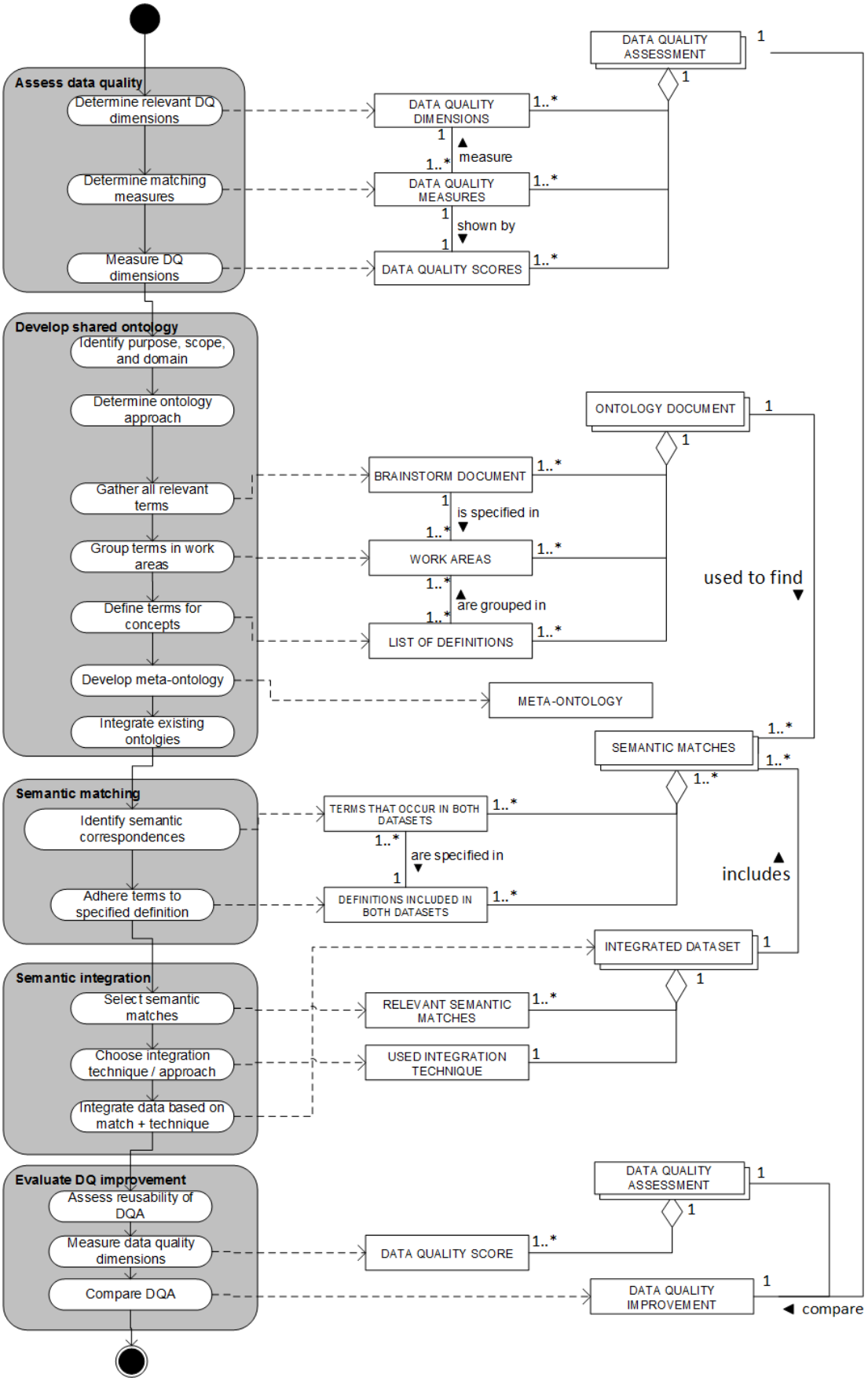

FIGURE 4.1: PDD of the method overview

The process-deliverable diagram in Figure 4.1 shows the deliverables corresponding to the processes from the method for this research. The figure includes the five main processes of the research method with their corresponding deliverables and the relations between the deliverables. The relation between the ontology document and the semantic matches is based on the fact that the shared ontology is used for all datasets to determine semantic matches between datasets. This relation is a 'one-to-many' relationship, meaning that all semantic matches are related to one ontology (this makes sense since there is only one ontology developed), and from this single ontology zero or more semantic matches can be discovered. The semantic matches are related to the integrated dataset in a 'one-to-many' relationship, all semantic matches are included to form the integrated dataset, and the integrated dataset includes one or more semantic matches (if there were zero semantic matches, the datasets cannot be integrated using the semantics). The integrated dataset also has relations with both the initial data quality assessments and the data quality assessment that evaluated the integrated dataset. The deliverable data quality improvement includes a comparison of the data quality assessments of the initial data sources and the data quality assessment of the integrated dataset. This comparison is required to evaluate whether the quality of the relevant data quality dimensions has improved by the semantic integration process.

4.1 Assess data quality

In order to assess the data quality of a set of data, a data quality assessment methodology will be used. The first step is to assess the available data quality methodologies (DQM) to find the methodology with the best fit, this will be based on the data quality dimensions that are measured in the DQM. Data quality dimensions can be measured using two distinct approaches: metrics or user surveys. DQ dimensions are grouped according to the conceptual data quality model by Wang and Strong (1996) to measure a specific aspect of the dataset.

According to the conceptual data quality model (Wang & Strong, 1996), four categories are used to describe data quality: intrinsic DQ, contextual DQ, representational DQ, and accessibility DQ. Each of the categories can be measured using the data quality dimensions used in data quality assessment methodologies. The categorization of quality dimensions can give an insight in to what aspect of data quality each dimension contributes. For the purposes of this research, where the data quality of individual datasets is assessed the importance of each category can be made to decide what data quality dimensions have to be measured. The most important data quality dimensions for this

research rely on the data itself; to be able to compare multiple datasets independent of their real world application.

- Intrinsic data quality: the data quality dimensions related to the data itself (accuracy, objectivity, believability, and reputation) this category is crucial to determine the quality of a dataset.
- Contextual data quality: data quality dimensions that assess quality based on the context of the data (relevancy, completeness, value-added, timeliness, and amount of data). Importance depends on the domain of the data; the measurement of these DQ dimensions determines how well the dataset fits the scope, not necessarily the quality of data.
- Representational data quality: dimensions that assess quality based on the representation of data to the data consumer, in both format of the data (concise and consistent representation), as well as meaning of the data (interpretability, and ease of understanding). The importance of representation of data is not crucial for the application of this research.
- Accessibility data quality: data quality dimensions that describe the availability of the data (accessibility, and access security). To the data quality assessment focused on the dataset itself, this category is of little importance.

Based on the categories for data quality and their importance, the DQ dimensions corresponding to these categories can be listed based on importance as well. The highest priority is data quality dimensions relating to the intrinsic data quality, followed by contextual data quality dimensions. The remaining two categories (representational DQ and accessibility DQ) are of lesser importance for the purpose of this research, single dimensions from these categories can be used if they are relevant to the situation.

The data quality dimensions for each of the four data quality categories are listed in the chapter on data quality (Chapter 2). For the two most important categories the relevant data quality dimensions are as follows.

- **Intrinsic data quality**: Accuracy, Objectivity, Believability, and Reputation.
- **Contextual data quality**: Relevancy, Value-Added, Timeliness, Completeness, and Amount of Data.

These dimensions have corresponding data quality measures that are used to provide a measurable data quality score for the related data quality dimensions. The data quality

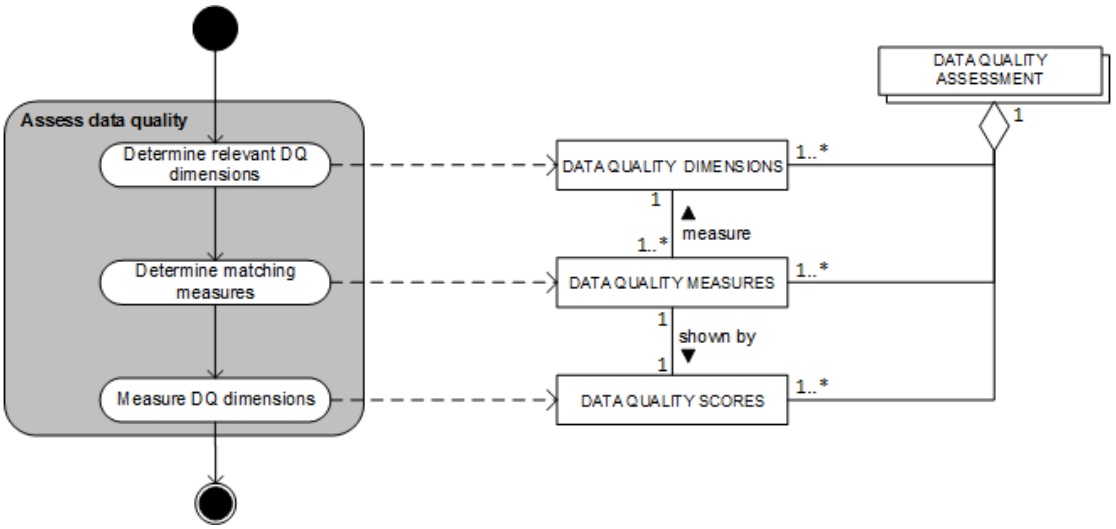

FIGURE 4.2: PDD for the 'Assess data quality'-step in the method

measures can be either objectively or subjectively measured, in Chapter 2 an overview of data quality measures used by several data quality methodologies is provided (Batini et al., 2009).

All the data quality measures relate to the data quality of the dataset as a whole (the data quality score for Completeness relates to the entirety of the dataset, not specified on a specific part of the dataset). In this research regarding integration of data, it is important to pinpoint specific weaknesses within a dataset that can be overcome by integrating an additional dataset. For this reason additional questions will be asked to experts on each dataset to determine specific weaknesses (for example: data based on trends or estimations, outdated data). Additional questions may refer to data quality dimensions that are not covered in the dimensions above, such as volatility regarding outdated data, and correctness for invalid data. Additionally, insights on completeness can be collected, additional information that regards to missing values rather than 'not null values', as is stated by the metric from the DQ methodologies described by Batini et al. (2009). Furthermore, the subjective measurement of completeness from the AIMQ methodology only regards to the dataset as a whole, it does not specify where data is incomplete or insufficient.

The process for assessing data quality, with the result of a data quality assessment for each dataset, is displayed in the form of a Process-Deliverable Diagram in Figure 4.2. The three main processes are: determining the relevant data quality dimensions, finding matching measures for these dimensions, and lastly measuring the data quality accordingly.

4.2 Develop a shared ontology

The first step for the development of an ontology is to determine the purpose and scope, and the domain of the individual datasets. The purpose and scope are defined to gain an insight on why the ontology is being built. The intended uses of the ontology are identified (finding semantic matches for this application) and it is useful to identify and characterize the users (Uschold & Gruninger, 1996).

The analysis of the domain of each dataset helps to choose an ontology approach. There are three ontology approaches proposed by Wache and his colleagues (2001): single ontology approach, multiple ontology approach, hybrid ontology approach. When the view on the domain is similar between the data sources the easiest approach can be applied: single ontology approach. This approach adopts a single ontology that provides a shared vocabulary that describes the semantics of all information sources. When there is little similarity in the domains of the information sources, the multiple ontology or hybrid ontology approach is more suited. Within the multiple and hybrid ontology approach each information source is described by its own ontology. The main difference between multiple and hybrid approach is the use of a shared vocabulary in the hybrid approach, whereas the multiple ontology approach only applies inter-ontology mapping.

The choice for multiple or hybrid ontology approach depends on the dependencies of the information sources. The inter-ontology mappings are very hard to define when there are different views on the domain the ontologies describe. The shared vocabulary prevents the difficult mappings by an overarching shared vocabulary, but the hybrid approach does not support the reusability of existing ontologies; they have to be redeveloped completely. When the best fitting ontology approach has been determined, the following step is the building of the ontology. For the building of an ontology the method from Uschold and Gruninger (1996) is followed. The building phase consists of three main steps: ontology capture, ontology coding, and integrating existing ontologies. These steps will be elaborated on into more detail.

Ontology capture

During the ontology capture step the important concepts and terms are identified and defined. This process starts with scoping to find all relevant terms to the ontology and structure these terms into work areas. Grouping techniques can be applied to determine work areas and to semantically cross-reference related work areas (Uschold & Gruninger, 1996). After identifying the important concepts and terms the main objective is to produce definitions, starting from the work areas where there is the most

overlap, to define terms with the most dependencies first. Uschold and Gruninger (1996) propose several guidelines towards clarity and coherence for definitions; this is explained into more detail later on. Producing definitions will be done by using a middle-out approach; first fundamental terms are defined, and after that abstract and more specific terms. A middle-out approach is the most preferred approach since the focus of the middle-out approach lies with defining the most fundamental terms first, which is useful for building an ontology. This approach will eventually result in a list of definitions and the relations with the information sources; this list helps to determine semantically overlapping concepts in the later phase where semantic matches are identified.

Ontology coding

The purpose of the coding step is to make the representation of the conceptualization explicit in some formal language. Two important steps are to choose a representation language and develop the meta-ontology (to show the relations between concepts in the ontology). In this research the goal is to compare two separate information sources, the meta-ontology helps to show the relationships within the ontology and can determine how concepts from the different information source relate to each other. The meta-ontology can be modelled as an Entity-Relationship diagram (ERD).

Integrating existing ontologies

The step of integrating already existing ontologies is only relevant when the data source has an ontology prior to this project. This ontology may be reused, although the concepts defined in this ontology have to be adapted for reuse where possible. In the research from Gruber (1995) there are several principles provided for the ontology design, these principles are summarized into design criteria by Uschold and Gruninger (1996): clarity, coherence, extensibility (minimal ontological commitment, and minimal encoding bias). The design criteria provide clear guidelines for the ontology building process and definition of terms.

- Clarity: Minimize ambiguity, motivate distinctions, and provide examples to help the user understand definitions that lack specifications.
- Coherence: Definitions should use the same logic, consistency in the used axioms, and coherent use of natural language in documentation and examples.
- Extensibility: the ontology should be designed to support shared vocabulary uses; extending and specializing the existing ontology. Supported by the criteria:

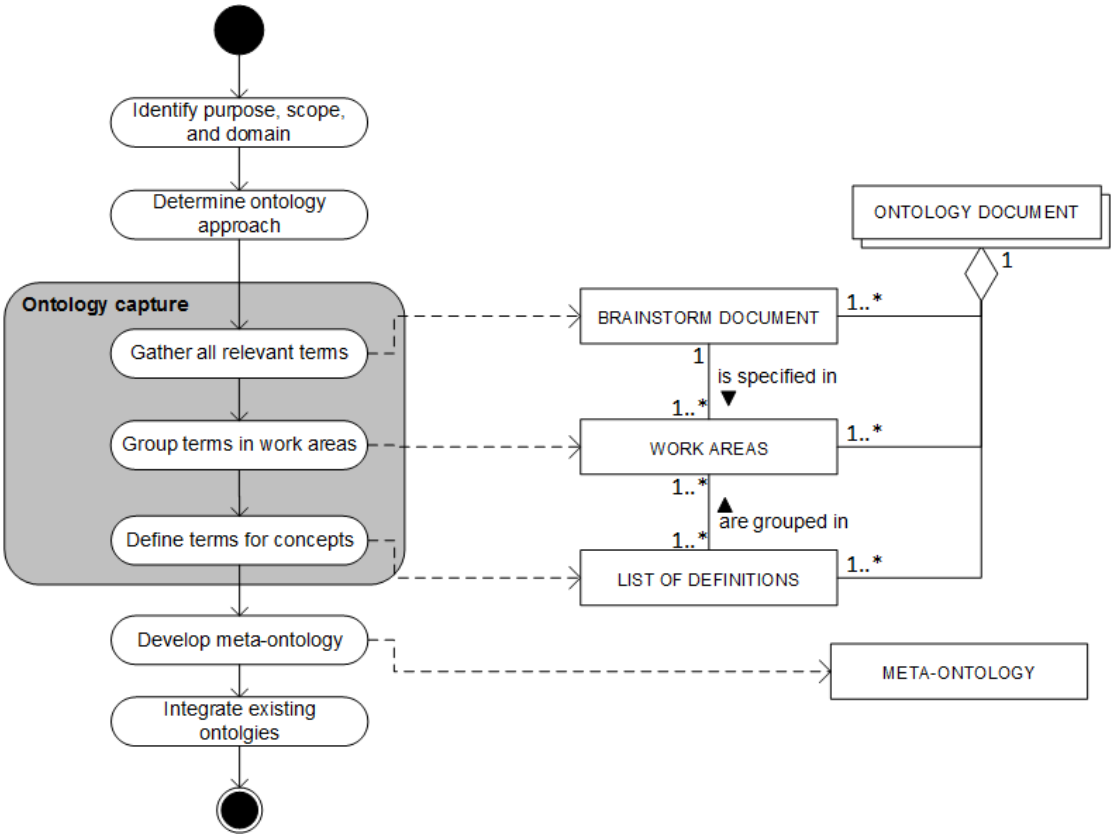

FIGURE 4.3: PDD for the 'Develop a shared ontology'-step in the method

- Minimal ontological commitment: an ontology is supposed to have a minimal ontological commitment required to support the intended knowledge sharing activities. The ontology should make as few claims as possible.
- Minimal encoding bias: the conceptualization should be specified on a knowledge-level not depending on a specific symbol-level encoding.

4.3 Semantic matching

The built ontology, more specifically the resulting ontology documents, provides a solid basis to determine semantic matches between the information sources. The list of definitions that is created during the ontology capture phase also provides the concepts from the dataset that relate to that definition. The concepts from the individual data sources do not always relate in a one-to-one relation to the definition, between some sources there is additional data modification required to reach semantic interoperability. For example the definition for the measurement of distance can be defined as kilometers (since this may be the standard for that application), in one dataset the distance is listed in kilometers in another dataset this can be listed in meters. Both concepts relate

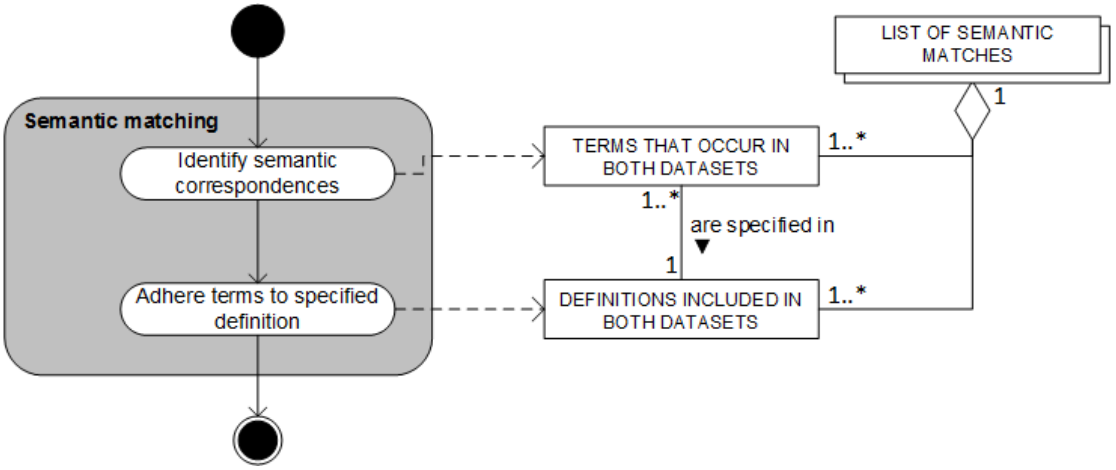

FIGURE 4.4: PDD for the 'Semantic Matching'-step in the method

to the same notion of distance but are not directly interchangeable (the second dataset, measured in meters, has to be modified into kilometers).

The deliverable from the semantic matching is a list of all definitions that relate to concepts from multiple information sources. This list includes the concepts of all datasets that relate to this definition and the additional data modification to adhere to the standard of this definition. For example values from a data source using meters as a measure have to be divided by 1000 to be usable for a standard measured in kilometers.

4.4 Semantic integration

The step of semantic integration applies the list of semantic matches to combine multiple data sources into a single integrated dataset. The deliverable from the 'semantic matching' activity includes what modifications are necessary for concepts to adhere to the set definitions. The list of semantic matches adhering to the definitions make sure that data values are semantically interchangeable, and are used to combine actual data values. The concepts and definitions do not necessarily refer to a single column from a database schema; sometimes multiple columns have to be adapted to make data integration possible. The resulting integrated dataset is the result of the selected semantic matches and the technique or approach that is chosen to integrate the datasets.

4.5 Evaluate data quality improvement

The final step, evaluation, is used to assess the data quality of the resulting, integrated dataset. The purpose of the method is to improve the data quality by applying semantic

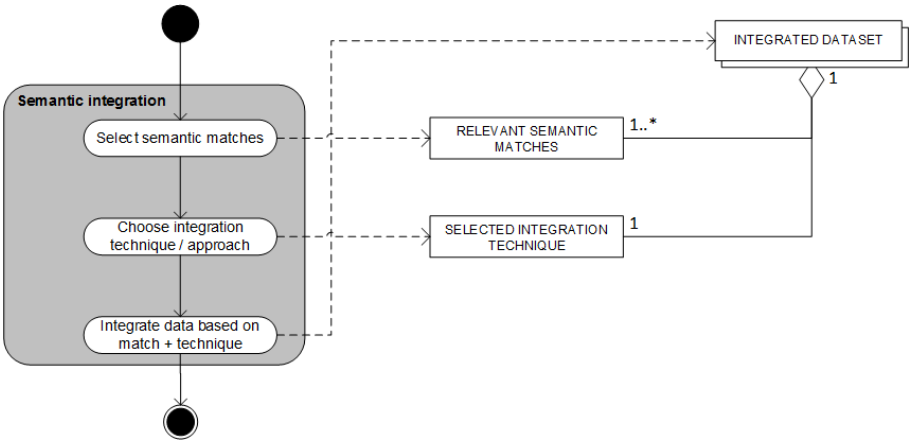

FIGURE 4.5: PDD for the 'Semantic integration'-step in the method

integration techniques, the evaluation step is used to check whether the data quality has improved in comparison to the individual datasets.

The data quality dimensions that are relevant for the project are determined in the step that assesses data quality for the individual datasets (Figure 4.2). The first process assesses the reusabiltiy of the data quality assessment that was used to evaluate the data quality of the initial data sources. The following step is to assess the data quality of the integrated dataset, if the same assessment of quality can be used as in the initial DQA, the data quality scores can be directly compared. When an alternative data quality assessment is applied the data quality requires additional analysis to evaluate data quality improvement. The final step of the evaluation of data quality improvement compares the data quality of the initial datasets with the data quality from the integrated data set to assess where the quality has been improved (or not). The process is visualized in a Process-Deliverable Diagram (PDD) in Figure 4.6, where the data quality dimensions and measures are derived from the 'Assess Data Quality' step.

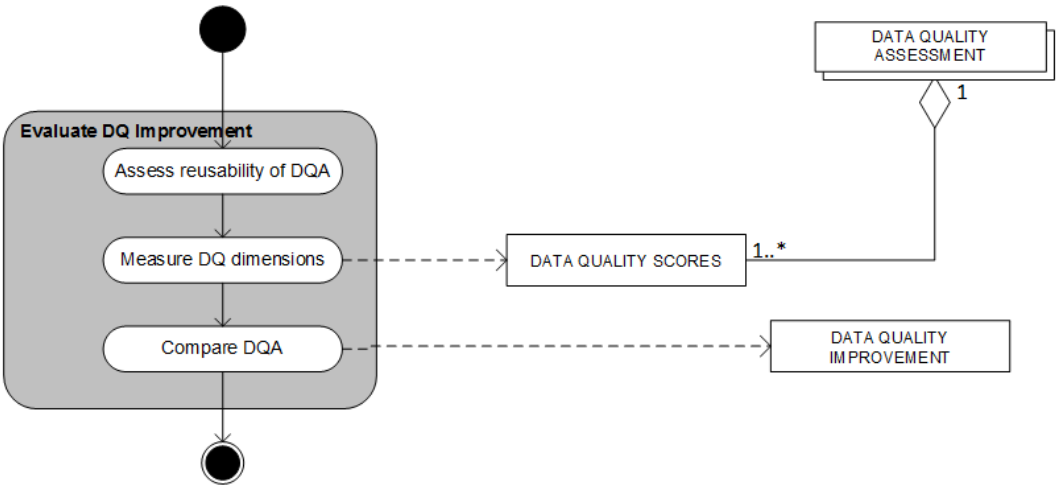

FIGURE 4.6: PDD for the 'Evaluation'-step in the method

Chapter 5

Research method evaluation

5.1 Introduction

This research strives to find a new method to measure carbon emission caused by traffic and to visualize their impact. Currently the carbon emission is calculated based on the total distance travelled using certain modes of transportation, this results in a yearly figure that can only be used to compare the emission to previous years to see the changes. The introduction of a method that can provide more specific information regarding carbon emissions can help to develop plans and policies to target specific sources of emission. This research focuses to collect and visualize carbon emission based on the location travel movements occur, this will result in an overview of carbon emissions per city; the so called mobility carbon footprint. Carbon footprints of cities can be compared and outliers can be identified (for example a small city that produces more emission than similar cities).

To correctly determine the carbon footprint of cities (or areas) in the Netherlands it is crucial to have access to a source of data which contains all travel movements. In this research two data sources are used, these datasets individually do not provide sufficient travel information to develop the desired carbon footprint. Therefore the proposed method will be applied to integrate the datasets and create a new information source which will provide a more complete insight in the travel movements. For the application of the method two information sources are used: a dataset from the company Mezuro, and a dataset by the Dutch statistics bureau (CBS) called OViN.

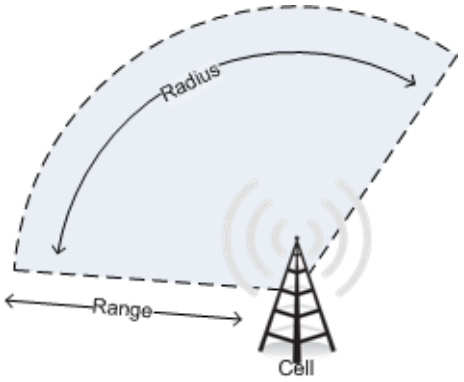

FIGURE 5.1: A cell tower and its coverage area

Mezuro

Mezuro is a company specialized in mobility using mobile phone location data, licensed from a major Dutch telecom provider. The telecom provider shares the event data from their subscribers with Mezuro. These events are extracted from call detail records (CDR), which occur when a user makes or receives a call, sends a text, or uses mobile data. These events are generated quite often, partially based on the frequency of the mobile phone usage, but the events are also periodically generated (roughly every 15minutes) by the provider to check the service quality.

The localization technique used is a network-based technique, information that is passively collected by the provider (active location data is collected by the user, for example GPS location). The big advantage of the passive location data is that a large population is available, in the instance of Mezuro the data from all subscribers of the partnered telecom provider. The location of a device is determined based on the cell tower the device is connected to during an event. A cellphone tower has a certain area of coverage which makes it possible to approximate the location of the user.

There are various techniques or tools available to make the location determination of the user within the cell coverage area more precise, but due to privacy restrictions these techniques cannot be used. A travel movement can be identified by using the cell towers a user connects to during the day, within the Mezuro data a destination is determined when a mobile device does not move for 30 minutes. 'Not moving' meaning that the device does not leave the coverage area of the cell tower it has connected to (even if the device connects to a different cell with an overlapping coverage). All location data is aggregated to make sure that very specific movements cannot be traced back to an individual. There is a minimum number of 15 travel movements between the area of origin and destination to prevent movements to be traced back to individual subscribers.

OViN

OViN stands for 'Research into movements in the Netherlands' and is a yearly conducted survey based research to the travel behavior of the Dutch inhabitants. The Dutch bureau of statistics (CBS) uses this research since 1999, in earlier years as the OVG (Research to travel movement behavior) and MON (Mobility research Netherlands). The OViN data is released on a yearly basis, the most recent version at this time is the version of 2013; this edition is based on 42.350 surveys.

The purpose of the OViN research is to gain an insight in the mobility of the Dutch inhabitants and use this insight for the development of traffic and transport policies, for example to improve road safety, reduce the amount of traffic jams, and improve environmental conditions caused by traffic.

5.2 Method application

The integration of both datasets will be done according to the method composed in the previous chapter, to shortly recap this method the five main steps are:

1. Assess data quality
2. Develop a shared ontology
3. Find semantic matches
4. Integrate datasets based on semantic matches
5. Evaluate quality of the integrated dataset

Step 1: Assess data quality

The method states that the data quality assessment phase consists of three main steps: first the relevant data quality dimensions have to be identified; these dimensions depend on the domain of the project. Secondly a list of data quality measures has to be drafted that corresponds to the chosen data quality dimensions. And lastly these data quality measures are used to assess the data quality of an information source. These three steps will be conducted for the research towards the carbon footprint of mobility.

Data quality dimensions

The important quality dimensions are chosen depending on the domain, the data quality dimensions regarding the data itself (intrinsic data quality dimensions) are always important to assess data quality. The intrinsic data quality consists of the dimensions: accuracy, objectivity, believability, and reputation. The contextual data quality dimensions are used to assess data quality based on the context of the data, the dimensions: completeness, timeliness, and amount of data will be included in the data quality assessment. These three dimensions measure important differences between both data sources, the two dimensions, relevancy and value-added are not included in the assessment. The main reason for the exclusion is the fact that both datasets (Mezuro and OViN) measure the same type of data regarding the context, which means both datasets are equally relevant and value-adding to the domain of this research.

Data quality measures

The data quality dimensions require measurements to express the data quality dimension into values and make the dimension quantifiable. In the previous section the relevant data quality dimensions for the research project are listed. Based on these dimensions a questionnaire is composed with data quality measures relevant to these dimensions. The complete questionnaire is listed in Appendix A, some of the questions are elaborated on shortly to give an impression.

The questionnaire is composed of two sections with their own purpose: the first section is used to gather additional information to specify the data quality problems, the second section contains only data quality measures to quantify the data quality for each of the datasets. The first part also consists of data quality measures for the respective dimensions.

The first part is considered with the data quality dimensions completeness and timeliness, in this part both the data quality score is gathered as well as more specific information on the data quality problem. For the data quality dimensions completeness this is listed in the questionnaire as follows:

1. "The information in the dataset is incomplete for te purpose of this research? (on a scale from 0 - 10)"
    (a) "In what aspects is the dataset lacking on the subject of completeness? (For example missing data on specific conditions)"

The second part of the questionnaire consists of more questions that are answered on a scale from 0 to 10 in order to form a data quality score for all dimensions that can be compared between multiple datasets. An example of the questions that are included is the following:

- **Objectivity:**
  1. "This information was objectively collected? (on a scale from 0 - 10)"
  2. "This information is based on facts? (on a scale from 0 - 10)"

The subjective measures for objectivity, believability, reputation, completeness and amount of data are gathered from the AIMQ methodology by Lee et al. (2002), these measures are commonly rated with a score from 0 to 10 where 0 corresponds to 'not at all' and 10 corresponds to 'completely'. In this research the surveys are focused to be more qualitative, mainly because the amount of experts on each of the datasets is quite low and will not result in significant statistical outcomes.

Data quality assessment

The data quality assessment is performed using expert interviews based on the questionnaire that is elaborated on in the previous section. For this research three expert interviews are conducted, each of the three participants is familiar with both the Mezuro mobility data as well as the OViN dataset, and are experts in the field of mobility. The participating experts are not directly related to either of the companies the data originates from, this allows for objective assessments of both data sources. In the table 5.1 the three participants are listed with their expertise on each of the datasets and the domain. The expertise is listed by the frequency each of the datasets is being used by the expert.

Experts:	Expertise Mezuro:	Expertise OViN:	Working domain:
Expert 1	Daily	Daily	Traffic management and mobility planning
Expert 2	Weekly	Weekly	Traffic and transport consultancy
Expert 3	(Almost) Daily	Daily	Economic decision making on mobility and infrastructure

TABLE 5.1: Overview of the expertise and working domain of the interviewed experts

The interviews follow the structure of the questionnaire, therefore the results of the interviews are used for two distinct purposes: firstly specifying the data quality problems in the dataset, secondly rating the data quality scores for all relevant dimensions on a scale from 0 to 10. An important note is that some of the questions measure the score

inverted to the data quality dimensions, meaning that a high score to the question actually means low quality on that specific dimension. This occurs in the questions 1 and 4, and data quality is measured inverted to have a distinct difference between that question and the second question related to the same data quality dimension (in this case question 1 differs from 2, and question 3 differs from 4). The table below (table 5.2) includes the average scores of all the data quality measures, in this overview the scores of questions 1 and 4 are adjusted to make sure the score 0 relates to low data quality and 10 relates to high data quality. The complete overview of the interview results and the data quality scores are listed in Appendix B, in this overview the data quality scores are lists as they were collected; not adjusted.

Data quality dimension:	Average score Mezuro:	Average score OViN:
Completeness	5.6	6.8
Timeliness	8	7
Accuracy	6	4.3
Objectivity	8.7	8.3
Believability	8.3	8.3
Reputation	7.2	7.6
Amount of data	9.3	5.3

TABLE 5.2: Averages data quality scores for the data quality dimensions of the Mezuro and OViN datasets

The table (table 5.2) shows some interesting insides in the data quality scores as they are judged by the three experts. The average data quality scores of most dimensions do not differentiate much from one another (less than a full point). However some data quality dimensions do differ a lot, in specific the dimension 'Amount of data' where the greatest difference in quality is measured. Besides 'Amount of data' the two dimensions 'Completeness' and 'Accuracy' show great differences.

The combination of the big differences in data quality on the dimensions Completeness and Amount of data define our data quality problem. The completeness data quality of the Mezuro dataset is noticeably lower than in OViN; this is due to the missing movements on low distances in the Mezuro data. On the other hand, Mezuro offers a much higher quantity in data (Amount of data), that helps to give a good insight in where specific movements are conducted.

This result is supported by the experts in their answers to specify the data quality problem. In the two questions that relate to the data quality dimension Completeness the data quality problem in the Mezuro data is emphasized: *"For the determination of the carbon footprint a precise determination of the types of transportation is important; besides, on the traffic within a GSM-area no statement can be made"*. Where the data quality for completeness on the Mezuro data was assessed more positively: *"OViN data*

gives the complete picture, but only for a limited group, so travel behavior parameters can be derived, but hardly any regional specifications".

Step 2: Develop a shared ontology

The first step towards a shared ontology is to determine the purpose, scope, and the domain. The purpose and scope are determined for the shared ontology based on the integration of all information sources. The domain is determined for each of the information sources to make it possible to compare them and decide on a fitting ontology approach.

The purpose of the shared ontology is to specify the semantics of multiple information sources in order to identify corresponding elements that allows these data sources to be integrated. In short the purpose of the shared ontology in this research is to support semantic integration. The scope of the shared ontology is to specify a vocabulary so that the purpose of the research can be met. In the case of the research on the mobility carbon footprint this means: identify sufficient semantic matches to create an integrated dataset that includes all travel movements in the Netherlands.

The domain of the two datasets is determined by their current application. The Mezuro dataset provides information on travel movements within the Netherlands (origin, destination, and number of travelers). The OViN dataset contains the information of people and their travel movements (for example: distance, mode of transportation, motivation). The common ground between both datasets, which is of interest for the integration of both information sources, shares the same domain: mobility. This domain can be described unambiguously (distance is defined as kilometers, and a travel movement is measured between the point of origin and the point of destination). This domain allows using a single ontology approach where one ontology is used to describe the vocabulary of both information sources.

Finding relevant concepts and terms: The first step to finding relevant terms is done by listing the column from both datasets that are relevant to the domain of mobility. In the table below (Table 5.3) the relevant terms are listed that are directly derived from the data sources. Some of the terms can be modified to be more consistent, for example the Mezuro dataset measures the total number of movements where OViN measure per single travel movement. The term 'movement' can easily be adapted to also measure the number of movements from the OViN dataset. Important to know is that these terms are not yet coherent, as both information sources may use different definitions of the terms, therefore terms are listed individually for each of the two data sources. An

Mezuro data	- Origin - Destination - Distance - Location (geometry) - Number of movements - Modality (mode of transport) - Urbanity - Weighting factor
OViN data	- User - Movement - Origin - Destination - Distance - Distance class - Modality - Urbanity - Weighting factor

TABLE 5.3: Terms relevant to the domain in the datasets

example of this incoherence is the term for 'Origin' or 'Destination': the Mezuro dataset uses a zoning based on roughly 1250 Mezuro-areas, whereas OViN uses the zip code for the place of origin and destination. Ensuring definitions of the terms are coherent between the information sources is done when all terms are listed in work areas, then the definitions for the relevant concepts are chosen. Choosing the definitions is an important part of creating a shared ontology, since the definitions determine how the datasets relate to eachother.

The following step is to group concepts into work areas. Since all terms that are identified relate to the concept 'travel movements' there will be a distinction made between the primary aspects that are critical to define a movement and the terms that provide additional information. The work area with primary terms for a movement:

- Origin
- Destination
- Distance (or distance class)

The second work area provides additional information regarding the number of movements between origin and destination, what modality is used and a weighting factor that makes sure the amount of measured movements represents the number of movements from the total population. Secondary work area:

- Number of movements

Terms:	**Definition:**	**Relation to information source:**
Movement	Travelling the route between a certain point of origin and a destination.	OViN: Movement
Origin	Starting point of a movement, described in terms of Mezuro areas	Mezuro: Origin
Destination	End point of a movement, described in terms of Mezuro areas	Mezuro: Destination
Distance	The length of a travel movement measured in kilometers.	Mezuro: Distance OViN: Distance / 10
Number of movements	Total amount of movements between a certain Origin and Destination for a specific period of time.	Mezuro: Movements * Weighting factor OViN: Movements * Weighting factor
Modality	Distinction between the use of road or rail travel for a movement, or undefined.	Mezuro: Modality OViN: Modality
Urbanity	Factor of urban density, measured by the number of home addresses per km2 (5 classes) in the area of origin of a movement.	Mezuro: Urbanity OViN: Urbanity

TABLE 5.4: Important terms and definition, and their correspondence to the information sources

- Modality
- Weighting factor

The two work areas are closely interconnected, mainly because the first work area describes the essentials of a movement (there is a point of origin, there is a destination, and the path in between has a certain distance). The second work area provides additional information for this 'movement'-work area; the modality provides extra detail about how the movement was made, and the number of movements and the weighting factor make the movement comparable. Comparability is achieved either by making it possible to compare the amount of movements within the dataset (number of movements), or by making it possible to compare it to external sources (weighting factor).

The fact that a concept is defined in both datasets does not mean the data can be integrated on that concept, there needs to be an identifying factor that can link the movements from the Mezuro data source to those from OViN. Because both datasets use a different division of areas for the origin and destination of movements the integration of data on these concepts is not possible. The challenge for semantic integration is to identify semantic matches that make it possible to combine movements.

Step 3: Find semantic matches

The list of definitions from both datasets includes all relevant terms to the scope of this research, the list of definitions includes in which of the datasets this definition is listed. The definitions that are listed in both datasets are semantically interoperable with each other; the definition of a term specifies that the semantic meaning of the term in both data sources corresponds. Within the scope of this research there are four semantic matches identified based on the definition of relevant terms:

- Distance: In both datasets the term for distance is used to describe the same concept. In OViN data distance is listed in hectometers and should be divided by 10 to correspond to the definition (and Mezuro data) which measures in kilometers.
- Number of movements: The number of movements is measured over a specific period of time; as long as the measured period is consistent the number of movements has a semantical match.
- Modality: In this definition the distinction for the modality is either road (car), rail (train) or undefined. OViN data actually includes more possible modalities, but the distinction between road and rail movements is the focus and makes it possible to relate both data sources.
- Urbanity: The level of urbanity is measured for the area of origin of a movement. The area division differs between Mezuro and OViN data this term does not directly relate both datasets; however the use of urbanity helps to identify where a certain type of movement (based on distance) occurs most.

The purpose of the semantic matching is to identify correspondences in the two datasets that allow the datasets to be integrated, to support integration these matches have to identify a specific movement. For example a correspondence solely on 'modality' does not provide sufficient information to integrate movements. The semantic match on "Urbanity" provides an additional insight in the environment of a movement, the urbanity of the area of origin of a movement does not allow integrating the datasets on this definition alone. However, the urbanity is a relevant factor to specify the surroundings of the movements.

The solution for this application will be to use the semantic match on distance to find the relative number of movements for distance classes. By making a classification of the movements by distance the distribution of movements along their distance can be identified in both datasets. The combination of the distance of a movement and the urbanity of the movements within this distance class allows to link specific movement

	<5,0 km	5,0 - 10 km	10 - 15 km	>15 km
OViN	52,63%	15,97%	7,40%	24,01%
Mezuro	4,39%	15,10%	14,88%	65,63%

TABLE 5.5: Percentage distribution of movements in distance classes

characteristics (distance and urbanity) to the more detailed movements in the Mezuro dataset (for example to link it to locations).

The table (5.5) clearly shows a difference in the distribution of movements according to the distance, this is where the data quality assessment can be applied. In this assessment is stated that the completeness of the Mezuro dataset is lacking, especially for movements on short distances. However on greater distances the completeness of Mezuro movements is very high and due to the flexible and timely manner of data collection should be of higher quality than the OViN data.

Step 4: Integrate datasets based on semantic matches

How are the semantic matches processed in order to actually integrate both datasets? For the main semantic match on distance there needs be a decision what data source will be used. The data quality assessment shows quality varies based on the distance in the Mezuro dataset, on low distance movements the quality is lower than OViN data, but on longer distances the quality is higher. Therefore the integrated dataset will choose to use input from both datasets, the distribution of movements per distance class is taken from OViN and is applied to the absolute number of movements from Mezuro. The minimum distance for Mezuro movements to be registered completely is set at 10 kilometers; this means that for the movements with a lower distance the data will be modified. The number of movements will be modified according to the distribution of movements per distance class from OViN. In Table 5.6 the distribution of movements by their distance is shown for both Mezuro and OViN, since movements with a distance below 10 kilometers are considered inaccurate within the Mezuro dataset these numbers will be adjusted according to the movement distribution from OViN. This means that movements below 5 kilometers distance are incremented from 4% to 52% of the total movements. This calculation is also performed for the distance class 5 to 10 kilometer movements, and the resulting number of movements are listed in the bottom row of the table (Table 5.6).

The increase of the low distance travel movements also impacts the total number of movements and the total distance in the Netherlands. The number of movements has been incremented according to the distribution from OViN on short distance movements;

	<5,0 km	5,0 - 10 km	10 - 15 km	>15 km	Total
Measured movements Mezuro (December 2014)	19.701.595	67.836.174	66.861.235	294.786.696	449.185.700
Distribution measured movements	**4,39%**	**15,10%**	14,88%	65,63%	
Distribution from OViN	*52,63%*	*15,97%*	7,40%	24,01%	
Number of movements using OViN distribution (<10 km)	*605.982.234*	*183.858.137*	66.861.235	294.786.696	1.151.488.302

TABLE 5.6: Increase in movements by combining Mezuro measured movements and OViN distribution

however these additional movements need to be assigned to a certain area to have any effect on the carbon footprint of a specific area.

To divide the extra movements over the existing Mezuro areas the urbanity of movements is applied. The urbanity class (a five class classification of the density of home addresses in an area) is used it both the OViN as the Mezuro dataset. The urbanity of the origin determines the frequency of low distance movement, based on the distance and the urbanity the modified movements can be allocated to an area of origin. The table below (Table 5.7) shows the distribution of movements per urbanity class on the distances below 5 kilometer and between 5 and 10 kilometer, this distribution is derived from the OViN dataset.

Urbanity is measured by the number of home addresses in a square kilometer and uses a classification with 5 classes:

- *Very high urbanity*: More than 2500 home addresses per km2
- *High urbanity*: Between 1500 and 2500 home addresses per km2
- *Moderate urbanity*: Between 1000 and 1500 home addresses per km2
- *Low urbanity*: Between 500 and 1000 home addresses per km2
- *Not urban*: Less than 500 home addresses per km2

The distribution of the movements along the urbanity shows the absolute number of movements there is in each urbanity class, this allows the number of incremented movements to be compared to the measured movements from the Mezuro data. Table 5.8

	Very high urbanity	**High urbanity**	**Moderate urbanity**	**Low urbanity**	**Not urban**
<5 km	21,74%	28,60%	20,95%	19,96%	8,75%
5 - 10 km	25,31%	28,91%	18,18%	18,72%	8,89%

TABLE 5.7: Distribution of movements in classes of urbanity

includes the number of movements per urbanity class based on the incremented number of movements and the OViN distribution of movements by urbanity.

	Very high urbanity	**High urbanity**	**Moderate urbanity**	**Low urbanity**	**Not urban**	**Total**
<5 km	131.750.090	173.288.755	126.966.054	120.981.470	52.995.865	605.982.234
5 - 10 km	46.532.623	53.154.479	33.417.585	34.410.223	16.343.227	183.858.137

TABLE 5.8: Incremented number of movements divided according to the distribution from OViN

The number of movements per urbanity class for the incremented data (5.8) can be compared to the measured data from Mezuro. This comparison will result in the difference between the measured data (Mezuro) and the incremented data (Mezuro adjusted using OViN movement distribution on short distances), this difference is listed for each urbanity class for both movements below 5 kilometers and movements between 5 and 10 kilometers. The difference can be calculated as a factor to increment movements based on the urbanity and distance. This factor is used to create the integrated dataset by incremented movements from Mezuro using an increment factor specific for the distance class and the urbanity. This integration step is the result of the identified data quality problem of low accuracy on movements with a distance below 10 kilometers.

The number of movements in each urbanity class in the Mezuro dataset is listed in Table 5.9, the number of movements per urbanity is listed for the two short distance classes. The increment factors for all urbanity classes based on the difference in movements between the measured movements from Mezuro and the incremented movements from Mezuro with OViN movement distribution are listed in table 5.10

	Very high urbanity	**High urbanity**	**Moderate urbanity**	**Low urbanity**	**Not urban**	**Total**
<5 km	8.595.897	5.756.164	2.813.332	1.759.466	776.736	19.701.595
5 - 10 km	18.996.388	17.008.139	14.168.680	12.729.117	4.933.850	67.836.174

TABLE 5.9: Measured number of movements (Mezuro) per urbanity class for the low distance movements

The integration of movements from Mezuro and OViN data is the first step towards an integrated dataset that can support the mobility carbon footprint. Besides knowing

	Very high urbanity	**High urbanity**	**Moderate urbanity**	**Low urbanity**	**Not urban**
<5 km	15,3	30,1	45,1	68,8	68,2
5 - 10 km	2,4	3,1	2,4	2,7	3,3

TABLE 5.10: Increment factors based on the difference in the number of movements between the measured and incremented data

all travel movements and their distances, an important factor is to know the type of transport that is used for that movement.

The second part of the semantic integration focuses on the semantic match of modality, with the purpose of assigning a modality to every travel movement.

Based on the data quality assessment the selection of data source for the modality is based on movement distance. For the Mezuro dataset is stated that movements with a distance below 15 kilometers have a data quality problem with assigning the right modality. Either the movement is based on only a few events that makes assigning a modality difficult. Or the events from a movement use cell towers with both road and train tracks within the coverage area, meaning no distinction between road or rail traffic can be made. For movements longer than 15 kilometers the modality distribution from Mezuro is selected, mainly because of the high quantity and more timely data. The distribution of movements per modality per distance class are listed in table 5.11 (OViN distribution of short distance) and table 5.12 (Mezuro distribution on longer distance).

	0 - 5 km	**5 - 10 km**	**10 - 15 km**
Road (OViN)	29,28%	58,83%	65,68%
Rail (OViN)	2,68%	12,25%	19,16%

TABLE 5.11: OViN distribution of modality by distance class (below 15 km)

	15 - 20 km	**20 - 30 km**	**30 - 40 km**	**40 - 50 km**	**>50 km**
Road (Mezuro)	67,62%	69,11%	70,35%	71,30%	74,49%
Rail (Mezuro)	27,61%	26,38%	25,51%	25,17%	22,79%

TABLE 5.12: Mezuro distribution of modality by distance class (above 15 km)

The two modalities together do not add up to 100%, especially on lower distance classes, because only car and train are addressed. On short distance modalities such as bicycle or pedestrian account for a substantial amount of movement, their share drops when movements become longer and on movements over 15 kilometers car and train traffic account for over 95% of all movements. The distribution of road and rail movements according to both Mezuro and OViN is depicted in Figure 5.1. The figure shows a relative low amount of road and rail traffic on the very short distances, the usage of both

modalities increases rapidly on longer distances, the railway traffic levels off around 25% whereas road traffic climbs to around 70% of the total movements.

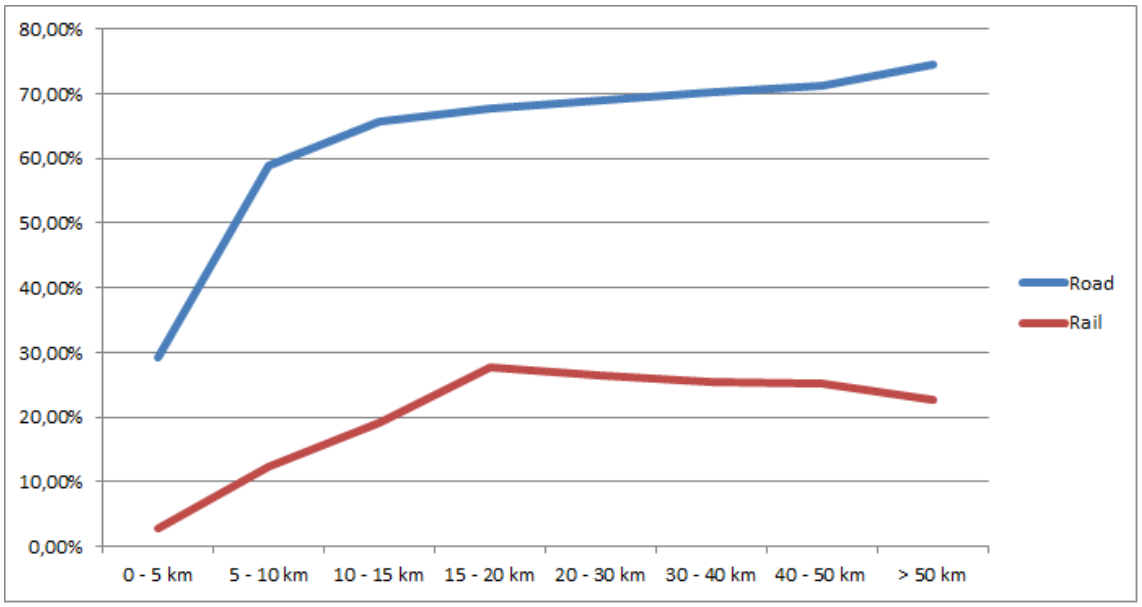

FIGURE 5.2: Distribution of both modalities by movement distance

When the absolute number of movements from the incremented dataset is applied to the distribution of modality the result is the number of movements per modality. These numbers represent the number total number of movement for each modality in each distance class for the month December 2014 and are listed below in table 5.13 and 5.14.

	0 - 5 km	**5 - 10 km**	**10 - 15 km**
Total	**605.982.234**	**183.858.137**	**66.861.235**
Road	177.444.142	108.166.353	43.917.187
Rail	16.259.958	22.525.012	12.812.043

TABLE 5.13: Number of movements per modality (for movements under 15 kilometer)

	15 - 20 km	**20 - 30 km**	**30 - 40 km**	**40 - 50 km**	**>50 km**
Total	**57.094.762**	**81.700.322**	**49.344.864**	**31.154.036**	**75.492.712**
Road	38.607.478	56.463.093	34.714.112	22.212.828	56.234.521
Rail	15.763.864	21.552.545	12.587.875	7.841.471	17.204.789

TABLE 5.14: Number of movements per modality (for movements over 15 kilometer)

With the integration of the Mezuro and OViN dataset to increment the total number of movement and the distribution of the modalities the data integration is completed. In the following section, evaluation, the integrated dataset will be tested on data quality to assess if the data quality has improved compared to the two data sources separately.

The chapter Results (Chapter 6) includes more detailed results from the integrated dataset that are used to compose the mobility carbon footprint.

Step 5: Evaluate data quality of the integrated dataset

The data quality assessment for the Mezuro and OViN datasets is conducted by interviewing experts on the respective datasets. The integration of these two data source into a single integrated dataset has created a new source of information, unfortunately there are no expert specifically for this set of data since it is only used in this research. This makes conducting interviews with experts to determine the data quality of the integrated dataset impossible. For this reason the data quality assessment of the integrated dataset is done by analyzing the data quality dimensions addressed in the assessment of the Mezuro and OViN datasets. Secondly, the results of the integrated dataset are evaluated using reference data from the CBS.

In this research the improvement of data quality is measured using data quality dimensions, for this research project dimensions from two of the four data quality categories are included in the data quality assessment: intrinsic data quality and contextual data quality.

The intrinsic data quality dimensions (accuracy, objectivity, believability, and reputation) are measured independently of the context of the data; the intrinsic data quality remains the same when the application of the data is changed. For example the dimensions objectivity and reputation, are DQ dimensions that relate to either the method of data collection (objectivity) or the source of the data (reputation). The intrinsic data quality dimensions that relate to the dataset as a whole (objectivity, believability, reputation) and not specific elements of the data (accuracy) are not greatly impacted by data integration; the data quality for these dimensions after integration is somewhere in between the quality level of both individual datasets. The dimension of objectivity, for example, is very high in the Mezuro dataset but of lower quality in the OViN dataset (the OViN survey is filled in by the participants and not measured as in Mezuro), the integrated dataset will be a combination of both. Since the data quality dimensions that relate to the complete dataset do not translate well to specific elements of a dataset that are integrated the data quality level for these dimensions is between that of the individual datasets (Mezuro and OViN).

The improvement of data quality can be found mainly in the contextual data quality category, consisting of the data quality dimensions relevancy, value-added, timeliness, completeness, and amount of data. The contextual DQ dimensions are measured based on the situation, the level of data quality is dependent on the application of the dataset. Since the application of the individual datasets, Mezuro and OViN, is in a specific domain the datasets were not completely intended for the data quality can be improved. The Mezuro dataset does not include short movements because these are difficult to measure

using cellphone towers, this relates to the data quality dimension 'completeness'. Using the distribution of movements by their distance from the OViN data source, the problem of missing values on low travel distance can be resolved.

On the other hand, the OViN dataset consists of data collected over the course of one year that are incremented using various factors to represent the travel behavior of the Dutch population. The data quality dimensions timeliness and amount of data are rated much higher in the Mezuro dataset, due to the continuous data collection and the quantity of raw data that is being processed to measure all movements. Integration of these dimensions will result in levels of data quality in between the values from Mezuro and OViN; both dimensions relate to the individual datasets as a whole and are not specific to elements that are being integrated, therefore the data quality for timeliness and amount-of-data is between the data quality of Mezuro and OViN.

A second approach to evaluating data quality improvement is done my comparing results from the integrated dataset to figures published by the Dutch bureau of statistics, CBS. Every year the CBS publishes figures regarding transportation in the Netherlands. One of the figures that is published is the total travel distance of all 'mobile sources' in the Netherlands, meaning the total distance traveled over the entire year. The integrated data consists of the total travel distance over one month (December 2014), since there is no alternative monthly CBS data available for evaluation the measures movement data from the integrated dataset is multiplied by 12. In the table (Table 5.15) below the total travel distance (incremented to yearly kilometers) is compared to the published total travel distance of 2014 from the CBS alongside the total travel distance from the initial Mezuro data (also incremented December 2014 data). The table shows that the integrated dataset calculates the total travel distance over a full year with just a 1% discrepancy to the reference data from CBS, the initial mobile phone data (Mezuro) differs much more and total 22% lower than CBS data.

	Total travel distance	Reference data (CBS)	Difference
Integrated data (Mezuro + OViN)	202,3	199,5	1%
Mezuro data	164.2	199,5	22%

TABLE 5.15: Comparison of total travel distance (in billion kilometers) of the integrated dataset (Mezuro + OViN) and initial Mezuro data to published figure from CBS

Chapter 6

Results

The semantic integration helped to create one dataset that contains all elements needed to calculate the carbon footprint and has a high data quality that allows the resulting footprint to be accurate. The integrated dataset includes all the necessary information to calculate the number of kilometers traveled from and to all Mezuro areas in the Netherlands (due to the integration based on the Mezuro area zoning the mobility carbon footprint can be composed for each area). The relevant elements for the carbon footprint are:

- Area of origin
- Area of destination
- Distance
- Modality
- Number of travels over the period (December 2014 in this case study)

To get an insight how travel movements are distributed over the Netherlands the figure below (Figure 6.1) shows the total number of kilometers travelled from each Mezuro area.

In this figure the areas with a high density of movement are easily distinguished as the major cities; generally the more inhabitants an area has, the higher the output of travel distance is. This result is not surprise, the higher the population the more people travel therefore the total traveled distance from the area is higher. It may also be interesting to investigate the travel behavior of an area calculated per inhabitant; in a big city there are a lot of people but all services are nearby, whereas in a small town the people may have

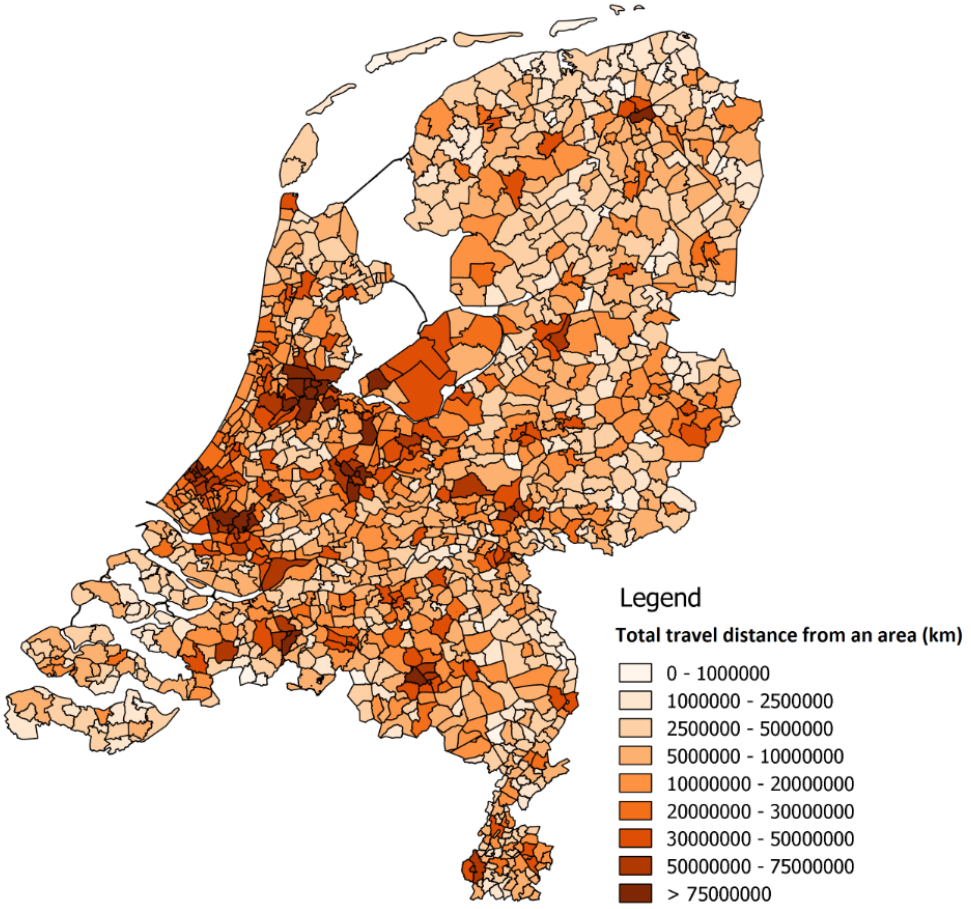

FIGURE 6.1: An overview of the total travel distance measured from the originating areas

to travel much further to get their groceries or see a doctor for example. To visualize this effect the total travel distance from an area will be divided by the population of that area, this results in the average total travel distance per capita (for December 2014).

The total traveled distance originating from an area can give an insight in the travel activity of an area but does not fully cover the distribution of carbon emission. Apart from distance travelled it is important to distinguish the modality that is used. There is a significant difference in carbon emission per kilometer between an average car and a train (with average occupation), therefore the distribution of total distance via road and rail is an important factor to take into account.

Firstly the average emission of each modality will be determined, after the emission per kilometer for both modalities has been found these numbers can be combined with the results from the integrated dataset in order to create the mobility carbon footprint. The calculation for the emission of a train depends on the occupation of the train, since the emission needs to be calculated per traveled kilometer for each passenger. Luckily the Dutch railways (NS) have included the carbon emission per traveler kilometer in their

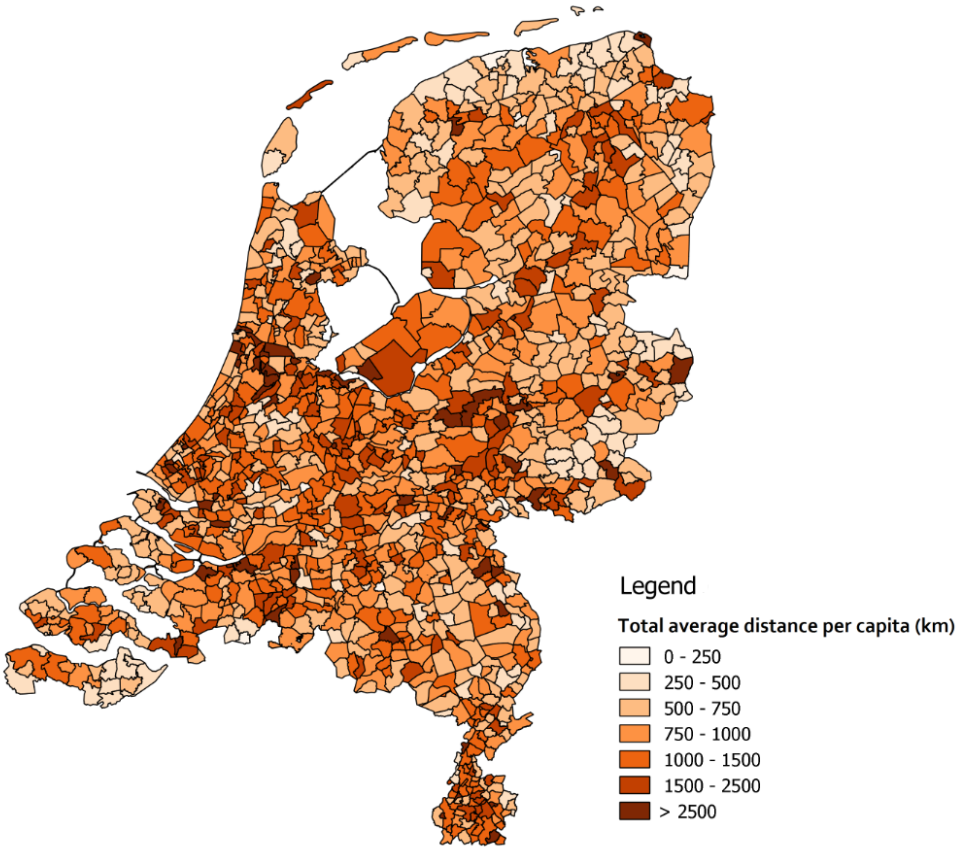

FIGURE 6.2: An overview of the average travel distance per capita from the area of origin

annual report. According to the 2014 annual report their trains emit an average of 31 grams of CO_2 per traveler kilometer (NS, 2014).

The calculation of the average carbon emission for a car has to be done more manually since this figure is not available from recent sources. Since cars become increasingly more fuel efficient it is important to calculate the average emission based on recent data to get the most accurate emission factor. The calculation of the average CO_2 emission for car traffic is based on CBS data from 2013 (CBS, 2013), this source includes the yearly traveler kilometers with cars and the total emission of carbon dioxide caused by traffic. These two numbers are used to calculate the average carbon emission of car traffic per kilometer. It is important to note that it is not the average emission of a car kilometer that is calculated, but the average emission per traveler kilometer using car; this takes into account car trips with more than one person in a single car.

Traveler kilometers using car:	145,5 bln km
Emission road traffic (excluding tractors)	25.786 mln kg CO_2

TABLE 6.1: CBS data regarding travel distance and emission for cars over 2013

The emission figure deliberately excludes tractors since these farm machines are used mainly on the field and not for road transport. The calculation to create the average emission per kilometer is the division of the total emission by the total distance. The two figures divided directly result in the average emission in grams of CO_2 per kilometer (10^9 gram / 10^9 kilometer).

$$25.786/145,4 = 177,3\ gram\ CO_2/km$$

With the calculation of the emission per traveler kilometer for both modalities the total emission for an area can be calculated based on the travelled distance with each modality. In table 6.2 the total traveled distance per modality and the total emission for each modality is listed (for the month December 2014).

	Total distance (mln km)	**Total emission (ton CO_2)**
Road traffic	10.936	1.938.961
Rail Traffic	3.361	104.180

TABLE 6.2: Total distance and emission per modality over December 2014

The total emission can also be calculated for a specific Mezuro area by measuring the traveled kilometers to and from a specific area. The problem that now arises has to do with assigning kilometers to a footprint. With the current data it is possible to measure the total distance originating from a specific area and distance towards a specific area (and the corresponding carbon emission based on the modality). The question is to what area a certain movement (with a specific emission) belongs; the emission from a movement from an area A to B can be allocated to either of the area depending on the definition from the carbon footprint. The solution to the issue to what area the emission from a movement belongs can be found in the Mezuro dataset, in the Mezuro data is included the elements 'homeorigin' and 'homedestination'. For all movements between two areas besides the total number of movements, the dataset also includes the number of people going towards their perceived area of residence (homeorigin) and the number of people leaving the area of residence (homedestination). The residence is determined per hashedID (device) based on the area the device is in during the night.

The number of movements that are originating from or returning to the area the user lives in is used to determine attraction and production of an area. The attraction and production are terms used to specify the purpose of a movement; if a person travels from his home area (A) to another area (B) this is caused by the attractiveness of the area B. Apparently area B has something to offer that makes the user leave area A and travel to B, for whatever reason. A movement can be identified as a movement caused by the attraction of a certain area when that area is the destination of a movement for a person

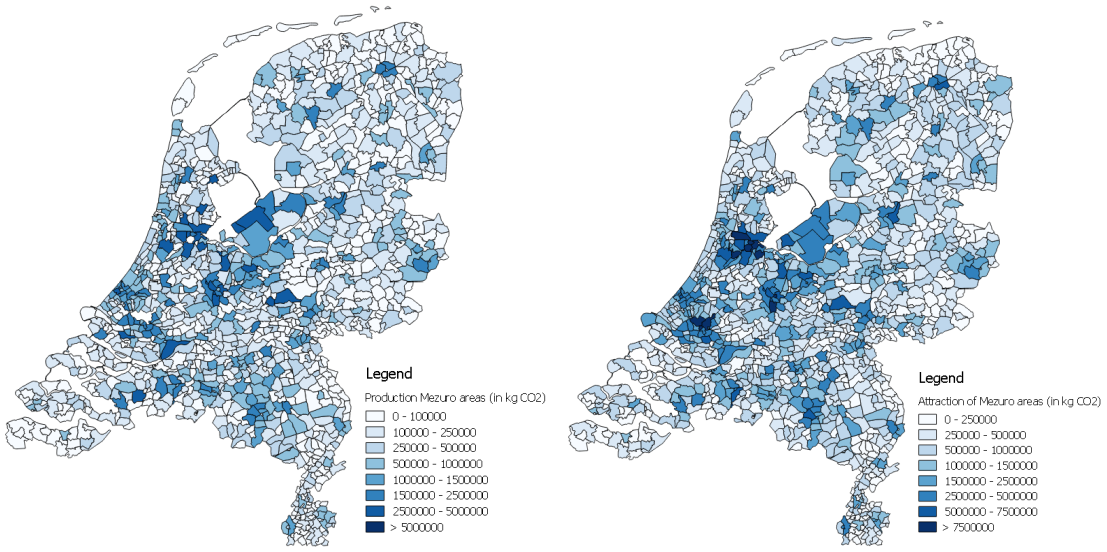

FIGURE 6.3: An overview of carbon emission measured by production and attraction over December 2014

not living in that area (residents returning home do not contribute to the attraction of an area). The production of an area is defined as all movements from residents leaving their residential area. To put this division of all movements into attraction and production in terms of the integrated dataset:

$$Production(A) = \text{Number of movements homeorigin}(A \rightarrow x)$$

Where 'x' is any other area (except A) and 'homeorigin' is the number of movements that are carried out originating from the area of residence. The production for an area A is described as: all movements origination from A, carried out by residents of area A.

$$Attraction(A) = \text{Number of movements}(x \rightarrow A) – \text{Number of movements homedest}(x \rightarrow A)$$

Where, same as for production, 'x' is any other area (except A) and 'homedest' is the number of movements that are carried out towards the area of residence.

The production and attraction can be visualized in order to emphasize areas that have a high attraction or production; this is shown in Figure 6.3 where the darkness of the blue color represents the order of emission for an area. The both figures regarding production and attraction visualize the emission in the absolute numbers per Mezuro area, the carbon dioxide emission is measured in tons over the month December 2014.

The figures only show the production and attraction on scale with seven classes, within the highest class (of over 7.500.000 kg CO_2) there are great difference that do not become apparent in the figures. The areas with the highest emission for both production and

attraction are shown in the tables 6.3 and 6.4, these tables present the 'Top 5 areas with the highest production and attraction'.

Top 5 Production	**Total carbon emission (in mln kg CO_2)**
Amsterdam South	5,66
Almere City	5,53
Rotterdam North	5,48
Rotterdam East	5,08
Amsterdam East	4,39

TABLE 6.3: The five highest emitting areas measured by production (monthly total for December 2014)

Top 5 Attraction	**Total carbon emission (in mln kg CO_2)**
Schiphol	13,72
Amsterdam South East	13,22
Amsterdam Center	12,68
Amsterdam South	10,62
Rotterdam Center	9,98

TABLE 6.4: The five highest emitting areas measured by attraction (monthly total for December 2014)

The tables above (6.3 and 6.4) show the top five in absolute numbers; with Amsterdam and Rotterdam being the two largest cities in the Netherlands it is not surprising that these cities are high on the list in absolute emission numbers. Schiphol scores high in terms of attraction, this makes sense since Schiphol is the main airport of the Netherlands and transports millions of passengers on a monthly basis. In the list for production areas there are some areas that are known for the high number of commuters. A good example is Almere, and the areas Rotterdam East and North also have a high number of residents that leave the area to work. The numbers from the attraction and production have a significant discrepancy due to the definitions of both concepts; attraction related to all travels excluding residents, production only measures residents.

The list of top five carbon emitting areas may present the areas that have the highest impact on the environment, but areas that produce a high amount of emission over a small amount of inhabitant may be more interesting to investigate. The areas with the highest average carbon emission per capita allow for the most improvement; a high amount of emission cause by a lot of people is harder to reduce than a high amount of emission caused by a small group of people. For this reason the production and attraction per capita is calculated and the top five emitters are listed in table 6.5 and 6.6. For both tables the minimum number of inhabitants for an area was set to 100, this is done to remove skewed results from areas that produce a lot of movements but do not have (many) inhabitants such as: Rotterdam harbor areas (Botlek) or airports (Maastricht airport, Schiphol).

Top 5 Production per capita	**Average emission per capita (in kg CO_2)**
Almere Hout	209
Arnhem Center	109
Almere Poort	97
Amersfoort Center	92
Eindhoven Strijp	81

TABLE 6.5: The top five areas based on production per capita (over December 2014)

Top 5 Attraction per capita	**Average emission per capita (in kg CO_2)**
Zevenbergschen Hoek	773
Velsen-Zuid	630
Rijsenhout	570
Almere Hout	524
Duivendrecht	516

TABLE 6.6: The top five areas based on attraction per capita (over December 2014)

In the examples that are presented in this chapter carbon emission has been described using two concepts; production and attraction. Both concepts cover a specific part of the total movements that are more concerned with the reason of a movement (attraction or production), the mobility carbon footprint is proposed to include all movements to give an insight in the absolute carbon emission for areas in the Netherlands. The complete overview of movements is both incoming and outgoing movements from an area. However, a single movement originates from one area and ends in another, which means that including all of the incoming and outgoing movements results in a carbon footprint that is twice as high as in reality. This problem will be solved by including half of the incoming and half of the outgoing movements of an area in the carbon footprint of an area. This results in the carbon footprint to be defined as the emission of half the incoming and half of the outgoing movements, as is shown in the equation below.

$$CarbonFootprint(A) = \frac{\text{Number of movements}(A \rightarrow x) + \text{Number of movements}(x \rightarrow A)}{2}$$

The five highest carbon footprints are listed in table 6.7, this list contains the mobility carbon footprint calculated per capita. The areas listed in this top five are derived from a high production (Arnhem Center), high attraction (Zevenbergschen Hoek) or a combination of both factors (Almere Hout). In the subsequent figure (Figure 6.4) the overview of carbon footprint per capita for all areas in the Netherlands is depicted, this shows a graphical distribution of carbon emission per inhabitant of an area over the month December 2014. In this figure more dark colored areas are shown than are listed

in the table (table 6.7), this is the result of areas with no, or very little, inhabitants. These skewed results are filtered for the table, but are shown as they are in the figure.

Carbon Footprint per capita	**Average emission per capita (in kg CO_2)**
Zevenbergschen Hoek	788
Almere Hout	724
Velsen-Zuid	645
Arnhem Center	571
Rijsenhout	560

TABLE 6.7: Mobility carbon footprint top five highest emitting areas per capita (over December 2014)

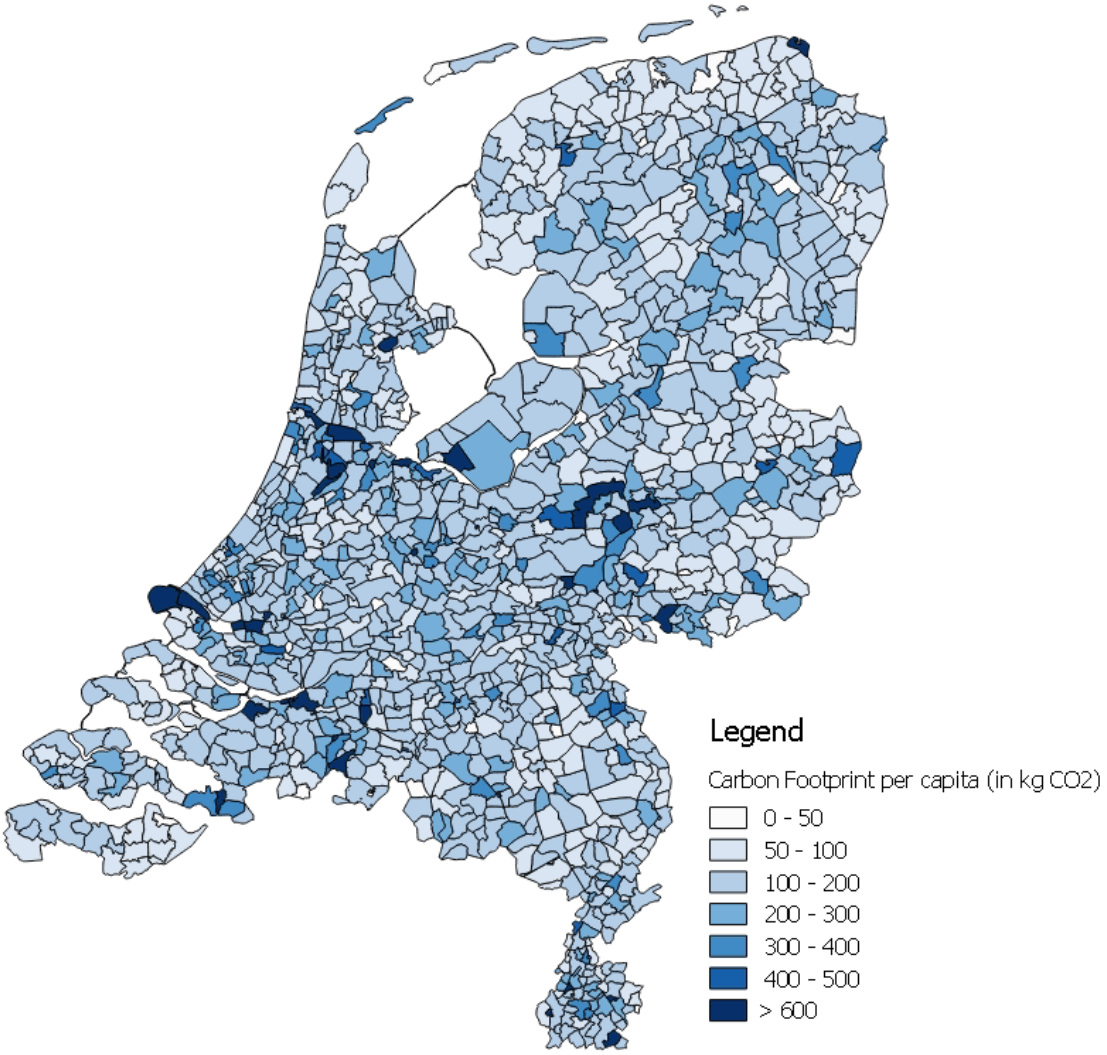

FIGURE 6.4: Overview of the carbon footprint per capita per area (over December 2014)

Chapter 7

Conclusion

This section is used to present the findings of the research that answer the main question of the research. This research question is as follows: *"How can semantic integration of multiple data sources be used to improve data quality?"*

The proposed research question is two-fold: (improving) data quality, and semantic integration. To evaluate data quality improvement, the initial data quality has to be measured. Measuring the data quality again after performing the semantic integration is done to determine data quality improvement.

Data quality can be measured on several data quality categories (intrinsic, accessibility, contextual, and representational) using specific measures for data quality dimensions. Applying the same method to assess the data quality of multiple data sources makes it possible to compare data quality between these datasets. The assessment of data quality can help identify weaknesses in specific aspects of the data quality (for example timeliness, or accuracy), the comparison of the data quality between multiple datasets determines on what dimension the data quality of one dataset is higher than the other. The following step is to combine the datasets in such a fashion that the dimensions with high data quality of a data source are highlighted; this process is called semantic integration.

Semantic integration is used to combine multiple datasets based on semantic correspondences, these correspondences can be determined when the same interpretation of a natural language is used (the same vocabulary). For multiple datasets to apply the same vocabulary it is necessary to develop a shared ontology; an ontology is used to define domain specific elements and interrelationships. The shared ontology consists of definitions of relevant terms and concepts of the domain of the datasets, this definition is used to make sure that the terms and concepts in the different datasets refer to the same

real world entity. Using the shared ontology the correspondences between the datasets are identified; the so called semantic matches. A semantic match defines how an element from one dataset relates to that from another dataset, using a semantic match multiple datasets can be integrated.

Semantic integration is used to integrate the datasets based on their strengths and weaknesses that are identified during the data quality assessment. The weaknesses in data quality are specific conditions in the shared ontology where one dataset contains better or more information than the other data source. In practice this means that based on the data quality assessments data is added or replaced using integration of semantic matches.

The data quality for the Mezuro and OViN data has been improved up to a point that an accurate mobility carbon footprint can be composed. The data quality assessment revealed a low accuracy in the Mezuro dataset on measurement of short distance movements (below 10 kilometers), and assigning a modality to short movements (below 15 kilometer movement distance). A shared ontology allowed Mezuro and OViN to integrate based on the distribution of movements depending on their distance. This meant that information from OViN is used to resolve the data quality problem on low distance movements (both the number of measured movements and their modality). The example from the case study illustrates the importance of specifying the data quality assessment; merely a score on a scale from 0 to 10 for the data quality dimensions does not provide sufficient information to determine how a data quality score relates to actual, specific elements from the data source.

Chapter 8

Discussion

In the previous chapters a method for semantic integration and the evaluation of the method using a case study is conducted. The purpose of the method is to improve the data quality of a dataset by complementing weaknesses in data quality with data from an additional dataset. The case study is used to evaluate the theoretical basis, posed in the method, in practice. The findings in the case study have shown that semantic integration techniques can be applied to multiple datasets from external sources, resulting in a dataset that solves a problem the individual datasets could not accurately do (compose the mobility carbon footprint). The data quality improvement occurs in the dimensions that are concerned with the actual data (accuracy) and dimensions that are closely related to the application of the data (completeness). Data quality dimensions concerned with the collection of data are less impacted since the integration of data only combines the data collection from both datasets into one, examples of data quality dimensions that are related to the data collecting are timeliness, objectivity, reputation, and believability.

The results from the case study indicate that the data quality dimension of completeness has improved with regard to the project (creating a carbon footprint). It is important to note that data quality has to be assessed from the perspective of the project at hand, since a dimension such as completeness is measured by the 'fit' of the dataset to the project. The purpose of the project, towards a mobility carbon footprint, is primarily focused on the dimension of completeness; gaining an insight in all movements (and their modality) in the Netherlands, this is a state of high data quality regarding completeness.

Another dimension that is crucial to this project is the accuracy of the data, this is partially related to the completeness of a dataset to the context of a project. When the completeness of a data source regarding a project is low, ergo there is a high amount of missing data. For a project where a high completeness is important (an overview of all movements in the Netherlands to determine an accurate mobility carbon footprint)

the accuracy of the integrated dataset defines whether the project will be a success. In the the Mezuro and OViN case study the accuracy regarding a complete insight of the movements was low, under a certain distance (roughly 10 kilometers) the movements were not accurately registered, meaning some short movements were measured but others were not (depending on the location the movement occurred).

8.1 Limitations

This method proposed in this research has been tested using a single case study, the application of semantic integration in the case study was done with two dataset with high data quality. The semantic matches between both datasets are easy to distinguish due to the similarities of the data; both data sources count movements with a corresponding distance and modality. The results of the method can differ when the application is in a domain that is less solid. The two major conditions for the limitations of the method are:

1. Applying the method to datasets where one dataset has (or both datasets have) a low data quality.
2. Applying the method in a domain where it is difficult to determine definitions for concepts from the individual datasets.

8.2 Future work

The application of semantic integration to improve data quality is a general concept that, in essence, only requires the availability of a related secondary dataset that can be used for integration. The research, in particular the case study, has shown that the method is applicable to data quality problems in a domain where external information is available. In the case study this external data source was the OViN data from the Dutch bureau of statistics (CBS). Agencies such as the CBS are a very useful source for the additional dataset to integrate with; data is tested for reliability and the availability of data is high.

Future research towards the application of semantic integration techniques to improve data quality can be focused on maturing the proposed method. The assessment of data quality in the current method is based on the dimensions that are relevant to the project at hand, in the future a more solid data quality assessment framework can be developed that consists of fixed data quality dimensions and measures that are most

important to data quality improvement using semantic integration. Additionally the data quality assessment requires a more sound approach for the integrated dataset; the individual datasets are assessed using experts for each of the datasets, the integrated data is a combination of both datasets and finding experts on both of the data sources is problematic.

The future of the application of the method in the case study depends on the demand for a mobility carbon footprint. In early stages of the project the possible applications were discussed with a government agency that is concerned with emissions from traffic and their impact on the environment. The possibilities for the mobility carbon footprint are most relevant to provinces and municipalities, an insight in the emissions from traffic can be used as a basis for policy makers to address measures to reduce carbon emissions.

The case study performed in this research can be improved on several aspects, these are changes that require more additional information in order to make the end product (the carbon footprint) more accurate. The first example is concerned with specifying the 'car' modality; in the current situation it is only possible to make the distinction between road traffic, rail traffic, or neither. However, the emission of a road vehicle greatly depends on the specific model, or type (for example a small car versus a truck), combining the travel movement with the specific type of vehicle (and therefore the specific emission) makes the carbon footprint more accurate. Another example of a future improvement to the carbon footprint, in this case to the application of the carbon footprint for policy makers, related to the motive for making a movement. When the incentive for the travel movement can be determined, it is possible to map (or at least gain an insight in) specific conditions and their corresponding carbon emission. For example: there might be a high density of movements by car between two areas, with the motive of these movements being 'going to work'. The emission of these commuters could be lowered by improving the public transport on this route.

Appendix A

Questionnaire

Questions regarding the data quality measures for the individual datasets. Data quality will be determined based on the purpose of the data integration; to create an overview of all movements in the Netherlands. Therefore the desired level of data quality is to come as close to the actual number of movements as possible.

15 questions have been drafted, partially based on the subjective measures from Lee and his colleagues (2002) that are used in the AIMQ methodology. For some of the questions it is important to view the data quality in compliance with the level of quality necessary to reach the research purpose (all movements in the Netherlands). This applies mainly to the dimensions accuracy, completeness, and amount of data, to a small extent it also applies to timeliness. The other three data quality dimensions are more focused on the intrinsic nature of the data (objectivity, believability, and reputation).

The questionnaire is divided into two sections, the first section (regarding the dimensions completeness and timeliness) is meant to gather additional information to specify data quality problems. The second section consists of questions that are rated on a scale from 0 to 10; these questions are derived from (Lee et al., 2002).

Completeness	1. "The information in the dataset is incomplete for the purpose of this research?" 2. "In what aspects is the dataset lacking on the subject of completeness?" 3. "The information covers the needs of our tasks?" 4. "For what tasks is the dataset not usable?"
Timeliness	5. "The information is up-to-date" 6. "When/over what period was the data collected?" 7. "The information is of volatile nature"

TABLE A.1: Data quality dimensions completeness and timeliness and subjective measures

The following data quality dimensions are rated on a scale from 0 to 10:

Accuracy	1. "The information is of sufficient precision for the purpose of this research"
Objectivity	2. "The information is objectively collected" 3. "The information is based on facts"
Believability	4. "The information is credible/believable" 5. "The information is trustworthy"
Reputation	6. "The information has good reputation" 7. "The information comes from good sources"
Amount of data	8. "The amount of information is of sufficient volume for our needs"

TABLE A.2: Data quality dimensions (accuracy, objectivity, believability, reputation, and amount of data) and subjective measures rated on scale from 0 to 10

The list of questions that is presented to data experts on the Mezuro and OViN dataset:

1. "The information in the dataset is incomplete for the purpose of this research? (on a scale from 0 – 10)"
   (a) "In what aspects is the dataset lacking on the subject of completeness? (For example missing data on specific conditions)"
2. "The information covers the needs of our tasks? (on a scale from 0 – 10)"
   (a) "For what tasks is the dataset not usable?"
3. "The information is up-to-date (on a scale from 0 – 10)"
   (a) "When/over what period is the data collected? (Collected once, or continuously?)"
4. "The information is of a volatile nature (on a scale from 0 – 10)"
5. "The information is of sufficient precision for the purpose of this research (on a scale from 0 – 10)"
6. "The information is objectively collected (on a scale from 0 – 10)"
7. "The information is based on facts (on a scale from 0 – 10)"
8. "The information is credible / believable (on a scale from 0 – 10)"
9. "The information is trustworthy (on a scale from 0 – 10)"
10. "The information has a good reputation (on a scale from 0 – 10)"
11. "The information comes from good sources (on a scale from 0 – 10)"

12. "The amount of information is of sufficient volume for our needs (on a scale from 0 – 10)"

Appendix B

Questionnaire results

B.1 Results Mezuro

Quantitative questions on a scale from 0 – 10,

Question number:	**Mezuro #1:**	**Mezuro #2:**	**Mezuro #3:**
1	8	2	5
2	5	8	-
3	8	9	9
4	5	1	-
5	5	8	5
6	8	5	10
7	10	9	10
8	9	8	8
9	9	8	8
10	6	6	6
11	8	8	9
12	10	9	9

TABLE B.1: Table with the results from the data quality assessment for Mezuro

Results to qualitative questions

1a. "In what aspects is the dataset lacking on the subject of completeness? (For example missing data on specific conditions)"

Korte ritten missen. Modaliteit is, zeker voor korte ritten, niet altijd juist.

#1: De vervoerwijze is zeer belangrijk voor berekening van de footprint

#2: Mezuro data does not involve characteristics of detected people, it is anonymous, so mobility parameters can only be derived implicitly. Regional specifications are available given the huge sample. Score 2

2a. "For what tasks is the dataset not usable?"

Movements within Mezuroareas. Hele gedetailleerde eigenschappen van een reis.

#1: Voor bepaling van de carbon-footprint is een adequate bepaling van de typen vervoerwijzen van belang; Daarnaast kan over het interne verkeer binnen een GSM-gebied geen uitspraak gedaan worden

#2: Mezuro data covers many tasks on a more rough scale, but with a huge sample, without knowing any characteristics of travelling people but following them over a whole month: 8.

3a. "When/over what period is the data collected? (Collected once, or continuously?)"

Continuously

#1: Continue per maand

#2: Mezuro data is collected and analyzed continuously, only delayed by process time: 9

B.2 Results OViN

Section for the data quality assessment for the OViN dataset

Question number:	**OViN #1:**	**OViN #2:**	**OViN #3**
1	2	3	5
2	7	7	-
3	8	7	9
4	5	4	-
5	5	3	5
6	6	9	10
7	10	5	10
8	10	8	8
9	8	8	8
10	8	8	8
11	8	7	7
12	5	6	5

TABLE B.2: Table with the results from the data quality assessment for OViN

Results to qualitative questions

1a. "In what aspects is the dataset lacking on the subject of completeness? (For example missing data on specific conditions)"

-

#1: OViN data zeer compleet voor het personenvervoer, geen vrachtvervoer

#2: OViN data gives the complete picture, but only for a limited group, so general travel behavior parameters can be derived, but hardly any regional specifications. Score 3

2a. "For what tasks is the dataset not usable?"

-

#1: Goederenvervoer valt buiten de scope van het OViN; voor kleienere gebieden kan geen statisch verantwoorde uitspraak gedaan worden

#2: OViN data covers many tasks on a detailed scale, but with a limited sample (no reliable geographical details) and travel information only covering one day per person (no insights in the effect of knowledge and experience): 7

3a. "When/over what period is the data collected? (Collected once, or continuously?)"

-

#1: Data wordt continue verzameld gedurende het jaar: per jaar wordt het resultaat gevalideerd en vrijgegeveven

#2: OViN data is collected each year (spread over time) and becomes available in April of the next year: 7

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, *41*(3), 16.

CBS. (2013). *Retrieved from http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=81126ned&D1=0&D2=a&D3=0,12-16&D4=a&HDR=T,G1&STB=G2,G3&VW=T*.

CBS. (2014). *Retrieved from http://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=70946ned&D1=a&D2=a&D3=a&HDR=G2&STB=T,G1&VW=T*.

Churchill, G. A., & Iacobucci, D. (2010). Marketing research: methodological foundations.

Deming, W. E. (1986). Out of the crisis, massachusetts institute of technology. *Center for advanced engineering study, Cambridge, MA*, *510*.

Doan, A., & Halevy, A. Y. (2005). Semantic integration research in the database community: A brief survey. *AI magazine*, *26*(1), 83.

Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). Methontology: from ontological art towards ontological engineering.

Fox, M. S., Barbuceanu, M., & Gruninger, M. (1996). An organisation ontology for enterprise modeling: Preliminary concepts for linking structure and behaviour. *Computers in industry*, *29*(1), 123–134.

Goh, C. H. (1996). *Representing and reasoning about semantic conflicts in heterogeneous information systems* (Unpublished doctoral dissertation). Citeseer.

Gómez-Pérez, A. (1998). Knowledge sharing and reuse. *Handbook of applied expert systems*, 10–11.

Gómez-Pérez, A., Fernández, M., & Vicente, A. d. (1996). Towards a method to conceptualize domain ontologies.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199–220.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, *43*(5), 907–928.

Hammer, M., & Champy, J. (1993). Business process reengineering. *London: Nicholas Brealey*, *444*.

Hammer, M., & Champy, J. (2009). *Reengineering the corporation: Manifesto for business revolution, a*. Zondervan.

IPCC. (2014). *Retrieved from http://www.ipcc.ch/report/ar5/wg3/*.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, *45*(4), 184–192.

Kriebel, C. (1979). Evaluating the quality of information systems. *design and implementation of computer based information systems*, 29–43.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). Aimq: a methodology for information quality assessment. *Information & management*, *40*(2), 133–146.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first acm sigmod-sigact-sigart symposium on principles of database systems* (pp. 233–246).

NS. (2014). *Retrieved from http://nsjaarverslag.nl/jaarverslag-2014/s1462_co/a1393_Energieverbruik-en-CO%E2%82%82-uitstoot*.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218.

Pottinger, R. A., & Bernstein, P. A. (2003). Merging models based on given correspondences. In *Proceedings of the 29th international conference on very large data bases-volume 29* (pp. 862–873).

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *the VLDB Journal*, *10*(4), 334–350.

Redman, T. C., & Blanton, A. (1997). *Data quality for the information age*. Artech House, Inc.

Shankaranarayan, G., Wang, R. Y., & Ziad, M. (2000). Modeling the manufacture of an information product with ip-map. In *Proceedings of the 6th international conference on information quality*.

Skuce, D. (1995). Conventions for reaching agreement on shared ontologies. In *Proceedings of the 9th knowledge acquisition for knowledge based systems workshop*.

Spruit, M. (2013). Selecting data quality dimensions: towards a business impacts assessment. *6th World Summit on the Knowledge Society*.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Tan, J. K., & Benbasat, I. (1990). Processing of graphical information: A decomposition taxonomy to match data extraction tasks and graphical representations. *Information Systems Research*, 416–439.

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, *11*(02), 93–136.

Uschold, M., King, M., Moralee, S., & Zorgios, Y. (1998). The enterprise ontology. *The knowledge engineering review*, *13*(01), 31–89.

van de Weerd, I., & Brinkkemper, S. (2008). Meta-modeling for situational analysis and design methods. *Handbook of research on modern systems analysis and design technologies and applications*, *35*.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hübner, S. (2001). Ontology-based integration of information-a survey of existing approaches. In *Ijcai-01 workshop: ontologies and information sharing* (Vol. 2001, pp. 108–117).

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 5–33.