MASTER’S THESIS

A DATA INTEGRATION DESIGN APPROACH FOR THE PLANNING PROCESS OF A PUBLIC TRANSPORT OPERATOR

Erik D. van der Kuil

PROGRAM
Master of Science Business Information Technology

SPECIALIZATION
IT Management & Enterprise Architecture

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE
Department of Services, Cybersecurity & Safety

FACULTY OF BEHAVIOURAL, MANAGEMENT AND SOCIAL SCIENCES
Department of Industrial Engineering & Business Information Systems

GRADUATION COMMITTEE

dr. ir. M.J. van Sinderen	University of Twente
dr. ir. J.M. Moonen	University of Twente
F.F. van de Velde, MSc	GVB Exploitatie BV

DOCUMENT CONFIDENTIALITY
Public

17 December 2020

MASTER'S THESIS

A DATA INTEGRATION DESIGN APPROACH FOR THE PLANNING PROCESS OF A PUBLIC TRANSPORT OPERATOR

17 December 2020

Master's thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science in Business Information Technology

AUTHOR

Name	Erik D. van der Kuil
Study program	Master of Science Business Information Technology
Specialization	IT Management & Enterprise Architecture
Institute	University of Twente, Enschede, The Netherlands
Faculty	Electrical Engineering, Mathematics and Computer Science (EEMCS) and Behavioural, Management and Social Sciences (BMS)
Email address	e.d.vanderkuil@alumnus.utwente.nl
Host institution	GVB Exploitatie BV, IT & Innovation, Information Management and Architecture (IMA)
Internship period	February 2020 – December 2020

GRADUATION COMMITTEE

dr. ir. M.J. (Marten) van Sinderen	Associate professor	University of Twente	Faculty of EEMCS
dr. ir. J.M. (Hans) Moonen	Assistant professor	University of Twente	Faculty of BMS
F.F. (Frank) van de Velde, MSc	Enterprise architect	GVB Exploitatie BV	IT & Innovation, IMA

PREFACE

Writing this preface of my master's thesis marks the end of my time as a student at the University of Twente. Slightly more than six years after starting the bachelor Business & IT, I am very proud to finish my master's study and grateful to become a Master of Science in Business Information Technology. Many fantastic years passed by, through which I was privileged to get high-level education, work on great projects, be a student assistant, organize the Bachelor Open Days, study and live in Budapest, work as a research intern at KLM Royal Dutch Airlines, graduate at GVB, meet many inspiring people, make friends for life and find my true self. It has been a great ride!

Everyone who knows me personally, knows that I have a passion for public transport and, especially, for buses. For this reason, graduating at GVB was a perfect combination of my personal interests and my field of study. I have learned a lot about both business and IT in the field of public transportation, especially within the planning domain, and I am proud of the results presented in this thesis.

When starting this graduation project in February 2020, nobody could have expected that the world was going to change drastically. The impact of Covid-19 was visible everywhere. Also at GVB, where in the first weeks of the pandemic only 10% of the passengers were left compared to a year before. From mid-March we started working from home, which resulted in a completely different graduation internship than I expected. Fortunately, I got a lot of understanding and my contract at GVB got extended.

The research presented in this master thesis could not have been carried out without the help of many people. I am grateful that I had a very interesting, understanding, supportive and intelligent daily supervisor from GVB. Thank you very much, Frank van de Velde! Without your help, opinion, advice and motivation I would by far not have achieved this result. Second of all, I want to thank Marten van Sinderen and Hans Moonen from the University of Twente for their critical view, feedback and answers to my questions, which helped to keep up the scientific level. Furthermore, I want to thank the respondents of the interviews, the experts who took part in the validation and my colleagues at GVB, especially from the Information Management and Architecture team. Many thanks also go to Paul, Merijn, Evelyn, Peter and Marcel for their feedback on and proofreading of my thesis. Last, but for sure not least, I would like to thank my friends and family. Without their support, positive attitude and independent insights, I could not have achieved this result.

I hope the provided insights and designed artifact will be used to improve the current planning process of public transport operators and that you will enjoy reading this master's thesis. Thank you for your interest and feel free to reach out to me in case you have any questions.

Erik van der Kuil
Amsterdam, 14 December 2020

EXECUTIVE SUMMARY

Public transport operators (PTOs), among many other companies nowadays, are experiencing an organizational shift towards an expanding role of data and IT in their business processes. Many embedded applications are dependent on each other's and external data sources. In addition, data-driven business is emerging and as such, many data integrations between business processes and departments are necessary. This number will very likely grow in the future, as business and technology are continuously developing.

The PTO's planning process is crucial for offering public transportation and is a very data-intensive process. It consumes and provides a vast amount of data from and to different providers and consumers, both internally as well as externally. It is necessary to carefully manage these vast amounts of data to reach the stakeholders' goals for this research: improve the planning process, offer better IT service quality to the business, be prepared for future integrations and save costs.

Data management is a comprehensive research area. This research focuses on the data integration aspect and provides a data integration design approach for the planning process of a PTO. The main research question is: 'What constitutes a good data integration design approach for the planning process of a public transport operator?'. The artifact consists of a target data architecture and an approach to move towards this architecture, and is developed by using TOGAF ADM.

The target data architecture for the planning process is based on literature and GVB's practice and accounts for data integration challenges, as well as for the planning process, its tasks and data requirements. The in-depth planning process analysis contains the data providers, input, output and data consumers of the planning process. Furthermore, data standards in the field of public transportation and common data integration methods are used.

Subsequently, the approach contains an implementation example, data authorizations, data quality aspects and data roles. Authorizations and roles are related to different planning process phases, as identified during the in-depth analysis explained above. The data quality aspects are used to mitigate the identified challenges in literature and practice.

The data integration design approach is validated by experts in the field of public transport planning (business and IT). The validation showed that the proposed artifact can help to reach a better data integration situation. Firstly, goals related to actual data are reached (decoupling, reuse, correct and up-to-date data and preparation for future integrations). Secondly, the approach was validated positively because it accounts for clarity, provision of data quality aspects and provision of data responsibility. Thirdly, business process improvements are enabled by the design. Examples are the integration of planning phases (for optimization purposes) and dynamic planning.

The conducted research has resulted in a comprehensive analysis of a PTO's planning process and its data requirements. The target data architecture is based on this analysis and accounts for all important data entities within the planning process, which can be used as a reference architecture for PTOs. Furthermore, the data architecture contributes to academic research as it provides a building block for optimizing the planning process by integrating planning process phases, which is a relevant study area in operations research. The approach turned out to be useful for PTOs to help them move towards the target architecture. Data roles are identified and made responsible for data quality aspects, which helps to overcome the identified data integration challenges from practice and literature. In conclusion, the data integration design approach contributes to reaching the stakeholder goals of this research.

TABLE OF CONTENTS

LIST OF ACRONYMS

Acronym	**Description**
AC	Assign Crew (planning phase; Section 4.3)
ADM	Architecture Development Method (Section 2.1.3)
AI	Artificial Intelligence
API	Application Programming Interface
AV	Assign Vehicles (planning phase; Section 4.3)
BISON KVs	Beheer Informatie Standaarden OV Nederland (Koppelvlakken) (Dutch platform for public transport data standardization and their actual standards (KVs))
BV	Besloten Vennootschap (Dutch; comparable to Limited Liability Company (LLC))
CAO	Collective Labor Agreement (original Dutch: Collectieve Arbeidsovereenkomst)
CDM	Canonical Data Model (Section 3.3.2)
CEN	European Committee for Standardization (original French: Comité Européen de Normalisation)
CIA	Confidentiality, Integrity, Availability (Section 2.2.2)
CISO	Chief Information Security Officer
CRUD	Create, Read, Update and Delete
CT	Control Transport (phase; Section 4.3)
DN	Design Network (phase; Section 4.3)
DSM	Design Science Methodology
EA	Enterprise Architecture (Section 1.1.3)
EMV	Europay, Mastercard and VISA (Section 1.1.2)
EPIP	European Passenger Information Profile (NeTEx profile)
ESB	Enterprise Service Bus
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
HR	Human Resources
IDSA	International Data Spaces Association (Section 2.5.2)
IoT	Internet of Things
IS	Information Systems
ISO	International Organization for Standardization
IT	Information Technology
ITxPT	Information Technology for Public Transport (Section 1.1.2)
NeTEx	Network Timetable Exchange (Section 1.1.2)
NV	Naamloze Vennootschap (Dutch; comparable to Incorporated (Inc))
OV	Openbaar Vervoer (Dutch for 'public transport')
PC	Plan Crew Rosters (planning phase; Section 4.3)
PL	Plan Lines (planning phase; Section 4.3)
PT	Plan Timetable (planning phase; Section 4.3)
PTO	Public Transport Operator (Section 1.1.2)
RPA	Robotic Process Automation
RQ	Research Question (Section 1.3.4)
SC	Schedule Crew Duties (planning phase; Section 4.3)
SIPOC	Supplier, Input, Process, Output, Consumer (Section 2.3.3)
SIRI	Service Interface for Real-time Information (Section 1.1.2)
SLA	Service Level Agreement
SLR	Systematic Literature Review (Section 2.1.1)
SV	Schedule Vehicle Blocks (planning phase; Section 4.3)
TA	Transport Authority (Section 1.1.2)
TAP/TAF TSI	Telematics Applications for Passenger/Freight Services Technical Specification for Interoperability
TOGAF	The Open Group Architecture Framework (Section 2.1)
UIC	Worldwide Railway Organization (original French: Union Internationale des Chemins de fer)
UITP	International Association of Public Transport (original French: L'Union Internationale des Transports Publics)
UT	University of Twente
UTAUT	Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003)
VRA	Vervoerregio Amsterdam / Transport region Amsterdam
XML	Extensible Markup Language
ZE	Zero Emission

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The role of IT has been increasing in many enterprises. Kirkpatrick (2011) confirms this statement in Forbes, mentioning that "every company is a software company", to which Microsoft's CEO Nadella (2018) added that "every company is a digital organization". This organizational change towards an increasing role of IT is also present at public transport operators (PTOs) which inherently become a more digitized business. In the past, public transportation was only about mass public transport (combining the aspects of public access and collective use) (UITP, 2019a). Nowadays, it goes far beyond this definition and includes many more aspects than the holistic system of transporting passengers from point A to point B. The boundaries of the definition are being pushed towards more digital, data-driven and people-driven (combination of different) services (UITP, 2019a).

The frontier is being pushed by the development of new mobility services, such as ride-hailing, moped-hailing and bike-sharing. These new services are slowly being incorporated in the definition of public transport (Arneodo, 2015; UITP, 2019a) by means of the Mobility-as-a-Service concept (UITP, 2020). Furthermore, technologies such as automated vehicles (Wray, 2020) and artificial intelligence (UITP, 2020) will very likely be introduced in the field of public transportation within several years. Another example of an important development is the crowdedness indicator of public transport vehicles (Metselaar, 2020) and the PTO's demand to better and real-time react to unforeseen circumstances in the real world.

All these new services and technologies produce a lot of data. In turn, these services require a lot of data to meet the demands of the customers. These vast amounts of data need to be managed somehow, which also gets a lot of attention from organizations such as The International Association of Public Transport (UITP, 2020). When having proper data integration in place, initiatives based on these vast amounts of data can be started. Examples of these range from improved network planning methods to predictive maintenance of vehicles (UITP, 2017).

Having vast amounts of data available for the business increases possibilities for improving their services. Data-driven public transportation services are slowly being introduced within the sector (Busscher, 2020). However, within PTOs, these vast amounts of data are most often not available to the entire enterprise, which refrains from the opportunities the data can offer (Busscher, 2020).

Increased use of IT, data, devices, technologies, third-party services and data-driven business lead to an immense need for integration between different applications, data, companies and stakeholders, including for PTOs (Arneodo, 2015). Without these integrations of processes, applications and data, the services as demanded by the customers cannot be offered sufficiently. Hence, the customer demands make it such that PTOs have to increase their IT skills, capabilities and maturity.

Data management is often stated to be the domain in which data integration takes place (DAMA International, 2014; Doan et al., 2017; Halevy et al., 2005; Hausladen & Schosser, 2020; Pratama et al., 2018). As the role of data is becoming more important, the role of data management increases accordingly. According to The Open Group (2018), a proper approach to data management "enables the effective use of data to capitalize on its competitive advantages". Hence, data management enables business value improvement.

This research is conducted at GVB, the PTO in the city of Amsterdam, and aims to contribute to the current data management situation for PTOs by developing a data integration design approach for the planning process of a PTO. The data integration design approach consists of a *target data architecture* and an *approach* of how to use this architecture in practice.

1.1 Background

In this section, background information that describes the context, assumptions, definitions and fields of research which are seen as the starting point of this research is provided. At first, GVB is introduced. This is Amsterdam's PTO and the company at which this master's thesis was conducted at. Secondly, background knowledge is provided and is grouped into sections concerning the following topics: public transportation with an emphasis on its planning process, enterprise and data architecture and data management and data integration.

1.1.1 GVB

GVB is a Dutch PTO offering services in the Dutch capital, Amsterdam, by means of four different modalities: bus, tram, metro and ferry. In the past, GVB was the acronym for 'Gemeente-vervoerbedrijf' (municipality's transport operator). However, since GVB's privatization in 2004, the acronym has become the company's full name. The holding company, GVB Holding NV, consists of six subsidiaries (private companies), see Figure 1-1. The municipality of Amsterdam is 100% shareholder of GVB Holding NV (GVB Holding NV, 2020a).

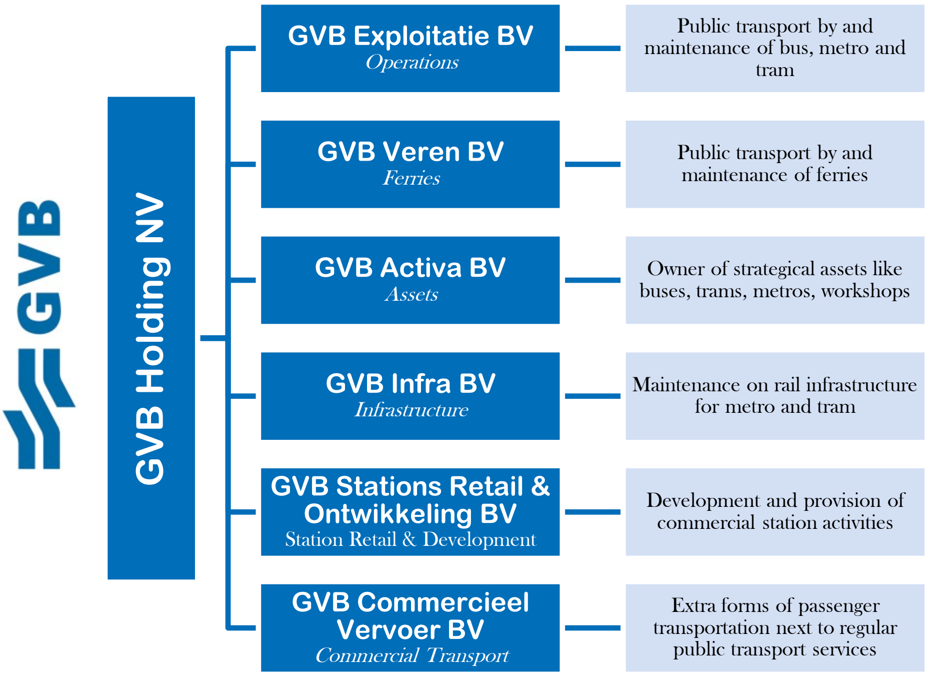
GVB
GVB Holding NV
GVB Exploitatie BV
Operations
Public transport by and maintenance of bus, metro and tram
GVB Veren BV
Ferries
Public transport by and maintenance of ferries
GVB Activa BV
Assets
Owner of strategical assets like buses, trams, metros, workshops
GVB Infra BV
Infrastructure
Maintenance on rail infrastructure for metro and tram
GVB Stations Retail & Ontwikkeling BV
Station Retail & Development
Development and provision of commercial station activities
GVB Commercieel Vervoer BV
Commercial Transport
Extra forms of passenger transportation next to regular public transport services

Figure 1-1 GVB Holding NV and its subsidiaries (adapted from GVB Holding NV (2020a))

The 2019 annual reports (GVB Activa BV, 2020; GVB Holding NV, 2020a) shed a light on the size of GVB. They own a total of 203 buses, 200 trams, 90 metros and 18 ferries. These vehicles and vessels cover a total of 33 bus lines, 10 night bus lines, 15 tram lines, 5 metro lines and 9 ferry lines. On average, in 2019, 938,000 passengers were transported every business day, which resulted in a total of 1,033 million passenger kilometers and a revenue of €462.5 million that year. This result was realized by 5,000 employees (3,557 FTEs).

The current license to operate the public transportation in Amsterdam (contract called 'Concessie Amsterdam') started in December 2013 and lasts until December 2024. The license is given to GVB Exploitatie BV by Vervoerregio Amsterdam (VRA). VRA is the transport authority (TA) of the Amsterdam region and is a partnership of fifteen municipalities in the region of Amsterdam. A subsidy is given by VRA to GVB Exploitatie BV, which is in turn funded by the municipality of Amsterdam

(GVB Holding NV, 2020a). This subsidy decreases every year, which means that GVB has to become a profitable company running the business while being fully dependent on passenger revenue. In 2014, GVB's cost coverage ratio was 73.6%, whereas this ratio grew to 99.4% in 2019 (GVB Holding NV, 2020a). This means that, in 2019, only 0.6% of the revenue came from subsidy. The global Covid-19 pandemic changed this situation (temporarily), since subsidy is given to every Dutch PTO because of the immense drop in passenger numbers (Rijksoverheid, 2020). The decreasing subsidy and any unforeseen circumstances emphasize the importance of having proper data management, since it will help to work more efficiently and to better and real-time react to unforeseen circumstances.

1.1.2 Public Transportation

Public transportation is "a system of vehicles such as buses and trains that operate at regular times on fixed routes and are used by the public" (Cambridge University Press, n.d.-b). This definition is too narrow for the scope of this research, since it does not consider demand-responsive transport and it only considers transport that takes place on land, implicitly on roads and rail tracks. In the case of GVB and the city of Amsterdam, ferries also belong to the public transport network. Amsterdam is not the only city in which ferries are a part of the public transport network. They can also be found in, among others, Budapest (BKK Budapest, n.d.), London (TfL London, n.d.) and Rotterdam (RET NV, n.d.). Therefore, the research will adhere to the following definition of public transportation:

Public transportation is a system of passenger transport modalities such as buses, trains and ferries that operate at regular times on fixed routes and are used by the public.

Every PTO has a planning process in place in order to offer these public transport services. This process is shortly described in the next sub-section, followed by an introduction of data standards that are commonly used within the public transportation domain.

1.1.2.1 Planning Process

The planning process of a PTO starts on a strategic level with the planning of bus stops, tram stops, bus lanes, metro stations, rail tracks and so on. Moreover, a transport authority (TA) often provides the PTO with a plan or contract, which defines high-level transport service requirements. These requirements, combined with the infrastructure, form the basis of the transport services. In several steps, the planning goes from contract to timetable and finally to individual journeys, vehicle schedules (so-called 'vehicle blocks') and crew schedules. As part of this research, the planning process of a PTO will be discussed in more detail in Chapter 4.

The planning steps mentioned above can only be carried out when all the necessary data is available. One can think of data about the infrastructure, vehicles and personnel. The planning process requires a lot of data from other business functions as well as from external parties. Moreover, it also provides vast amounts of data to other business functions within the enterprise. These complex relationships between data sources constitute the data dependencies of this system, which will be discussed in Chapter 5.

1.1.2.2 Data Standards

In modern public transportation environments, the European NeTEx and SIRI standards are used for exchanging data. These are based upon the conceptual model Transmodel. This model offers a clear overview of all main data objects and their relationships (CEN, 2019). Transmodel, NeTEx and SIRI are all developed by the European Committee for Standardization (CEN). NeTEx is responsible for the data exchange concerning the network and schedule (long-term), whereas SIRI is developed for the real-time data exchange about individual journeys (short-term). More details about Transmodel, NeTEx and SIRI are provided in Section 3.2.

A framework in the field of public transportation is proposed by Information Technology for Public Transport (ITxPT), which is a non-profit organization consisting of 121 members (in 2019-Q4 (ITxPT, 2020)) within the (IT) public transport business: IT suppliers, transport authorities, PTOs and vehicle manufactures. ITxPT defines its interoperability framework on three different levels:

hardware level, communication protocol level and service level (ITxPT, n.d.). To reach interoperability and offer the possibility to share data, ITxPT endorses the importance of using widely accepted standards in their framework. For this reason, it is based on Transmodel, NeTEx and SIRI (ITxPT, n.d.).

1.1.3 Enterprise Architecture, Data Architecture and Approach

As the goal of this research is to propose a data integration design approach consisting of a target architecture and an approach, the definitions of these concepts in the context of this research are given.

1.1.3.1 Enterprise Architecture

Saint-Louis et al. (2019) have shown recently that many differences and divergences between the definitions of enterprise architecture (EA) are present in literature. This finding was based on a systematic literature review of 102 journal articles containing 160 definitions of EA. As the goal of this research is not to provide a deep and unified understanding of EA, the EA definition provided by The Open Group (2018) is used for this research, because their framework is applied to this research (as explained later in Section 2.1.3):

Enterprise architecture is "an architecture which crosses multiple systems, and multiple functional groups within the enterprise" in which the enterprise is "the entire enterprise, encompassing all of its business activities and capabilities, information, and technology that make up the entire infrastructure and governance of the enterprise, or to one or more specific areas of interest within the enterprise" (The Open Group, 2018).

1.1.3.2 Data Architecture

Data architecture is part of the broader enterprise architecture. For consistency purposes, the definition from The Open Group (2018) is also used for data architecture:

Data architecture is "a description of the structure and interaction of the enterprise's major types and sources of data and logical data assets (...)" (The Open Group, 2018).

To further operationalize the data architecture definition as provided above, its objective is to "enable the business architecture" (The Open Group, 2018). In other words, the data architecture should be an enabler of the business process defined in the business architecture. The missing part in the definition of data architecture is the description of the structure and interaction of 'data management resources' (The Open Group, 2018). This part is covered in the approach proposed by this research, as explained hereafter.

1.1.3.3 Approach

The usage of an approach differs in research, which leads to different interpretations. The definition of the designed approach in this research is related to the Cambridge University Press (n.d.-a) definition: "a way of considering or doing something".

The **approach** in the context of this research is the way of doing something and is based on considerations regarding the implementation of the proposed target data architecture.

1.1.4 Data Management and Data Integration

Data integration can be seen as part of data management. According to DAMA International (2014) data management is "an overarching term that described the processes used to plan, specify, enable, create, acquire, maintain, use, archive, retrieve, control, and purge data. These processes overlap and interact with each data management knowledge area". One of these knowledge areas is data integration and interoperability. All knowledge areas identified by DAMA International (2014) can be found in Figure 1-2.

Data governance is seen as the middle knowledge area and thereby interacting with every other area. This is because data governance is about "planning, oversight, and control over management of data and the use of data and data-related resources" (DAMA International, 2014) and is therefore governing every knowledge area.

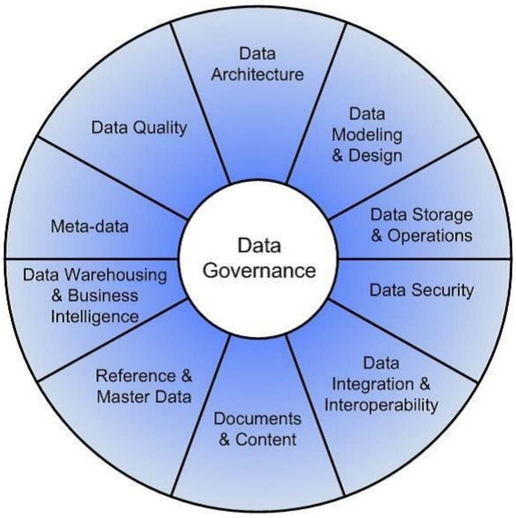

Figure 1-2 DAMA DMBOK2 data management framework (DAMA International, 2014)

Within an IT landscape, it is very common that one or more applications are using functionality and/or data from one or more other applications. This integration of functionality or data means that applications become dependent on each other's outcomes. To realize this dependency, the integration should be established in such a way that both the providing and consuming application function properly, taking into account both data and interaction understanding. For this research, we will focus on the understanding of data, whereas the interaction understanding is considered to be a more technical aspect, which is out of scope.

The definition of data integration is not entirely agreed upon in the literature. Some authors restrict data integration to combining database schemas into a unified schema (Batini et al., 1986) or creating a data warehouse based on different sources that is used for data analytics (Salguero et al., 2008). Integrating data can take place with enterprise's data, but also with external data (Doan et al., 2012), structured and unstructured data (Bahga & Madisetti, 2015) and vast data volumes for big data purposes. Bernstein and Haas (2008) add to this that data integration should be seen as all kinds of information reuse in (a) system(s) different than the source system. For this research, we adhere to DAMA International's (2014) data integration definition as proposed in their data management framework:

"**Data integration** uses both technical and business processes to merge data from different sources, with the goal of accessing useful and valuable information, efficiently" (DAMA International, 2014).

1.2 Problem Statement

Data management is important for every organization, since data "is a valuable asset which must be managed properly to ensure success" (DAMA International, 2014). The data management situation at GVB turned out to be sub-optimal after exploratory research was carried out. This is shown through duplicate data storage, contradicting data, no reuse of data integrations (point-to-point connections), often transformed data (high risk of data loss) and no clear governance about data ownership. Furthermore, real-time data exchange is hardly used within GVB, because data is often integrated through batch processing performing file transfers only on scheduled moments of a day or week. This means that processes are rather inflexible, which is a drawback because quickly adapting the IT to the business' needs becomes more difficult.

Batch processing in public transport is also recognized by Scholz (2016) as a downside. This is because any change that occurs after the file is transferred, is not taken into account by the consuming application. This one-way street means that consuming applications are not aware of any changes, which might lead to the use of different data values for the same object. Moreover, data integration challenges are already known for many years within the world of data management and are hard to address, mostly due to the different understandings and interpretations from stakeholders (Jarke et al., 2014).

The planning process of a PTO is a very data-dependent business function for the enterprise (as explained in Section 1.1.2 and Section 3.1.1). The process depends on vast amounts of diverse data, but at the same time also serves many other business functions that are provided with data generated in the planning process. Furthermore, the planning process is one of the primary business functions of a PTO and is crucial for offering public transportation. Also, it is expected that the ongoing development of new initiatives for public transportation will continue, which results in more (requests for) data.

Current data integration technologies and methods lack the specified context of the PTOs' planning process and are focused on the technology only. They provide technical solutions to the integration problem, but do not provide a solution for integration on data-level. On the contrary, operations research studies such as Ceder (2016), Qiu et al. (2018) and Scholz (2016) clearly describe processes and data, but do not provide a solution to the data management and data integration problem.

In conclusion, due to the complex data dependencies (both consuming and providing) and vast amounts of data within the planning process, proper data management including a clear data integration solution is highly necessary. This allows the planning process to use and provide correct, up-to-date and high-quality data, which can result in a more optimal planning. Additionally, it enables improvements of the planning process by also taking into account other kinds of data (e.g. maintenance planning, weather and congestions). Currently, no data management and data integration solution is available which suits a PTOs' planning process, which complicates the data exchange and hinders innovation.

1.3 Research Design

This section aims to explain the research design, which includes the objective, scope, stakeholders, research questions, relevance, process and a research overview.

1.3.1 Research Objective

Following the problem statement presented in Section 1.2, it becomes clear that data management for the planning process of a PTO must be carefully addressed. Since data management is a very complex and broad field of study, the focus of this research is on the data integration aspect (Figure 1-2). The primary objective for this research is to provide a PTO with a data integration design approach such that the data management situation can be improved. GVB's situation is used as the starting point for this research.

The data integration design approach consists of a target data architecture and an approach on how to use this data architecture in practice. This latter part accounts for considerations while implementing the target data architecture. Together, they are supposed to save costs, improve the planning process, offer better service quality and to prepare the enterprise for future integrations in the ever-changing world of public transportation and IT.

The research objective for this research can be applied to the design problem template as proposed by Wieringa (2014), see Figure 1-3. This template ensures a clear view of what the problem context, artifact, requirements and stakeholders' goals are (Wieringa, 2014).

Improve	<a problem context>
by	<(re)designing an artifact>
that satisfies	<some requirements>
in order to	<help stakeholders achieve some goals>.

Figure 1-3 Design problem template (Wieringa, 2014)

Applying this pattern to the problem statement, the design problem can be formulated as presented in Figure 1-4.

Improve	the data management situation of the planning process of a public transport operator
by	proposing a data integration design approach
that satisfies	the need for data within the planning process and for adjacent business areas and data quality aspects from the field of data management
in order to	save costs, plan public transportation more efficiently and flexibly, offer better IT service quality and be prepared for future data integrations.

Figure 1-4 Design problem of this research according to Wieringa's (2014) template

1.3.2 Research Scope

The scope of this research is limited to the planning process of a PTO (the definition is provided in Section 1.1.2). This is because the planning process comes with a vast number of data dependencies, both consuming and providing. Furthermore, the planning process is one of the key activities within a PTO's business and provides a key role in many other business functions. The actual transport operations are not seen as part of the planning, yet the operations can trigger changes in the planning process. For this reason, the data exchange between the transport operations and the planning is accounted for. More details about this are provided in Section 4.3.

As explained in Section 1.3.1, GVB's situation is used as a starting point for this research. Their data integration situation is assessed, important aspects for the planning process are derived and stakeholder goals are formulated. However, the goal of this research is to design a data integration design approach that is valid for any PTO, thus no GVB-specific solution will be provided. How this research accounts for the general PTOs' operations research perspective and input from GVB's situation is explained in Chapter 2. In Chapter 9, the applicability of the proposed artifact to GVB's situation is addressed.

Scoping the data management improvement to data integration, data architecture and data quality (based on DAMA International's (2014) data management framework shown in Figure 1-2) is done because, otherwise, the research would become too comprehensive. Since data management and data integration problems are already present for a long time (Jarke et al., 2014), it was decided to focus on a smaller set of data management principles.

1.3.3 Stakeholders and their Goals

According to Wieringa (2014), the identification of stakeholders and their goals is important since they are the source of the goals and constraints, and thus requirements, for the solution. The stakeholders and their goals for this research are partly set beforehand and partly added and adapted during the research process. This iterative process took place because of new insights and findings during the research. An overview of stakeholders and their goals is presented next. For this overview, the stakeholder taxonomy – especially the stakeholder types (in *italic* hereafter) – as proposed by Alexander (2005) is used. More details about this taxonomy can be found in Appendix A, which includes an explanation per stakeholder type.

Normal operators can be seen as the end-users of the data integration design approach. These are the employees responsible for the IT, data and enterprise architecture within the PTO and their development teams, for (IT) projects and the enterprise's strategy. They are supported by *maintenance operators* and *operational support*, which is expected to be provided by a specialized group of IT/data architecture employees.

The *functional beneficiary* of the artifact is the PTO's business (especially the planning department) and IT. IT offers services to the PTO's business in order to offer the demanded services. Hence, the business is offered a better solution for data usage and data integration, which ensures more flexible and quicker development. The IT department benefits from the artifact as well, since reuse of data integrations is enhanced and therefore the maintenance and development tasks will become easier.

The *interfacing systems* consuming public transport data also benefit functionally from the data integration solution as proposed by the data integration design approach.

The financial benefits of the data integration design approach are expected to become apparent after some time (when actual reuse is going to take place). The PTO is expected to be the *financial beneficiary*, as both business (quicker and more flexible IT) and IT (reuse) benefit financially from the artifact. *Political beneficiaries* and *threat agents* have not been identified. Possible *negative stakeholders* are IT employees and business stakeholders who do not see the added value of data integration, reuse of (data) integration interfaces and a proper data integration situation.

Stakeholders involved during the development of the artifact (*sponsor*, *purchaser*, *developer*, *consultant* and *supplier*) are highly necessary for the development of the artifact, since they also are an important source of goals and requirements (Wieringa, 2014). These are GVB, University of Twente, the researcher, supervisors and all interview and validation respondents from other PTOs and public transport-related enterprises and organizations who took part in this research.

As already briefly touched upon during the identification of the stakeholders above and added and altered during the research process, the stakeholders' goals are formulated as follows:

- Improve the IT situation which leads to less maintenance and development resulting in cheaper and more flexible IT:
  - Always use the correct, most up-to-date and high-quality data;
  - Reuse data and data integrations;
  - Be better prepared for future (data) integrations by developing more easily, flexibly and against lower costs.
- Improve business processes when having a proper data integration situation:
  - Integrate planning phases in order to reach global planning optimization;
  - Plan more dynamically (real-time);
  - Better align with other planning tasks such as maintenance planning;
  - Introduce depot management practices;
  - Better schedule battery-equipped vehicles/vessels (opportunity charging);
  - Introduce self-rostering for crew members.

During the validation of this research (Chapter 8), the abovementioned goals are validated by experts in the field who can be identified as *normal operators*, see Section 2.6.1.

1.3.4 Research Questions

To achieve the research objective (Section 1.3.1) and the stakeholders' goals (Section 1.3.3), the following main research question is defined:

> What constitutes a good data integration design approach
> for the planning process of a public transport operator?

The main research question is answered through several sub research questions (RQs):

1. What is the current data integration situation within public transportation and what data standards, data challenges and data integration methods exist?
   Rationale: Having a proper and realistic knowledge base about the subject is key in proposing a solution. Hence, the data integration situation of a PTO is assessed and literature about this topic is studied.

2. What is the best-practice planning process of a public transport operator?
   Rationale: The planning process determines which decisions and in which sequence they have to be made in order to offer public transportation services. It will provide a grounded base for the artifact.

3. What changes in the planning process that are highly dependent on data (integration) are foreseen?
   Rationale: Designing an artifact for the as-is situation is not future-proof. Based on the answers to these questions, the artifact also accounts for upcoming changes foreseen in the planning process.

4. What data requirements (both internally, to adjacent business areas, and externally) does each step in the planning process have?
   Rationale: Based on the planning process and the decisions to be made (RQ2) and future improvements (RQ3), data requirements can be set up for every step in the planning process.

5. What data quality aspects are important to address when defining a data integration design approach for the planning process?
   Rationale: From the domain of data management, data quality aspects are determined. These aspects play an important role in data management to reach enterprise-wide goals and benefits.

6. How can all insights be combined into a data integration design approach in order to ensure cost savings, offer better service quality and be prepared for future data integrations?
   Rationale: The actual development of the data integration design approach, accounting for the requirements identified in RQ3 – RQ5 in line with the method explained in Chapter 2.

7. How well does the proposed data integration design approach contribute to the data integration situation of a public transport operator's planning process?
   Rationale: The validation of the proposed data integration design approach in RQ6, which accounts for the stakeholders' goals as identified in Section 1.3.3.

8. Is the proposed data integration design approach generalizable to other public transport operators and other industries?
   Rationale: To increase the applicability of the proposed (RQ6) and validated (RQ7) data integration design approach, the generalization is studied.

The relations between the research questions as introduced above can be found in Figure 1-5.

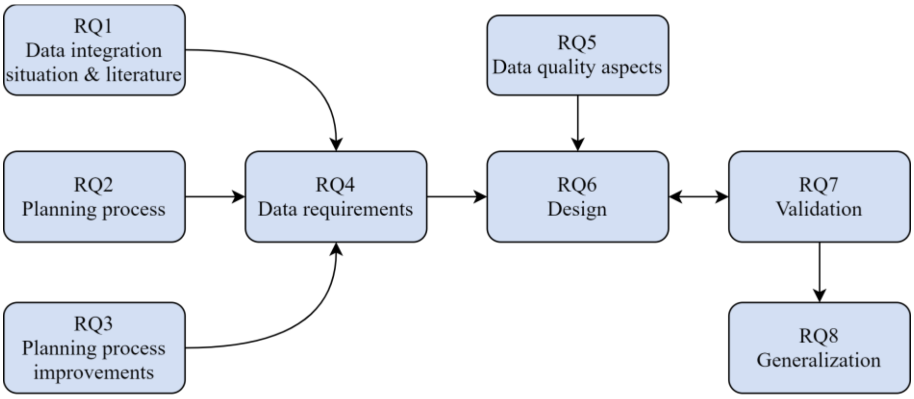

Figure 1-5 Relations between research questions

1.3.5 Research Relevance

The outcomes of this research are expected to be relevant because of several aspects. This research focuses on the data integration from a business process point of view. This means that – in addition to providing a data integration design approach – also the planning process is analyzed thoroughly. PTOs can use this outcome to verify their planning process based on a vast number of data sources. Additionally, it establishes a connected view between business and IT departments, which in practice often operate independently.

Using this outcome as a starting point for the data integration design approach leads to the next relevant aspect, which is the actual design of an optimal data integration situation. Moving towards such a design ensures that possibilities for improving the planning process and incorporating new technologies, which modern-day PTOs are facing due to the aforementioned developments in customer demands, are taken into account. Doing this without any considerations and guidance would probably not lead to success, due to the data management complexity. For this reason, the approach is useful. It points out important aspects to consider and applies this to the practice of a PTO. The target data architecture together with the approach is expected to be relevant for PTOs, because it can improve the data integration situation of the data-driven planning process, and thus increase the data management maturity.

1.3.6 Research Process

To solve the design problem as stated in Section 1.3.1, the Design Science Methodology (DSM) by Wieringa (2014) is used. This methodology provides guidelines for doing design science research in the field of information systems and software engineering. These guidelines lead from the problem investigation to the actual design and validation of an artifact in context. When dealing with real-world problems, the design process can be summarized in three steps: problem investigation, treatment design and treatment validation (Wieringa, 2014). These steps together are called the design cycle, which is visualized in Figure 1-6.

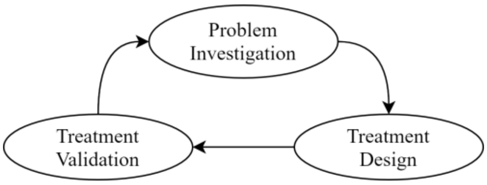

Figure 1-6 Design cycle (adapted from Wieringa (2014))

The goal of the problem investigation step is to "prepare for the design of a treatment by learning more about the problem to be treated" (Wieringa, 2014). This step is about identifying stakeholders and their goals, studying phenomena leading to the problem statement including their effects. The next step, treatment design, is the design of an artifact that should interact with a problem context, which is then called a treatment (Wieringa, 2014). In this step, requirements that contribute to the stakeholders' goals are set, available treatments are taken into consideration and a new treatment is designed. The last step in the design cycle is called treatment validation. This is the justification that the treatment would contribute to stakeholders' goals. Based on this validation, a design theory could be developed. This theory is a substantiated prediction of what would happen if the artifact is implemented, i.e. if it is transferred to the problem context (Wieringa, 2014).

For this research, the problem investigation step is mostly based on the results of RQ1. This research question is answered in a standalone research by the same author and led to the problem statement and research objective as stated in Section 1.2 and 1.3.1, respectively. The results of this standalone research are important for the current research and are therefore seen as the problem investigation step. To further investigate the problem, also the planning process is accounted for during the problem investigation (RQ2).

The treatment design covers the requirements, existing treatments and the actual design of an artifact. The requirements for the artifact to be designed in this research are retrieved from literature and semi-structured interviews. These requirements consist of future planning process changes (RQ3), data requirements arising from the planning process (RQ4) and data quality aspects from the field of data management (RQ5). Finally, all requirements are included in the design of the data integration design approach (RQ6).

After designing the treatment, it should be validated (RQ7). Validation is carried out through expert opinions from experts within GVB but also from other PTOs. After this, the generalizability of the artifact is studied (RQ8).

As discussed above, the research questions introduced in Section 1.3.4 are part of the different steps of the design cycle (Figure 1-6). The relation between the research questions and the design cycle steps can be found in Figure 1-7. This figure also contains the different research methods for answering the previously identified research questions. At the bottom of the figure, one can also find the global deliverables of the different steps.

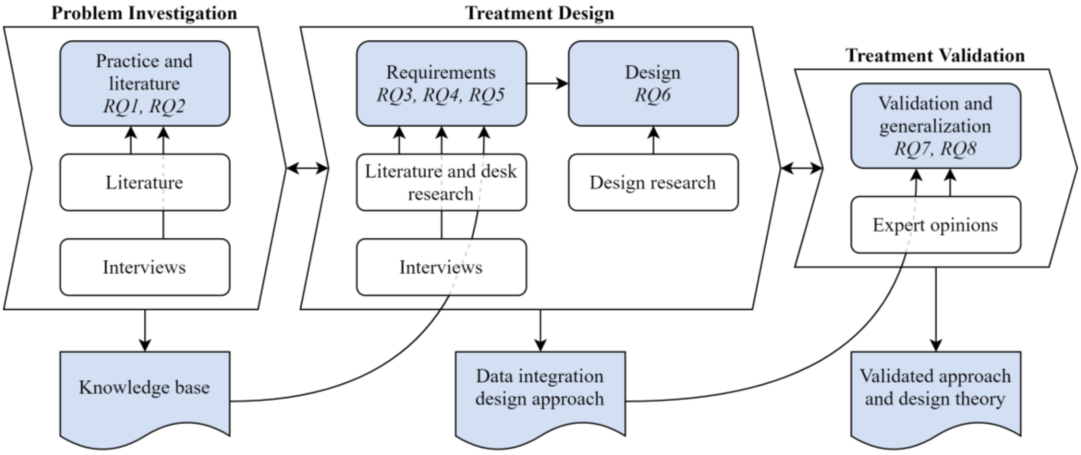
Problem Investigation
Practice and literature *RQ1, RQ2*
Literature
Interviews
Treatment Design
Requirements *RQ3, RQ4, RQ5*
Design *RQ6*
Literature and desk research
Design research
Interviews
Treatment Validation
Validation and generalization *RQ7, RQ8*
Expert opinions
Knowledge base
Data integration design approach
Validated approach and design theory

Figure 1-7 Research process and methods overview

1.3.7 Research Overview

An overview of research questions, methods, chapters in which they are discussed and the expected outcomes are presented in Table 1-1. A detailed explanation of the research methods is presented in Chapter 2.

Table 1-1 Research overview

Research Question		Method	Chapter	Outcome
1	What is the current data integration situation within public transportation and what data standards, data challenges and data integration methods exist?	Literature review; Interviews	3	Base and requirements for the design
2	What is the best-practice planning process of a public transport operator?	Literature review; Interviews	4	
3	What changes in the planning process that are highly dependent on data (integration) are foreseen?	Interviews		(Data) requirements for the design
4	What data requirements (both internally, to adjacent business areas, and externally) does each step in the planning process have?	Literature review; Interviews; Results RQ1	5	
5	What data quality aspects are important to address when defining a data integration design approach for the planning process?	Desk research	6	
6	How can all insights be combined into a data integration design approach in order to ensure cost savings, offer better service quality and be prepared for future data integrations?	Design	7	Data architecture and approach
7	How well does the proposed data integration design approach contribute to the data integration situation of a public transport operator's planning process?	Expert opinions	8	Validated approach and design theory
8	Is the proposed data integration design approach generalizable to other public transport operators and other industries?			

1.4 Thesis Outline and Reading Guide

This thesis is structured as follows:

- Chapter 1 introduces the subject, provides the reader with necessary background information and explains the problem statement and research design.
- Chapter 2 provides a detailed description of the applied research methods.
- Chapter 3 provides the content for the problem investigation (RQ1) by means of interviews and a systematic literature review.
- Chapter 4 explains the planning process from an operations research perspective (systematic literature review), practical perspective (interviews) and proposes the best-practice planning process (RQ2). It also considers foreseen changes in the planning process, including its requirements (RQ3).
- Chapter 5 provides an in-depth analysis of the planning process and a consolidated list of data requirements, containing both consuming as well as providing data requirements (RQ4).
- Chapter 6 is about the data quality aspects as part of data management and explains which aspects are important to consider in the context of this research (RQ5).
- Chapter 7 contains the design of the data integration design approach (RQ6).
- Chapter 8 shows the validation (RQ7) and generalization (RQ8) of the proposed artifact.
- Chapter 9 presents the discussion, conclusion, contributions, limitations and future research recommendations for this research.

For reading this thesis, it is important to realize that everywhere where 'vehicles' is written, also 'vessels' or 'ferries' are meant, except when explicitly stated that this is not the case. This is in line with the public transportation definition provided in Section 1.1.2.

Furthermore, Chapter 3 until Chapter 8 contain a visual research overview based on Figure 1-5. This overview is presented at the beginning of every chapter and aims to show the reader which part of the research will be discussed. Moreover, Chapter 3 until Chapter 6 end with a chapter summary, in which a summary of the findings is presented and their function related to the research is given.

2 RESEARCH METHOD

The goal of this research is to design a data integration design approach. This consists of a target data architecture and an approach that considers important aspects of working towards such a target data architecture. As introduced briefly in Section 1.3.7, several research methods are used to reach the goal of this research. In this chapter, these methods are explained in detail and the relationship to every research question is given, as visualized in Figure 2-1.

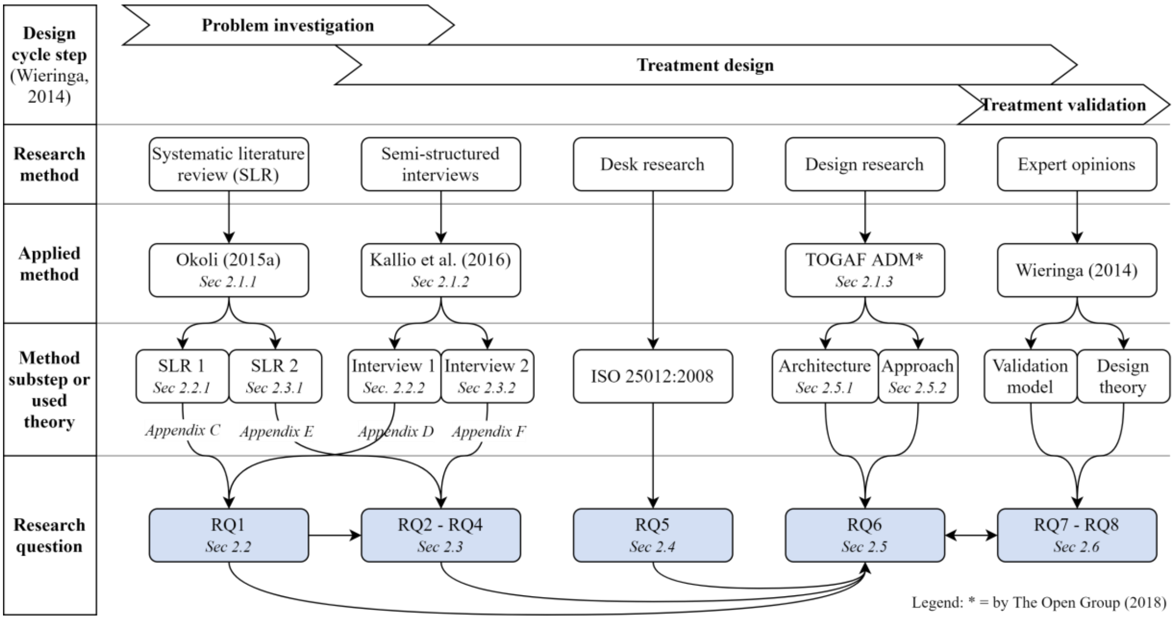

Figure 2-1 Research process and methods in detail

The overall research methodology is presented in Section 2.1 and consists of the method for conducting interviews, conducting a systematic literature review and the overall design method. Subsequently, the method to answer RQ1 is explained in Section 2.2, whereas the method to answer RQ2 – RQ4 is explained in Section 2.3. Section 2.4 presents the method for including data quality aspects in the design (RQ5). The method for the design of the data integration design approach (RQ6) is presented in Section 2.5. Finally, Section 2.6 contains the method for both the validation and generalization of the artifact designed in this research.

2.1 Overall Research Methodology

As explained in Section 1.3.6, the design cycle proposed by Wieringa (2014) is used for this research (Figure 1-6). Within this cycle, the problem investigation, treatment design and treatment validation steps are carried out. For the problem investigation and treatment design, two systematic literature reviews and multiple interviews were conducted. The applied methods are presented in Section 2.1.1 and Section 2.1.2, respectively. The details of the methods to answer the different research questions (as can be seen in Figure 2-1) are presented in Section 2.2 (RQ1) and Section 2.3 (RQ2 – RQ4), yet these are both based on the applied method explained in this section. The selection and explanation of the method used for the treatment design is elaborated in Section 2.1.3.

2.1.1 Literature Review

A systematic literature review (SLR) is important to gain a comprehensive overview of the available literature in a field of study. Doing this systematically enhances the clarity, validity and auditability of the review (Booth et al., 2016). Clarity is enhanced by executing the review in several clear and predefined steps. This makes it easier to read, understand and interpret the literature review. By defining clear steps and setting up criteria for inclusion, exclusion and resources, it becomes easier to determine the validity of the literature review. Lastly, the auditability of the review is enhanced by having the aforementioned properties of a literature review in place.

In this research, the 'Standalone Systematic Literature Review Guide' as proposed by Okoli (2015a) is used. For developing this guide, the author used a total of 23 journal papers in the field of Information Systems (IS). The author states that "no uniformity in methodology or structure among these studies exist, even between those published in the same journal" (Okoli, 2015a). For this reason, Okoli (2015a) consolidated the methods of all 23 journal papers into one guide of conducting an SLR, illustrated with examples based on several literature review methodologies in the field of IS. Hence, the proposed guide is widely accepted, can be used within the field of IS and is not a standalone methodology but rather a comprehensive guide taking into account best practices of SLRs conducted in the field of IS.

The model of the standalone SLR guide as proposed by Okoli (2015a) is visualized in Figure 2-2. The identified steps are used throughout the literature review. Okoli (2015a) supposes that the literature review is conducted as a standalone systematic literature review by a group of reviewers. However, the literature review of this research is carried out by one reviewer. To comply with this situation, some of the proposed steps of the SLR guide are slightly adapted. This is explained for the applicable steps.

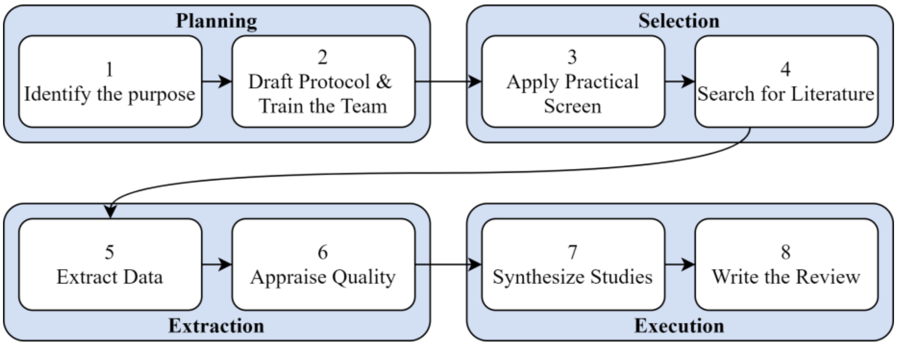

Figure 2-2 Steps of a standalone SLR (adapted from Okoli (2015a))

This research contains two separate systematic literature review processes. The first review answers RQ1 and is about the data integration in general, whereas the second review answers RQ2 – RQ4 and is about the planning process, its tasks and data requirements. Both review processes are set up according to Okoli's (2015a) guide and are explained in more detail in Section 2.2.1 and Section 2.3.1. A high-level explanation of the four phases is presented next.

2.1.1.1 Planning

The planning phase of the standalone SLR guide by Okoli (2015a) consists of identifying the purpose of the SLR and defining the protocol and training the teams. Since the literature review in this report is conducted by one researcher only, training the teams is left out of scope. The protocol contains information such as the sources, criteria for in- and exclusion and search terms.

2.1.1.2 Selection

Applying the search protocol as defined in the previous phase results in a set of literature results. A selection of this set is used for retrieving the intended information. This selection is often carried out in several steps in which the results are filtered based on their title, abstract, introduction, conclusion and, finally, after reading the entire text. In this selection phase, also forward and backward referencing is applied, as suggested by Webster & Watson (2002).

2.1.1.3 Extraction

The previous phase ends with a list of literature sources. These sources need to be analyzed in more detail in order to extract data for the literature review. Okoli (2015a) suggests extracting data by using a data extraction form and proposes the methodology of Bandara et al. (2015) by setting up a form that contains questions about the origin of the literature, its content, validation and limitations.

2.1.1.4 Execution

For the qualitative synthesis and execution of the SLR, the concept-centric approach of Webster & Watson (2002) can be used. This approach maps concepts found within the literature to the different sources and authors. The other way around (describing the author's view and map to the concepts) is not recommended, since that method fails to synthesize the literature (Webster & Watson, 2002).

2.1.2 Interviews

Interviews hold a lot of information for the researcher and the research and several methods are present. Therefore, it is important to choose the correct method of interviews. Possible interview methods include structured interviews, unstructured interviews and semi-structured interviews. The best interview method should be chosen based on how well the data generated by each interview method matches the research goal (Doody & Noonan, 2013).

When conducting a structured interview, the researcher is completely in control of the questions. For the respondents, there is no possibility to elaborate their answer as long as it is not a predefined question. This limits the diversity of data, which can be a disadvantage in terms of gathering information, but also an advantage in terms of time (Doody & Noonan, 2013). Structured interviews is a quantitative research method since every respondent answers the same closed-ended questions. This makes it is easier for the researcher to analyze and compare the outcomes of the interviews.

On the contrary, unstructured interviews generate a lot of rich data. This is because the researcher only calls a subject, on which the respondent is asked to elaborate on and share their thoughts and interests. This is a qualitative research method. The data outcome is very high, however, this data is – due to its diversity – hard to analyze and compare (Doody & Noonan, 2013).

Lastly, semi-structured interviews are a combination of the two previously explained methods and allow the researcher to ask for clarification while respondents answer predefined questions. This enriches the data and makes the interview more flexible (Doody & Noonan, 2013). It is in line with Adams (2015) and Rowley (2012), who both state that open-ended questions sometimes require follow-up questions to sufficiently answer the main question. Similar data types in the answers are created by first setting up an interview guide (Doody & Noonan, 2013). This enables the comparison of the outcomes.

Semi-structured interviews are chosen as the best option for this research, since it allows for gathering data both quantitatively and qualitatively. This is useful because there are multiple goals for using interviews: exploring a field of study, retrieving knowledge about processes (both qualitative) and analyze the differences between interview respondents (quantitative). Moreover, semi-structured interviews allow the researcher to ask in-depth questions in case further exploration (of unknown topics prior to the interviews) is necessary.

For designing the semi-structured interviews, the practical tool for developing an interview guide for semi-structured interviews developed by Kallio et al. (2016) is used. It is based on ten theoretical and methodological papers about developing semi-structured interviews which were published between 1994 and 2015. Kallio et al. (2016) propose five phases, which can be found in Figure 2-3.

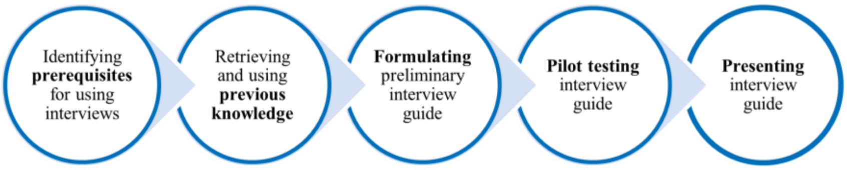
Identifying prerequisites for using interviews
Retrieving and using previous knowledge
Formulating preliminary interview guide
Pilot testing interview guide
Presenting interview guide

Figure 2-3 Phases in developing a semi-structured interview guide (Kallio et al., 2016)

This research contains two different semi-structured interviews. The first interview answers RQ1 and is about the general data integration situation at GVB, whereas the second review answers RQ2 – RQ4 and is about the planning process, its tasks and data requirements. Both interviews are set up according to Kallio et al.'s (2016) guide and the five phases are explained in more detail per interview in Section 2.2.2 and Section 2.3.2. A high-level explanation of the five phases is presented next.

2.1.2.1 Identifying the Prerequisites for using Semi-structured Interviews

During this phase, it is important to identify the goals of the interview sessions and justify whether the semi-structured interview method is the best way to reach these goals.

2.1.2.2 Retrieving and using Previous Knowledge

The researcher needs to get a comprehensive and adequate understanding of the subject. This can be reached by conducting a literature review, consulting experts or by using empirical knowledge.

2.1.2.3 Formulating the Preliminary Semi-structured Interview Guide

In this phase, the first version of the interview guide is made. Questions have to be set up on two different levels: main themes and follow-up questions (Adams, 2015; Kallio et al., 2016; Rowley, 2012). The main theme questions are asked to every respondent, whereas the follow-up questions are there to direct the respondent in the direction of the research. Follow-up questions can be predefined and spontaneous. Predefined questions increase the consistency and spontaneous questions most often lead to more in-depth answers or examples. Before the main questions are asked, it is recommended to first include some easy introduction questions, to make the interviewee feel more comfortable (Adams, 2015; Doody & Noonan, 2013).

2.1.2.4 Pilot Testing of the Interview Guide

This fourth phase aims to confirm the coverage and relevance of the preliminary interview guide. By testing the guide, it is possible to make changes in the questions and improve the quality of data the interview will lead to. Testing is possible by conducting internal testing (with fellow researchers), expert assessments (validation by experts in the field of research) or field-testing (validation with potential respondents; simulation of a real interview) (Kallio et al., 2016).

2.1.2.5 Presenting the complete Semi-structured Interview Guide

The last phase is the presentation of the interview guide in the research and the execution of the interviews. Presenting the interview guide is important for the clarity and validity of the interview and its results (Kallio et al., 2016).

2.1.3 Architecture Framework and Method

For the treatment design step in the design cycle, an enterprise architecture (EA) framework and methodology were selected for the design of the data integration design approach. Several EA frameworks were developed over the past years. Cameron & McMillan (2013) compared five major EA frameworks: Zachman, the TOGAF Standard, FEA, DoDAF and Gartner. The selection of these five EA frameworks was based on a survey with 271 participants "whose job roles and responsibilities directly reflected working in EA within their organizations" (Cameron & McMillan, 2013). The comparative analysis of the five frameworks led to the conclusion that the TOGAF Standard addresses most of the criteria with the highest score[1]. The TOGAF Standard is, among other reasons, mostly used because of its Architectural Development Methodology (ADM).

Since the comparative analysis shows that the TOGAF Standard is the favorable EA framework and the author of this research is already familiar with this framework, it is chosen to be used for this research. In the next subsections, the TOGAF Standard, TOGAF ADM and the EA modeling language are explained.

[1] Scores are derived from Cameron & McMillan (2013, p. 69): TOGAF: 40, FEA: 33, DoDAF: 25, Gartner: 23, Zachman: 22.

2.1.3.1 TOGAF Standard

The TOGAF Standard is a framework designed for EA. The first version was published in 1995 and since then maintained and improved by the members of The Open Group Architecture Forum (The Open Group, 2018). The standard consists of six parts, which are shown in Table 2-1.

Table 2-1 The TOGAF Standard parts (The Open Group, 2018)

Part	Name
I	Introduction
II	Architecture Development Method (ADM)
III	ADM Guidelines & Techniques
IV	Architecture Content Framework
V	Enterprise Continuum & Tools
VI	Architecture Capability Framework

As the TOGAF Standard is expected to be customized by its users and organizations (e.g. by choosing core elements, customize some, exclude some) (The Open Group, 2018), it is especially useful for this research. As the goal of this research is to design a data integration design approach – consisting of an architecture and approach – the TOGAF ADM (Part II) is selected to be focused on.

2.1.3.2 TOGAF ADM

The TOGAF Architecture Development Method (ADM) "describes a method for developing and managing the lifecycle of an EA, and forms the core of the TOGAF Standard" (The Open Group, 2018). The basic structure of the ADM is a cycle that consists of multiple phases (see the orange circles in Figure 2-4). An explanation of every phase is provided in Appendix B.

Similar to the TOGAF Standard, the ADM cycle is also subject to be tailored to the users' and organizations' needs (The Open Group, 2018). For the design in this research, we focus on one ADM cycle phase C as shown in Figure 2-4: Information Systems Architectures. It is important to design the data integration design approach based on the business' needs, otherwise it would be of no reason to design the architecture and approach. For this reason, also Phase B (Business architecture) is assessed. This is done by defining the business' needs for a data integration design approach and answered by RQ1 – RQ4.

2.1.3.3 ArchiMate

In order to design business and technology artifacts and their relationships properly, the ArchiMate language is used. ArchiMate is a modeling language for Enterprise Architecture and is described as: "a visual language with a set of default iconography for describing, analyzing, and communicating many concerns of Enterprise Architectures as they change over time" (The Open Group, n.d.). ArchiMate consists of five modeling layers: business, application, technology, strategy & motivation and implementation & migration. Throughout this research, several figures are presented which are using the ArchiMate language. The language is capable of showing the relationship between business needs and application architecture clearly. The relation between the ArchiMate layers and the TOGAF ADM phases can be found in Figure 2-4.

2.2 Data Integration in Public Transport (RQ1)

This section contains the methods for answering RQ1. This research question is about the data integration situation within public transportation and to which requirements this leads for the design of the data integration design approach. It is answered through a systematic literature review and semi-structured interviews. The methods are explained in Section 2.2.1 and Section 2.2.2, respectively.

Since the research presented in this section was carried out earlier by the same author (as unpublished research), it originally had some different research questions and scoping boundaries. For this reason, only a part of the original SLR and interview results is used for this research. To ensure the validity of the method and research, the entire method is explained. The only difference is that not all results are used in this research. The selection criteria for this process are presented as well.

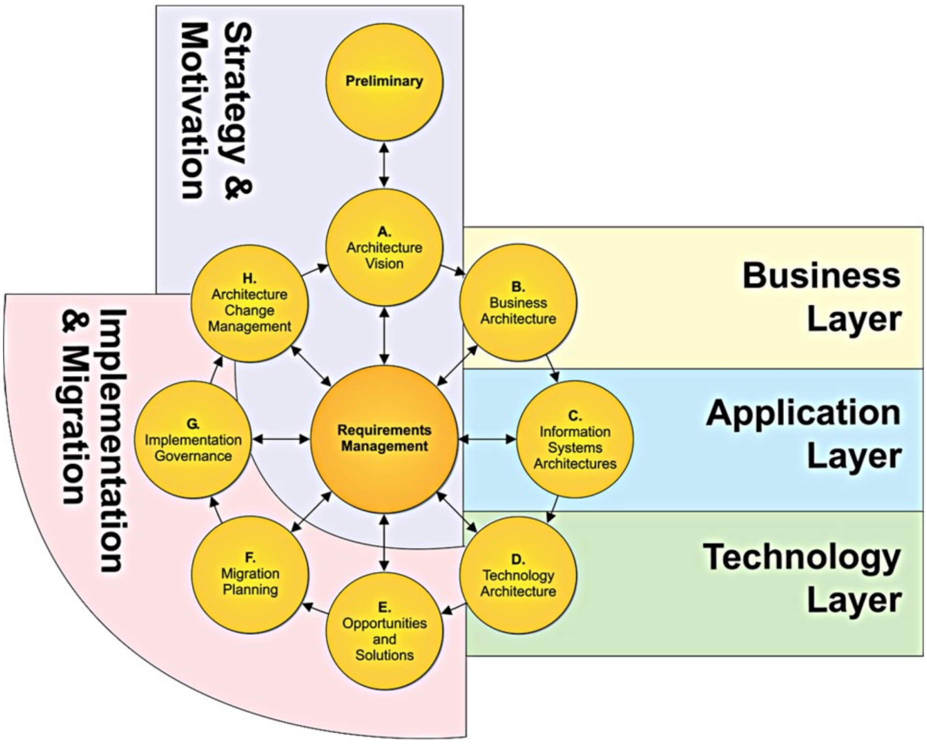
Strategy & Motivation
Preliminary
A. Architecture Vision
H. Architecture Change Management
B. Business Architecture
Business Layer
Implementation & Migration
G. Implementation Governance
Requirements Management
C. Information Systems Architectures
Application Layer
F. Migration Planning
E. Opportunities and Solutions
D. Technology Architecture
Technology Layer

Figure 2-4 TOGAF ADM and ArchiMate (The Open Group, n.d.)

2.2.1 Literature

Due to the larger scope of the original SLR, only some SLR topics and findings are used within the current research. The entire original SLR process and the resulting set of literature sources can be found in Appendix C. The final set consisted of 27 literature sources, of which 23 are used for the current research (indicated in Appendix C). Data was extracted from the original set of sources through mapping the content to the different concepts which were defined in the scope of this literature review, as proposed by Webster & Watson (2002). The concepts can be found in Table 2-2, in which the concepts applicable to the current research are colored blue. This selection was based on the scope of this research and should therefore be about data integration and/or public transportation. Data integration levels were considered to include but turned out to be too detailed for the current research. The other concepts were not focused (enough) on data and are therefore not reused in the current research.

Results from the literature were connected to one of the concepts in Table 2-2, which gave a total overview of information, statements, conclusions and views on the different concepts. For answering RQ1, the concepts *data integration definition*, *data integration challenges*, *data integration methods* and *data integration in public transportation* are used within this research. Subsequently, the information about these concepts was combined into a concept-centric literature review, which is presented in Chapter 3.

Table 2-2 Subjects for data extraction

Data integration definition	Interoperability layers	Data integration methods
Interoperability definition	Integration levels	Data integration in public transportation
Enterprise Application Integration	Data integration levels	
Enterprise Information Integration	Data integration challenges	

2.2.2 Interviews

The interviews presented in this section should provide answers to RQ1. As explained in Section 2.1.2, the interviews are designed with the help of the practical tool proposed by Kallio et al. (2016). The design process and final interview guide can be found in Appendix D. In the next subsection, the interview participants, data collection and data analysis are explained.

2.2.2.1 Interview Participants

Since the main goal of the interviews is to get an overview of the current data integration situation within GVB, it means that stakeholders from different business departments should be interviewed to ensure a holistic view of the current situation. On IT-side, all GVB business departments are supported by a so-called product or infra team (Table 2-3). Each of these teams has one or more application consultants. At GVB, application consultants are – among other things – responsible for the interfaces and data integrations between applications. Therefore, these application consultants are asked to take part in the semi-structured interviews.

Table 2-3 GVB's IT teams

Team	Type	Original Dutch name
Business Intelligence (BI)	Product team	Business Intelligence (BI)
Passenger information (PI)	Product team	Reizigersinformatie (RI)
Information provision (IP)	Product team	Informatievoorziening (IV)
Payment	Product team	Betalen
Operations	Product team	Exploitatie
Planning	Product team	Planning
Assets	Product team	Assets
Compute	Infra team	Compute
Connectivity	Infra team	Connectivity
Digital Workplace (DW)	Infra team	Digital Workplace (DW)
Generic IT Vehicle Architecture (GIVA)	Infra team	Generieke ICT Voertuigarchitectuur (GIVA)

2.2.2.2 Data Collection

The first part of the interview contains introductory questions about the respondent, their function and working experience. To get a clear understanding of the different responsibilities of each team and the domain in which they are working, some introductory questions about their team are asked as well.

After the introductory questions, the main questions are included in the interview guide. This is according to the structures proposed by Adams (2015) and Doody & Noonan (2013). The main questions are divided into two categories: applications and data. Several questions are stated around these topics. These are further specified in so-called follow-up questions, which direct the interview towards the research questions and maintains the flow of the interview (Kallio et al., 2016).

The interview guide was used for eight out of twelve interviews. The other four interviews were held after the initial interviews in order to get a better understanding of some subjects. Since the questions for these respondents were set up after collecting information from the initial set of interviews, the questions were too specific and tailored to the business areas the four respondents were working for and/or closely connected to. Due to the anonymization of the interview results, these four unique interview guides cannot be revealed.

2.2.2.3 Data Analysis

The data collected from the interviews could be analyzed relatively easily since the interview guide ensured that the answers could be compared. This is for example the case for the applications within the IT landscape and the data dependency between the different business functions.

The interviews were recorded and the statements made by the respondents were summarized per question of the interview guide. These interview summaries were sent back to the respondents to verify whether the researcher interpreted the respondents correctly. Some minor changes and additions took place. After writing the summaries, some extra questions came up. These questions were sent to the respondents as well and the answers were included in the summaries afterward. In one case, the number of questions that arose was relatively high. Hence, a second interview with this respondent took place.

To answer the research questions about the demand and exchange of data objects, all interview results were mapped to GVB's business functions. The reason to explain the findings on the level of business functions is that they are more stable than business processes and the IT components serving the processes (the latter two are subject to change and improvements). Furthermore, this method shows the data integration situation of the planning process related to other business functions. For data ownership and responsibilities, the CIA Triad aspects were used: Confidentiality, Integrity and Availability (Samonas & Coss, 2014).

2.3 Planning Process and its Data Requirements (RQ2 – RQ4)

This section contains the methods for answering RQ2, RQ3 and RQ4. An overview of the methods per research question and the sections in which they are explained, can be found in Table 2-4 and are explained thereafter.

Table 2-4 Overview RQ2, RQ3 and RQ4

	Method	Section
RQ2	Literature	2.3.1
	Interviews	2.3.2
RQ3		
RQ4		
	Literature	2.3.1
	RQ1	2.2

RQ2 seeks to understand the standard planning process of a PTO as defined in the field of operations research. This research question is answered through a systematic literature review, which is explained in Section 2.3.1. To ensure the applicability of the planning process in practice and to answer RQ3 about future planning process improvements and RQ4 about the data requirements, interview sessions were held as well. Information about the interview method is presented in Section 2.3.2.

Next to the interviews, RQ4 is also answered by combining the interview results with literature and the answers on RQ1. This means that for answering RQ4, several sources need to be combined. The method for triangulating them is explained in Section 2.3.3.

To answer RQ4, planning tasks were derived from multiple sources and mapped to the planning phases (RQ2). Based on these planning tasks, data requirements were identified. Both upstream and downstream data requirements (the difference is visualized in Figure 2-5) from the planning process itself and adjacent business areas and external consumers are considered. The process from the qualitative research to planning tasks to data requirements is visualized in Figure 2-6.

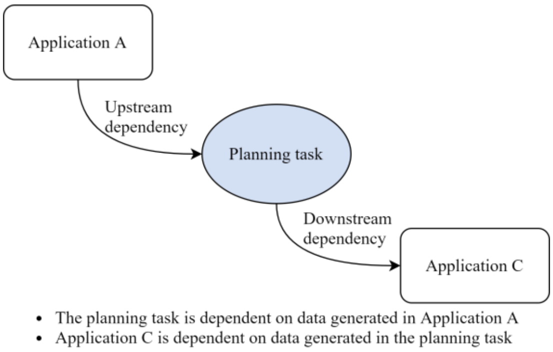
Application A
Upstream dependency
Planning task
Downstream dependency
Application C
• The planning task is dependent on data generated in Application A
• Application C is dependent on data generated in the planning task

Figure 2-5 Upstream and downstream dependencies

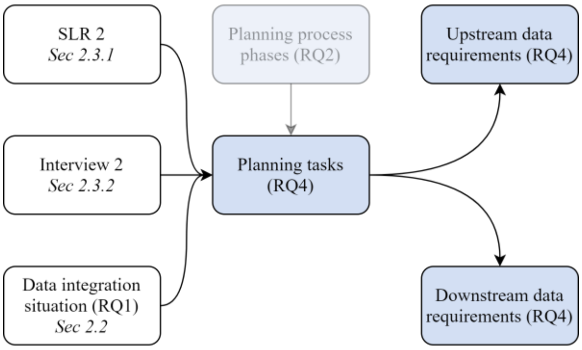
SLR 2
Sec 2.3.1
Planning process phases (RQ2)
Upstream data requirements (RQ4)
Interview 2
Sec 2.3.2
Planning tasks (RQ4)
Data integration situation (RQ1)
Sec 2.2
Downstream data requirements (RQ4)

Figure 2-6 From planning tasks to data requirements (RQ4)

2.3.1 Literature

The final set of sources to identify the planning process and its data requirements consisted of 18 literature sources. The entire SLR process and the resulting set of literature sources can be found in Appendix E. As proposed by Webster & Watson (2002), the sources were mapped to different concepts. This SLR focuses on three different concepts for answering RQ2 and RQ4:

- Planning process phases *for answering RQ2*
- Planning decisions/planning tasks *for answering RQ4*
- Data requirements *for answering RQ4*

Appendix E also contains the mapping of the literature to the above-mentioned three concepts. A visualization of this mapping is shown in Figure 2-7.

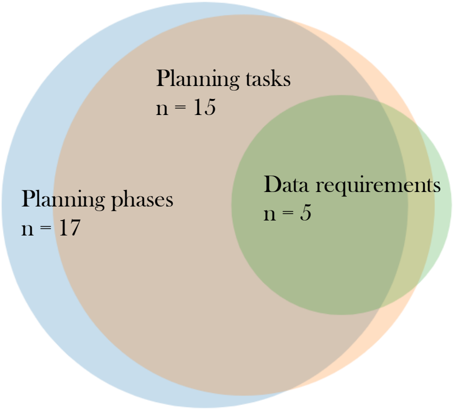

Figure 2-7 Subjects in literature sources SLR 2

The size of the circles represents the number of literature sources mapped to the particular concept. As can be seen in the figure, most sources are about the different planning phases within the planning process and the planning tasks associated with them. Only a few literature sources are about the actual data requirements for the planning process.

As indicated before, this SLR is used for answering two different research questions (RQ2 and RQ4). This leads to two different processes for answering them. RQ2 is answered through the identification of the different planning phases within the planning process. RQ4 is answered by using these phases and combining them with the planning tasks and data requirements as identified in the literature. Both processes are explained hereafter.

2.3.1.1 Planning Process (RQ2)

Concepts for the planning process phases are identified during the data extraction and execution of the SLR, in line with Webster & Watson's (2002) approach. The list of concepts was formed iteratively during the literature review, since new concepts came up and similar concepts were combined and renamed. The identified concepts led to the identification of the planning phases, of which the results are presented in Chapter 4.

2.3.1.2 Planning Tasks and Data Requirements (RQ4)

For identifying the planning tasks, first, a selection of literature sources was made based on the results of the concept grouping. Since the goal of RQ4 is to identify data requirements for the planning process, literature that includes both 'planning tasks' and 'data requirements' is selected to answer this question. This resulted in a set of five sources. Two of these sources are textbooks and the other three are scientific journal papers. One of these papers is from the same author as the book and is used as a reference in that book. For this reason, this paper is left out. This resulted in a set of four sources, which are listed in Table 2-5.

Table 2-5 Literature sources for planning tasks

Literature
Ceder (2016)
Friedrich et al. (2016)
Nagy & Tick (2019)
Scholz (2016)

Identifying the planning tasks within the abovementioned set of sources is done by carefully reading through the literature, coding the use cases and decisions to be made and copying them into a total list of planning tasks. This outcome is further explained and described in Chapter 5.

2.3.2 Interviews

The method addressed in this section was carried out to provide answers to RQ2, RQ3 and RQ4. The interviews are designed with the help of the practical tool proposed by Kallio et al. (2016), as explained and introduced in Section 2.1.2. The design process and final interview guide can be found in Appendix F. The interview participants, data collection and data analysis are explained next.

2.3.2.1 Interview Participants

The planning personnel within GVB is focused on a small part of the planning process (e.g. line planning, timetabling, rostering, scheduling vehicles, scheduling crew duties and scheduling personnel) and many of these employees are specialized in only one transport modality (bus, tram, metro, ferry). Interviewing these employees would result in a high number of interviews with a level of detail that is too specific for the goals of this research. For these reasons, it was chosen not to interview the specific business employees within the planning process. Instead, the business analyst focused on the planning process and product owner of Team Planning (Table 2-3) were chosen as participants for these interviews. They are assumed to be very familiar with the planning process and to be able to provide sufficient information for this research. Table 2-6 lists the respondents.

Table 2-6 Respondent's functions

Respondent's Function
Business Analyst
Product Owner Team Planning

2.3.2.2 Data Collection

The first part of the interviews contained questions about the respondent and their professional career, to get an idea of the context in which the respondent answers the questions. The second and third part of the interviews were about the current (RQ2) and future planning process (RQ3). In these two parts, the respondents were asked about the planning phases, decisions to be made within these phases (i.e. planning tasks), the data necessary for making these decisions (RQ4) and, lastly, future changes in the planning process. The last part of the interviews was about general questions to ensure the completeness of the interviews.

The interviews were conducted through Microsoft Teams calls, due to the Covid-19 pandemic. The interviews were recorded with the permission of the respondents.

2.3.2.3 Data Analysis

The interviews are used to answer three different research questions, which leads to three different data analysis processes. The first questions are about the different planning phases within the planning process. After these questions, more details are asked about the (future) improvements, decisions to be made, planning tasks within the planning process and data requirements for these decisions and tasks (see interview guide in Appendix F). These latter questions give answers to RQ3 and RQ4, whereas the first questions answer RQ2.

The interviews are not transcribed but interview summaries were written. Transcribing was not expected to be necessary, since the purpose of the interviews was to explore the scope of the planning process and its data requirements in practice. Moreover, the findings serve as practical validation and completeness-check of the literature study. To ensure the validity of the findings and to check the interpretation of the author, the interview summaries and findings were sent to the respondents for validation. In case it was necessary, they were adapted based on the respondents' input and sent back for approval.

2.3.3 Triangulation

For combining the results from literature, interview and RQ1, the triangulation method is used. Using multiple research methods is less vulnerable than only using one. This is because it results in a study that is less vulnerable to errors related to one particular research method, even if not all results point in the same direction (Patton, 1999). Such inconsistencies lead to a deeper insight into the relationship between the sources and the studied planning process (Patton, 1999). For this reason, triangulation of data sources also enhances the quality and credibility of the qualitative analysis (Patton, 1999).

According to Patton (1999), four different kinds of triangulation contribute to verification and validation of qualitative analysis: methods triangulation, triangulation of sources, analyst triangulation and theory/perspective triangulation. Each of them is described in Table 2-7.

Table 2-7 Four kinds of triangulation (Patton, 1999)

Triangulation	**Description (Patton, 1999)**
Methods triangulation	Checking out the consistency of findings generated by different data collection methods
Triangulation of sources	Examining the consistency of different data sources within the same method
Analyst triangulation	Using multiple analysts to review findings
Theory/perspective triangulation	Using multiple perspective or theories to interpret the data

Within this research, the method and source triangulation is applied for consolidating findings from literature (Section 2.3.1), interviews (Section 2.3.2) and the results from RQ1 (Section 2.2), as visualized in Figure 2-8. In essence, "triangulation (...) is a form of comparative analysis" (Patton, 1999). This comparison remains subject to the interpretation of the author. For this reason, the data requirements leading from the triangulation are also validated externally, as explained in Section 2.6.

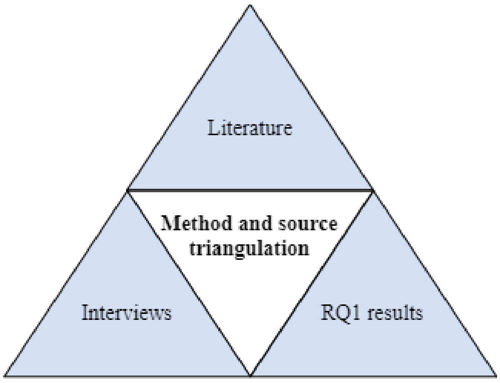

Figure 2-8 Method triangulation process

Triangulating the three qualitative data sources led to the identification of the planning tasks. Details of this triangulation process are key in the research and are based on the concept-centric approach of Webster & Watson (2002) and presented in Chapter 5.

After the planning tasks were identified, the method for determining the total set of data requirements was based on Dingemans' (2019) recommendation for a method for modeling input and output data objects for business processes, called the SIPOC method. SIPOC stands for Supplier, Input, Process, Output and Consumer. This method leads to clear process mapping and shows the necessary input and outputs for each process (Carey & DeLayne Stroud, n.d.).

The planning tasks and data requirements for the planning process, which are the result of the triangulation process, are validated by a functional application manager of GVB. This officer has almost 20 years of experience within the planning domain and is part of Team Planning (Table 2-3). Moreover, this employee is also a member of the project team responsible for implementing a new software package for the entire planning process within GVB, which means that knowledge about planning tasks, data requirements and integrations is assumed to be present and valid.

2.4 Data Quality Aspects (RQ5)

Data quality is one of the knowledge areas of the DAMA International's (2014) Data Management Framework (Figure 1-2) and turned out to be a difficult data aspect during the problem investigation of this research. For this reason, data quality aspects are taken into account for the design of the data integration design approach. The method explained in this section aims to answer RQ5.

The data quality aspects from the ISO standard ISO/IEC 25012:2008 are used. This is a standard for software engineering specialized on the software product quality requirements and evaluation and contains a data quality model (ISO, 2008). The standard is widely accepted in the field of software engineering and its last periodical review dates from June 2019, after which it was confirmed (ISO, n.d.). Therefore, we assume that it is a valid standard for this research. The model consists of 15 different data quality aspects, which are shown in Figure 2-9. The standard divides data quality into inherent data quality, system-dependent data quality aspects (left and right side in Figure 2-9, respectively) and aspects in between these categories.

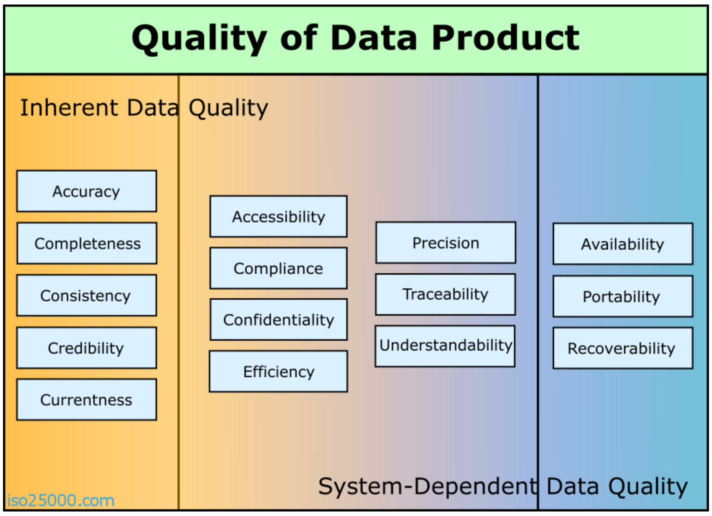

Figure 2-9 Data quality model adapted from ISO/IEC 25012:2008 (iso25000.com, n.d.)

RQ5 aims to indicate which data quality aspects are important to address in the data integration design approach. Firstly, the 15 quality aspects are mapped to previously identified data integration challenges from practice and literature (RQ1) which provides a view of which data quality aspect can help to address which data integration challenge to ensure the applicability.

Secondly, to identify which aspects should be accounted for in the data integration design approach, they were divided into four groups based on two planning scenarios (static and dynamic planning, explained in detail in Section 6.3):

- Less important data quality aspects for this research (assuming dynamic planning is the goal);
- Data quality aspects included in the data integration design approach;
- Data quality aspects enabled by the data integration design approach;
- Inherent data quality aspects.

The categorization of the data quality aspects is explained and their operationalization (i.e. how to apply these in the design) is presented.

2.5 Data Integration Design Approach (RQ6)

As introduced in Section 1.3.1, the data integration design approach as designed in this research consists of two deliverables: a target data architecture and an approach with considerations on how to use this data architecture in practice. The method for the design of the target data architecture and the approach are presented in Section 2.5.1 and Section 2.5.2, respectively. Both are based on the TOGAF ADM, which was introduced in Section 2.1.3.

2.5.1 Data Architecture

The main task during the treatment design step of this research is to consolidate all previously identified results into a target data architecture. TOGAF ADM's phase C is focused on the development of a data architecture and its steps are visualized in Figure 2-10. The performed steps within this research are highlighted with blue circles.

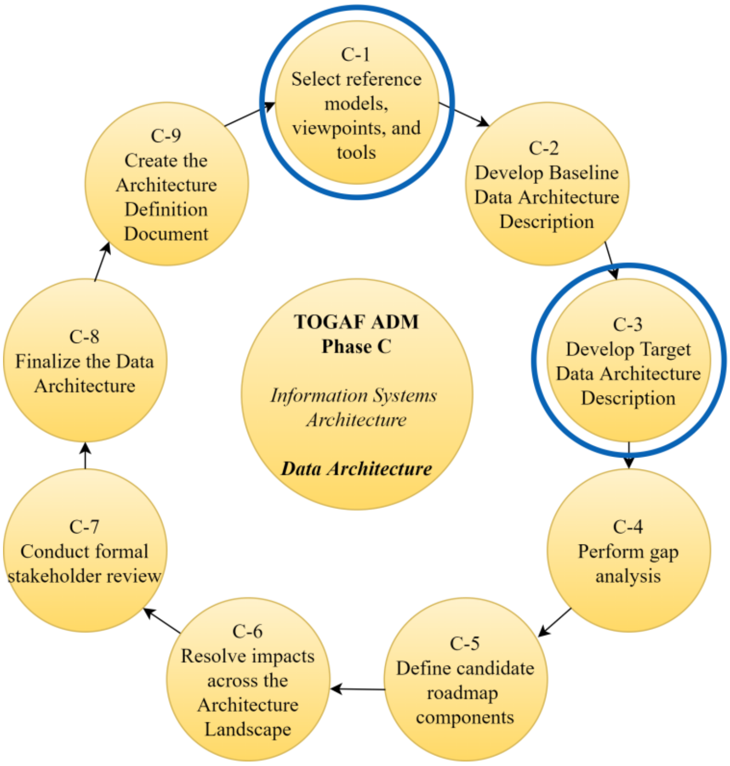
C-1
Select reference models, viewpoints, and tools
C-2
Develop Baseline Data Architecture Description
C-3
Develop Target Data Architecture Description
C-4
Perform gap analysis
C-5
Define candidate roadmap components
C-6
Resolve impacts across the Architecture Landscape
C-7
Conduct formal stakeholder review
C-8
Finalize the Data Architecture
C-9
Create the Architecture Definition Document
TOGAF ADM
Phase C
Information Systems Architecture
Data Architecture

Figure 2-10 TOGAF ADM phase C (adapted from The Open Group (2018)) indicating design scope

The two highlighted steps include selecting reference models, viewpoints and tools and the actual development of the target architecture. According to the developing method (step 3), some assets (defined in step 1) are required. As indicated before (Section 2.1.3), TOGAF is a framework (and ADM a method) that should be adapted to the situation it is applied to. For this reason, some of the required architecture assets are out of scope. An overview of the assets and whether or not they are included in this research is listed in Table 2-8.

Table 2-8 Required catalogs, matrices and diagrams (The Open Group, 2018)

Required asset	Provided by this research
Data entity catalog	Yes: Table 5-3
Data entity/business function matrix	Yes: Table 5-3
Business service/information matrix	Yes: Appendix L
Application/data matrix	No: PTO-specific (see Section 2.5.2)
Conceptual data diagram	Yes: Appendix N
Logical data diagram	No: not in scope
Data dissemination diagram	Yes: in matrix form (Table 5-3)
Data lifecycle diagram	Provided implicitly in Table 5-3
Data security diagram	No: not in scope
Data migration diagram	No: PTO-specific (see Section 2.5.2)

As can be seen from the requirements, a strong relationship to the business architecture is necessary. This research has sought to provide this by answering RQ1 – RQ4. Developing the architecture is done based on the planning process and data requirements as defined in this research. The data quality aspects as defined in RQ5 are included during the design of the architecture.

The data requirements had to be scoped into a set of data that is owned by the planning process, since only that data can be managed by the PTO's planning process. Ownership of data entities was given by using the data entity/business function matrix as proposed by The Open Group (2018).

The architecture is designed application-independently, in order not to focus the design on a specific PTO. To better explain the context in which the target data architecture can be used, the architecture's context is also designed through a data access service within a fictive application landscape. This tries to decrease the gap which is made since no application-specific relationships are defined during the requirements phase.

2.5.2 Approach

TOGAF's approach for data architectures consists of two different aspects: key considerations for data architecture and the architecture repository. These two aspects are important for the approach of developing, analyzing, implementing and using a data architecture (The Open Group, 2018). The key considerations, their relevance and application within this research are discussed hereafter. They are grouped (as proposed by TOGAF) in data management, data migration and data governance. The proposed data integration design approach is supposed to be included in the architecture repository, which is therefore not further explained.

2.5.2.1 Data Management

TOGAF's data management considerations focus on the identification of the system of record, compliance to standards, utilization of data entities, data authorization, data quality and data roles (The Open Group, 2018). As the main goal of this research is to contribute to the improvement of the data management situation, this part is the focus of the designed approach. Hence, the method for defining the approach is based on this part of TOGAF's considerations.

To address the considerations mentioned before, the approach is divided into three categories: adoption of the target data architecture, data authorizations and data quality and roles. The inclusion of these considerations and the used methods are as follows:

- For the adoption of the target data architecture, an implementation example from a fictive PTO is given. This is also one of the proposed TOGAF ADM (The Open Group, 2018) steps when implementing a new data architecture.
- The data authorizations are derived from the in-depth data requirements analysis explained in Chapter 5. Based on the identification of the data requirements (RQ4), a clear overview of data authorizations is proposed using a CRUD-matrix (create, read, update, delete), as recommended by Dingemans (2019).
- The quality aspects based on the ISO 25012:2008 standard (ISO, 2008) and defined in Chapter 6 are put in the context of several data roles. These data roles are based on the International Data Spaces Association (IDSA, 2019b), which is an international association that aims to standardize data exchange between different parties included in their network. IDSA is chosen as it is a well-known association within the field of data exchange, consisting of research institutes (e.g. Fraunhofer) and over 125 members worldwide (IDSA, n.d., 2019a).

2.5.2.2 Data Migration

Data migration is out of scope for this research, since it highly depends on the PTO's implementation and applications landscape. It is, however, an important aspect when using the data integration design approach in practice.

2.5.2.3 Data Governance

Having the necessary dimensions in place to enable transformation on data level is categorized as data governance (The Open Group, 2018). These dimensions are divided into three categories, as listed and defined in Table 2-9.

Table 2-9 Data governance dimensions (The Open Group, 2018)

Dimension	Definition (The Open Group, 2018)
Structure	Pertains to whether the enterprise has the necessary organizational structure and the standards bodies to manage data entity aspects of the transformation.
Management System	An enterprise should have the necessary management system and data-related programs to manage the governance aspects of data entities throughout its lifecycle.
People	Addresses what data-related skills and roles the enterprise requires for the transformation.

For this research, we assume that the dimensions *structure* and *management systems* are in place. The *people* dimension is included in the approach, as was already discussed before.

2.6 Validation and Generalization (RQ7 – RQ8)

The methods for validation and generalization of the outcomes of this research are presented in this section. In Section 2.6.1, the design of the validation is addressed. This includes the respondent selection. Lastly, Section 2.6.2 shows the method for the generalization.

2.6.1 Validation

The goal of validation is to "develop a design theory of an artifact in context that allows us to predict what would happen if the artifact were transferred to its intended problem context" (Wieringa, 2014). Since the validation takes place before the actual implementation, validation models are used to simulate the implementation. Such a validation model consists of a "model of the artifact interacting with a model of the problem context" (Wieringa, 2014).

Multiple methods exist to validate a model of an artifact in a model of the problem context. For this research, expert opinions are chosen as the validation method, as they allow gathering information from a diverse panel of experts, and thus from a diverse set of companies and businesses. This is especially useful for the validation of the artifact within a broader public transport scope and can be used for generalization purposes as well, as explained in Section 2.6.2.

When asking experts for their opinion about the validation model, the expert is asked to imagine the validation model and 'observe' how it functions in their imagination. Due to the experts' experience, this is expected to be a useful validation method. However, expert opinions only work if the experts understand the validation model perfectly, imagine realistic problem contexts and make reliable predictions about the effects of the validation model (Wieringa, 2014). This problem is addressed in this research by providing extensive explanations and specific questions.

The validation model of this research is specified as follows:

<u>The model of the artifact</u> is the data integration design approach consisting of the target data architecture and approach, which are based on the best-practice planning process (RQ2), its future improvements (RQ3), its data requirements (RQ4), data quality aspects from the field of data management (RQ5) and TOGAF's data architecture approach considerations (Section 2.5.2).

<u>The model of the problem context</u> is the PTO's planning process including data-providing and data-consuming applications, processes and external parties.

2.6.1.1 Validation Instrument

To be able to collect and analyze the expert opinions, a validation instrument in the form of a Google Forms survey was designed. This validation survey consists of six different parts: business content validation, IT content validation, approach validation, validation model imagination, use and acceptance validation and personal information. Respondents are not obliged to validate every part (except for the mandatory part requesting consent and personal details), since not every respondent has enough business and/or IT knowledge (due to the diverse set of respondents). Moreover, the total validation instrument is quite time-consuming. In order not to discourage the respondents, they were in control of whether or not to answer a specific validation part (a time indication was provided). Each part is explained hereafter, whereas the entire validation survey can be found in Appendix G. The validation survey started with six compulsory questions regarding the respondent's consent.

Business Content Validation

The business content validation is focused on validating the design of the planning process and its tasks (RQ2). Since these lead to the data requirements (RQ4), it is important to check for their validity. Validation was about the proposed planning process as the base for this research, whether every phase is relevant for the company they work for and whether or not processes were missing. The validation questions for this part can be found in Appendix G: questions B1 – B12.

IT Content Validation

The IT-related validation is to validate the outcomes of RQ3, RQ4 and RQ5. Questions are about the data object (categorization), the data access service and data quality aspects. To mitigate the bias regarding data quality aspects, the respondents were asked to score data quality based on two scenarios (as presented in Table 6-3). The validation questions for this part can be found in Appendix G: questions I1 – I53.

Approach Validation

The validation of the approach contains questions about data ownership, data roles and a mapping of data objects to planning process phases. Respondents were asked for the relevance of these parts and which data roles are responsible for the data quality aspects. The validation questions for this part can be found in Appendix G: questions A1 – A8.

Validation Model

The respondents were asked to which extent the goals of this research were reached by imagining the validation model. Furthermore, they were asked which other goals could be reached and which negative impacts the artifact might have. The validation questions for this part can be found in Appendix G: questions V1 – V29.

Use and Acceptance

For the validation of use behavior and acceptance of IT artifacts, quite some research has been carried out. For this research, the Unified Theory of Acceptance and Use of Technology by Venkatesh et al. (2003) is used, of which the model can be seen in Figure 2-11. This theory is chosen because it is based on eight different IT artifact acceptance theories and it reaches a higher R squared value than the eight individual theories, which implicates that it represents the user adoption the most successfully (Venkatesh et al., 2003).

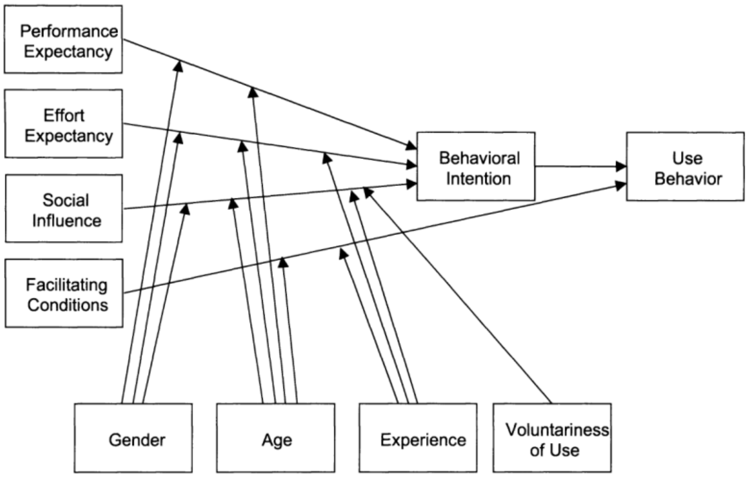

Figure 2-11 Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003)

The model is based on several other models and consists of four constructs (independent variables): ‘performance expectancy’, ‘effort expectancy’, ‘social influence’ and ‘facilitating conditions’. These constructs influence the ‘use behavior’ of a technology artifact. The relations between the four constructs and ‘use behavior’ are moderated by four moderating variables: gender, age, experience and voluntariness of use. For every construct and moderating variable, questions are included in the validation survey (questions U1 – U24, P1 – P6, P11 and P12 in Appendix G).

Personal Information

To better interpret the results, some more personal information than required by the UTAUT model was asked as well. Respondents were asked about their company, company type, offered modalities and some other details, especially for generalization purposes (RQ8). The validation questions for this part can be found in Appendix G: questions P1 – P15.

2.6.1.2 Validation Respondents

The respondents for the expert opinions are selected carefully, since in the perfect scenario they should be familiar with both the business side as well as the IT side of the planning process. All respondents can be classified as *normal operators* within the stakeholder identification (Section 1.3.3). They originate from three groups, which are GVB, national PTO architecture workgroup and rail transport business. From these groups, experts with both business and IT knowledge are asked to take part in the validation. To ensure completeness and not to leave out any important information, the experts were also asked to distribute the validation survey among their colleagues in case they think their feedback would be useful.

2.6.1.3 Validation of the Validation Instrument

The validation survey has been tested several times by a colleague of the author. Some minor changes were incorporated and the survey was iteratively improved. The last validation of the validation survey was carried by the graduation committee, as they were asked to approve it. Some small adaptations were made after which the validation survey was sent to the selected experts.

2.6.2 Generalization

Generalizability can be seen as the application of the findings to other contexts (Noble & Smith, 2015). By defining this generalizability, we seek to answer RQ8. Generalization can be done either case-based or sample-based (Wieringa, 2014). For this research, the case-based generalization is carried out, since no statistical method has been used to generalize from a sample to a population. This is not done because it would be unrealistic to reach a number of experts in the specific business & IT field for the planning process of a PTO which is high enough to apply statistical methods to. The case-based generalization is based on the validation carried out to answer RQ7.

For generalizability, conditions are presented which should be satisfied by other contexts. This is expected to be a useful method, since it is clear to see which aspects make the data integration design approach unique and at the same time decides the applicable contexts. Subsequently, these defined conditions are mapped to the PTO contexts of the experts who took part in the validation survey to propose a clear generalization overview of PTOs’ and broader contexts.

3 DATA INTEGRATION IN PUBLIC TRANSPORT PLANNING

This chapter aims to answer the first research question about the data integration situation at a public transport operator (PTO) and the data integration literature, as visualized in Figure 3-1. The data integration situation at GVB is explained in Section 3.1, followed by data standards in Section 3.2 and data integration literature in Section 3.3. It should be noted that every topic discussed in this chapter is part of a non-published stand-alone research by the same author. Only the relevant topics of that research are included in this current research, in line with the method explained in Section 2.2.

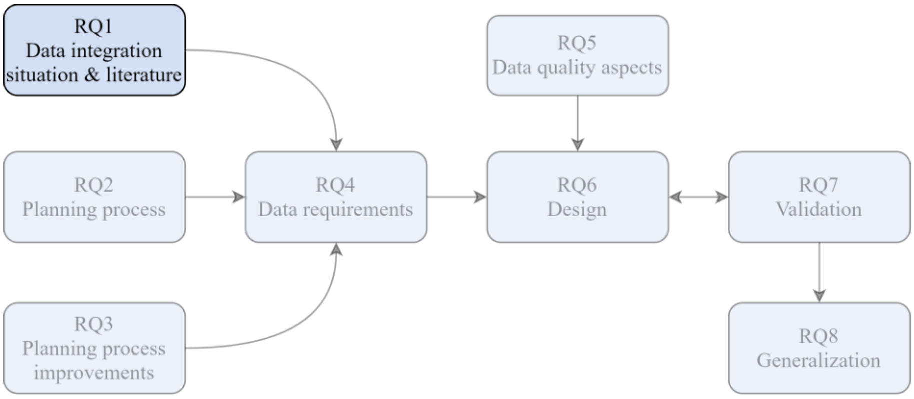

Figure 3-1 Content of chapter 3 (based on Figure 1-5)

3.1 Situation at GVB

As GVB is used as the perspective from which this research started, its data integration situation is assessed. A total of twelve interviews with eleven respondents were conducted. An overview can be found in Table 3-1. The interview method and participant selection are explained in Section 2.2.2.

Table 3-1 Respondent details

Respondent	**Function**	**Duration** (min)
R1	Application consultant	57
R2	Application consultant	96
R3	Application consultant	59
R4	Application consultant	61
R5	Application consultant	162
R6	Application consultant	44
R7	Application consultant	40
R8	Application consultant	53
R9	Program manager	53 + 60
R10	Solution architect	57
R11	Functional application manager	44

As can be seen in Table 3-1, the interviews with R5 and R9 took longer than the others. This was because R5 digressed from the main topic often and that R9 was interviewed twice since major questions arose after the first interview.

In the following subsections, the data integration situation, its challenges and incentives for improving the data integration situation are presented. All of these topics lead to requirements for the design of the data integration design approach.

3.1.1 Data Integration Situation at GVB

Assessing the data integration situation was carried out for the entire enterprise, in order to understand the data integration necessity of the planning process compared to other business functions. It was found that business functions depend on internal (from other business functions) and external data. This resulted in a complex data-dependency matrix of business functions. A simplified version of this matrix can be found in Table 3-2. In this table, only business functions that have an upstream and/or downstream data dependency are included. Therefore, it does not contain all of GVB's business functions. The explanation of upstream and downstream data dependencies can be found in Figure 2-5, whereas the marking of business functions in blue is explained later in this section.

Table 3-2 GVB's business functions and their data dependencies

GVB's business function	**Downstream** *(count)*	**Upstream** *(count)*
Risk management	0	1
Business monitoring	0	11
Transport planning	8	3
Staff scheduling	4	2
Vehicle and vessel scheduling	3	1
Transport operations	5	5
Rail infrastructure management	3	1
Vehicle and vessel management	3	1
Stations management	1	0
Social safety assurance	1	1
Travel information composition	2	2
Customer relation management	2	3
Marketing	2	2
Sales & collection	4	2
Security	0	1
HR management	13	3
Financial administration	2	9
IT services	0	1
Facility services	1	1
Purchasing	1	1

A total of 19 out of 20 business functions with data dependencies are dependent on internal data (upstream data dependencies). This data is provided by 16 business functions, which means that data is integrated between different business functions. One out of 20 business functions with data dependencies do not depend on data from other business functions, but only provides data. The other way around – being dependent on data but not providing data – is valid for 4 out of 20 business functions. The number of actual data integrations is significant higher than the number of data dependencies as shown in Table 3-2. This has two reasons:

- Business functions are realized by more than one application. There is more than one integration possibility/necessity per business function (note that in Table 3-2 an integration between two business functions is counted as one).
- Applications within one business function also integrate with each other. This is not accounted for in Table 3-2.

The actual number of integrations is very difficult to count, and documentation of integrations is not kept up to date. However, a well-educated guess based on an analysis of the GVB enterprise architecture repository was done by Van de Velde (2019), who stated that there are a total of at least

1700 main integrations within GVB. More details about the data integrations are discussed on planning and external level next.

3.1.1.1 Planning Data

To indicate the data integration situation for the planning process within GVB, business functions that realize the planning process are identified. From Table 3-2, the following business functions are the responsibility of the planning process: *transport planning*, *staff scheduling*, *vehicle and vessel scheduling*, *travel information composition* (partly) and *sales & collection* (partly). These latter two are only included partly, as the planned timetable and fare information is part of the planning process' responsibility. This decision is validated after the planning process study (Chapter 4) and validation (Chapter 8).

As can be seen in Table 3-2, the aforementioned five business functions have a vast number of data dependencies. The planning process can thus be seen as a data-intensive process, as it consumes and provides a lot of data. Since every business function is supported by one or more applications, the number of integrations is even higher than the number shown in Table 3-2. A main finding from the interviews was that all integrations are established point-to-point. Hence, a lot of redundancy is likely to be present in the integrations, especially for the business functions which have a vast number of upstream data dependencies.

3.1.1.2 External Data Dependencies

External parties are parties that are a consumer of GVB's data, without being within the responsibility of any of GVB's business function(s). These external parties use GVB's data for their own business functions and customers. An overview of external parties and the (planning) data they consume, can be found in Table 3-3.

Table 3-3 External parties consuming GVB's planning data

External party	Data
9292	Planned and real-time schedule information, including announcements, cancellations, delays, etc.
OpenOV	
DOVA	
Translink	Check-ins/outs, price
MaaS platform	Same as all of the above, including real-time ticket information (e.g. ticket validity)
VRA	Reporting about planning, execution, tickets, passengers, fines, etc.

The MaaS (Mobility-as-a-Service, explained in Section 3.1.2) platform is not in use yet, which means that no data is consumed yet. However, since this development is ongoing, it is already taken into account. A growth of external integration parties is expected, both for upstream and downstream data dependencies. This is also endorsed by several experts after validation of this research. More information is presented by means of running and upcoming projects in Section 3.1.2.

3.1.2 Data Integration Challenges at GVB

During the interviews, challenges related to data management came to light. These can be divided into the following five subjects: data standards, sources, integrations, ownership and new projects. All of these subjects are briefly explained hereafter.

3.1.2.1 Data Standards

GVB provides schedule and real-time information to external parties in BISON KV[2] and – for some parts – SIRI standard. In the near future, the BISON KV standard will be replaced by the European NeTEx (explained in Section 3.2.2) and SIRI standard (explained in Section 3.2.3) entirely. Data is transformed into these data standards only at the end of GVB's process. Within and in between the business functions, processes and applications, data regarding planning, scheduling and operations is

[2] Dutch data standard for public transportation: https://bison.dova.nu/

transferred using other standards (e.g. VDV452[3]) or sometimes even without any standard (application-specific). As explained in Section 3.1.1, the business functions realizing the planning have a vast number of upstream data dependencies. This means that data is transformed very often to different applications while using different data standards. Using one data standard for a specific data domain would improve this process and would lead to lower costs for development and maintenance.

3.1.2.2 Data Sources

Some data entities within GVB are maintained within several applications and have therefore more than one source where data about these entities is stored. A consequence is that redundant data is present across the IT landscape and no single point of truth for the data entity can be found. This can lead to situations in which outdated data is used, which is an undesirable situation. Working in a more real-time manner is even impossible in such a situation, since real-time decisions require real-time and always correct data. For data integrity and confidentiality reasons, data should be maintained at only one location, preferably as close as possible to the source.

3.1.2.3 Data Integrations

Interfaces used for data integration are mostly made point-to-point. This means that it might take long before an integration with a data source is established, because no integration can be reused. Furthermore, the time spent on development, maintenance and changes increases with every new integration.

3.1.2.4 Data Ownership

It became apparent that data ownership is not officially assigned and employees do not have a common view on this topic. One respondent answered that data is often stored in application databases. Since applications are not developed by GVB, sometimes data does not end up in a GVB-owned database. This is tricky, since the external development party then actually 'owns' the data. Since data is stored in application databases and applications serve departments, the respondent stated that the ownership of data is on the level of GVB's departments. Yet, sometimes applications are shared between different departments, which means that ownership is difficult to assign.

Since there is no data owner, there is also not someone who can make decisions on data-level. This results in changes executed by vendors if the majority of the product team agrees (sometimes without input from the business). It is even a bigger challenge if the application database or application is not allocated to an IT team. Another consequence of not having clear data ownership is that sometimes documentation about data is not present, outdated and/or incorrect. This means that – in the case of changes and projects – it is hard to estimate the impact of changes.

To get a more detailed insight into the data situation, the CIA (confidentiality, integrity and availability) triad is used. The CIA triad provides practitioners a straightforward way to understand and address problems (for data owners) that relate to information security (Samonas & Coss, 2014). It can be seen as a method to explain and assure confidentiality, integrates and availability of data.

According to the respondents, no corporate overview of the CIA aspects of data or applications is maintained by GVB. However, often the CISO (Chief Information Security Officer) is involved regarding the confidentiality of data and that availability is often determined in service level agreements with vendors.

Confidentiality Problems

No confidentiality problems came up during the interviews.

Integrity Problems

Data provided by *Transport operations* is not always correct. An example is that in the case of a tram ride, operations data shows that a tram passed tram stop 1 and tram stop 3, but not tram stop 2.

[3] German data standard for public transportation: https://www.vdv.de/oepnv-datenmodell.aspx

However, since a tram is running on a rail track, the tram simply should also have passed tram stop 2. This is an undesirable situation, since it could lead to fines from the VRA. Right now, this situation is solved by including extra business rules at the side of the data consumer (in this case the Business Intelligence IT team). This means that a business rule is necessary to interpret the data, which, in this case, is not located at the source of the data. This can lead to different interpretations of data among different applications, especially when using different sources of the same data.

Availability Problems

Sometimes data is unreachable, since vendors of applications do not allow GVB to read the data to integrate within other services. An example is a mechanism that automatically checks the tire pressure of buses. This data is stored in an application and database from a specific vendor. For some reasons, this vendor does not let GVB access the raw data, which means that GVB cannot use the data for their own purposes, but only read the data through the vendor's application (Information Manager, personal communication, March, 2020). This is an undesirable situation, since GVB cannot incorporate the data in their maintenance process and an extra separate application needs to be used and placed in the IT landscape (completely stand-alone) to access the data.

3.1.2.5 Projects and Innovations

During the period of conducting this research, some ongoing projects within GVB have a clear demand for vast amounts of data. Moreover, this data should be of high quality and should be exchanged in a real-time manner. Many different sources of data have to be accessed and data needs to be shared and combined to fulfill the goals the projects aim for.

Examples of such projects are the introduction of new payment methods (banking card and barcode), the introduction of Mobility-as-a-Service (in which different transport modalities will be offered as a complete package to the passenger), the development of a new GVB application and the improvement of the planning process and depot management by introducing a completely new planning software package. For the future, more data integration will likely have to take place. This ensures the ongoing increase of integration of data and processes, implying higher complexity and more maintenance.

All aforementioned projects are incentives for a better data integration situation, since having that in place would ease the project process. More details about the projects, the demanded data objects and requirements can be found in Appendix H.

3.2 Data Standards

Several data standards in the field of public transport planning exist. These are developed for sharing schedules and real-time planning information. These were identified during the interviews (see Section 3.1) and desk research. Most of them are based on the same conceptual model, of which the first steps were set back in 1989 (CEN, 2019). This conceptual model is called Transmodel and has been updated continuously since then (CEN, 2019).

Currently, the common standard in The Netherlands is BISON KVs. A new standard to which the PTOs must adhere, are the Dutch profiles of the European standards NeTEx and SIRI. Therefore, BISON KVs will be phased out in the future. Each of the mentioned standards ensures that IT systems from any stakeholder can communicate with each other and is able to understand the exchanged data. The pyramid structure of the standards based on Transmodel is visualized in Figure 3-2.

Next to data standards which are based on Transmodel, other standards that are based on other conceptual models also exist. These are for example the VDV452[3] and GTFS[4] standards. Since PTOs do not have to provide their data in these standards, they are left out of scope for this research. Besides this, the BISON KVs standard will not be discussed, as it will be phased out. Transmodel, NeTEx and SIRI are explained in the next subsections.

[4] International data standard for public transportation: https://developers.google.com/transit/gtfs

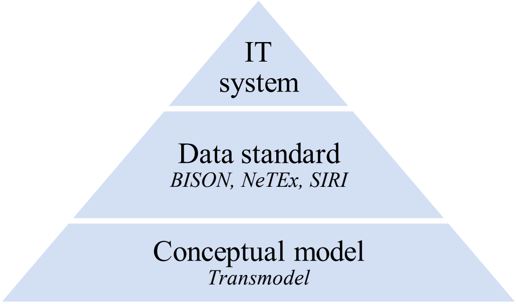

Figure 3-2 Conceptual model, data standards and IT systems

3.2.1 Transmodel

Transmodel is a conceptual model and aims to establish a clear concept about all main data objects and their relationships within the public transportation sector. This is necessary because passengers demand schedule and real-time information from any PTO for any transport modality to plan their journeys. Hence, data of different operators and modalities need to be integrated, which means that a standard for communication is desirable (CEN, 2019).

An example of standards according to Transmodel is the definition of (and differences between) a *service journey* and *PT trip*. A *service journey* is related to the movement of a vehicle, whereas a *PT trip* is related to the movement of a passenger (CEN, 2019). This means that a *PT trip* can consist of (parts of) one or more *service journeys*. All parties involved in the data exchange must have the same meaning for these terms, otherwise no valuable information can be extracted from the data.

Transmodel is organized in eight different parts, each with its own scope regarding exchanging public transportation data. An overview of all parts can be found in Table 3-4.

Table 3-4 Transmodel parts (adapted from CEN (2019))

Part	**Scope**
Part 1	Common Concepts
Part 2	Public Transport Network
Part 3	Timing Information and Vehicle Scheduling
Part 4	Monitoring and Control
Part 5	Fare Management
Part 6	Passenger Information
Part 7	Driver Management
Part 8	Management Information and Statistics

3.2.2 NeTEx

NeTEx is a data standard for exchanging public transport networks and schedules which is developed by the European Committee for Standardization (CEN) and stands for Network Timetable Exchange. It strives for systemized and harmonized European passenger information data (Arneodo, 2015). The standard is based on Transmodel and some national standards like VDV452 (DE), Neptune (FR) and TransXChange (UK) (CEN, n.d.-a).

NeTEx is divided into three functional parts. Part 1 describes the network of the PTO (e.g. stops, routes and lines). Part 2 is about timetables and part 3 covers fare data (ticketing and pricing) (Arneodo, 2015). NeTEx does not cover real-time data exchange. The SIRI standard is used for the exchange of data to modify the planned network and timetable (in NeTEx) in a real-time manner. More information about SIRI can be found in Section 3.2.3.

Different profiles of the NeTEx data standard exist within Europe. Many countries have their country-based profile of the standard. The reason for this is because every country has its specialties regarding public transportation, in the field of data but also regarding business processes. Consequently, this means that exchanging information between countries becomes impossible, even though all profiles are based on NeTEx. For this reason, NeTEx also proposed the European Passenger Information

Profile (EPIP), which is a small subset of the entire NeTEx profile. Some data elements have been left out, e.g. dead runs (journeys without passengers, most often from a depot to a starting point or vice versa) and operational information like vehicle blocks (Reynolds, 2019). The relationships between Transmodel, NeTEx, NeTEx EPIP, NeTEx NL's and other country's profiles can be found in Figure 3-3.

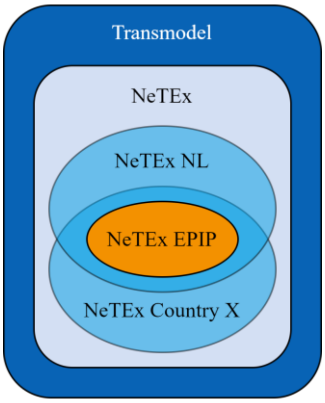

Figure 3-3 Transmodel, NeTEx and NeTEx profiles (adapted from Reynolds (2019))

3.2.3 SIRI

The Service Interface for Real-time Information (SIRI) standardizes exchanging information about planned, current and projected performance of real-time public transport operations between different stakeholders and systems (CEN, 2005). It allows the exchange of structured real-time information about schedules, vehicles and connections and provides the publisher with the possibility to share general information messages related to the operations of the public transportation services (CEN, n.d.-b). Examples are real-time departure information, real-time progress information and the movement of vehicles (CEN, n.d.-b). SIRI is divided into five different parts, which can be found in Table 3-5.

Table 3-5 SIRI parts (adapted from CEN (n.d.-b))

Part	Scope
Part 1	Context and framework
Part 2	Communications infrastructure
Part 3	Functional service interfaces
Part 4	Functional service interfaces: Facility Monitoring
Part 5	Functional service interfaces: Situation Exchange

3.3 Data Integration Literature

For providing an overview of the current data integration literature, a systematic literature review (SLR) was carried out. The method of this SLR is described in Section 2.2.1. Topics that are covered in this SLR are data integration challenges (Section 3.3.1) and data integration methods (Section 3.3.2). Both are important to consider while designing the data integration design approach. This section is not PTO planning-specific, as that is covered by Chapter 4 and Chapter 5.

3.3.1 Data Integration Challenges

Challenges around data integration were already discovered in the time that databases came up. Data conversion was necessary to store data in databases (Shu et al., 1977). This means that data integration challenges are already known for over 40 years (Jarke et al., 2014). However, there is still no perfect solution for this problem. To understand the complex task of data integration and to find solutions, it is necessary to identify all possible data integration challenges.

Both Halevy et al. (2006) and Doan et al. (2012) divided data integration challenges into different categories. Whereas the first classified the problems into 'social' and 'complexity', the latter classified the difficulties into 'system reasons', 'logical reasons' and 'social and administrative reasons'. A classification of categories is useful for overview purposes and understanding. For this reason, the categories from Halevy et al. (2006) and Doan et al. (2012) are combined into four different categories: ‘systems’, ‘logic’, ‘social’ and ‘administrative’, see Table 3-6.

Table 3-6 Categorization of data integration challenges

Category	Halevy et al. (2006)	Doan et al. (2012)
Systems	Complexity	Systems reasons
Logic		Logical reasons
Social	Social	Social and administrative reasons
Administrative	-	

In total, 52 data integration challenges were found in 16 studies. Some of them overlap and others extend each other. The complete list can be found in Appendix I. All challenges are consolidated into a list of 12 different data integration challenges and can be found in Table 3-7. The main findings are discussed per category in the next sections.

Table 3-7 Consolidated list of data integration challenges

Category	Challenge	Literature
Systems	Ongoing development of (new) data sources	Bernstein & Haas (2008); Halevy et al. (2005)
	Performance issues while querying over multiple systems	Doan et al. (2012); Halevy et al. (2005); Jarke et al. (2014)
	Technical differences between systems	Bahga & Madisetti (2015); Doan et al. (2012); Giachetti (2004); Lemcke et al. (2012); Zhang et al. (2018)
	Vast number of data sources: more sources for one data element which might lead to data inconsistencies	Bernstein & Haas (2008); Evgeniou (2002); Golshan et al. (2017); Halevy et al. (2006); Lenzerini (2002); Salguero et al. (2008)
Logic	Agreements on and poor definitions of data semantics	Batini et al. (1986); Giachetti (2004); Golshan et al. (2017); Halevy et al. (2005); Lemcke et al. (2012); Zhang et al. (2018)
	Different inter-schema relationships with other objects	Batini et al. (1986); Giachetti (2004)
	Semantic heterogeneity of data sources, schemas, models, definitions, perspectives, concepts and meanings due to human nature and usage in different contexts of the business	Bahga & Madisetti (2015); Batini et al. (1986); Bernstein & Haas (2008); Doan et al. (2012); Evgeniou (2002); Giachetti (2004); Halevy et al. (2005); Halevy et al. (2006); Jarke et al. (2014); Lemcke et al. (2012); Peng et al. (2019); Sezgin et al. (2019); Zhang et al. (2018)
Social	Convincing people that sharing data contributes to a higher-level goal (by offering incentive(s))	Doan et al. (2012); Halevy et al. (2006)
	Integration takes away the exclusive right of the data owner	Doan et al. (2012)
Administrative	Finding the correct data source	Doan et al. (2012); Halevy et al. (2005); Halevy et al. (2006)
	Privacy issues and anonymization for legitimate legal reasons	Doan et al. (2012); Halevy et al. (2006)
	Security to prevent unauthorized users from access	Doan et al. (2012); Halevy et al. (2005); Halevy et al. (2006)

3.3.1.1 Systems

In this category, challenges are related to the technical aspects of data integration. The challenge which was mentioned the most in the assessed literature is the challenge regarding the multiple data sources. Especially the challenge of having multiple values/possibilities for the same data object in different systems. This leads to more than one truth. Furthermore, technical differences and the ongoing development of (new) sources lead to physical difficulties when integrating data. Lastly, performance is a key challenge, since querying over multiple data sources decreases the performance compared to querying over only one data source.

3.3.1.2 Logic

The data integration challenge which was present in almost every study, is the challenge of the semantic heterogeneity of data. This means that different interpretations for the same data exist. These interpretations are most often related to a specific smaller business context in which they are seen as the only definition. However, since this is defined only within a smaller business context, it is not always valid for the entire business (Doan et al., 2012; Giachetti, 2004; Golshan et al., 2017; Halevy et al., 2006). Furthermore, even as soon as there are a common understanding and definition of the semantics, challenges still arise with the inter-schema relationships of data, which proofs that this challenge is rather complex.

Trying to mitigate the problem of semantic heterogeneity leads to the problem of reaching agreements of data semantics which are often poorly defined. In both cases, the business is still confronted with the problem of semantic heterogeneity or, even worse, no data semantics at all.

3.3.1.3 Social

Only a small number of studies mentioned challenges related to the social aspect of data integration. The reason for the limited presence of social aspects is most likely due to the more technical approaches of the assessed literature.

Two social aspects which were identified in the literature are that (1) data owners need to be convinced of the added value for the business of sharing 'their' data and (2), as a consequence, data will become more publicly available which might result in issues regarding ethics (e.g. employee's efficiency data) (Doan et al., 2012).

3.3.1.4 Administrative

In accordance with the social challenges, administrative challenges were also less present in the assessed literature, presumably also because of the more technical approach of the studies. Nevertheless, three studies mentioned difficulties regarding finding the right data sources, two studies mentioned that privacy issues are challenging (including anonymization when necessary according to the law) and three studies addressed the importance of security of the data used in data integration. The latter one is especially challenging since integrated data might not always be under the control of the data owner anymore (Doan et al., 2012).

3.3.2 Data Integration Methods

Standardization of data and transactions is highly necessary for proper data integration (Giachetti, 2004). While working with standards, the users and developers of systems work towards the agreed standard, which avoids high costs associated with integration. However, this only works for smaller parts of enterprises, since not all domains could be combined into one standard. This means that data standards and other data integration methods are still necessary. Moreover, no single technology approach can achieve data and enterprise integration, since the technology and demands within an enterprise are diverse. Rather more integration methods should be combined to achieve the data integration of an enterprise (Giachetti, 2004).

Data integration techniques (such as API, XML file transfer, ESB, RPA) are left out of scope for this research, since these are the technologies working according to a data integration method. Which technology suits the best differs per data integration case and depends on aspects such as data latency, data volume, communication methods (e.g. event-based or push/pull) and many more.

Broadly speaking, data integration methods can be classified into two categories: virtual integration and warehousing (Doan et al., 2012). Whereas virtual integration offers solutions based on the operational databases, the latter ensures that data is physically stored in another database. These two main categories can be further specified in specific methods. The following methods are covered and discussed widely in literature:

- Global schema
- Canonical Data Model (CDM)
- Ontology matching
- Data warehouse

As data warehousing is focused on analytics (Doan et al., 2012; Sezgin et al., 2019) by providing a historical vision of the data in different operational systems (Salguero et al., 2008) and does not contain real-time information (Katasonov & Lattunen, 2014), it is out of scope for the design in this research. The other three methods are introduced in the next subsections.

3.3.2.1 Global Schema

A global schema (Batini et al., 1986; Golshan et al., 2017; Halevy et al., 2006; Jarke et al., 2014; Lenzerini, 2002), also referred to as mediated schema (Bernstein & Haas, 2008; Doan et al., 2012; Golshan et al., 2017; Halevy et al., 2005, 2006; Lenzerini, 2002), global view (Giachetti, 2004; Halevy et al., 2005) or unified schema or view (Giachetti, 2004; Katasonov & Lattunen, 2014; Lenzerini, 2002), provides a uniform query interface for the user (or application) for multiple heterogeneous data sources. It is sometimes called the (query) mediator (Bernstein & Haas, 2008; Jarke et al., 2014). The method is seen as virtual data integration, since data is not physically stored in the global schema (Bernstein & Haas, 2008; Doan et al., 2012; Halevy et al., 2005; Jarke et al., 2014). This is also referred to as the federated database principle (Giachetti, 2004; Zhang et al., 2018).

The advantage is that users pose queries to only one global schema (having fewer connections because of the 'mediator', see also Figure 3-4) (Golshan et al., 2017), that data remains in the operational databases (Doan et al., 2012) and that it increases the visibility of data on global level due to the centralized approach (Evgeniou, 2002). However, a drawback is that flexibility on the lower levels in the organization is reduced (Evgeniou, 2002), because they need to adhere to the global schema. A typical architecture based on a global schema can be found in Figure 3-4.

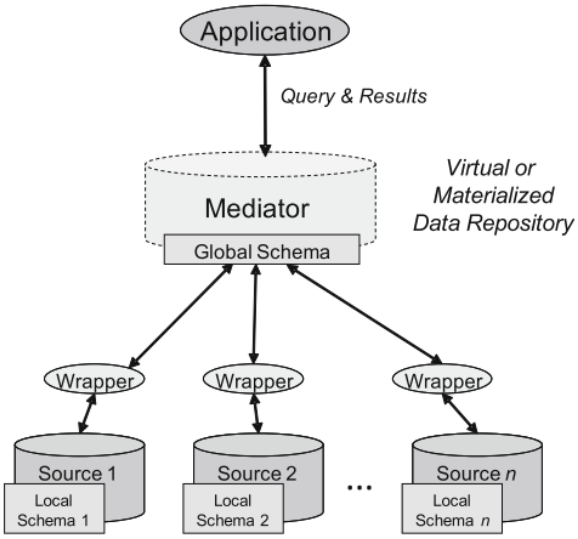

Figure 3-4 Architecture of an information integration system (Jarke et al., 2014)

A global schema is seen as a solution for data integration for some time already and some disadvantages and difficulties have been identified. It turns out to be very difficult to agree upon a global schema, even when data standards are in place (Lemcke et al., 2012). This becomes even more difficult when the number of data sources within the enterprise is enormous, which is most often the case in modern enterprises (Golshan et al., 2017).

Practically seen, an issue is that it might be impossible to have a global schema or model that satisfies all usage scenarios or use cases defined by the business (Peng et al., 2019). Another issue is the rate at which new data sources emerge within an enterprise (Halevy et al., 2005). All new data sources always need to be mapped to the global schema. Modifying and/or extending the global schema might be difficult (Katasonov & Lattunen, 2014) and could result in completely refactoring the schema.

When you cannot build a global schema, the question should arise whether it is really necessary to integrate all available data (Golshan et al., 2017). This is in line with Halevy et al. (2005), who stated that a one-size-fits-all approach is often unsuitable and therefore suggested: "to develop an integration approach that is significantly more nimble and adaptable to the needs of each integration (...)".

3.3.2.2 Canonical Data Model

A canonical data model (CDM), also referred to in literature as a global domain model (Bahga & Madisetti, 2015), shared data schema (Giachetti, 2004) or common data model (Salguero et al., 2008), is a model which ensures a single view of data for multi-enterprises, enterprises, departments or processes and can be independently used by any system or user (Lemcke et al., 2012). It ensures linking and combining data, which facilitates semantic interoperability among different applications (Singh et al., 2017). By doing this, the CDM should be able to integrate differently and possibly even conflicting structures (Lemcke et al., 2012).

A CDM ensures that data can be used in one standard data format. It uses a global schema (global in this case can be on (multi-)enterprise, department or process level) to which all data is transformed. This is different compared to a global schema, which does not physically store the data.

The advantage of a CDM is that applications can be integrated via a mediator, as explained for a global schema, which will result in fewer integrations. However, the disadvantages mentioned for a global schema are also applicable to the construction of a CDM. It should take into account many different stakeholders, with all of them having their own opinion and meaning for the data objects. Even though Peng et al. (2019) mention that CDMs are still trending, the disadvantages and challenges should be considered very carefully when opting for a CDM.

3.3.2.3 Ontology Matching

Ontologies are a representation and specification of a knowledge domain based upon a controlled and standardized vocabulary for describing entities (Salguero et al., 2008; Zhang et al., 2018). The semantic relationships between them are described formally and have so-called 'grammar' for using the vocabulary to express meaningful information based on the data (Salguero et al., 2008; Zhang et al., 2018).

Ontologies can be used to facilitate data integration (Zhang et al., 2018). It keeps away the user from matching on data-level (e.g. for global schemas and CDMs), but lets the user focus on modeling the problem itself (Salguero et al., 2008). This is a shift from local data view to higher concepts, which results in higher semantic interoperability (Giachetti, 2004).

By using ontologies instead of lower-level data mapping, a softer integration view is created which is much easier to modify and extend later (Katasonov & Lattunen, 2014). Furthermore, ontology matching considers semantic knowledge, which results in integrating data based on the actual meaning of a data element (Zhang et al., 2018). This knowledge comes from meta-data.

Meta-data is data about data, often a vast amount of information that describes the characteristics of the data (Salguero et al., 2008) depending upon data types, sources and intended use (Singh et al., 2017). Next to the reason that meta-data is important for the understanding of the actual data (Bahga & Madisetti, 2015; Batini et al., 1986; Halevy et al., 2005; Jarke et al., 2014; Salguero et al., 2008), it is also very important for data extracting, loading and refreshing processes (Salguero et al., 2008) and the automatic schema matching based on ontologies (Jarke et al., 2014; Salguero et al., 2008; Zhang et al., 2018). In other methods, meta-data is often stored in a separate database (Singh et al., 2017).

An interesting difference in the application of ontology matching was found in the assessed literature. Whereas Salguero et al. (2008) state that ontology matching can be used for defining a CDM or schema for a data warehouse, Zhang et al. (2018) claim that the use of ontologies goes way further than the 'traditional approaches' such as creating a CDM. From the process of using ontology matching as explained in this section, it can be stated that ontology matching could be used for developing a CDM. However, by using a CDM for data integration, tasks such as meta-data representation, automatic data verification and global conceptualization are not taken care of (Zhang et al., 2018).

3.4 Chapter Summary

The data integration situation, challenges, standards and methods are discussed in this chapter. These are all important for the design of the data integration design approach.

It turns out that GVB's planning process consumes and provides a lot of data, from and to different (internal and external) parties. Some challenges around data integration are identified as well, which are related to data standardization, data sources, data integrations and data ownership.

Data standards in the field of public transportation planning are explained. The important European conceptual model within the public transportation is the Transmodel. Based on this model, several European standards are developed, which include NeTEx and SIRI.

The last topic in this chapter was about generic data integration literature. Within literature, data integration challenges are already known for 40 years (Jarke et al., 2014). A literature study to identify these challenges was performed and led to a consolidated list of data integration challenges. Furthermore, to solve the challenges, the following three data integration methods were studied and their similarities and differences were presented: developing a global schema, canonical data modeling and ontology matching.

4 PLANNING PROCESS

This chapter aims to answer the research questions about the planning process and its improvements, as visualized in Figure 4-1. The planning process is studied based on the operations research domain (Section 4.1) and GVB's practice (Section 4.2). These two perspectives are combined into a best-practice planning process which is presented in Section 4.3. The expected innovations of the planning process are presented in Section 4.4.

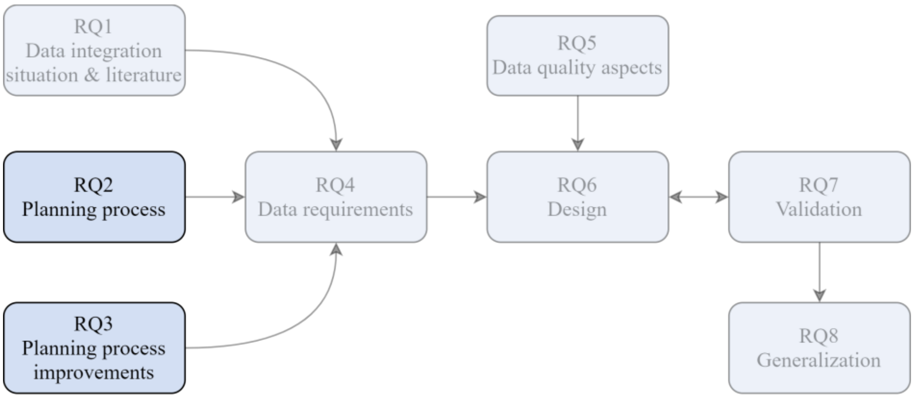

Figure 4-1 Content of chapter 4 (based on Figure 1-5)

4.1 Operations Research Perspective

The planning process of a PTO is a widely-studied research area in both operations research and mathematical research. The planning process was defined for the first time in the late 60s (Lampkin & Saalmans, 1967) and was based on the bus planning process. It turned out that this planning process could be generalized to other public transport modalities, for example for tram, metro and train services. The most recent literature about this topic originates from 2016 (Ceder, 2016) and still contains the main steps as were defined by Lampkin & Saalmans (1967).

In order to identify the planning process, the concept-centric literature review approach from Webster & Watson (2002) as proposed by Okoli (2015a) was used, in line with the method explained in Section 2.3.1. The public transport planning phases were identified and the literature sources were analyzed to identify the presence of the planning phases. In Figure 4-2, an overview of the concepts within the planning process is given (details can be found in Appendix E). The size of the circles represents the number of literature sources in which the particular planning phase was present.

As can be seen in Figure 4-2, a vast overlap exists between the phases described and explained by each literature source. Many authors state that the planning process for public transportation consists of several sequential phases, since the overall planning problem is not tractable as a whole (Békési et al., 2009; Ceder, 2016; Ceder & Wilson, 1986; Desaulniers & Hickman, 2007; Lampkin & Saalmans, 1967). These phases can be mapped to the three commonly known organizational planning phases: strategical, tactical and operational (Desaulniers & Hickman, 2007). The output from a particular phase becomes the input for the next phase in the sequence (Ceder, 2016; Ceder & Wilson, 1986).

Figure 4-3 shows the typical planning process. When *t = day of operation*, the process starts at *t – some years* and ends on day *t*. The process is explained per phase in the remainder of this section.

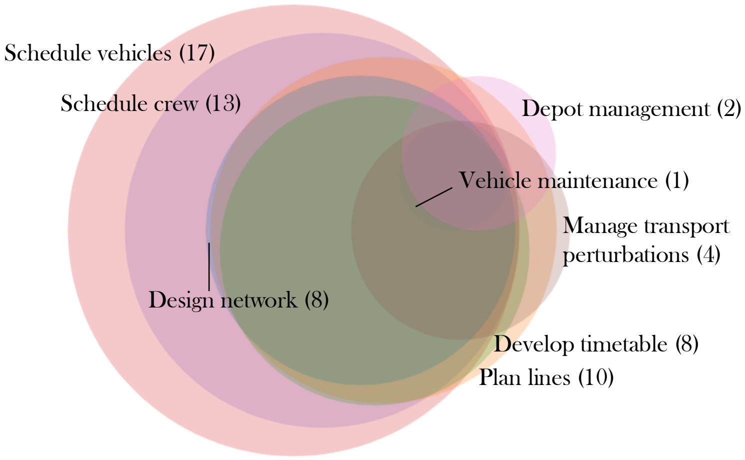

Figure 4-2 Planning phases concepts

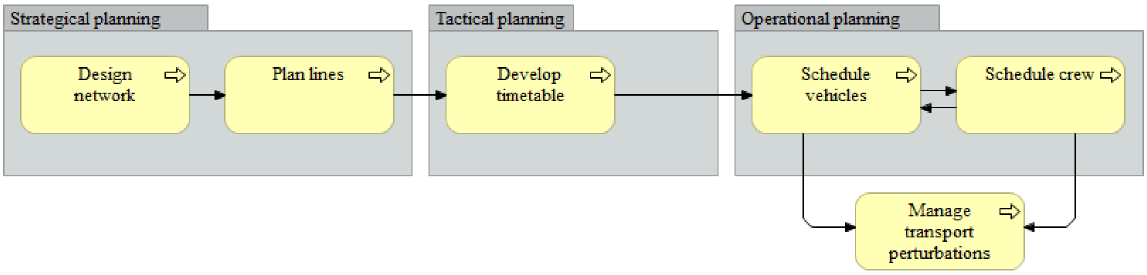

Figure 4-3 Planning process for public transportation[5]

Design network takes into consideration strategic infrastructural planning such as the locations of bus and tram stops, metro stations, rail tracks and bus lanes. It provides a layout for the public transport system.

Plan lines ensures that the network designed in the previous step is amplified with the actual lines (see Figure 4-4) and their frequencies. In this step, public transport supply and demand is tried to be optimized.

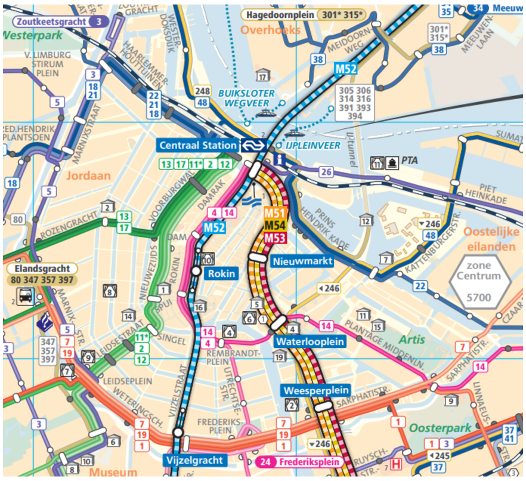

Figure 4-4 Excerpt from GVB's line planning (GVB Holding NV, 2020b)

[5] Combined overview of the planning process based on Békési et al. (2009), Ceder (2016), Ceder & Wilson (1986), Desaulniers & Hickman (2007) and Lampkin & Saalmans (1967).

Develop timetable contains activities to transform the given frequencies per line into a timetable for a particular line. This phase takes into consideration the travel times between stops, the time necessary at start and end locations for, respectively, beginning and finalizing the journey and the time necessary in between ending a journey and starting a new one. In case it is important to allow passengers a transfer from one line to another (from and to any modality), these transfer times are also taken into account, together with the synchronization of the journeys.

Schedule vehicles is the first operational planning phase. In this phase, the timetables are changed into vehicle schedules. These vehicle schedules determine what a vehicle's tasks are for a given day. These tasks consist of so-called 'service journeys' and 'dead runs' (CEN, 2019). A service journey is a journey that is meant to carry passengers, whereas a dead run is meant to not carry passengers. Examples of when a dead run is necessary, are to travel from and to the garage, to switch lines, to park and to turn around. The ultimate goal for the vehicle scheduling phase is to execute the timetable with a minimum number of vehicles. Important to note that this phase is not responsible for the exact vehicle, but only for the so-called 'vehicles blocks': "the working of a vehicle from when it leaves the depot to when it returns back" (CEN, 2019).

Schedule crew is the phase that is responsible for the scheduling of crews, i.e. drivers and/or conductors. This phase can be split into *Crew scheduling* and *Crew rostering*, as proposed by Békési et al. (2009):

- *Crew scheduling* is the phase that is responsible for the creation of crew schedules based on the vehicle schedules of the previous step. While creating these schedules, several constraints need to be taken into account. Examples are maximum working times, maximum driving times, minimum break times, locations of breaks and extra working time necessary for starting/ending duties, changing vehicles and other non-driving activities.
- *Crew rostering* is the phase in which the crew schedules are divided among the personnel. It highly depends on the PTO how this phase takes place. What does not differ are the several constraints that need to be taken into account. For example regulations concerning working times, working days, contract hours and granted leave.

The difference and dependency between vehicle scheduling and crew scheduling are shown in Figure 4-5. Even though the case in the figure is very simplified, it still shows the main differences and dependencies between the two scheduling activities. In the figure, it is assumed that the vehicles start and end at the same location, to let the driver change and take a break.

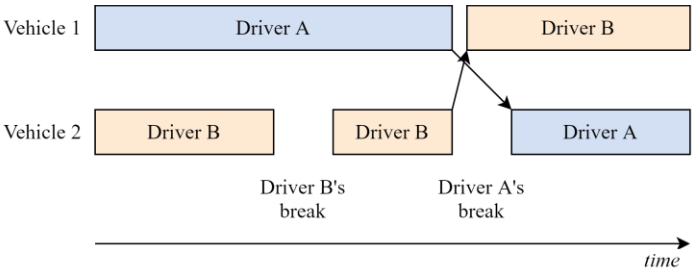

Figure 4-5 Vehicle and crew scheduling example (simplified)

Manage transport perturbations is the activity that should solve problems during operations, taking into account the minimization of passenger inconvenience (Desaulniers & Hickman, 2007). Harbering (2017) and Shen et al. (2016) refer to this phase as delay management and on-site operation, respectively. Delays occur unavoidably within a public transport network (Harbering, 2017). Unexpected delays should be managed real-time during the operation. Expected delays are the ones that can be planned and scheduled upfront. This means that they are not part of this real-time control phase but are included in the overall public transport planning prior to this (Harbering, 2017).

Desaulniers & Hickman (2007) added two extra phases to the planning process: 'bus parking and dispatching in garages' and 'maintenance scheduling'. Since these tasks do not always take place in sequential order in the process as shown in Figure 4-3 and the other sources do not mention these phases as part of the planning process, these are not incorporated in the figure. However, these two planning tasks can be seen as tasks that are dependent on the outcomes of the planning process, but can also influence the planning itself because dispatching is dependent on maintenance scheduling (David et al., 2018).

4.1.1 Optimization of the Planning Process

Already since the beginning of modeling the planning process, it was noted that optimizing each individual step might not lead to an overall optimized solution (Lampkin & Saalmans, 1967). More recently, this statement is still endorsed (Desaulniers & Hickman, 2007). In science, this optimization of every separate step to reach an optimized situation is called suboptimization. The principle states that "suboptimization in general does not lead to global optimization" (Machel, 1965). Hence, optimizing the public transportation planning process by optimizing the several steps does not lead to an optimized global planning.

To reach a more optimized planning, the integration of planning activities is necessary (David et al., 2018; Desaulniers & Hickman, 2007). More specifically, the combination of strategic and tactical planning is advised (Canca et al., 2016) as well as the combination of vehicle and crew scheduling (Nagy & Tick, 2019; Shen et al., 2016; Weider, 2007).

As shown by Weider (2007), in some situations it is even impossible to first schedule vehicles followed by the scheduling of drivers. This is, for example, the case when a vehicle is on-route for a longer period than the driver is allowed to drive, without the possibility for a driver to take a break. Furthermore, both Steinzen (2007) and Scholz (2016) showed that scheduling vehicles and drivers independently can lead to higher costs.

4.1.2 Complexity of the Planning Process

It turned out that every step in the public transport planning process is a difficult computational problem to solve, due to the many different possible solutions. This is proven by Lenstra & Kan (1981) and Bertossi et al. (1987) for the vehicle scheduling problem, stated by Weider (2007) for the crew scheduling problem and proven by Winter & Zimmermann (2000) for the vehicle parking and dispatching problem. If one phase is already difficult to be solved, a combination of phases will be even more difficult to solve.

Furthermore, the complexity arises due to the dependency between the different planning phases (Ceder, 2016; Ceder & Wilson, 1986). Since every phase in the planning process requires data from the previous phase, every phase assumes that the input is the only possible truth and bases its optimal planning on this input. However, it might be possible that minor changes in the phase before would lead to higher efficiency in the phases thereafter. This aspect has its origin in the optimization problem as explained in Section 4.1.1 and is not taken into account when optimizing each step separately.

4.2 Practice-based Perspective

Complex planning phases have taken place before an actual public transport vehicle or vessel shows up at a stop or station to transport passengers from point A to point B. This entire process is explained in Section 4.1, which is based on literature in the field of operations research. For this research, a commonly accepted planning process from both literature and practice should be found, because the applicability of the final design at PTOs should be ensured. Consequently, next to the literature study, empirical evidence is gathered through semi-structured interviews to ensure the completeness of this study.

The interview method is explained before in Section 2.3.2 and the interview guide used for the interviews can be found in Appendix F. In this section, an overview of the actual interviews is given and the interview results are presented. The summaries of the interviews can be found in Appendix J.

4.2.1 Interview Overview

In this phase of the research, a total of three interviews took place with two respondents. These two respondents are both GVB employees and their job responsibilities are related to the planning process of the PTO. The first respondent (R1) is a business analyst for GVB's planning process. Since after the first interview session with R1 still some questions were unanswered because of the vast amount of data collected, a second interview was planned with this respondent. The second respondent (R2) is working within the agile product team 'Planning' as product owner. Both respondents are involved in the implementation of a new software package for the planning process. Hence, both respondents are aware of both the current planning process as well as the improvement to be made for the new planning process. An overview of the interview respondents can be found in Table 4-1.

Table 4-1 Overview of interview respondents

Respondent	**Function**	**Duration** (min)
R1	Business Analyst	61 + 61
R2	Product Owner Team Planning	75

Interviewing more respondents in this phase of the research was considered to be unnecessary, since the broader view of data integration within public transportation is considered in RQ1 (Chapter 3). This includes interviews with a broader sample of respondents. Furthermore, interviewing a broader sample would include interviews with the actual planning personnel. This was expected to result in an overload of detailed information (see Section 2.3.2), which is unnecessary for this research.

4.2.2 Interview Results

The results of the interviews are grouped by subject and research question, as explained in the method in Section 2.3.2. The results related to the planning process are presented in the next subsections.

4.2.2.1 Planning Process

In GVB's documentation, the public transport planning process is defined as shown in Figure 4-6. As can be seen, the planning process starts with the *Transport plan* (with the input of the *Transport vision*) and goes via *Schedule*, *Vehicle blocks* and *Crew duties* to *Operation*.

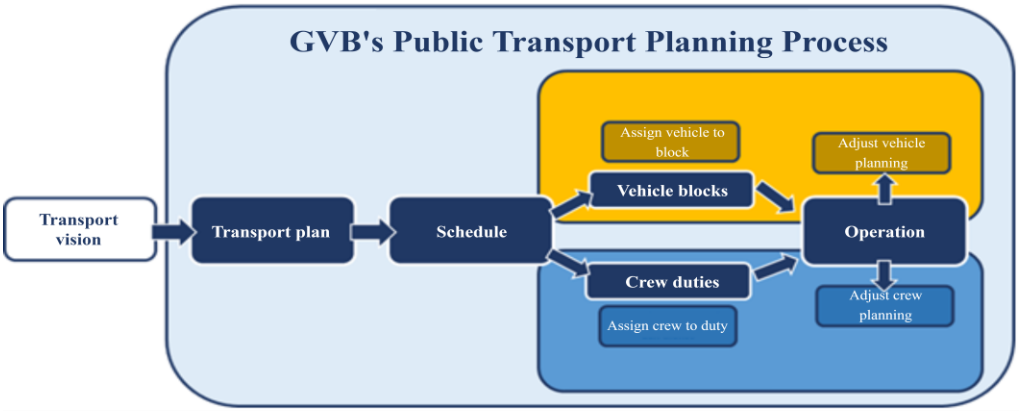

Figure 4-6 GVB's Public Transport Planning Process (GVB Holding NV, 2020c)

During the interviews, it turned out that the high-level overview as presented in Figure 4-6 is not entirely consistent with the actual planning process. For this reason, the process phases identified during the interviews are described hereafter and shown in Figure 4-7. Every phase of the planning process is shortly explained and the relationship with Figure 4-6 is given. The interview respondents agreed, independently from each other, on the adjusted high-level planning process phases.

Both respondents start with a phase in which the routes, lines and stops of the public transport network are defined. This phase is called the conceptual phase and definition phase by R1 and R2, respectively. The end of this phase results in a contract which the PTO needs to adhere to. This phase is similar to *Transport plan* in Figure 4-6 and is called *Determine transport concept* in Figure 4-7.

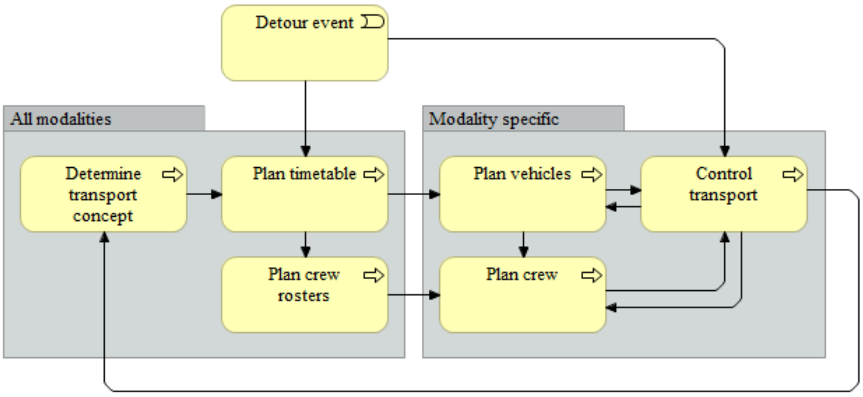

Figure 4-7 Planning process according to interview respondents

The next sequential phase is the one in which the actual timetable is formed. This phase is based on the conceptual phase, in which the basics such as routes and lines are determined. In the timetabling phase, the exact departure, driving and arrival times of every journey is planned (R1, R2). In Figure 4-7, this phase is called *Plan timetable* and is similar to *Schedule* in Figure 4-6.

After the timetable is determined, both vehicles and crew can be scheduled. Most often crew schedules are based on the vehicle blocks (R1), which thus are the input for the crew duties (R2). This relationship between vehicle scheduling and crew scheduling is not present in Figure 4-6 and was also questioned by one respondent. However, during the interviews, it became apparent that there exists an order between these two phases. Hence, the relationship is included in Figure 4-7. At the time of conducting the interviews, vehicles are scheduled on vehicle type-level, whereas crew is scheduled on the exact crewmember level. The level of vehicle planning will change in the future, since new software enables GVB to also schedule on specific vehicle-level, taking into account depot management and vehicle maintenance (R1, R2). Both planning tasks are incorporated in the planning process overview as *Plan vehicles* and *Plan crew*.

Another important planning phase is the planning of crew rosters. These rosters need to be determined in advance, such that crewmembers know when they have to work and what kind of duty it is about (early, day, evening, night). This needs to be known at least 9 months in advance (at GVB, this is a company-specific rule). Since these rosters need to be known far in advance, they cannot be based on the actual crew duties. Therefore, this planning phase follows after *Plan timetable* and is based on the expected timetable in the future (R1, R2). The phase is called *Plan crew rosters* in Figure 4-7.

After the vehicles and crew are planned, public transport can be commenced. The public transport operation is controlled by the control center and is called *Control transport*.

Detour events need to be taken into account within the planning process. Detours can be scheduled (proactive detour management (R1)), but can also happen unexpectedly (reactive detour management (R1)). In the first case, the detour is managed through the entire planning process and schedules are adjusted if necessary. In the latter case, detours are not managed beforehand and need to be managed real-time by the control center. Scenarios exist for unexpected situations which happen more often. The provision of these scenarios is the responsibility of the planning process.

According to R2, the planning process within public transportation can be seen as a continuous process. When public transport is commenced, it is always changed based on the findings during the operation. For this reason, the last phase (*Control transport*) is connected with the very first phase (*Determine transport concept*) allowing a feedback loop.

4.2.2.2 Future Changes in the Planning Process

Both respondents stated that the planning process is going to change soon, mostly due to the implementation of a new software suite for the planning process. This new software offers new possibilities to plan more efficiently and real-time. This also has an impact on the planning process. Even though the high-level planning process as shown in Figure 4-7 remains the same, the content of the phases changes slightly.

The first future impact on the planning process comes from the integral planning software. It ensures that optimization is not only done per phase, but also between phases. This reduces the impact of the problem of sub-optimization as explained in Section 4.1.1.

Two other impactful future changes are regarding depot management and vehicle maintenance. The first can be included in the scope of the planning process and determines which specific vehicle will drive according to which vehicle block and which way the vehicles should be parked to be ready for the next vehicle block (most often the next day). Important input for this phase is the vehicle maintenance planning. When combining these two aspects and incorporating them in the planning process, vehicles can be planned more specifically, predictive maintenance can be carried out at the best moment and vehicles will have a higher availability for operation.

4.3 Proposed Planning Process

The results from both the literature study and the interview sessions are combined into one best-practice planning process. This combination ensures that insights from both qualitative data sources (literature and semi-structured interviews) are taken into consideration in the design of the planning process and that consistency between the two sources is tested.

As explained in Section 4.1, the different phases of the planning process are most often solved individually, which means that the planning process is optimized per phase. This does not necessarily mean that the overall planning is the most optimal. Research has proven and suggested that it is more beneficial to integrate (some of) the planning phases (Canca et al., 2016; David et al., 2018; Desaulniers & Hickman, 2007; Huisman, 2016; Lampkin & Saalmans, 1967; Nagy & Tick, 2019; Shen et al., 2016; Weider, 2007). However, to propose a clear and detailed enough data integration design, one should have a clear planning process to identify data requirements per phase. This does not mean that the algorithm in the planning software should also handle these phases separately. Hence, the planning process proposed in this research does not prescribe to handle each phase separately, yet it shows a holistic view of the process and its phases. In Chapter 5, these phases are extended with data requirements. In algorithms, phases and data can (and are suggested to) be combined, e.g. as proposed by Nagy & Tick (2019) for the integrated vehicle and crew scheduling problem.

The combination of the literature and the interviews led to several minor changes in the planning process as defined in the literature. All proposed changes can be found in Table 4-2 and an explanation per proposed change is given after, which starts with providing the proposed planning process.

Table 4-2 Triangulation of planning phases

In literature (Section 4.1)	Proposed phase	Practice-based (Section 4.2)
Design network	*Out of scope for this research*	N/A
Plan lines	Plan lines	Determine transport concept
Develop timetable	Plan timetable	Plan timetable
Schedule vehicles	Schedule vehicle blocks	Plan vehicles
	Assign vehicles	N/A
Schedule crew	Plan crew rosters	Plan crew rosters
	Schedule crew duties	Plan crew
	Assign crew	
Manage transport perturbations	*Out of scope for this research*	Control transport

The above-mentioned changes and the changes stated in Table 4-2 are incorporated in a new overview of the planning process, which can be found in Figure 4-8. Business events are included in the figure to show where the process starts when business events take place. These business events are based on the interview results as presented in Section 4.2.

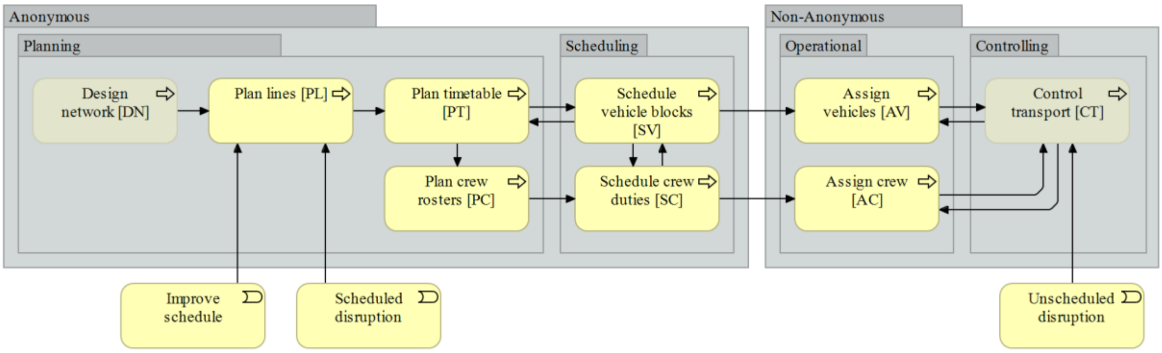

Figure 4-8 Proposed planning process

The naming of the stages of the planning process is changed from *strategical*, *tactical* and *operational* (as defined in literature) into *planning*, *scheduling* and *operational*, respectively. This is done to be more precise on the responsibilities for each high-level stage and to better suit the goal of this research. The definitions from the Cambridge dictionary (Cambridge University Press, n.d.-b) are taken:

To plan	"to decide what you are going to do or how you are going to do something"
To schedule	"to plan an event for a particular day or time"
To operate	"if a machine operates, it works, and if you operate it, you make it work"

Planning public transportation is about finding the correct relationships between three factors: journeys, crew members and vehicles. This triangle can be seen in Figure 4-9. A combination of the three aspects is only known close to the operational day. Qiu et al. (2018) call this combination the basic data of a PTO. To stress the difference in time in which this combination gets known, two high-level stages are introduced in the planning process overview as well: *anonymous* and *non-anonymous*.

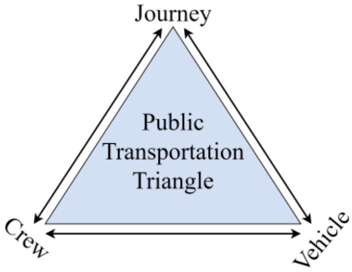

Figure 4-9 Public transportation triangle

Every phase in the *anonymous* stage of the planning is not making any combination of the three factors explained above, i.e. the phases do not combine journeys, crew members and vehicles. On the contrary, in the *non-anonymous* stage of the planning process, actual combinations between journeys, crew members and vehicles are made. Scholz (2016) also emphasizes the difference on this higher level and identifies the stage as either being connected to a physical object (vehicle or crew member; *non-anonymous*) or not (*anonymous*).

It was decided not to include the difference between modality-wide planning and modality-specific planning, while this is a very company-specific choice. Moreover, this split between responsibilities does not change the planning process itself and does not provide the process with a clear distinction such as *(non-)anonymous*. An explanation per planning phases is presented in the next subsections.

4.3.1 Design Network

Design network is excluded from the scope of this research. This was decided because of several reasons. First of all, the respondents of the interviews did not consider these phases at all. Furthermore, this phase is dependent on a lot of negotiations between different parties, such as the PTO, the transport authority, the government and possibly even road authorities and construction companies.

Moreover, these negotiations most often continue for many years before a decision can be made. An example of a project in this phase is the Noord/Zuidlijn in Amsterdam. The first proposal for this new metro line and its infrastructure was done in 1964, the start was in 2002 and the delivery was only in 2018, which means a throughput time of more than 50 years (Claus, 2018). Another example is the development of a new neighborhood. This normally takes years, in which also public transport possibilities are considered.

Based on the interview, the complex negotiations, the different parties involved and the throughput time of this phase it is justified to exclude this phase for the data integration design approach for a PTO. Yet, it can be seen as important input for the *Plan lines* phase.

4.3.2 Plan Lines

The phase *Plan lines* is a combination of phases identified both in literature and during the interviews. It is about developing a plan of routes, lines, stops and more specifications for the public transport network. According to literature and interviews, more than one modality is possibly combined in this plan, to ensure consistency, transfers and coverage. In this plan also requirements from transport authorities are accounted for.

4.3.3 Plan Timetable

Having the public transport concept in place, the actual timetable can be designed in *Plan timetable*. This phase is responsible for the development of actual journeys, taking into account the network consistency, transfers and coverage, which are a result of the previous phase *Plan lines*. At the end of the timetable phase, every line consists of journeys. This forms the input for both the vehicle and crew scheduling phases.

4.3.4 Plan Crew Rosters

The phase in which crew rosters are set up is not covered individually in literature, yet it is seen as part of the crew scheduling in general. However, as became apparent during the interview sessions, the crew rostering cannot be part of the crew scheduling, since the crew rosters need to be defined much more in advance. This is because of company-specific rules (see Section 4.2.2) and Dutch labor agreements in the field of public transportation (*CAO Multimodaal*, 2019; *CAO Openbaar Vervoer*, 2020). In literature, crew duties are mapped on crew rosters. This also implies that the crew roster needs to be known prior to the crew duties. Moreover, Békési (2009) also splits the roster and duty phases.

For the above-mentioned reasons, *Plan crew rosters* is incorporated as a separate phase in the planning process. After validation, it turned out that this planning phase is recognized by all experts, which also substantiates the choice to include it as a separate phase.

4.3.5 Schedule Vehicle Blocks

In literature, often no clear distinction exists between vehicle scheduling and vehicle assigning (see Section 4.1). A split between the scheduling and assigning of vehicles is proposed, since these two tasks take place within different departments and require different data inputs. Furthermore, in the interviews it turned out that depot management and predictive maintenance will be applied to the planning process soon (Section 4.2), which also substantiates the split of these two phases. After validation, it turned out that both of the phases are recognized by the experts and the split was understood. In *Schedule vehicle blocks*, the goal is to develop vehicle blocks based on a timetable. In *Assign vehicles* the actual vehicles will be assigned to vehicle blocks, in which maintenance and other data can be taken into account. This planning phase can also trigger changes in the timetable in case this would be beneficial for the vehicle blocks.

4.3.6 Schedule Crew Duties

Similar to *Schedule vehicle blocks*, also for *Schedule crew duties*, in literature, no distinction exists between crew duty scheduling and crew assigning. A split between these two phases is made, since the scheduling of crew duties is not connected to a specific crewmember in any way. Furthermore, by splitting the phase into two phases, it is more consistent with the vehicles-phases as can be seen in Figure 4-8. This means that in *Schedule crew duties* the duties for crewmembers are made, but not connected to any specific crewmember yet.

4.3.7 Assign Vehicles

In *Assign vehicles* the actual vehicles are assigned to the vehicle blocks (generated in *Schedule vehicle blocks*). The assignment of vehicles should take into account vehicle types and data regarding vehicle maintenance. This phase also covers the functionalities of depot management, as was mentioned in both literature and interviews.

4.3.8 Assign Crew

In *Assign crew* the duties are mapped to crew rosters. Since crew rosters are made prior to the crew duties, and thus prior to the moment you know exactly how many crewmembers you need, this process always results in either crewmembers without a duty or duties without a crewmember. In the first case, the PTO still needs to pay the crewmember. In the latter case, the PTO needs to find extra crew members (possibly via employment agencies).

4.3.9 Control Transport

Control transport is responsible for controlling and managing public transport operations during operations. Whatever happens during operations is managed by the transport control center. This controlling task is not seen as part of the planning process by most of the studied literature and both interview respondents and is therefore excluded from the scope of this research. However, control transport uses data generated in the planning process and also triggers planning process phases whenever incidents happen (e.g. crew or vehicle disruptions). These data dependencies and triggers are accounted for in this research. Hence, the planning process is closely connected to the controlling process, which is the reason that the phase is incorporated in Figure 4-8, but the phase itself is not part of the scope of the planning process in this research.

4.3.10 Process Triggers

In Figure 4-8, three different business events are shown that can trigger the planning processes. These business events are identified in literature and during interviews. The explanation of these business events can be found next.

4.3.10.1 Improve Schedule

PTOs improve their schedule based on experiences from the past. Within the Netherlands, these bigger schedule changes happen typically once a year, in the second week of December. The improvement of the schedule starts at the designing phase, in which the lines are planned.

4.3.10.2 Scheduled Disruption

A scheduled disruption is a disruption that is communicated to the PTO in advance. In this way, the PTO can adapt the planning in case this is necessary. Adaptations can be made based on several levels. Lines, routes, schedules, vehicle blocks and duties can be changed to be able to offer public transport during the disruption.

4.3.10.3 Unscheduled Disruption

Unscheduled disruptions are not planned and happen during operations. The control center manages these disruptions and can trigger the planning process for crew and vehicles in case this is necessary (i.e. when crew and vehicle disruptions happen). Scenarios developed in the planning process can be used to minimize the passenger inconvenience and costs for handling the disruptions.

4.4 Innovations in the Planning Process

During the assessment of the literature and interviews, new processes and technologies for the planning process which are highly dependent on data integration became apparent. These future changes are innovations for the planning process and are likely to be implemented soon. For this reason, they are requirements for the data integration design approach.

Firstly, we present a short overview of these future improvements in table-form, as can be found in Table 4-3. In this overview, the improvements are listed including the adjacent business areas and/or external sources that are necessary to cooperate with to implement the innovations. Thereafter, every future improvement is shortly described individually.

Table 4-3 Future improvements for the planning process

Innovation	**Intervening business area**
Better alignment between planning and maintenance	Vehicle management & maintenance
Support depot management solutions	Vehicle management & maintenance
	Operations
Scheduling and monitoring battery-equipped vehicles	Vehicle management & maintenance
	External (e.g. weather)
Self-rostering for crew members	HR
Integration of planning process phases	Within the planning process
Support for continuous planning	Within the planning process

4.4.1 Planning and Maintenance Alignment

To decrease costs for maintenance, have a more uniform mileage and decrease vehicle failure, it is important to better align the public transport and maintenance planning. This was endorsed by literature (Scholz, 2016) and both interview respondents. Alignment of both planning activities requires communication between the transport planning and maintenance planning.

Requirement for the data integration design approach: the vehicle forecast and actual vehicle planning should be easily accessible by the maintenance department of the PTO. Also, the maintenance planning should be considered as input for planning and assigning vehicles.

4.4.2 Depot Management

Depot management is about all activities which take place in the depots. One can think of parking vehicles, combining vehicles, carrying out small maintenance tasks, refueling and washing. This is a separate optimization problem, which takes into consideration the necessary number of vehicles for the operation, but also maintenance planning (see Section 4.4.1) and other tasks such as instruction rides. Moreover, most often this task ensures the vehicle distribution to crew members. Many software companies specialized in public transport planning already provide solutions for this (GIRO, n.d.; INIT, n.d.; IVU, n.d.; PSI, n.d.).

Requirement for the data integration design approach: the design approach should take into consideration the tasks which take place within depot management. Data should be accessible for performing the right planning tasks, but also the right data should be incorporated (e.g. maintenance planning) in the planning.

4.4.3 Battery-equipped Vehicles

Lately, more and more vehicles become electric. Advantages are that electric engines are often quieter, have better acceleration and offer zero-emission transport. However, the disadvantage compared to combustion engines is that battery-equipped vehicles need to be charged several times a day, whereas combustion engines are most often only refueled once a day.

This results in a more complex scheduling problem, since vehicles need to be charged during operation. Charging should not lead to any disruption or delay for the passengers, which means that charging needs to take place before or after normal service journeys. These charging moments should be scheduled carefully, because charging too often results in a higher number of vehicles necessary for offering public transport and thus higher costs. Optimizing scheduling of these vehicles also means that, every day, the outside temperature, driver and battery age should be considered. This is because heating and cooling the inside temperature of the vehicle drains the battery, as well as the driving style of some drivers may do. The battery's age is important because its capacity decreases over time.

Next to the more complex scheduling problem, electric vehicles should be monitored carefully during operation. Since they need to charge more often than a combustion engine vehicle needs to be refueled, the chance of having an empty battery is higher.

Requirement for the data integration design approach: both scheduling and monitoring the electric vehicles require a lot of data. This data should be real-time accessible and planning last-minute should be possible.

4.4.4 Self-rostering for Crew Members

From the interviews it became apparent that the PTO has the goal to offer self-rostering for crew members. This would be beneficial for employees' happiness and health, as is also proven by Garde et al. (2012). The biggest issue for this is that it introduces more dependencies on new data objects, which should be taken into account while planning crew rosters and assigning crew members to duties. Next to this data-aspect, proper software should be acquired or developed to combine crew preferences and crew demand into an optimal crew roster for every crew member.

Requirement for the data integration design approach: self-rostering seems to become an important aspect in the future. The data integration design approach should take into consideration the extra input of desired crew rosters.

4.4.5 Integration of Planning Process Phases

As explained in Section 4.1.1, the optimization of the planning process most often happens at the level of the process phases. However, "suboptimization (...) does not lead to global optimization" (Machel, 1965). For this reason, the planning process phases should be integrated to define the most optimal planning. This is acknowledged in both the literature (Canca et al., 2016; David et al., 2018; Desaulniers & Hickman, 2007; Huisman, 2016; Lampkin & Saalmans, 1967; Nagy & Tick, 2019; Scholz, 2016; Shen et al., 2016; Steinzen, 2007; Weider, 2007) as well as during the interviews. This integration of planning process phases is out of scope for this research, however, the data integration design approach should allow the integration of several or all phases.

Requirement for the data integration design approach: even though the integration of the different processes is out of scope for this research, the design approach should not be based on individual planning process phases. The design should allow the integration of process phases in implementation.

4.4.6 Support Continuous Planning

Offering good public transport means that the public transport offer should be perfectly aligned with the demand. This can only be done when the schedule is tied to the daily situation, which can be very different because of road works, accidents, weather, etc. In other words, the planning process is a continuous process, which should – in the best case – take place daily. By the interview respondents, this phenomenon is called *continuous planning*. This way of planning demands that data and processes are perfectly integrated and that changes in one particular planning process phase are visible in any other phase immediately. Also, downstream dependencies towards other (also external) applications should be updated real-time, since only then the actual public transport can be realized and communicated to the passengers. This underlines the importance of the integration of planning process phases, as explained in Section 4.4.5.

Requirement for the data integration design approach: continuous planning asks for real-time access to all necessary data, and, also the real-time provision of data to all other consuming applications, business areas and external parties.

4.5 Chapter Summary

The best-practice planning process is identified to form a grounded basis for this research. The process was derived from the field of operations research through a systematic literature review and GVB's practice by conducting interviews. Combining these two sources led to a newly proposed planning process. Two planning phases were placed out of scope for this research, since they are not (entirely) carried out by the PTO or are not part of the PTO's planning department. This led to the identification of seven planning phases, as presented in Figure 4-8.

It is expected that the planning process will be improved in the future. These improvements are identified in this chapter and it turns out that they come with an increase in data (integration) requirements. To support the innovations, some extra data requirements have to be taken into account and correct and up-to-date data should be accessible easily and in a real-time manner. These are important inputs for the design of the data integration design approach.

5 DATA REQUIREMENTS

This chapter aims to identify the data requirements for the planning process and therefore answers RQ4, as visualized in Figure 5-1. In line with the method described in Section 2.3, the data requirements are identified on the level of planning tasks of the planning phases identified in Chapter 4. First, the planning tasks are explained (Section 5.1), after which the data requirements per task are presented in Section 5.2. Subsequently, in Section 5.3, all data requirements are consolidated.

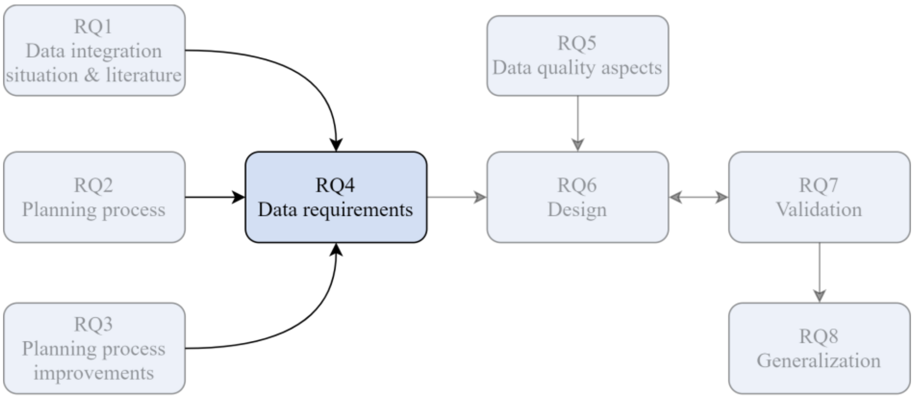

Figure 5-1 Content of chapter 5 (based on Figure 1-5)

5.1 Identification of Planning Tasks

Every phase in the planning process as proposed in Section 4.3 (see also Figure 5-2) consists of several decisions that have to be made and tasks that have to be carried out. These decisions and tasks are carried out based upon the experience of employees in the planning department, algorithms in software, contracts, historical data, present data, data from previous phases in the planning process and, preferably, data generated in the next sequential phase (planning by iterations). In this section, these decisions and tasks are formulated in planning tasks. As explained in Section 2.3, these tasks and their data requirements are identified based on literature, interviews and the analysis of the results on RQ1, according to the triangulation method. Innovations for the planning process as explained in Section 4.4 are taken into account as well.

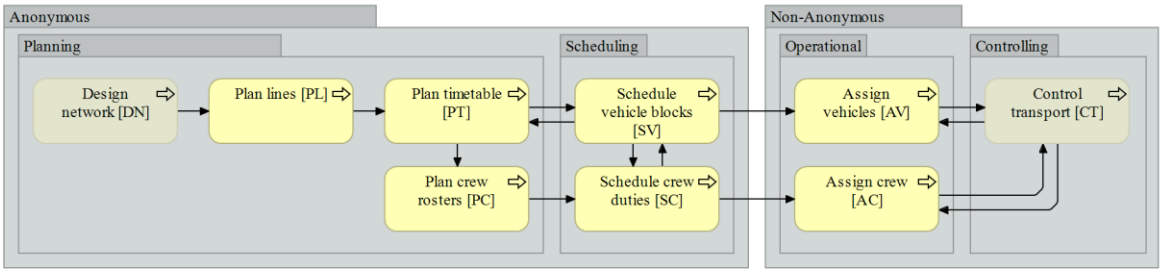

Figure 5-2 Planning process (from Figure 4-8)

An overview of the consulted sources and the number of resulting planning tasks per source can be found in Table 5-1. In total, 214 tasks were derived. More details about the method for identifying the planning tasks and their data requirements can be found in Section 2.3.3.

Table 5-1 Sources and the number of identified planning tasks

Type	Source	Identified planning tasks
Literature *Section 4.1*	Ceder (2016)	31
	Friedrich et al. (2016)	2
	Nagy & Tick (2019)	3
	Scholz (2016)	75
Interviews *Section 4.2*	Business Analyst (GVB)	41
	Product Owner Team Planning (GVB)	40
Results RQ1 *Chapter 3*	Practical information from GVB	22
	Total	**214**

The triangulation resulted in a set of many identical and overlapping planning tasks. For this reason, the list of tasks is brought back to a consolidated list of 23 planning tasks. This consolidation process was an iterative filtering process over the entire list of tasks and is illustrated in Figure 5-3. The total list of tasks was first consolidated based on the task description, followed by their upstream data dependencies. The last step was the consolidation and filter based on the planning responsibility. An example of this last step is that 'informing passengers' is not seen as part of the planning process. Whereas it is evident that passengers should be informed about the outcomes of the planning process, the actual communication towards the passengers is assumed not to be the responsibility of the planning process.

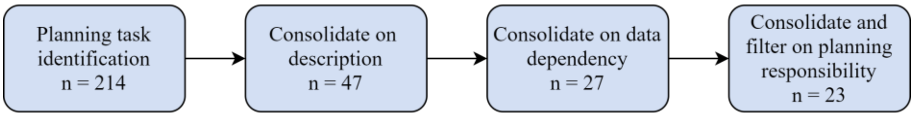
Planning task identification n = 214
Consolidate on description n = 47
Consolidate on data dependency n = 27
Consolidate and filter on planning responsibility n = 23

Figure 5-3 Process of planning task consolidation

Table 5-2 lists the planning tasks after the consolidation process was finished. Every task is mapped to the previously identified planning process phases (in Chapter 4 and visualized in Figure 5-2). The original tasks are referred to as 'Reference' and refer to the planning task IDs in the total list of tasks as can be found in Appendix K. From the total set of 214 planning tasks, 13 tasks were left out. This is done because these tasks represent the tasks of the operational control (task no. 60, 65, 67, 68, 95, 189) or are high-level goals for improving the planning process (task no. 117, 118, 120, 132, 164, 165, 192). For the downstream dependencies, only the dependencies outside the scope of the planning process in this research are included in the table, since there are too many downstream dependencies towards other planning process tasks. These dependencies between planning tasks are still taken into account for this research but are presented in Section 5.2.

The overlap of the planning tasks within the different sources is illustrated in Figure 5-4. As can be seen, most of the tasks are retrieved from literature. Most of them were also identified during the interviews and in the analysis of GVB's situation for RQ1. The total set also contains planning tasks that were only identified in either literature, interviews or RQ1. These were considered to be important and are therefore incorporated. The process underlines the success of triangulation, while a possible bias from using only one source and method is mitigated.

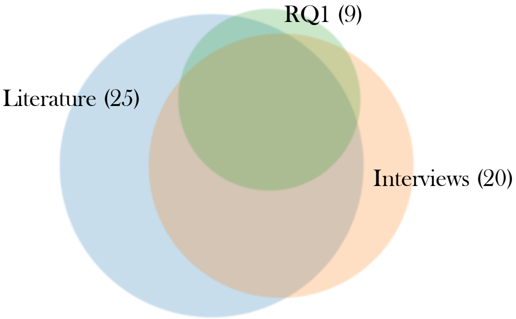

Figure 5-4 Planning task sources and their overlap

Table 5-2 Consolidated list of planning tasks within the planning process

Planning phase	ID	Planning task	Reference (Appendix K)	
			Original planning task	Dependency on the planning task
Plan lines	PL1	Create route network	1, 2, 3, 4, 5, 6, 10, 11, 12, 69, 70, 71, 72, 75, 76, 83, 93, 94, 121, 134, 141, 142, 143, 174, 175, 185	127, 155, 160, 173, 178, 179, 181, 193, 196, 197, 198, 199, 200, 201, 209, 210, 211
	PL2	Determine relief points	8	
	PL3	Provide fare information	7, 115, 146, 197, 213	160, 161, 180, 197, 198, 213
	PL4	Develop scenarios	64, 166	57, 58, 59, 62, 63, 123, 162
	PL5	Plan detours	112, 122, 145, 146, 163, 203	43, 173, 178, 210
Plan timetable	PT1	Determine timetables	9, 13, 77, 84, 87, 88, 96, 97, 135, 182, 175, 186, 193	55, 56, 61, 66, 115, 116, 139, 140, 115, 157, 158, 160, 173, 176, 177, 178, 179, 181, 196, 197, 198, 199, 200, 201, 202, 208, 209, 210, 211, 212
	PT2	Adjust/improve timetables	14, 73, 85, 159	
	PT3	Design frequency changes	64, 78, 86	57, 58, 59, 63, 162
Plan crew rosters	PC1	Create roster layouts	15, 53, 82, 98, 128, 144, 204	127, 173, 208
	PC2	Assign crew to roster layouts	16, 204	127, 173, 208
	PC3	Plan training and holidays	17, 18, 102, 128	127, 173, 214
	PC4	Arrange self-rostering	133, 170	
Schedule vehicle blocks	SV1	Determine layover times	19	
	SV2	Plan dead runs	21, 89	
	SV3	Create vehicle blocks	20, 22, 74, 79, 80, 89, 90, 106, 107, 136, 147, 148, 167, 175, 183, 187, 190, 205	116, 173, 178, 208, 210
Schedule crew duties	SC1	Create crew duties	23, 53, 81, 91, 99, 129, 138, 151, 152, 171, 175, 184, 188, 191, 194, 204	127, 208
	SC2	Assign crew duties to roster	44, 92	
Assign vehicles	AV1	Assign vehicle to block	24, 26, 27, 30, 31, 32, 33, 34, 35, 36, 37, 110, 111, 137, 149, 168, 169, 207	25, 29, 37, 116, 155, 168, 173, 178, 208, 210
	AV2	Manage vehicle disruptions	49, 51, 54, 108, 113, 119, 156	178, 210
	AV3	Plan vehicle parking	27, 28, 109, 126, 150	
Assign crew	AC1	Assign crew to duties	39, 41, 42, 43, 45, 46, 52, 53, 100, 130, 195, 206	114, 115, 125, 127, 131, 155, 173, 208, 209, 214
	AC2	Manage crew disruptions	38, 40, 50, 51, 54, 103, 104, 105, 113, 119, 124, 154, 156	114, 127, 131, 214
	AC3	Manage leave, duty swap and requests	47, 48, 101, 130, 153, 172	114, 127, 131, 173, 214

Legend: [PL] = Plan lines; [PT] = Plan timetable; [PC] = Plan crew roster; [SV] = Schedule vehicle bocks; [SC] = Schedule crew duties; [AV] = Assign vehicles; [AC] = Assign crew

5.2 Data Requirements per Planning Task

After identifying the planning tasks as presented in Section 5.1, we have derived data requirements to fulfill these tasks. First, in Section 5.2.1, reporting purposes on data objects used within the planning process are explained. From Section 5.2.2 onwards, every planning phase is explained by its planning tasks and data requirements. These data requirements are grouped into four different categories: upstream non-planning, upstream planning, up- and downstream and downstream data objects. The meaning of the latter two is evident. The difference between the first two is that the first category is about data objects not coming from the PTO's planning process, whereas the second category is about data objects originating from the PTO's planning process. A general description of the planning phases is not provided in this section, since these are already presented in Section 4.3.

Whereas in this section the planning tasks are shortly described and their data requirements are addressed, the in-depth planning task data requirements analysis can be found in Appendix L. In line with the method as explained in Section 2.3.3, the SIPOC (Suppliers, Inputs, Processes, Outputs and Consumers) method is used. The planning tasks are seen as the process ('P' in SIPOC) in the middle of every table in Appendix L, for which the suppliers' inputs and the outputs for consumers are identified. This leads to the identification of upstream and downstream data requirements, which are used for the data requirements on planning phase level as presented in this section. The definition of every data object can be found in Appendix M.

Upstream and downstream dependencies might lead from and to the *Design network* and *Control transport* planning phase. However, as explained before, these are not part of the scope of this research. For this reason, the data objects coming from these phases are not identified as such, but referred to as 'PTO' and 'TA' (transport authority) as supplier and 'Operations' as consumer in Appendix L. It is also important to note that some planning tasks have both upstream and downstream dependencies for the same data objects. This is possible when planning tasks are adding more information to a data object (data enrichment) or are simply changing the data object.

5.2.1 Reporting

Reporting on data objects used and generated within the planning process turned out to be an important task for the PTO (based on planning tasks 29, 37, 95, 115, 121, 122, 142, 143, 157, 158, 159, 166, 177, 178, 181, 200 and 209 in Appendix K). This reporting function is demanded at almost every step in the planning phase. Reporting enables the PTO to check their efficiency, costs and compliance, but is also – very often – demanded by the TA. The TA demands these reporting facilities to check the compliance of the PTO concerning the contract. Later, in Section 5.3, the reporting purposes on the level of data objects is presented. Since reporting is an important factor, it also stresses the importance of a good data integration situation. Only with correct and real-time data, reporting purposes can create added value and the business can become more data-driven.

5.2.2 Plan Lines

During the identification of planning tasks for *Plan lines [PL]*, a total of 30 tasks were found. These are consolidated into five planning tasks, which are presented in Figure 5-5 and explained thereafter.

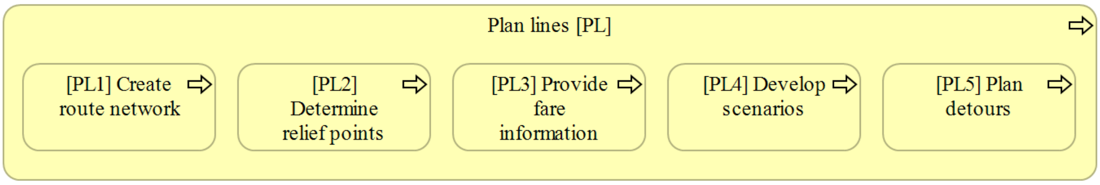
Plan lines [PL]
[PL1] Create route network
[PL2] Determine relief points
[PL3] Provide fare information
[PL4] Develop scenarios
[PL5] Plan detours

Figure 5-5 Planning tasks in Plan lines [PL]

[PL1] Create route network is the most important planning task of this planning phase. It creates the route network and includes the infrastructure, routes, lines and line plan and information such as headways, first/last journeys, frequency, vehicle types and many more. This is the base of the public transport planning and is distributed to many consumers. It accounts for vehicle/crew supply and available physical places at depots, stations and sidings.

[PL2] Determine relief points determines the relief points ("points where a relief is possible, i.e. a driver may take on or hand over a vehicle" (Transmodel, 2019)) within the public transport network. The relief points should cover many lines, as that increases the planning efficiency (more relief possibilities). Hence, this planning task is based on the route network as defined in *[PL1]*.

[PL3] Provide fare information provides fare information for passengers. This information contains prices, plans, subscriptions, etc. The input for this task might originate from the government, TA and the PTO's commerce/sales department.

[PL4] Develop scenarios makes scenarios that can be used by the control center. Scenarios are developed beforehand to save time during the operation in case unexpected situations occur and are based on history and experience. Examples are metro/tram-replacement buses and detours for bridge/tunnel closures.

[PL5] Plan detours is about the planning of changes to the normal schedule. Reasons are for example road/track/water works, station/stop works, events and bridge openings. This planning task might have consequences for the original planning as defined in *[PL1]*, and thus also for every following planning task.

Figure 5-6 shows all data requirements for the planning tasks as explained above. The in-depth analysis which led to these data requirements can be found in Appendix L.

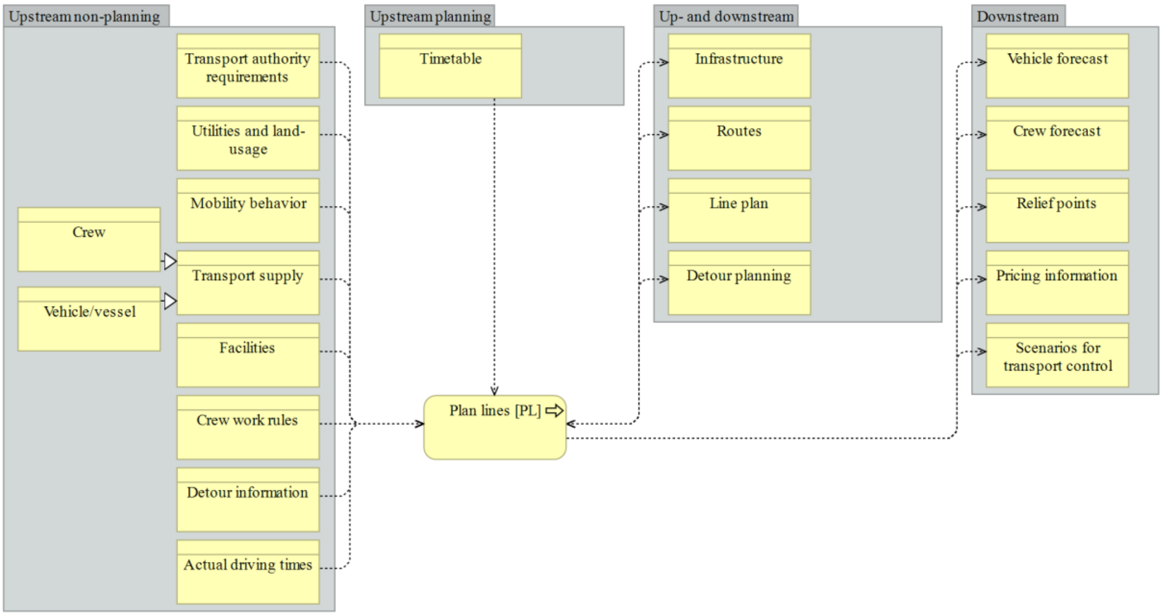

Figure 5-6 Data dependencies Plan lines [PL]

5.2.3 Plan Timetable

During the identification of planning tasks for *Plan timetable [PT]*, a total of 18 tasks were found. These are consolidated into three planning tasks, which are presented in Figure 5-7 and explained thereafter.

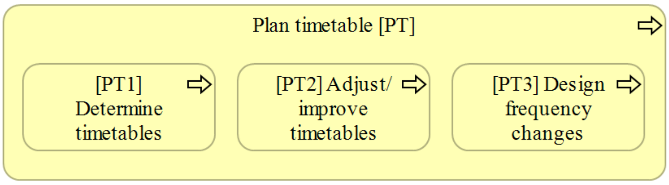
Plan timetable [PT]
[PT1] Determine timetables
[PT2] Adjust/ improve timetables
[PT3] Design frequency changes

Figure 5-7 Planning tasks in Plan timetable [PT]

[PT1] Determine timetables can be seen as the most important planning task within this phase. The actual planning from the previous phase (*Plan lines [PL]*) is transformed into a timetable that includes all individual journeys. In the timetable planning, the frequencies, first/last specifications, headway determination, connections and other specifications from *[PL1]* are accounted for.

[PT2] Adjust/improve timetables minor adjustments and improvements take place often. These can be based on runtime analysis (check the actual driving time and adjust), events (adjust frequencies/vehicle types) and other circumstances. It changes the original timetable from *[PT1]*.

[PT3] Design frequency changes is related to *[PL4]*. For frequent changes and disruptions during operations, frequency changes should be present. These can be used to adjust the frequencies of existing lines to better match the transport demand.

Figure 5-8 shows all data requirements for the planning tasks as explained above. The in-depth analysis which led to these data requirements can be found in Appendix L.

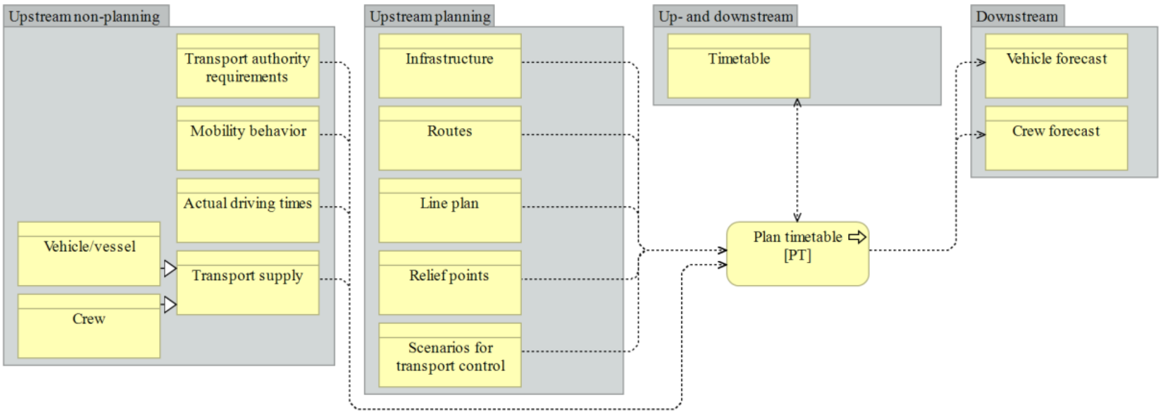

Figure 5-8 Data dependencies Plan timetable [PT]

5.2.4 Plan Crew Rosters

During the identification of planning tasks for *Plan crew rosters [PC]*, a total of 11 tasks were found. These are consolidated into four unique planning tasks, which are presented in Figure 5-9 and explained thereafter.

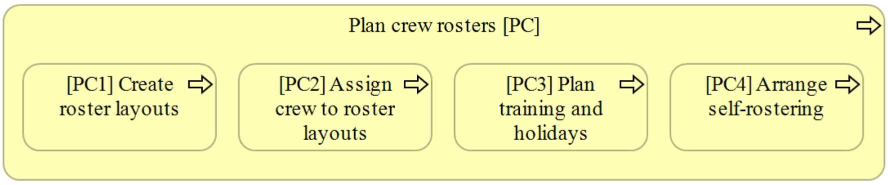
Plan crew rosters [PC]
[PC1] Create roster layouts
[PC2] Assign crew to roster layouts
[PC3] Plan training and holidays
[PC4] Arrange self-rostering

Figure 5-9 Planning tasks in Plan crew rosters [PC]

[PC1] Create roster layouts creates the high-level rosters for crew members. These rosters contain information about the type of duty (early, day, late, night, spare, stand-by, etc.) and do not contain any details about a specific duty. Rosters are based on historic figures, HR data and the pool (number and contract hours) of employees.

[PC2] Assign crew to roster lay-outs assigns crew members to the different roster layouts. This process takes into account duty types, contract hours and any other reasons/agreements between the PTO and their crew members.

[PC3] Plan training and holidays training and holidays are often planned in the long-term. For this reason, this task is situated in this planning phase. This also enables a proper crew forecast.

[PC4] Arrange self-rostering offers crew members the possibility to roster themselves (see Section 4.4.4). By offering this possibility, crew members can request a desired crew roster.

Figure 5-10 shows all data requirements for the planning tasks as explained above. The in-depth analysis which led to these data requirements can be found in Appendix L.

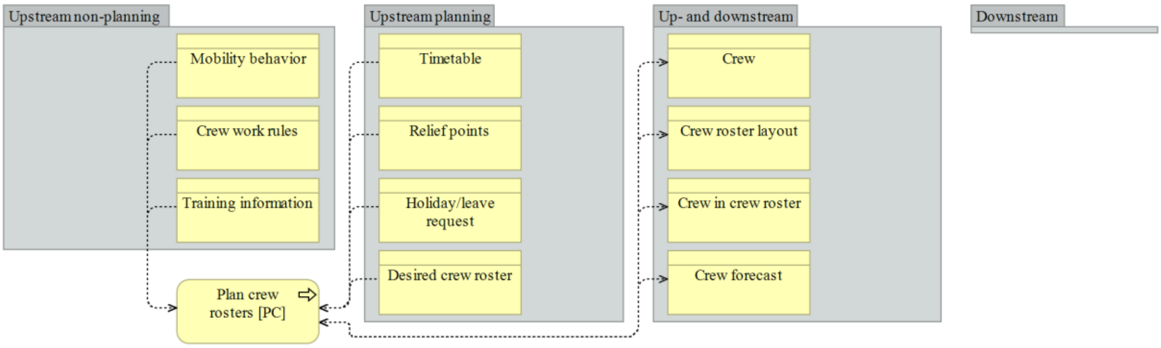

Figure 5-10 Data dependencies Plan crew rosters [PC]

5.2.5 Schedule Vehicle Blocks

During the identification of planning tasks for *Schedule vehicle blocks [SV]*, a total of 18 tasks were found. These are consolidated into three unique planning tasks, which are presented in Figure 5-11 and explained thereafter.

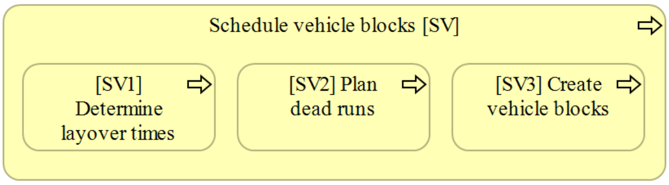
Schedule vehicle blocks [SV]
[SV1] Determine layover times
[SV2] Plan dead runs
[SV3] Create vehicle blocks

Figure 5-11 Planning tasks in Schedule vehicle blocks [SV]

[SV1] Determine layover times determines how much time should be allocated to offer a stable timetable. Reserving layover times ensures – among other things – that delays from journey *x* are not passed through to journey *x + 1*.

[SV2] Plan dead runs plans journeys without passengers, so-called dead runs. These are necessary to offer public transport and – for example – reach from the depot to the starting point, but also from an ending point to (another) starting point. For rail equipment this is called shunting.

[SV3] Create vehicle blocks creates vehicle blocks based on the timetable, layover times and dead runs. These vehicle blocks contain all information for what vehicle *x* has to do on a specific day. Note: no specific vehicle is assigned yet. This takes place in *Assign vehicles [AV]*.

Figure 5-12 shows all data requirements for the planning tasks as explained above. The in-depth analysis which led to these data requirements can be found in Appendix L.

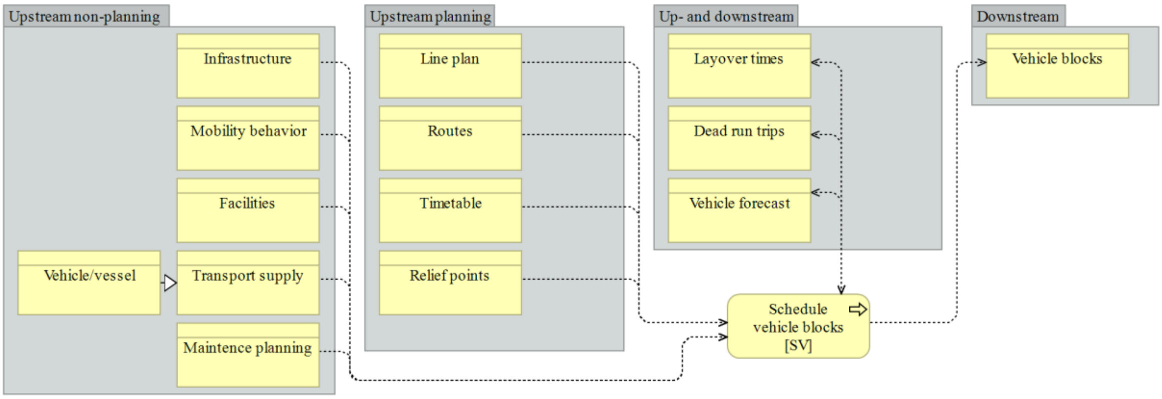

Figure 5-12 Data dependencies Schedule vehicle blocks [SV]

5.2.6 Schedule Crew Duties

During the identification of planning tasks for *Schedule crew duties [SC]*, a total of 15 tasks were found. These are consolidated into two unique planning tasks, which are presented in Figure 5-13 and explained thereafter.

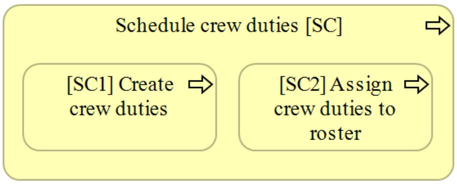

Figure 5-13 Planning tasks in Schedule crew duties [SC]

[SC1] Create crew duties creates the crew duties. These duties contain journeys, dead runs, personal time (often according to law and/or company-specific regulations), break-time and any other activity necessary for offering public transport services.

[SC2] Assign crew duties to roster assigns the duties to the roster layout (and thus duties to crew members, since crew members are assigned to a roster layout). The optimal solution should be found to match all crew duties to the specific rosters, taking into account the difference in duty length, depots and their specific crew supply.

Figure 5-14 shows all data requirements for the planning tasks as explained above. The in-depth analysis which led to these data requirements can be found in Appendix L.

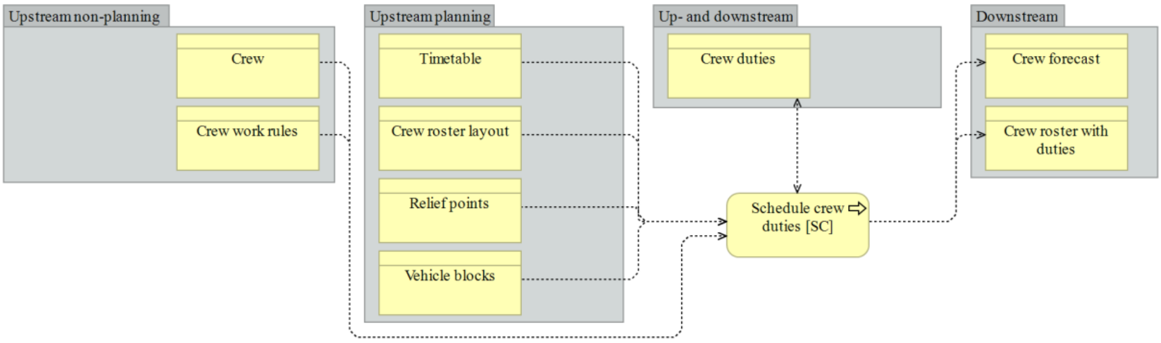

Figure 5-14 Data dependencies Schedule crew duties [SC]

5.2.7 Assign Vehicles

During the identification of planning tasks for *Assign vehicles [AV]*, a total of 25 tasks were found. These are consolidated into three unique planning tasks, which are presented in Figure 5-15 and explained thereafter.

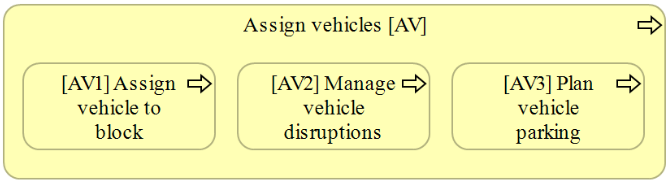
Assign vehicles [AV]
[AV1] Assign vehicle to block
[AV2] Manage vehicle disruptions
[AV3] Plan vehicle parking

Figure 5-15 Planning tasks in Assign vehicles [AV]

[AV1] Assign vehicle to block assigns a specific vehicle to a specific vehicle block for a particular day. This task takes into account vehicle type restrictions for the vehicle blocks and, preferably, maintenance planning and any other vehicle-demanding activity (e.g. crew training).

[AV2] Manage vehicle disruptions manages disruptions in case something happens with a vehicle right before or during (triggered by the control center) operation. This might lead to a vehicle change and, in the worst-case scenario, canceled journeys. At some PTOs this planning task is carried out under the responsibility of the control center.

[AV3] Plan vehicle parking accounts for depot management solutions (Section 4.4.2) and ensures that depot activities are running smoothly. The main task is to ensure that vehicles are parked in the right way to easily start the next operational day.

Figure 5-16 shows all data requirements for the planning tasks as explained above. The in-depth analysis which led to these data requirements can be found in Appendix L.

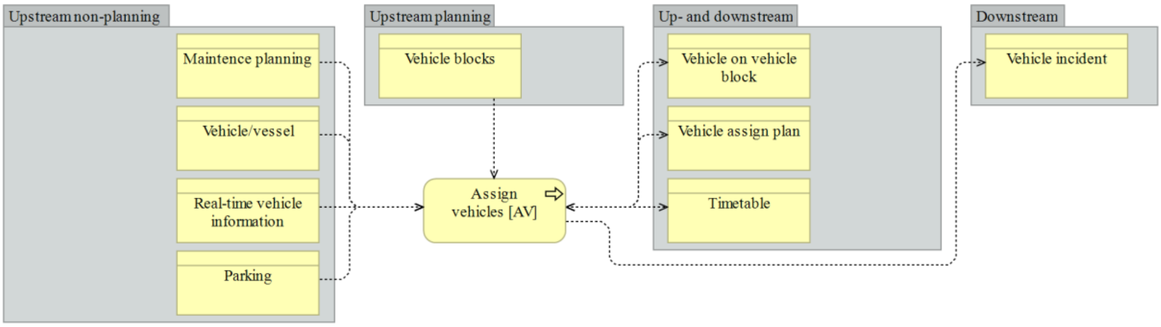

Figure 5-16 Data dependencies Assign vehicles [AV]

5.2.8 Assign Crew

During the identification of planning tasks for *Assign crew [AC]*, a total of 27 tasks were found. These are consolidated into three unique planning tasks, which are presented in Figure 5-17 and explained thereafter.

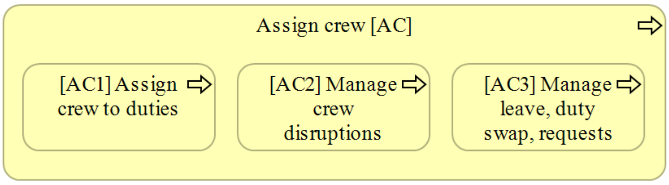
Assign crew [AC]
[AC1] Assign crew to duties
[AC2] Manage crew disruptions
[AC3] Manage leave, duty swap, requests

Figure 5-17 Planning tasks in Assign crew [AC]

[AC1] Assign crew to duties assigns crew members to duties. Most of them are already assigned, but right before operations changes are very likely. Furthermore, spare duties can be given a specific duty (in case of sickness or leave).

[AC2] Manage crew disruptions manages any disruption on crew-level right before or during (triggered by control center) operations. This might lead to crew changes (on the day itself, but possibly also for the next days due to working time regulations), the deployment of crew members with a spare or standby duty or, in the worst case, cancelled journeys because no crew member can be found. At some PTOs this planning task is carried out under the responsibility of the control center.

[AC3] Manage leave, duty swap and requests manages all crew requests regarding leave, duties and duty swaps (with colleagues). This task accepts or denies requests based on planning and forecasts.

Figure 5-18 shows all data requirements for the planning tasks as explained above. The in-depth analysis which led to these data requirements can be found in Appendix L.

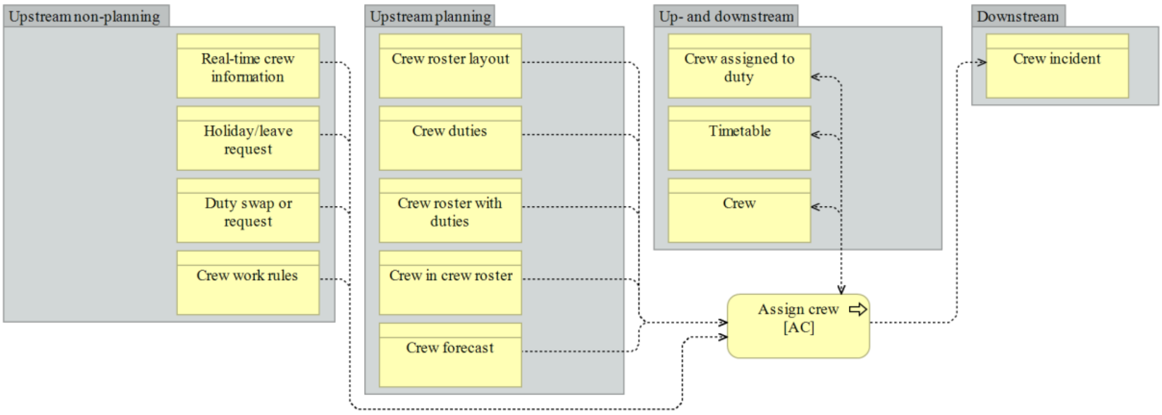

Figure 5-18 Data dependencies Assign crew [AC]

5.3 Data Requirements for the Planning Process

In the previous section, we have identified all data requirements per planning phase by identifying the data requirements per planning task. Combining all data requirements into one figure would result in an incomprehensible figure, this is why the total overview of data requirements is presented as a list in Table 5-3. A description of every data object can be found in Appendix M.

Table 5-3 Data requirements (data catalog) for the planning process

Data object	Upstream *(consumed by)*			Downstream *(provided by)*			Reporting
	C1	C2	C3	C1	C2	C3	
Actual driving time		PL, PT					Y
Crew		PL, PT, PC, SC, AC			PC, AC		Y
Crew assigned to duty	AC			AC	AC		Y
Crew duties	SC, AC			SC	SC		Y
Crew forecast	PC, AC			PL, PT, PC, SC	PL, PT, PC, SC		Y
Crew in crew roster	PC, AC			PC	PC		Y
Crew incident		AC			AC		Y
Crew roster layouts	PC, SC, AC			PC	PC		Y
Crew roster with duties	AC			SC	SC		Y
Crew work rules		PL, PC, SC, AC	PL, PC, SC, AC				-
Dead runs	SV			SV			Y
Desired crew roster	PC						Y
Detour information			PL				-
Detour planning	PL			PL	PL	PL	Y
Duty swap or request	AC						Y
Facilities		PL, SV					-
Fare information		PL					-
Holiday/leave request	PC, AC						Y
Infrastructure	PL, PT, SV	PL	PL	PL	PL	PL	Y
Layover times	SV			SV			Y
Line plan	PL, PT, SV			PL	PL	PL	Y
Maintenance planning		SV, AV					Y
Mobility behavior		PL, PT, PC, SV	PL, PT, PC, SV				Y
Parking		AV					-
Pricing information					PL	PL	Y
Real-time crew information		AC					Y
Real-time vehicle information		AV					Y
Relief points	PT, PC, SV, SC			PL			Y
Routes	PL, PT, SV			PL	PL	PL	Y
Scenarios for transport control	PT			PL	PL		Y
Timetable	PL, PT, PC, SV, SC, AV, AC			PT	PT	PT	Y
Training information		PC	PC				-
Transport authority requirements			PL, PT				-
Transport supply		PL, PT, SV					Y
Utilities and land-usage			PL				Y
Vehicle assign plan	AV			AV	AV		Y
Vehicle blocks	SC, AV			SV	SV	SV	Y
Vehicle forecast	SV			PL, PT, SV	PL, PT, SV		Y
Vehicle incident		AV			AV		Y
Vehicle on vehicle block	AV			AV	AV		Y
Vehicle/vessel		PL, PT, SV, AV					Y

Legend: The labels PL, PT, PC, SV, SC, AV and AC refer to the planning process phases; Y = yes

In this table, the data requirements are indicated on the level of data objects, as they were identified in Section 5.2. It is indicated whether the data requirements have upstream, downstream or both dependencies concerning the planning process and the particular phase. Moreover, the identification of reporting purposes on the level of data objects is added as well (see Section 5.2.1).

We have identified three different contexts for all data object dependencies. This is useful because of the different contexts within a PTO and their external parties, which might lead to inconsistent understandings of data objects (see data integration challenges in Section 3.3.1). This research is written within the context of a PTO's planning process, which means that all data objects have a meaning related to the planning process. Within the larger enterprise, however, one single data object may have different meanings, or, the other way round, multiple data objects exist with the same meaning. The three contexts defined for this research are visualized in Figure 5-19 and are as follows:

- **C1** Business objects generated and/or used within the PTO's planning process.
- **C2** Business objects generated and/or used within any other business area of the PTO.
- **C3** Business objects generated and/or used externally of the PTO.

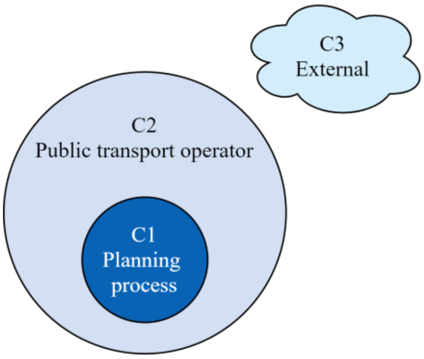

Figure 5-19 Different contexts for data objects

As data integration challenges have shown, every data object can have (slightly) different meanings within different contexts of a business. It should be noted that different contexts within the previously identified contexts (as presented in Figure 5-19) might still exist, however, these are limited to a smaller context due to the provided split into three different contexts.

To verify the outcomes of the data objects in Table 5-3, some checks have taken place. First of all, it was checked whether every data object has at least one up- or downstream dependency. Secondly, the relation between up- and downstream dependencies was checked. An upstream dependency from within the planning process means that this dependency is realized by another planning process phase or planning task. For this reason, it is checked whether every upstream dependency has its corresponding downstream dependency. The same check was carried out the other way around (downstream dependency to its corresponding upstream dependency).

Most of the data dependencies in Table 5-3 are upstream dependencies. Hence, more data objects are demanded by the planning process than provided. This is – among other reasons – because routes, line plans and timetables are dependent on many different data objects and result in only one object: route, line plan and timetable, respectively. However, often the size and context of these latter data objects are bigger than their individual input data objects.

Another finding which can be derived from Table 5-3 is that most of the downstream dependencies are realized by only one planning phase. Hence, the responsibility of the provision of data objects by the planning process is clearly stated. Data objects that have downstream dependencies from more than one planning phase (*Crew*, *Crew forecast* and *Vehicle forecast*) can be explained by the fact that these data objects are changed by the different planning phases, because, in these cases, more (accurate) data is put downstream.

Reporting is an important task for the business. By doing so, the enterprise exactly knows how it performs. This performance can be measured by comparing the actual public transport with the planning, but also by comparing the actual transport with the contract the PTO has with a transport authority (TA). For these reasons, it is important to be able to access many data objects. This is why almost every data object in Table 5-3 has reporting purposes.

As the results in Section 5.2 and Table 5-3 show, the planning process is interacting with many different business areas of a PTO to retrieve and provide all different kinds of data objects. From the identified data requirements, the business area dependencies between the planning process and other business areas can be deducted. Figure 5-20 shows the associations between the public transport planning and the other business areas of a PTO.

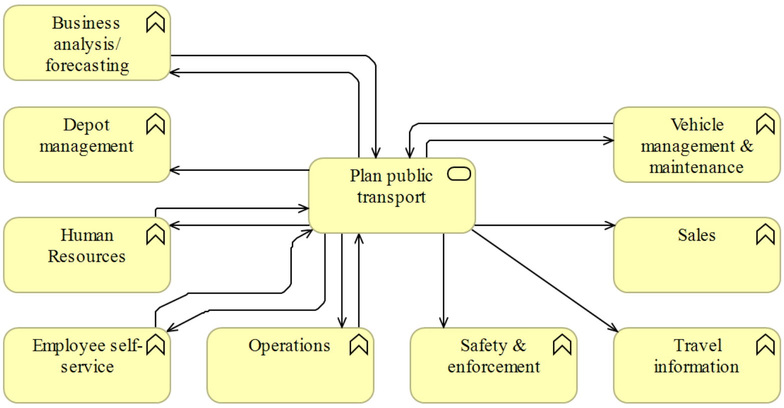

Figure 5-20 Downstream dependencies towards other business areas

5.4 Chapter Summary

For the design of the data integration design approach, it is important to account for all data requirements. These data requirements are identified on the level of planning tasks, which have been presented in this chapter. These planning tasks cover both internal and external data integrations. In the identification process of the data objects, all upstream and downstream data objects (both consumed and provided internally as well as externally) are derived. A total list of data requirements is presented and the responsible planning phase (Chapter 4) and the context of every data object are specified (Table 5-3).

6 DATA QUALITY ASPECTS

This chapter aims to introduce data quality aspects, put these in the context of this research and defines data quality aspects for the design of the data integration design approach. It hereby aims to answer RQ5, as visualized in Figure 6-1. Firstly, the ISO 25012:2008 data quality standard is introduced (Section 6.1), in line with the method explained in Section 2.4. Secondly, in Section 6.2, the relevance of the data quality aspects is indicated by mapping the aspects to the earlier identified data integration challenges in Chapter 3. This is followed by the identification of scenarios for determining the data quality (Section 6.3). Lastly, in Section 6.4, the relevant data quality aspects for the data integration design approach and their operationalization are presented.

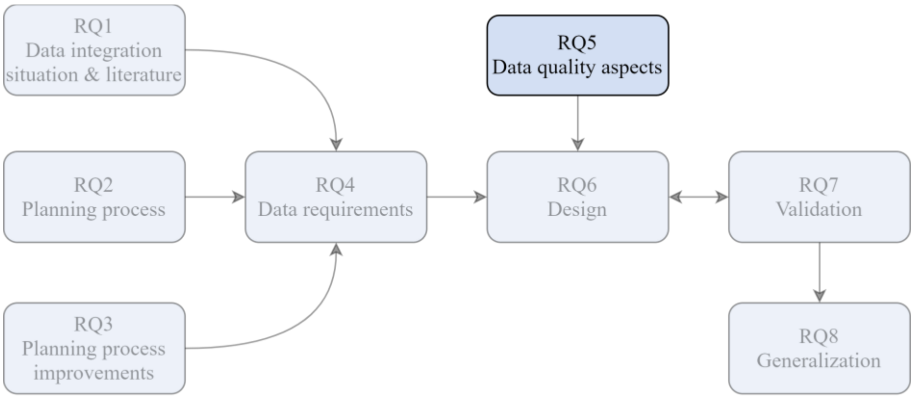

Figure 6-1 Content of chapter 6 (based on Figure 1-5)

6.1 Data Quality Standard

To ensure data quality, data quality aspects as defined in the ISO/IEC 25012:2008 standard are used. This standard proposes fifteen data quality aspects which are part of a so-called ‘data quality model’ (ISO, 2008). The data quality aspects and their definitions can be found in Table 6-1.

When accounting for every data quality aspect, the overall data quality will be increased. According to the ISO standard (ISO, 2008), a division between *inherent data quality* and *system-dependent data quality* can be made. The first category is about quality aspects which are inherent to the actual data value, relationship and metadata (ISO, 2008). Aspects that belong to this category are: accuracy, completeness, consistency, credibility and currentness. The second category contains quality aspects that depend on the technological implementation of the systems. Availability, portability and recoverability belong to this second category. The seven other data quality aspects (accessibility, compliance, confidentiality, efficiency, precision, traceability and understandability) are not categorized into one of the two categories. These are placed in between inherent and system-dependent data quality division.

6.2 Relationship with Data Integration Challenges

Many of the data quality aspects presented in Section 6.1 can be related to the data integration challenges as encountered in practice (Section 3.1.2) and found in literature (Section 3.3.1). The mapping of the data quality aspects to the data integration challenges can be found in Table 6-2.

Table 6-1 Data quality aspects and definitions (ISO, 2008)

Data quality aspect	Definition (ISO, 2008)	Type
Accessibility	The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.	I/S
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attributes of a concept or event in a specific context of use.	I
Availability	The degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use.	S
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.	I
Compliance	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.	I/S
Confidentiality	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.	I/S
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.	I
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.	I
Currentness	The degree to which data has attributes that are of the right age in a specific context of use.	I
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.	I/S
Portability	The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.	S
Precision	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.	I/S
Recoverability	The degree to which data has attributes that enable it to maintain and preserve a specified level of operation and quality, even in the event of failure, in a specific context of use.	S
Traceability	The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.	I/S
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.	I/S

Legend: I = inherent data quality; S = system-dependent data quality; I/S = combination I and S

Table 6-2 Mapping ISO 25012:2008 to data integration challenges from practice and literature

	Data integration challenge	
ISO 25012:2008	**Practice** *(Section 3.1.2)*	**Literature** *(Table 3-7)*
Accessibility	-	-
Accuracy	Multiple data truths	Logic
Availability	Availability	Systems
Completeness	Integrity	Logic
Compliance	Standardization	Systems; Logic
Confidentiality	-	Administrative
Consistency	Multiple data truths	Logic
Credibility	Multiple data truths	Systems
Currentness	Multiple data truths	Logic
Efficiency	-	-
Portability	No reuse possible	Systems
Precision	-	Logic
Recoverability	-	-
Traceability	-	Social
Understandability	-	Logic; Administrative

Most of the data quality aspects contribute to solving data integration challenges from both practice and literature. This shows the applicability of the ISO standard to this research and explains the relevance of the data quality aspects in terms of challenges found in practice and literature.

6.3 Planning Scenarios

The expected and minimal data quality necessary for data entities within the planning process is important to include while designing the data integration solution. Data quality requirements are defined based on the business processes and set by officers which are using the data, since they demand a particular level of data quality for their processes to reach business goals.

Even though every data quality aspect is important, it is impossible to focus on all quality aspects while designing the data integration design approach. The decision of which data quality aspects are taken into account for the design in this research is based on a process in which two planning scenarios were defined. This approach was chosen to scope the data quality aspects and to improve the applicability of the mapping of data quality aspects to the planning process of a PTO. The next step is the actual categorization of data quality aspects, which is presented in Section 6.4.

The two scenarios differ in the way the public transport planning is created. Also, adaptations closer to and possibly even during the operational day are managed differently. Hence, in the first scenario, the public transport planning is made and provided to the operations department as static data with a frequency of only several times per year (excluding detour changes, which happen more often and result in 'deltas' on the original planning). In the second scenario, the public transport planning is executed closer to real-time to better respond to the real-time conditions and optimize based on the real-time environment. The frequency of the data integration in this latter scenario is way higher compared to the first scenario. The scenarios are briefly explained in Table 6-3.

Table 6-3 Data integration scenarios

Scenario	Name	Explanation
1	Static planning	The public transport planning is made only several times per year (including scheduled detours) and the data is provided as static data to other processes, business areas and external consumers. It cannot perfectly respond to changes in the environment other than planned detours.
2	Dynamic planning	The public transport planning is made based on more real-time input and the data is provided to other processes, business areas and external consumers whenever it changes. It can perfectly respond to changes in the environment, since planning phases and tasks can be executed in a real-time manner.

According to Huisman (2016), the speed of planning is even more important than finding the most optimal planning, since the reality changes very quickly and determining the ultimate optimum takes too much time. Moreover, the goal of this research is to provide a data integration design approach to better facilitate data integration, reuse data and its integration and always use correct and up-to-date data. This means that scenario 2 (dynamic planning) is more applicable to the design of this research. However, to distinguish between the importance of data quality aspects, it is still important to consider scenario 1 (static planning). Since only then we can decide whether or not a data quality aspect gains importance.

6.4 Applicable Data Quality Aspects

All data quality aspects as shown in Table 6-1 are important to consider, implement and (continuously) measure to have and keep a proper and high-quality data environment. However, some have higher importance than others, especially when we take into account the two different scenarios as explained in Section 6.3. As stated before, the design in this research fits scenario 2 (dynamic planning) better than scenario 1 (static planning). We will therefore focus on the data quality aspects for dynamic planning. Moreover, for dynamic planning, an increase of some data quality aspects is a prerequisite, which is not the case for static planning (e.g. currentness).

We have defined a categorization based on the ISO 25012:2008 data quality standard. This categorization aims in scoping on data quality aspects for the design in Chapter 7. An explanation about the process of categorizing the aspects and the actual categorization is discussed hereafter. In the next subsections, a justification per category and data quality aspect is given.

First of all, we filter on the data quality aspects which can be included in the actual design of the data integration design approach. In other words, the data quality aspects which can be explicitly accounted for in the design. The second category contains the aspects which can benefit from the data integration design approach and are thus enabled by the design. So far, these categories are in the scope of the data integration design approach.

The third and fourth category are placed outside the scope of this research. The third category is about inherent data quality, which is the actual data value quality. Since this quality is highly dependent on the business processes (PTO-specific), it is out of scope for this research. The fourth category contains aspects that have been observed to be less important for dynamic planning compared to static planning, and are therefore place out of scope. A justification for these decisions and the explanation of every data quality aspect in relation to the research follows in the next subsections. The categorization is presented in Table 6-4.

Table 6-4 Data quality aspect categories based on dynamic planning

Data integration design approach		**Inherent data quality**	**Out of scope for dynamic planning**
Included	**Enabled**		
Compliance	Accessibility	Accuracy	Confidentiality
Consistency	Availability	Completeness	Recoverability
Portability	Currentness	Credibility	
Understandability	Efficiency	Traceability	
	Precision		

6.4.1 Covered by the Design of the Data Integration Design Approach

The data integration design approach should take into account the data quality aspects compliance, consistency, portability and understandability, especially when defining the target data architecture.

Compliancy increases when industry standards are used. As found in the literature and during interviews, Transmodel and NeTEx are widely adopted standards and developed by the European Committee for Standardization (CEN, n.d.-a, 2019). These standards are explained before in Section 3.2.

The *consistency* of data can be increased when one consistent meaning, content and relationship exist for a data entity. This is possible by using the industry standards as mentioned before. Furthermore, offering a global data architecture for the planning process enhances the consistency of the shared data objects. This is because data providers and consumers will be obliged to use the proposed data architecture, which provides consistent data entities.

Portability of data is increased when data can easily be installed, replaced or moved from one system to another while preserving its quality (ISO, 2008). When using the same industry standards, data entities can be transferred without any problem. This is closely related to the data quality aspect compliance.

Related to compliancy, *understandability* also increases when standards are used. In this case, Transmodel is the most important, since that is the conceptual model. Also, the technical standard NeTEx is based on this model. Understandability is ensured as long as the target data architecture uses Transmodel as the conceptual model.

6.4.2 Enabled by the Design of the Data Integration Design Approach

By proposing a data integration design approach, the following quality aspects will be enabled and can therefore be increased: accessibility, availability, currentness, efficiency and precision.

Accessibility increases by the design of the data integration design approach. When the design is provided with a clear target architecture and service, it will allow data consumers to access the data more easily.

Availability can be enabled because the data integration design approach aims at increasing the reuse of data. A solution should be provided which lets any data consumer (with authorization) consume the demand data.

Currentness of data increases when the data is of the right age (ISO, 2008). This quality aspect is highly important in dynamic planning and (near) real-time data exchange. The data integration design approach should aim for providing a clear architecture, including responsibilities, which can ensure that data is always up-to-date.

Efficiency can only be partly enabled by the design in this research, since one part of the efficiency is out of control for the scope of this research. However, the design should enable efficient data exchange by designing efficient data services. Since this is also closely related to the actual implementation and technologies, this research does not cover this aspect entirely.

Precision is related to data quality aspects consistency and compliance. As the latter two might specify precision requirements, then these increase the precision of data.

6.4.3 Inherent Data Quality

All data quality aspects categorized in this category are in line with the ISO 25012:2008 categorization (right column of Table 6-1). The standard also categorizes accuracy, completeness and credibility as entirely inherent data quality, whereas traceability is put in between inherent data quality and system-dependent data quality (ISO, 2008). As these quality aspects are about the actual data values, they are under the responsibility of the data owner and depends on the business process. This is discussed in the approach in Section 7.3.3.

6.4.4 Out of Scope for Dynamic Planning

Confidentiality of data decreases when more and more data is shared, which is the case for a dynamic and real-time planning process. This is because one always wants to have the most up-to-date data. Furthermore, to further optimize the dynamic planning, many data sources are probably going to be added to the planning process. One can think of a driver's behavior when scheduling charging moments. To be able to use this kind of information, data confidentiality for some data entities has to decrease. However, personal data regarding employees should always stay confidential, as GDPR should always be respected.

Since data is available in a real-time manner, *recoverability* becomes less important. Whenever data gets lost, newer data is already available. When this is not the case, the quality aspect availability should be looked at. A validation expert agrees that recoverability becomes less important due to the real-time character. However, the expert stated that for reporting purposes this might not hold entirely. Still, we decided to keep this data quality aspect out of scope for this research.

6.5 Chapter Summary

The data quality aspects including their categorization into *inherent* and *system-dependent* aspects of the ISO 25012:2008 standard (ISO, 2008) are presented. Moreover, their relevance is addressed in terms of data integration challenges from practice and literature, which were identified in Chapter 3.

Since the list of data quality aspects is rather long, it is impossible to account for all aspects of the design of the data integration design approach. For that reason, the quality aspects are categorized into four categories. The categorization is made based on data integration challenges and the importance of quality aspects in a dynamic (real-time) planning scenario.

We focused on dynamic planning while defining the aspects important for the design, since enabling more dynamic planning is one of the goals for the design in this research. This resulted in one category (consisting of compliance, consistency, portability and understandability) which should be included in the design and one category (consisting of accessibility, availability, currentness, efficiency and precision) which is expected to be enabled by the data integration design approach.

7 DATA INTEGRATION DESIGN APPROACH

This chapter provides the data integration design approach and its design process, as visualized in Figure 7-1. Firstly, an overview of the requirements for the design is presented in Section 7.1. Secondly, in Section 7.2, the target data architecture is presented and the context in which it should be used is provided. Thirdly, the approach is explained in Section 7.3. For both parts of the artifact, the design process is explained and the goals and requirement contributions are given. Improvements after validation (Chapter 8) are also taken into account.

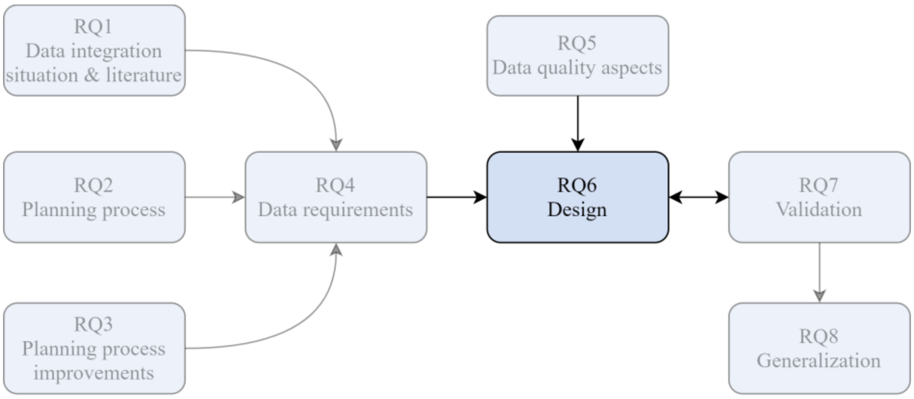

Figure 7-1 Content of chapter 7 (based on Figure 1-5)

7.1 Artifact Requirements

A vast number of goals for the design of the data integration design approach for the planning process of a PTO are identified. These are iteratively adapted during the research process, see Section 1.3.3. For recap purposes, an overview of these goals is repeated in Table 7-1.

Table 7-1 Goals for the data integration design approach

Goal
Always use the correct, most up-to-date and high-quality data.
Be better prepared for future integrations.
Integrate planning phases in order to optimize the global planning (instead of sub-optimization).
Plan more dynamically (real-time).
Align the public transport planning better with maintenance planning (and vice versa).
Introduce depot management solutions.
Improve the scheduling of ZE/battery-equipped vehicles.
Introduce self-rostering for crew members.
Improve the reporting facilities.

To reach the defined goals, requirements for the artifact were determined in the previous chapters. These requirements should – when satisfied by the artifact and applied to the problem context – contribute to the goals as defined before. This is called the contribution argument (Wieringa, 2014). The requirements for the design of the artifact are listed in Table 7-2. This table provides an overview of all requirements and the sections in which these requirements were formulated. Furthermore, it is indicated whether the requirement applies to the data architecture or the approach.

Table 7-2 Requirements for the data integration design approach

	Requirement	Section(s)	Architecture	Approach
1	Reuse data and integrations	1.3; 3.1.1; 3.3.2	✓	✓
2	Ensure data decoupling	1.3; 3.1.1	✓	✓
3	Account for data integration challenges	3.1.2; 3.3.1	✓	✓
4	Consider data integration methods	3.3.2	✓	✓
5	Include planning process phases	4.1 - 4.3	✓	
6	Include planning process improvements	4.4	✓	
7	Include all required data (providing and consuming)	5.1 - 5.3	✓	
8	Comply to data standard	3.1.2; 3.2	✓	✓
9	Include data quality aspects	3.1.2; 6.4.1; 6.4.2	✓	✓
10	Define data responsibility and ownership	3.1.2; 6.4	✓	✓
11	Include expert opinions feedback	8.1 - 8.5	✓	✓

The proposed data integration design approach is presented in the next sections. The contributions of the artifact to each requirement and goal are provided as well.

7.2 Target Data Architecture

Designing the target data architecture can be seen as the largest and most important design part of this research. In the previous chapters, multiple sources were combined and data requirements were formed, data standards were studied, quality aspects are incorporated and data integration challenges were accounted for (Table 7-2). Furthermore, a proposed data access service architecture is presented as well, in which the data architecture should be incorporated. This is decided to include to provide more context to the target data architecture. The architectures should comply with the requirements listed in Table 7-2. In the next subsections, the scoping and ownership of data objects are given (Section 7.2.1) and the target data architecture (Section 7.2.2) and data access service (Section 7.2.3) are presented. Lastly, in Section 7.2.4, the requirement satisfaction is presented.

7.2.1 Data Entities in Scope

For designing the target data architecture, it is not possible to include all data objects as identified in Section 5.3. This is because some of the data objects are out of control for the PTO or that data objects belong to different business areas within the PTO, not being the planning process (e.g. HR). Furthermore, ownership of data is important as became apparent during the problem investigation of this research (Section 1.2 and 3.1.2), in literature (DAMA International, 2014; IDSA, 2019b; The Open Group, 2018) and during the validation of this research and the proposed artifact (Chapter 8).

According to the TOGAF standard (The Open Group, 2018), ownership of data can be given to data entities using a data entity/business function matrix. This matrix is already presented in Table 5-3 in Section 5.3. The data entities which are owned by the planning process can be filtered from this matrix of data entities by the following two criteria:

- Data entities that have a downstream dependency on data originating from the planning process.
- Data entities that have upstream dependencies only within the planning process.

These two requirements together ensure that the resulting data entities are owned by the planning process. The first criterion provides all data entities which are generated within the boundaries of the planning process and are shared with other planning process phases, business areas and/or external consumers. The second criterion ensures that all data entities which are a specific input for the planning process, yet not shared with others, are included as well.

Using these two criteria, a complete overview of data entities and their ownership is made and shown in Table 7-3. After validation (Chapter 8), it turned out that some data entity categorizations are somewhat debatable. Especially data about HR turns out to have different ownership among the PTOs.

The data entities owned by the planning process (left column in Table 7-3) are used for the design of the target data architecture, which is presented in Section 7.2.2.

Table 7-3 Data entities and their ownership (derived from Table 5-3)

Data owned by planning *(input and output data)*	**Data not owned by planning** *(input data)*
Crew assigned to duty	Actual driving time
Crew duties	Crew
Crew forecast	Crew incident
Crew in crew roster	Crew work rules
Crew roster layouts	Detour information
Crew roster with duties	Facilities
Dead runs	Maintenance planning
Desired crew roster	Mobility behavior
Detour planning	Parking
Duty swap or request	Pricing information
Fare information	Real-time crew information
Holiday/leave request	Real-time vehicle information
Infrastructure	Training information
Layover times	Transport authority requirements
Line plan	Transport supply
Relief points	Utilities and land-usage
Routes	Vehicle incident
Scenarios for transport control	Vehicle/vessel
Timetable	
Vehicle assign plan	
Vehicle blocks	
Vehicle forecast	
Vehicle on vehicle block	

7.2.2 Data Architecture

After identifying the data entities which are owned by the planning process, the next step is to design a target data architecture. Designing a data architecture based on business data requirements is not a one-to-one copy of the data requirements. It contains logic and takes into account the Transmodel standard. The target data architecture is presented in Figure 7-2 and can be seen as a global schema based on the Transmodel ontology (see Section 3.2.1 and Section 3.3.2).

Almost every data entity in the presented data architecture is owned by the planning process, as this was the outcome of the data scoping carried out in Section 7.1. Two entities are, however, not owned by the planning process: *Vehicle data* and *Crew data*. We assume that these two data entities are owned by the vehicle and HR department, respectively. This means that the data entity as shown in the data architecture refers to these original data entities. As the architecture cannot be visualized correctly without these entities, they are included.

As can be seen in Figure 7-2, the data is grouped into the following domains: journey data, vehicle data and crew data. This grouping is made to give a clear view of the data entities, but also to already define ownership and responsibilities (described more specifically in Section 7.3). Furthermore, the domain grouping is based on the public transportation triangle as introduced in Figure 4-9.

In the next subsections, decisions regarding integration methodology and compliance, the data architectures on domain-level and information about meta data are presented. The domain-level architectures include the data requirements from the business layer, on which the data architecture is based (left column of Table 7-3). They also address some choices which have been made during the design of the data architecture. A total overview of the data architecture and the data entities on which the architecture is based can be found in Appendix N.

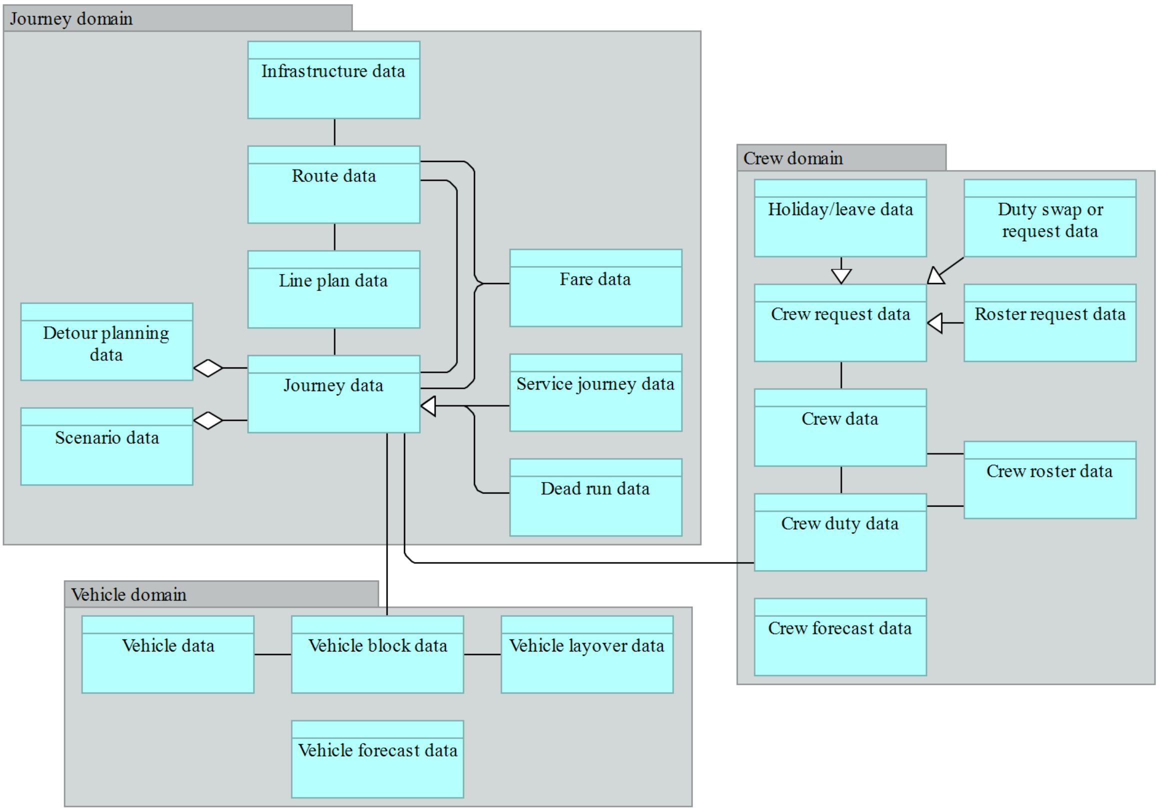

Figure 7-2 Proposed data architecture

7.2.2.1 Integration Methods and Data Standards

In Section 3.3.2, three data integration methods were presented and explained: global schema, canonical data model (CDM) and ontology matching. For the development of the artifact presented in Figure 7-2, both ontology matching and global schema methods are used. For ontology matching, the Transmodel conceptual model (Section 3.2.1) is used. This ensures that the design of the target data architecture is implementation-agnostic, as that improves its applicability to different PTOs. Combining every data requirement from the planning process, its adjacent business areas and external parties, the global schema methodology was used. This schema indicates the data objects and their relations.

Using a CDM is possible for data standardization purposes. Since the actual implementation is out of scope for this research, this part is not provided. However, since the proposed architecture is based on Transmodel, the NeTEx standard (Section 3.2.2) can be used. Using this European standard would increase the possibility for the reuse of data and data integrations and enhances the portability data quality aspect. NeTEx can be used for the greater part of the journey domain. Unfortunately, NeTEx does not cover the crew domain and a great part of the vehicle domain, as it is intended for the network and schedule data (CEN, n.d.-a). The applicability of NeTEx for only a small part of the data architecture (journey and vehicle domain) is also recognized after validation by experts. Standardization for the missing entities using alternative standards is an important input for future research (Section 9.6).

7.2.2.2 Journey Domain

The data in the journey domain contains all data entities related to the public transport timetable. This includes the entire network, routes, lines, prices and detour information. The data architecture of this domain including the realized data objects (on the business layer) is presented in Figure 7-3.

Most of this data is published to passengers and provided as open data. Exceptions are *Detour planning data*, *Scenario data* and *Dead run data*, these data entities are not shared externally and only used within the boundaries of the PTO. Furthermore, when we dive into more details, most likely not every single attribute of the included data entities is shared.

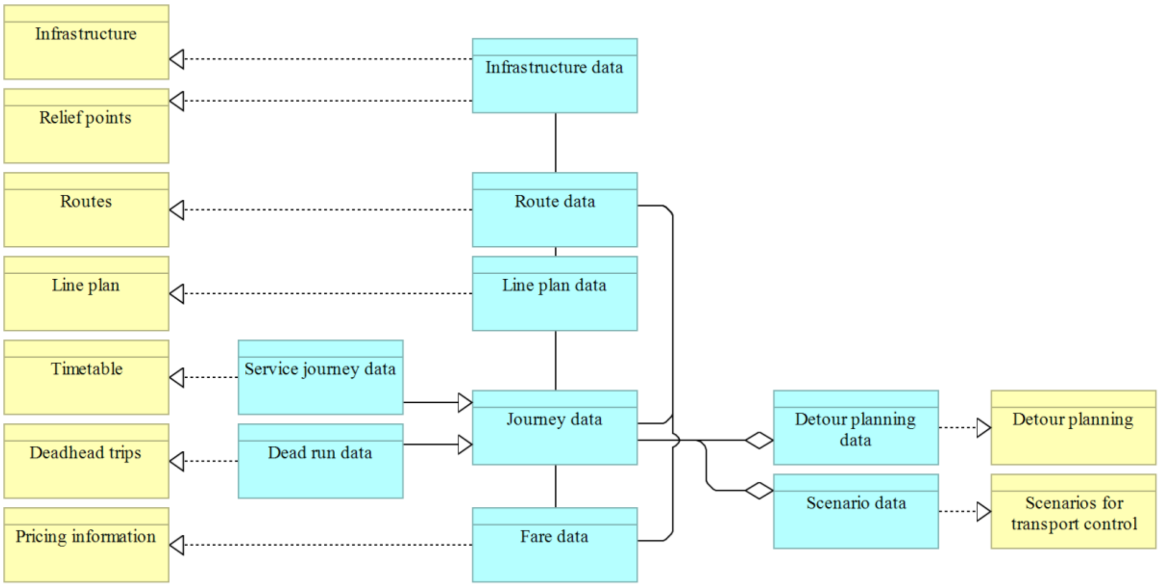

Figure 7-3 Proposed data architecture – journey domain

On IT abstraction level, a dead run and service journey are the same, since a vehicle needs to drive from point A to point B. For this reason, these two are considered to be of the type *Journey data*. Furthermore, as the *Detour planning data* and *Scenario data* contain information about changed journeys (compared to the 'standard' schedule), they both aggregate *Journey data*.

7.2.2.3 Vehicle Domain

In the vehicle domain of the data architecture, all data about vehicles is accounted for. The proposed data architecture for this domain is presented in Figure 7-4 and includes the realized data objects from the business layer.

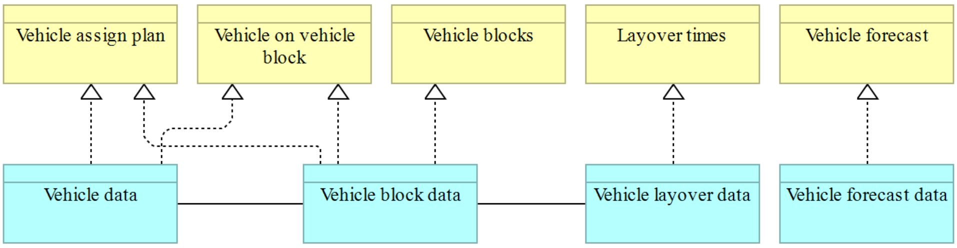

Figure 7-4 Proposed data architecture – vehicle domain

As explained before in this section, the *Vehicle data* in this domain is not owned by the planning process, but refers to the data from the vehicle department. The implementation of that integration is out of scope for this research, but only included as a necessary data entity for the data architecture. As visualized in the architecture, the combination of *Vehicle data* and *Vehicle block data* covers three required data objects. This is possible since a combination of these two data objects forms *Vehicle on vehicle block*.

7.2.2.4 Crew Domain

The data architecture of the crew domain contains all data regarding crew members, their duties and rosters. The relationship of the data objects with the data requirements can be seen in Figure 7-5.

Similar to *Vehicle data* in the vehicle domain, *Crew data* is not owned by the planning process. Yet, it refers to the data about crew members which is owned by HR. Similar to *Vehicle data*, the implementation of the integration between *Crew data* in the planning domain and HR is not in the scope of this research.

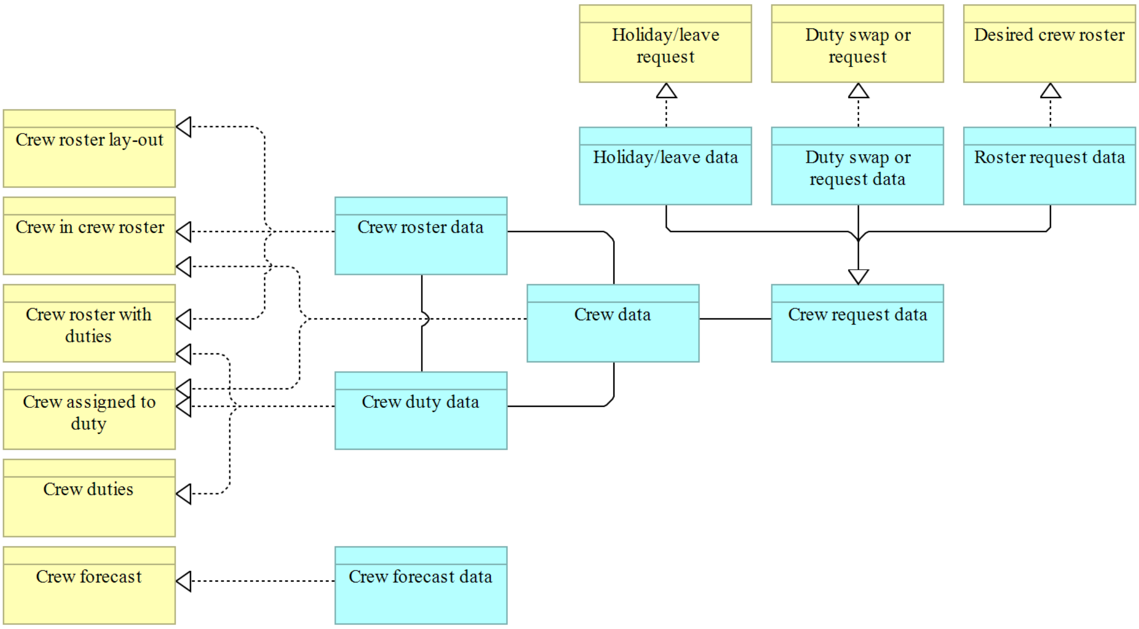

Figure 7-5 Proposed data architecture – crew domain

As becomes apparent from Figure 7-5, five demanded data objects about crew rosters and crew duties, can be modeled with three data entities in the application layer: *Crew roster data*, *Crew data* and *Crew duty data*. This is possible since the relationships between these data entities also represent data requirements (*Crew roster data* and *Crew data* represents *Crew in crew roster*). Many request possibilities exist on crew-level, which are consolidated into one data entity called *Crew request data*.

7.2.2.5 Meta Data

As became apparent during the literature review (Section 3.3.2), meta data is important for integrating data based on the actual meaning of a data element (Zhang et al., 2018). For automatic schema matching, meta data is even crucial (Jarke et al., 2014; Salguero et al., 2008; Zhang et al., 2018). Since the actual mapping of data elements to the proposed architecture is out of scope for this research, meta data is not accounted for in the design. The operational data entities (as all data entities are in Figure 7-2) should, however, also have meta data stored. Often this is done in a separate database (Singh et al., 2017), which contains the meta data about all data entries.

7.2.3 Data Access Service

The data architecture presented in Section 7.2.2 needs to be realized somehow to use the data architecture and thus to fully decouple and to be able to reuse data and data integrations. The planning process phases as described in de proposed planning process in Section 4.3 are visualized again in Figure 7-6. We assume that every planning process phase is served by an application function, which in this case has the same name as the planning process. Application functions abstract from the way they are implemented (The Open Group, n.d.), which is chosen on purpose because every PTO's application landscape is different (i.e. the applications assigned to the application functions may be different and may differ in number). This makes the figure PTO-independent.

As can be seen in Figure 7-6, the data access service is connected to the application functions through data flows. This also abstracts from the actual data integration implementation (The Open Group, n.d.), as we do not provide any implementation-related plans (e.g. a data (web-)service, batch processing, a data application with federated databases or one central database).

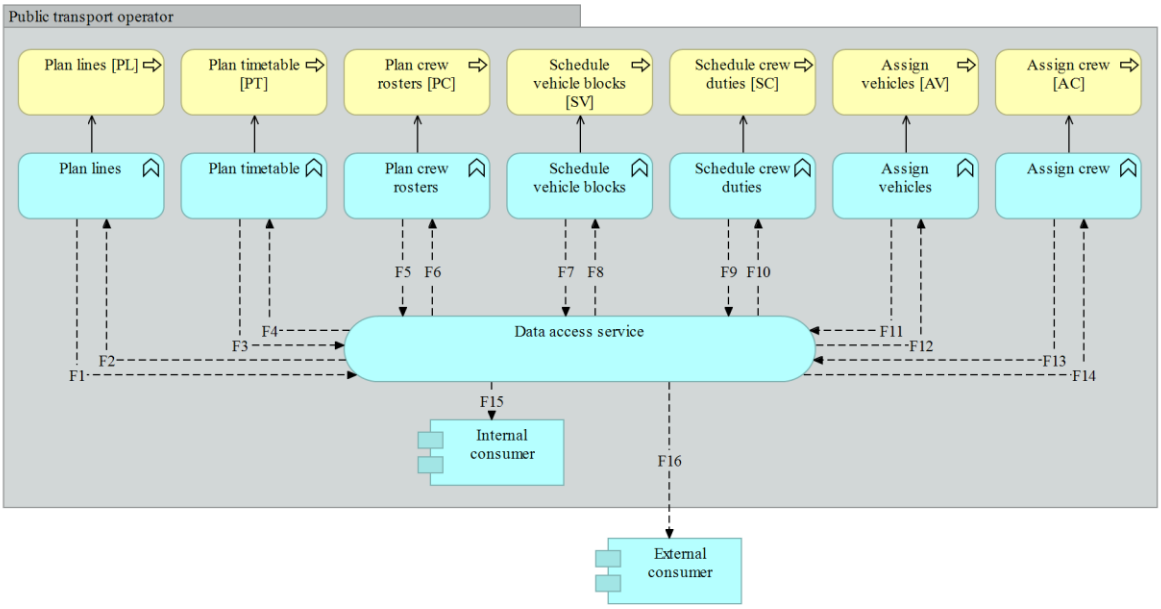

Figure 7-6 Data access service for the proposed data architecture

The data flow relationships from the application functions to the data access service provide data to the service, which can be consumed by any consumer who is authorized to consume the data. The data flow relationship from the data access service back to the application functions ensures that the application functions can access the data as well. This is because a lot of data is reused, enriched and changed during the planning process.

Within application functions (and thus within application components), data may be exchanged directly, without going through the data access service. This addition is included after validation of the artifact. It increases performance for the processes within this application function. However, it should always be secured that the most up-to-date and correct data is available via the data access service.

During implementation, one might choose to split the service into separate services. This could be better for its lifecycle management and responsibilities. Since it is about the way of implementing the data service, it is out of scope for this research. When choosing the modular approach, it should be taken into account that transactions and authorizations become more difficult, since a lot of data is reused, enriched and changed by many roles during the planning process.

When consuming applications are consuming data via the data access service, applications are loosely coupled. This means that changing the providing side of the data access service does not influence the communication between the consuming application and the data access service. This suffices the other way around as well. The consuming applications can change without changes in the communication between the providing application functions and the data access service. This is one of the main requirements and contributes to achieving the goals of this research.

The actual content of the data flows in Figure 7-6 was identified during the data requirements identification in Chapter 5. The data objects which are transferred on every data flow relationship (F1 – F16) can be found in Appendix L. Since the appendix is rather complex, an overview of the data flows to the related tables in Appendix L can be found in Table 7-4.

Table 7-4 Data flows and their data specifications

Data flow *(Figure 7-6)*	Data specification
F1; F2	Table A-14 SIPOC [PL1] Create route network Table A-15 SIPOC [PL2] Determine relief points Table A-16 SIPOC [PL3] Provide fare information Table A-17 SIPOC [PL4] Develop scenarios Table A-18 SIPOC [PL5] Plan detours
F3; F4	Table A-19 SIPOC [PT1] Determine timetables Table A-20 SIPOC [PT2] Adjust/improve timetables Table A-21 SIPOC [PT3] Design frequency changes
F5; F6	Table A-22 SIPOC [PC1] Create roster lay-outs Table A-23 SIPOC [PC2] Assign crew to roster layout Table A-24 SIPOC [PC3] Plan training and holidays Table A-25 SIPOC [PC4] Arrange self-rostering
F7; F8	Table A-26 SIPOC [SV1] Determine layover times Table A-27 SIPOC [SV2] Plan dead runs Table A-28 SIPOC [SV3] Create vehicle blocks
F9; F10	Table A-29 SIPOC [SC1] Create crew duties Table A-30 SIPOC [SC2] Assign crew duties to roster
F11; F12	Table A-31 SIPOC [AV1] Assign vehicle to block Table A-32 SIPOC [AV2] Manage vehicle disruptions Table A-33 SIPOC [AV3] Plan vehicle parking
F13; F14	Table A-34 SIPOC [AC1] Assign crew to duties Table A-35 SIPOC [AC2] Manage crew disruptions Table A-36 SIPOC [AC3] Manage leave, duty swap, requests
F15	Data consumers within the PTO's enterprise (Figure 5-20).
F16	Data consumers external to the PTO's enterprise (Table 3-3).

7.2.4 Requirements Satisfaction

The satisfaction of the requirements listed in Table 7-2 is presented in Table 7-5.

Table 7-5 Target data architecture requirement satisfaction

	Requirement *(Table 7-2)*	Satisfied by
1	Reuse of data and integrations	Target data architecture is modeled for reuse purposes which can be realized by the data access service and fed by application functions.
2	Ensure data decoupling	Providing a global schema within a data access service that connects provider and consumer and is based on Transmodel.
3	Account for data integration challenges	Categories (Table 3-7) accounted for: systems (development of data sources, technical differences and one truth), logic (agreements on definitions, clear relationships and semantic heterogeneity) and administrative (right source).
4	Consider data integration methods	Providing a global schema based on Transmodel ontology.
5	Include planning process phases	Defining data entities of the data architecture based on the triangulation process of Chapter 4 and Chapter 5.
6	Include planning process improvements	
7	Include all required data (providing and consuming)	
8	Comply to data standard	Application of Transmodel and the possibility to apply NeTEx.
9	Include data quality aspects	Compliance, consistency, portability and understandability are taken into account by using industry standards such as Transmodel and NeTEx. Accessibility, availability, currentness, efficiency and precision can be more easily increased with the help of the target architecture.
10	Define data responsibility and ownership	Data entities owned by the planning process are identified and mapped to specific domains (journey, vehicle, crew).
11	Include expert opinions feedback	Feedback is included as indicated in the text.

7.3 Approach

The approach presented in this section aims to clarify and touch upon several considerations that are important to take into account while implementing a data architecture (The Open Group, 2018). Furthermore, "(...) no single technology approach can achieve data and enterprise integration, since the technology and demands within an enterprise are diverse" (Giachetti, 2004), which means that the considerations presented here should be adapted to every specific enterprise. For this reason, these are presented on a high-level and implementation-agnostic. The approach should account for the requirements presented in Table 7-2.

For the proposed approach we have used TOGAF's method for approaching a data architecture (The Open Group, 2018). This is in line with the method described in Section 2.5.2. TOGAF lists several considerations concerning data architectures. The considerations of data management are used for the approach. An overview of the considerations and the proposed solutions are listed in Table 7-6.

Table 7-6 Approach considerations and proposed solutions

TOGAF's consideration *(The Open Group, 2018)*	**Proposed solution**
A clear definition of which application components in the landscape will serve as the system of record or reference for enterprise master data.	This highly depends on the application components which are in use at the PTO. An approach for this is presented in Section 7.3.1.
Will there be an enterprise-wide standard that all application components, including software packages, need to adopt?	Compliancy to Transmodel, NeTEx and the ISO 25012:2008 data quality standard, as explained in Section 7.3.3.
Clearly understand how data entities are utilized by business functions, processes, and services.	The design process of the target data architecture includes these understandings and is explained in Section 7.2.
Clearly understand how and where enterprise data entities are created, stored, transported, and reported.	The proposed approach for this is presented in Section 7.3.2 (data authorizations) and Section 7.3.3 (data roles).
What is the level and complexity of data transformations required to support the information exchange needs between applications?	These two considerations highly depend on the application components which are in use at the PTO. The data architecture and data access service as presented in Section 7.2 help to account for this consideration. The approach regarding data quality is presented in Section 7.3.3.
What will be the requirement for software in supporting data integration with the enterprise's customers and suppliers (e.g. use of ETL tools during the data migration, data profiling tools to evaluate data quality, etc.)?	

In the next subsections, all aspects and considerations which are part of the approach are presented. The requirement satisfaction is presented lastly.

7.3.1 Implementation Example

Moving towards the target data architecture as proposed in Section 7.2 requires more steps to be carried out as described in the TOGAF ADM Phase C. As explained in Section 2.5.2, these steps are not provided in this research. However, an example is provided of how moving towards the target data architecture can be accomplished. This is done to clarify the intentions of the target data architecture and the data access service. Figure 7-7 shows part of the data access service (only two out of seven application functions) and is based on an example in which a fictive PTO uses GIRO Hastus[6] for public transport planning and SAP[7] as enterprise resource planning software.

[6] Public transport planning software provider: https://www.giro.ca/en-ca/our-solutions/hastus-software/

[7] Enterprise application software provider: https://www.sap.com/products.html

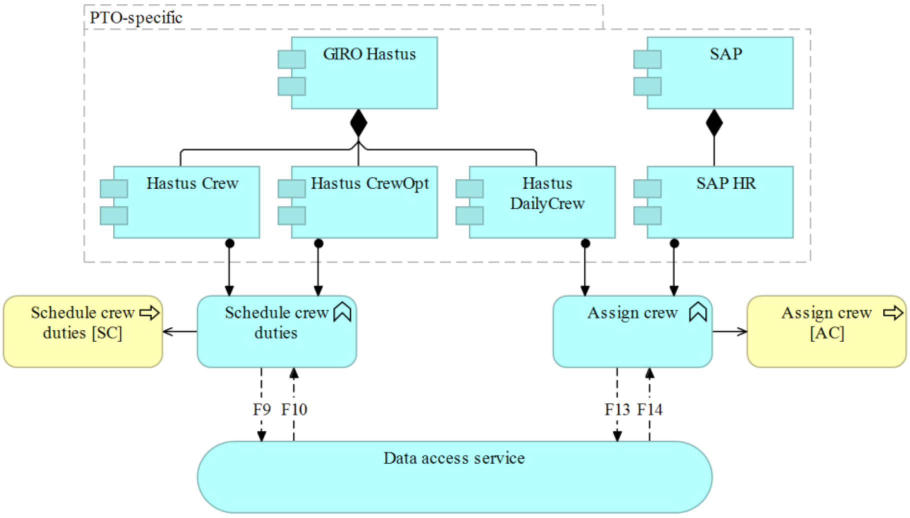

Figure 7-7 Data access service applied to a PTO's context (example, based on Figure 7-6)

Different application components are assigned to the application functions. To move towards the proposed data architecture, the enterprise should identify all applications which are either providing and/or consuming the data. One of the first considerations would be to determine which system is the system of record for the data (The Open Group, 2018). Taking into account data authorizations (Section 7.3.2), data quality and roles (Section 7.3.3) and the requirement of only having one truth of the data (use of up-to-date and correct data), the enterprise can determine this.

7.3.2 Data Authorizations

When providing a target data architecture, it is important to provide a clear overview of who is allowed to create, store, transport and report data entities (The Open Group, 2018). This can be done by providing a data entity/business function matrix (The Open Group, 2018). Such a matrix is already provided during the data requirements identification process and is presented in Table 5-3. Since this matrix does not provide any detailed information about the actual CRUD-actions (create, read, update and delete) on data entities, it is used as input for a so-called CRUD-matrix. This matrix is presented in Table 7-7 and helps a PTO to have a clear understanding of which data entities from the target data architecture may be created, updated, edited and deleted by which planning process phase.

Table 7-7 CRUD matrix for data objects within the planning process

Planning phase Data object	Plan lines	Plan timetable	Plan crew rosters	Schedule vehicle blocks	Schedule crew duties	Assign vehicle	Assign crew
Crew assigned to duty							CRUD
Crew duties				R*U*	CRUD		RU
Crew forecast	CRUD	RU	RU		RU		R
Crew in crew roster			CRUD		RU		R
Crew roster lay-outs			CRUD		R		R
Crew roster with duties					CRUD		R
Dead runs				CRUD			
Desired crew roster			CRUD				
Detour planning	CRUD						
Duty swap or request							CRUD
Fare information	CRUD						
Holiday/leave request			CRUD				CRUD
Infrastructure	CRUD	R		R			
Layover times				CRUD			
Line plan	CRUD	RU*		RU*			
Relief points	CRUD	RU*	RU*	RU*	RU*		
Routes	CRUD	RU*		RU*			
Scenarios for transport control	CRUD						
Timetable	R*U*	CRUD	RU*	RU*	RU*	RU*	RU*
Vehicle assign plan						CRUD	
Vehicle blocks				CRUD	RU*	RU*	
Vehicle forecast	CRUD	RU		RU		R	
Vehicle on vehicle block						CRUD	

C = create; R = read; U = update; D = delete; * = only if planning phases are integrated (applicable to the letter it is placed after)

7.3.3 Data Quality and Roles

In Chapter 6 we have identified which data quality aspects from the ISO 25012:2008 standard (ISO, 2008) can be included in the target data architecture (as already shown in Section 7.2.2) and which aspects are enabled by the design of the data integration design approach. Furthermore, we stated that some data quality aspects are inherent to the data values. Also, some aspects appear to be less important for dynamic planning.

To ensure the data quality and to control the data authorizations (Section 7.3.2), it is necessary to clearly define these responsibilities to data owners within a PTO. During the assessment of the data integration situation at a PTO as presented in Section 3.1 and the validation in Chapter 8, it became apparent that defining these responsibilities is necessary.

The reference architecture of the International Data Spaces Association (IDSA) is used (IDSA, 2019b), in line with the method described in Section 2.5.2. Their proposed roles and interactions architecture and the definition of every role can be found in Appendix O. Based on this reference architecture, the important data roles for a PTO are defined. This is done by accounting for the necessary data interactions and scoping the reference architecture to the context of this research. To better separate concerns regarding the other data quality aspects and responsibilities, a split of the data owner and data provider as proposed by IDSA (2019b) is made into data owner, data custodian and data steward. This split is widely discussed in literature and separates concerns regarding ownership, technical responsibility and content responsibility (Firican, 2018, 2019). These roles are necessary to ensure data quality and data authorizations. They should cover the proposed approach in this section, ensure data quality and be responsible for their data and its provision. In that way, they contribute "to the development of innovative business models and digital, data-driven services to be used (...)" (IDSA, 2019b). An overview of the roles and their interactions is presented in Figure 7-8. An explanation of all data roles is listed in Table 7-8.

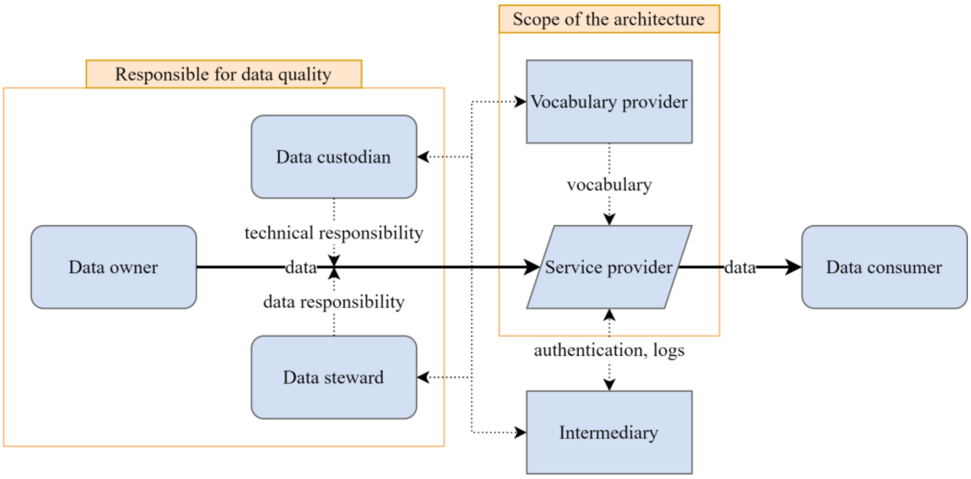

Figure 7-8 Roles and interactions in the PTO's data space

As visualized in Figure 7-8, the scope of the target data architecture proposed in this study is limited to the vocabulary provider and service provider. These play an important role in the data quality aspects which are included in the target data architecture (compliance, consistency, portability and understandability) and are mostly focused on the application of Transmodel and NeTEx.

Table 7-8 Roles in the PTO's data space

Role	Description
Data owner	Has the legal right for creating data and/or executing control over it. Defines usage contracts and policies, provides access to data and defines payment models. Makes data available for being exchanged between a Data owner and Data consumer (IDSA, 2019b). Consists of IDSA's Data owner and Data provider.
Data steward	Defines, implements and enforces the accountability and responsibility of the organization's data stakeholders (Firican, 2018).
Data custodian	Is responsible for the IT on a technical level: maintaining, archiving, recovering, backing up data, preventing data loss/corruption, etc. (Firican, 2019).
Data consumer	Receives the data from the Data owner and has the legal right to use the data of a Data owner as specified by the usage policy (IDSA, 2019b). Consists of IDSA's Data consumer and Data user.
Intermediary	Is related to the Service provider and is responsible for the authentication of the identities and logging of data exchange (IDSA, 2019b). Consists of IDSA's clearing house and identity provider.
Vocabulary provider	Manages and offers vocabularies (i.e. ontologies, reference data models or meta data elements) that can be used to annotate and describe datasets. Provides the information model (IDSA, 2019b). In this research, the target data architecture and ontology (Section 7.2.2) is meant.
Service provider	Hosts the required infrastructure to make data available. Receives data from the Data owner and offers the data within the organization (IDSA, 2019b). Also stores and manages information about the data sources available. Consists of IDSA's service provider and broker service. In this research, the provision of the data access service (Section 7.2.3) is meant.

7.3.4 Requirements Satisfaction

The satisfaction of the requirements listed in Table 7-2 is presented in Table 7-9.

Table 7-9 Approach requirements satisfaction

	Requirement (Table 7-2)	**Satisfied by**
1	Reuse of data and integrations	Proposing the data access service, an implementation example, data responsibilities and authorizations.
2	Ensure data decoupling	
3	Account for data integration challenges	Categories (Table 3-7) accounted for: systems (new and vast number of data sources), logic (relationships and semantic heterogeneity), social (data owner rights) and administrative (authorization).
4	Consider data integration methods	The design of a target data architecture within the data access service, of which the service is the mediator.
8	Comply to data standard	ISO 25012:2008 (ISO, 2008) data quality aspect compliance and the use of Transmodel.
9	Include data quality aspects	Quality aspects are assigned to the data owners and responsible parties.
10	Define data responsibility and ownership	The data entity responsibilities are defined on the level of planning process phases and assigned to data roles.
11	Include expert opinions feedback	Feedback is included as indicated in the text.

8 VALIDATION AND GENERALIZATION

This chapter aims to validate the designed artifact and generalize about its applicability, as visualized in Figure 8-1. By validating an artifact, the goal is "to build a theory of the implemented artifact in a real-world context, based on a study of validation models" (Wieringa, 2014). For this research, the validation model is validated using expert reviews, in line with the method in Section 2.6. The validation consists of several parts and is presented in Section 8.1 until Section 8.5. The generalization of the artifact is explained in Section 8.6.

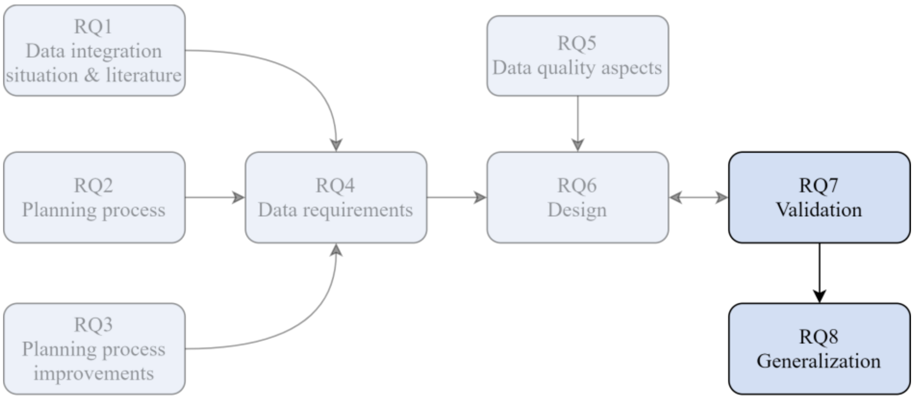

Figure 8-1 Content of chapter 8 (based on Figure 1-5)

Important note for reading this chapter: when talking about *experts*, we refer to the *experts who participated in the validation survey* (as introduced in Section 8.2), not to experts in general.

8.1 Validation Overview

Expert opinions are used for the validation of this research. As explained in the method in Section 2.6.1, a validation model is needed to validate the designed artifact. Such a validation model consists of a model of the artifact and a model of the problem context which represent their target (implemented artifact and intended context) by similarity (Wieringa, 2014). The validation model is introduced in Section 2.6.1, and repeated here:

The model of the artifact is the data integration design approach consisting of the data architecture and implementation approach, which is based on the best-practice planning process, its future improvements, its data requirements and data quality aspects from the field of data management.

The model of the problem context is the PTO's planning process including data-providing and data-consuming applications, processes and external parties.

When asking experts for their opinion about the validation model, the expert is asked to imagine the validation model and 'observe' how it functions in their imagination (Wieringa, 2014). Given the experts' experience, this is expected to be a useful validation method.

In line with the method explained in Section 2.6.1, the validation instrument is a Google Forms survey hosted at the Google workspace of the University of Twente. The validation survey consists of seven parts. Two of these parts are compulsory to be filled out by every expert, whereas for the other five parts the experts can choose whether they feel familiar enough with the topic to answer the questions. Table 8-1 lists an overview of these parts and in which section every validation part is discussed.

Table 8-1 Overview of responses per validation part

Validation part	Section
Consent*	8.2
Business content	8.2
IT content	8.3.2 – 8.3.4
Approach	8.3.5
Goals	8.4
Use and acceptance	8.5
Respondent details*	8.2

* = compulsory part

The validation survey including all questions and answer possibilities can be found in Appendix G, the results for all closed questions are provided in Appendix P. Next to the closed questions, the experts were asked to provide their reasoning for answering every question. Most of the experts provided this explanation, which is very useful for validation, especially for the improvements of the artifact. These answers are not provided in Appendix P due to anonymity reasons. The validation results are discussed in the next sections grouped by the individual validation parts, as indicated in Table 8-1.

For many questions, the experts were asked to score on a 5-point Likert scale: Strongly agree; Somewhat agree; Neither agree nor disagree; Somewhat disagree; Strongly disagree; Undecided. For visualization purposes, box plots are often used in the upcoming sections. These indicate the maximum and minimum values (lines), the mean (X), the first and third quartiles (the box) and outliers (dots).

8.2 Validation Experts

In total, 36 experts were asked to take part in the validation of this research. These experts were selected within the BISON Architecture working group. This is a Dutch group consisting of experts working at PTOs, integrators, public transport standard providers and other consultancies working with the Dutch public transport standards. Next to this group of experts, two other experts from the railway industry were asked individually to take part in the validation.

Due to the complexity of the validation survey and the diversity in covered subjects, we have asked to share the validation survey among the experts' colleagues in case they think their colleagues could provide useful feedback. In this process, we relied on the experience and honesty of the invited experts. Since we do not know exactly how often the validation survey was shared among colleagues, the actual response rate cannot be calculated. However, what we know is that 36 experts were invited, which led to 15 responses (not necessarily from these 36 experts). More details about the experts who participated in the validation survey can be found in Table 8-2.

Every expert has answered the consent questions (C1 – C6 in Appendix G) positively. This means that they stated they were informed about the research, could ask questions, consented voluntarily to be a participant, understood the process, understood the data collection method, agreed on the level of anonymity and understood that no (financial) compensation is provided to the participants.

The experts' gender and age are presented in Figure 8-2 and Figure 8-3, respectively. 80% of the research group (12) identified their gender as 'male', whereas 13% (2) identified themselves as 'female'. One expert preferred not to state their gender.

Table 8-2 Experts' roles/functions, experience and filled out validation parts

				Filled out validation parts *(from Table 8-1)*						
	Role/Function	**Working at a PTO***	**Experience** *(in years)*	**Consent****	**Business content**	**IT content**	**Approach**	**Goals**	**Use and acceptance**	**Respondent details****
E01	Advisor IT development	✓	> 20	✓	-	✓	✓	✓	-	✓
E02	Application manager	✓	> 20	✓	✓	✓	✓	✓	✓	✓
E03	Business analyst	✓	> 20	✓	✓	✓	-	-	-	✓
E04	Chairman	-	10 - 20	✓	✓	✓	✓	✓	✓	✓
E05	Consultant, functional manager, data architect	-	> 20	✓	✓	✓	✓	✓	✓	✓
E06	Domain architect	-	> 20	✓	✓	✓	✓	✓	✓	✓
E07	Enterprise architect	✓	> 20	✓	✓	✓	✓	✓	✓	✓
E08	Enterprise architect	-	> 20	✓	✓	✓	✓	✓	-	✓
E09	Information manager	✓	> 20	✓	✓	✓	✓	✓	✓	✓
E10	Manager	✓	> 20	✓	✓	✓	-	-	-	✓
E11	Planning consultant	✓	10 - 20	✓	✓	✓	✓	✓	✓	✓
E12	Product owner	✓	10 - 20	✓	✓	✓	-	-	-	✓
E13	Solution architect	✓	> 20	✓	✓	✓	✓	✓	✓	✓
E14	Specialist	-	> 20	✓	-	✓	✓	✓	-	✓
E15	Transport engineer	✓	0 - 3	✓	✓	✓	-	-	✓	✓
			Total count	15	13	15	11	11	9	15

* = according to the definition provided in Section 1.1.2; ** = compulsory part of the survey

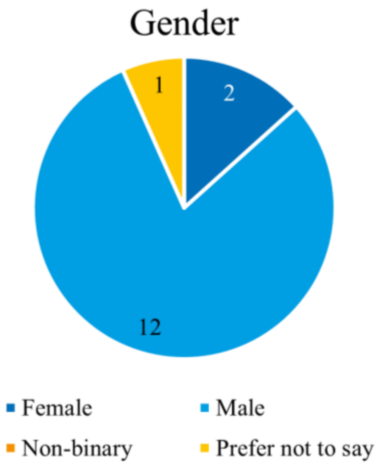

Figure 8-2 Experts' gender

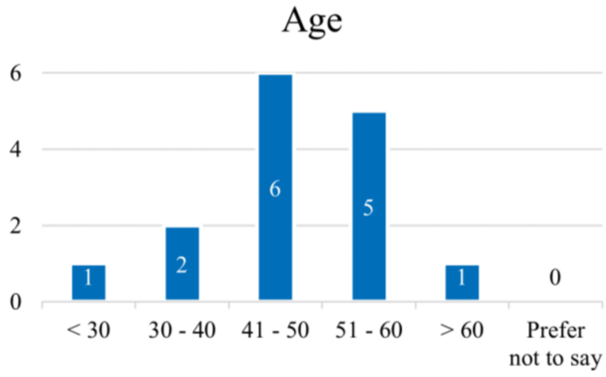

Figure 8-3 Experts' age

Details about the experts' role/function and their years of experience are presented in Table 8-2 and are therefore not repeated here. However, within the validation survey, the experts were also asked to identify their levels of experience within the fields of public transportation planning, enterprise architecture, data architecture and data integration (P3 – P6 in Appendix G). Scoring the level of experience is done on a 5-point Likert scale ranging from 'unfamiliar' to 'expert'. The results are visualized in Figure 8-4.

On average, we can state that all experts are relatively experienced in the field of PTO's planning process, data architecture and data integration. Unfamiliarity is the most present within the enterprise architecture domain. Since enterprise architecture knowledge is the least important (out of the four given domains) for this research, we do not consider this as a problem.

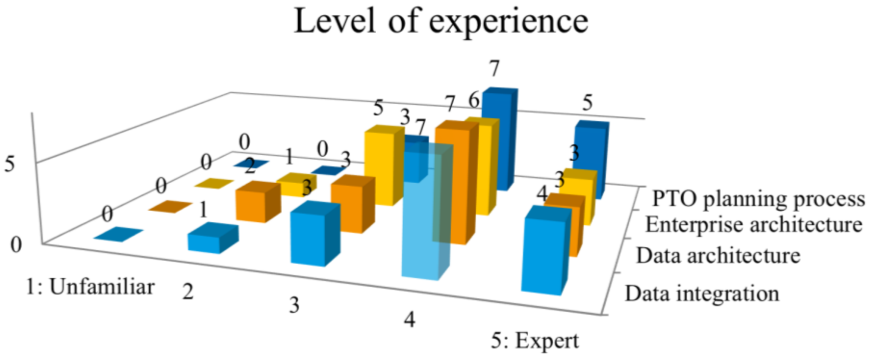

Figure 8-4 Experts' levels of experience

Some experts indicated that the name of the company they work for is allowed to be published in this research. Validation partners for this research are (in alphabetical order): BISON, DOVA, GVB, Stichting OpenGeo and Qbuzz. The description of these companies can be found in Table 8-9.

Table 8-3 Validation partners

Validation partner	Company description
BISON	Developer and manager of the Dutch public transport data standard for the exchange of (dynamic) travel information.
DOVA	Dutch partnership of the 12 provinces and three transport authorities (Amsterdam, Rotterdam The Hague and Groningen Drenthe) consisting of two clusters: public transport network and public transport data.
GVB	Urban public transport operator in Amsterdam
Stichting OpenGeo	Facilitates and stimulates initiatives that strive for public availability of geographic data, including open real-time public transport data.
Qbuzz	Urban and regional public transport operator in Utrecht, Zuid-Holland, Groningen and Drenthe.
Anonymous	Some experts declined to present their company names in this research. They have shared this in confidentiality with the researcher (except for one).

Experts working for other PTOs than the ones mentioned in Table 8-3 are included in the validation as well, which can be seen in Figure 8-5. As they did not agree on providing the company names publicly, they are not listed. However, they are known by the researcher (except for one expert, who did not provide the company name).

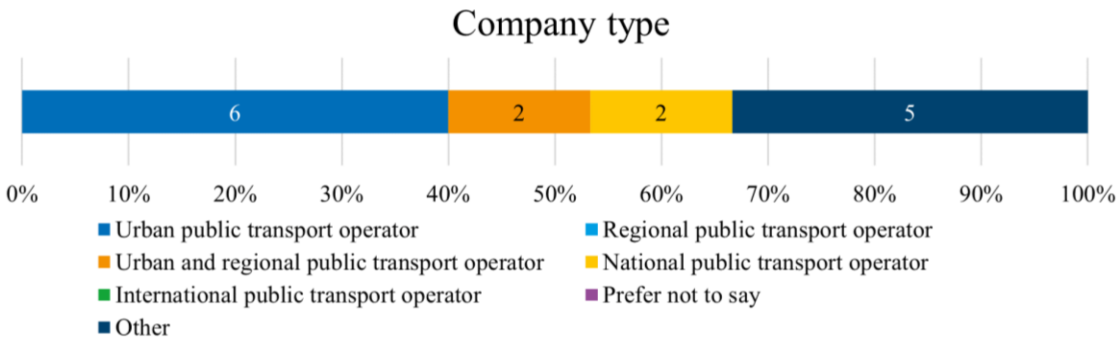

Figure 8-5 Company types experts are working at

As can be seen in Figure 8-5, 67% of the experts is working at a PTO (urban, urban and regional or national). One expert indicated 'Urban with a small amount of regional public transport also'. This answer is manually changed into 'Urban public transport operator', since the focus is on urban public transport. The different modalities offered by the companies are visualized in Figure 8-6.

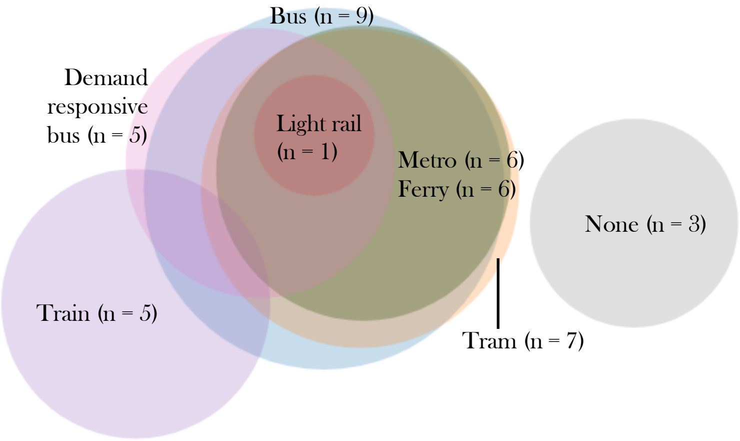

Figure 8-6 Company modalities

Combining the information from Figure 8-5 and Figure 8-6 shows that from the five respondents who stated not to be a PTO, only 3 stated that they do not offer any transport modality. From the results it can be seen that the two companies offer train services, however, one of them is not considered a PTO. For compliance with anonymity, we cannot provide more information about this.

8.3 Business and IT Content Validation

The content validation contains the validation of the business content (Section 8.3.1), IT content (Section 8.3.2 – 8.3.4) and the approach (Section 8.3.5).

8.3.1 Best-practice Planning Process

The relevancy of the proposed planning process phases is validated using a 5-points Likert scale. The planning process (Figure 4-8) and a detailed description of every planning phase were provided to the experts, after which they were asked how relevant the process phases were for the company they are working for (B1 – B9 in Appendix G). From the total of 15 validation results, 13 experts answered the questions in this part of the validation survey. As four of them are not working at a PTO, these are excluded from the answers to these questions (their expertise and feedback on the planning process are included in Figure 8-8 and its explanation). The results of the planning process phase relevancy are visualized in Figure 8-7.

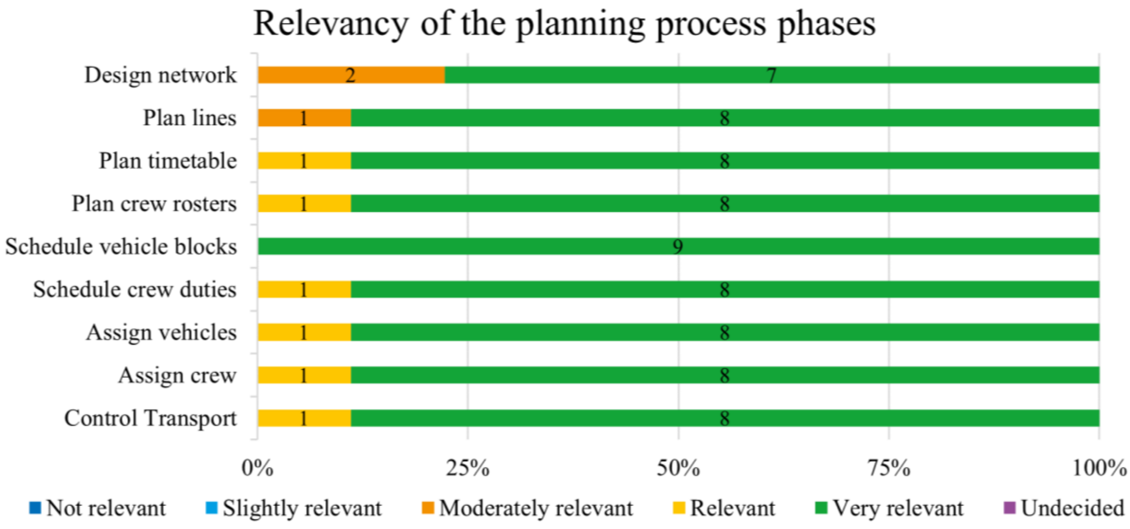

Figure 8-7 Relevancy of planning process phases

As can be seen in Figure 8-7, every planning process is – on average – scored as 'very relevant'. The planning phase 'Plan crew rosters', which is added to the operations research planning process (see Section 4.3.4), turned out to be (very) relevant according to every expert. This proofs that the planning phases included in the proposed planning process are a valid base. However, this relevancy does not mean that the phases are complete. For this, a question about the completeness of the planning process phases is included (B10 – B11 in Appendix G), which is visualized in Figure 8-8 and explained after. In this figure and explanation, all 13 opinions are taken into account.

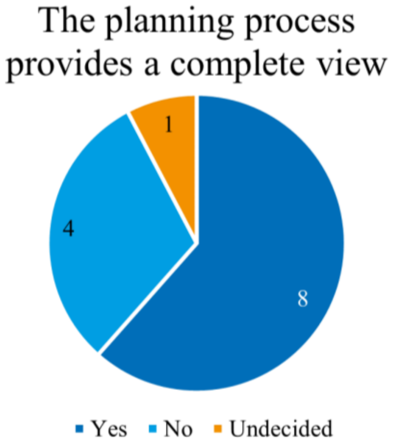

Figure 8-8 Completeness of the planning process

As can be seen in Figure 8-8, 62% of the experts indicated that the proposed planning process phases were complete and contain all necessary phases which are carried out during the planning of public transportation. Affirmative answers included but were not limited to the following explanations: "(...) required for matching the intended timetable with available resources", "all aspects (...) can be included in the steps and flows as presented" and "without one of the (...) processes the product (...) cannot be managed and controlled".

More interesting are the dissenting opinions on the completeness of the planning process phases. Table 8-4 lists the expert's suggestions, of which some are used for improving the artifact (green), some were already included in the artifact (blue) and one was not used for the improvement (purple). The improvements of the planning phases are included in Section 4.3.

Table 8-4 Expert's suggestions on the planning process phases

Expert's suggestion		**Explanation on improvement or not**
Feedback from *Schedule vehicle blocks [SV]* to *Plan timetable [PT]*.	🟩	A data flow is added between *Schedule vehicle blocks [SV]* and *Plan timetable [PT]*. The planning task *[PT2] Adjust/improve timetable* was already in place and covers the suggestion.
Include planning of bridge openings, decommissioning and depot/station/siding capacity.	🟩	The planning of detours was already included: *[PL5] Plan detours*. The scope of this planning task and its phase is broadened based on the suggestion.
Include the process of planning local (shunt) activities.	🟩	This process is added to the description and planning tasks of *Schedule vehicle blocks [SV]*.
Assigning vehicles in planning phases as well (i.e. specification of vehicle types).	🟦	This process is already covered in *Plan lines [PL]*.
Exclude network design from the planning process.	🟦	In this research, the *Design network [DN]* phase was already placed out of scope for the planning process (see Section 4.3.1).
Include opportunity scheduling.	🟦	Assuming that the expert meant opportunity scheduling e.g. vehicle charging purposes: this is already covered in *Schedule vehicle blocks [SV]*.
Include the process of generating cyclic crew rosters.	🟦	This process is already covered in *Plan crew rosters [PC]*.
Include the process of long-term demand planning for assets.	🟦	This process is already covered in *Plan lines [PL]*.
Integrate duty and vehicle planning.	🟦	The planning process phases are not integrated visually, however, data flows that imply the integration of these phases are present.
The planning process is considered as different domains, while in practice they are not.	🟦	
The deadhead trips need to be fed back to the timetable for capacity purposes (especially for rail transport).	🟦	For capacity planning, the dead runs can already be accessed. Since the timetable only contains passenger journeys in our case, we decided not to include them.
For bidding (tenders) the possible timetables should be optimized within the budgetary constraints.	🟪	We do not see this as part of this research, as it is more a feature of an application/software package rather than a common data architecture requirement. However, we believe that the data architecture contains all necessary data objects.

8.3.2 Data Requirements

The data requirements were validated on the level of their categorization (Table 7-3) and the correctness of the data objects within these categories (I1 – I4 in Appendix G). The questions about the data requirements were answered by all 15 experts. The usefulness of the categorization is measured on a 5-points Likert scale, whereas the correctness is measured by Yes/No. The results are visualized in Figure 8-9 and Figure 8-10, respectively.

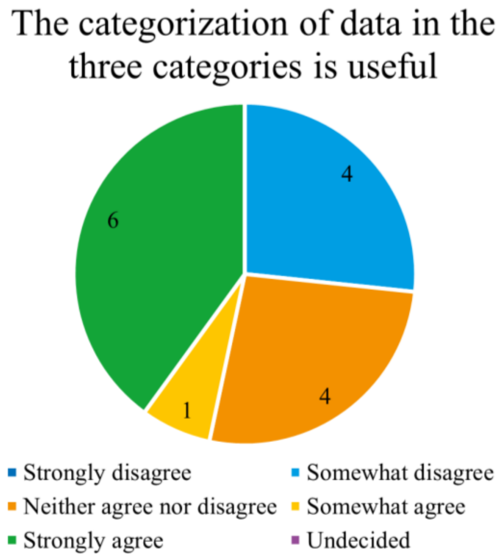

Figure 8-9 Usefulness of the data categorization

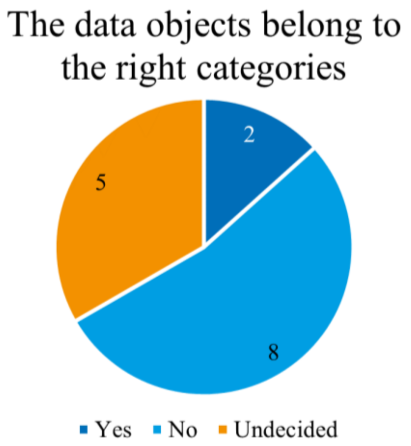

Figure 8-10 Correctness of the data categorization

Six out of 15 experts scored ‘strongly agree' and one scored ‘somewhat agree' on the usefulness of the data categorization. This means that 47% of the experts agrees that the data categorization is useful. They added to their answers that “data per domain and its responsibility is necessary for efficient organizations and systems”, “a clear distinction between input and output data is good”, “ownership helps responsibility to be taken where it should be” and “it makes dependencies on external data explicit and stabilizes the data for consumers (...)”.

Four experts (27%) stated to ‘neither agree nor disagree', whereas another four experts stated to ‘somewhat disagree' (27%) on the usefulness of the data categorization. Their suggestions are listed in Table 8-5 of which one is used for improving the artifact (green), two were already included in the artifact (blue) and another two were not used for the improvement (purple).

Table 8-5 Expert's suggestions on the data categorization usefulness

Expert's suggestion		Explanation on improvement or not
Change to two categories: input and output. Input data owned by the planning is considered ‘internal' to planning and (maybe) as output to the environment.	🟩	This suggestion is used to improve the overview of data objects (Table 7-3). Especially because the original division does not provide any extra information.
Some data objects owned by planning can also be output.	🟦	This was already stated in the design of the artifact.
Specify the data owner in case the data is not owned by the planning process.	🟦	This is already done and can be found in Appendix L.
Owned or not owned by planning is arbitrary.	🟪	No explanation of this statement was provided by the expert. The study shows that data ownership is not arbitrary.
The primary goal of the planning process is to optimize within constraints (input data). The distinction between owned/not owned by the planning process is not relevant. More relevant is whether new insights can arise during the planning process.	🟪	The business goal is to optimize the input data. However, for the design of the data integration design approach, it is important to scope which data objects are owned by the planning process and which not (see positive answers in Figure 8-9).

Many different opinions exist on the categorization of the 41 data objects to the proposed categories, as can be seen in Figure 8-10. However, most of the differences are assumed to be present due to the different data objects interpretations. A description of every data object was provided to the experts. As can be seen from the number of clicks on the link to this document, we know that not every expert has checked the definitions. For this reason, some of the suggestions are rendered invalid after careful consideration and are therefore not included. All other suggestions are listed in Table 8-6, of which some are used for improving the artifact (green), some were already included in the artifact (blue) and others were not used for the improvement (purple).

Table 8-6 Expert's suggestions on the data categorization

Expert's suggestion		Explanation on improvement or not
'Fare information' (output) and 'Pricing information' (input) should change places.	🟩	This was overlooked and is changed in the final design (Table 7-3).
Infrastructure should also be an input category now owned by the planning process.	🟩	The data object infrastructure is included as input.
Runtime analysis is something that planning should own and execute by heart.	🟩	This is correct. The intended data object embodies the operational information on runtimes (actual driving times). The data object name is changed into 'Actual runtimes'.
For the input not owned by planning and output you provide an enormous range of (vague) options. Maybe it would be better to focus this on the essentials.	🟦	This research has focused on the essentials by staying at a higher level. We believe that the data objects taken into account in this research provide the best high-level overview and are therefore the essentials.
Spread in driving times on road sections should be added as a data object.	🟦	This is considered to be part of the line plan (definition in Appendix M).
Under input data owned by planning you should also put the available crew.	🟦	The available crew is considered to be known by the combination of crew (assumed to be owned by HR) and crew rosters.
Doesn't relief point also partially belong to the not owned by planning process category (regarding the physical geo location)?	🟪	For this research, we have chosen to identify the planning process as the owner of this data object, since they make the physical location a relief point (definition in Appendix M). One can say that geographical data needs to be input for this, which is covered in facilities in this case.
For train operators the categorization is incorrect and incomplete.	🟪	Unfortunately, no explanation for this statement is given by the expert.
Infrastructure is not owned by the planning, but is a result of network design.	🟪	The data object as presented in the validation survey was intended to contain PTO-specific information, and is therefore categorized as owned by the planning.
There are two different types of crew: 'own' crew and extra crew hired via an employment agency	🟪	This is true but on the level of the proposed artifact this difference is not significantly important.

8.3.3 Data Access Service

The design of the data access service was shown and explained to the experts (Figure 7-6). The questions about this service (I6 – I11 in Appendix G) were answered by all 15 experts. Three questions were provided and the experts were asked to indicate their agreement on a 5-point Likert scale. Questions were about the application functions, responsibilities of the application functions and the solution in general.

The first question was answered positively (except for one expert, who was not familiar with the ArchiMate language), which means that the different application functions of the data service are clearly shown. The second question was less positive, yet still, it was relatively clear to the experts where the responsibilities are located (i.e. the business process (planning) is responsible for the IT functions, they are the owner). The third question was answered the most divergently by the experts, yet still more positive than negative answers. The results are visualized in Figure 8-11.

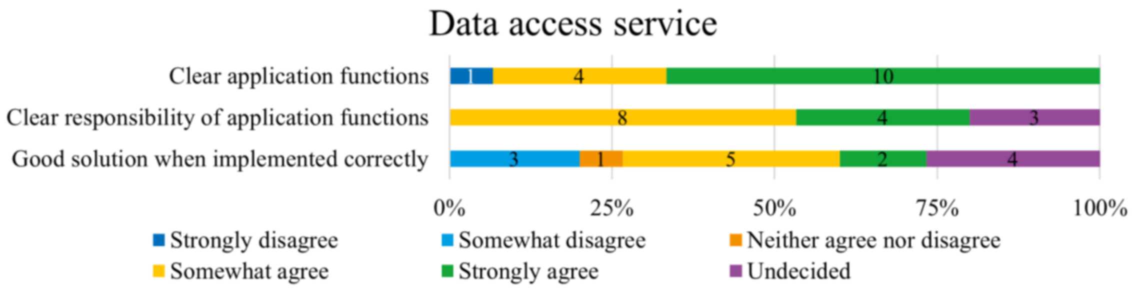

Figure 8-11 Data access service validation

Expert's suggestions related to the data access service are listed in Table 8-7, of which one is used for improving the artifact (green), some were already included in the artifact (blue) and two were not used for the improvement (purple).

Table 8-7 Expert's suggestions on the data access service

Expert's suggestion		Explanation on improvement or not
For performance reasons, it is better not to communicate via the data access service internally within the same software package.	🟩	Within the planning phases and/or software packages it is indeed proposed to communicate directly within the same package, which is added.
An integral data layer would make application life cycle management more difficult (e.g. replacing applications/application components). Integration on services level looks more appropriate.	🟦	A more modular approach has its advantages. However, since the degree of reuse, changes and enrichments of the data objects is rather high and takes place by several roles, a modular approach is more difficult regarding transactions and authorizations. Since the proposed design can still be split into several services, we added this to the artifact explanation (Section 7.2.3) and consider is it an implementational choice. It is included in future research recommendations (Section 9.6).
More service-oriented on the level of responsibilities instead of one data access service layer.	🟦	
Now you have a single point of failure and a direct vendor lock-in. A decentralized approach that standardizes interfaces is likely to be more future-proof.	🟦	
A modular approach with e.g. entity-group data services. The relationships between these entity-groups are managed centrally (i.e. vehicles, crew, lines, infrastructure).	🟦	
Consumers should access data via an application layer and not directly via a database.	🟦	This is already included in the design, however, not visualized as this is highly PTO-dependent.
All Dutch PTOs use Hastus[6], which can combine this in one application.	🟪	Besides the fact that this statement is not true, the data access service design can still be applied to an application such as Hastus[6], see Section 7.3.1.
Individual data layer per application. This makes lifecycle management easier.	🟪	This might be an implementation choice, but can't be provided in this research since every application landscape is different and combines different planning phases and tasks.

8.3.4 Data Quality Aspects

The data quality aspects are validated by asking the experts to score the relevancy of every data quality aspect from the ISO 25012:2008 standard (ISO, 2008) to the two scenarios as identified in Section 6.3 (I13 – I44 in Appendix G). These two scenarios are – shortly described – static planning and dynamic planning. Details about the two scenarios can be found in the aforementioned section.

It turned out that most of the data aspects are stated to be more relevant in a dynamic planning environment than they are in a static planning environment, as can be seen in Figure 8-12. This is in line with the assumptions in Section 6.4.

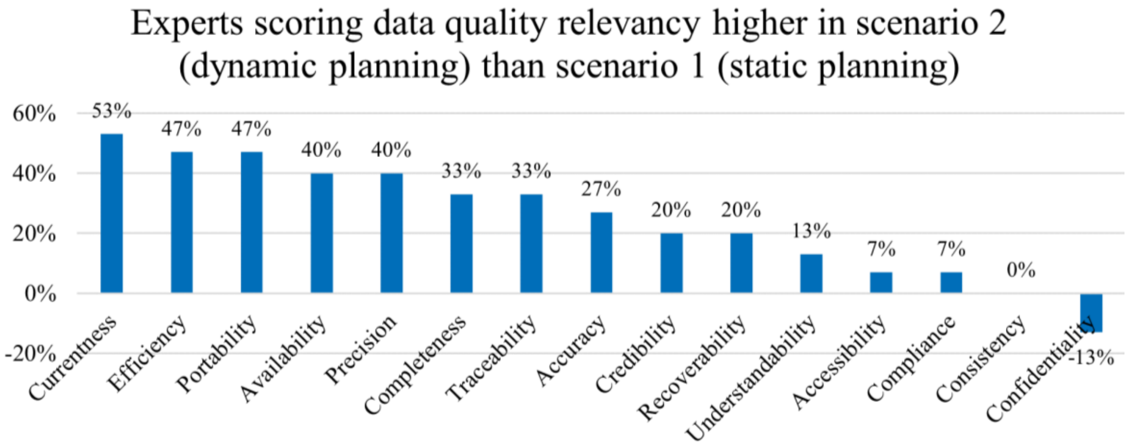

Figure 8-12 Data quality relevancy differences between two scenarios

According to the experts, currentness is the data quality aspect that needs the most attention if you compare scenario 2 (dynamic) with scenario (1): more than half of the experts (53%) answered a higher relevancy for scenario 2 than they did for scenario 1. Interestingly, consistency remained equally relevant for both scenarios, whereas confidentiality decreased in relevancy. When the percentages are lower, it does not mean that the data quality aspect is not relevant, it means that the relevancy of the data quality aspect is more equal for both scenarios. The proposed data integration design approach accounts for most of the aspects which are scored highly relevant, which matches the experts' view on data quality relevancy.

As explained in Section 7.2.2, the target data architecture is divided into three domains: journey, vehicle and crew. For data compliance purposes, it is proposed to use the NeTEx standard for the data exchange within the journey domain. Experts were shown the data architecture (Figure 7-2) and asked whether this is possible, and were also asked whether the standard is applicable to the other two domains of the data architecture (I46 – I52 in Appendix G). Only the answers of the experts who stated to know NeTEx – which are 9 out of 15 – are included in the results in Figure 8-13 and thereafter.

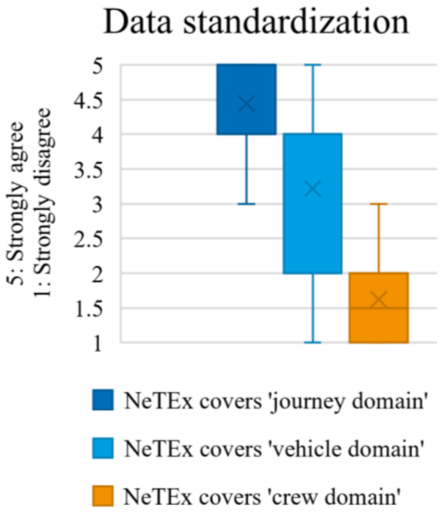

Figure 8-13 Data standardization in NeTEx

As can be seen in Figure 8-13, the NeTEx standard is only applicable to the journey domain. Parts of the vehicle domain are possibly covered, but experts disagree about the applicability of NeTEx on the crew domain. This is in line with the proposed artifact. An important finding is that within the railway business other standards than NeTEx are used: TAP TSI (passenger) and TAF TSI (freight). Experts were also asked to think about a solution for the two domains which are not (entirely) covered by NeTEx. Their reactions are listed in random order below:

- Transmodel contains (almost) all objects of the three domains, NeTEx does not (vehicle and crew). Start a workgroup to define the gaps between Transmodel and NeTEx/SIRI regarding the crew domain.
- The vehicle domain is partly included in SIRI.
- Internally within the planning domain: data model of the software.
- Externally outside the planning domain: develop and agree on a standard.
- Define the data model based on Transmodel.
- Why do you want to exchange this information? I am not going to exchange crew data.
- Request extensions to NeTEx. In any case stay on a Transmodel compliant standard.
- Important to agree upon a standard for now and the future.
- An alternative can be to look at the German VDV standard.
- Public transport is a slow-changing 'vehicle' because of heavy national interests. That hinders international standardization and availability of commercial off-the-shelf (COTS) systems.
- Data standardization cannot be reached through IT, but must be reached through cooperating organizations/businesses.

8.3.5 Approach

From the total of 15 validation results, 11 experts answered the questions in this part of the validation survey. Experts were asked to their agreement on two statements about the CRUD-matrix (A1 – A4 in Appendix G). In this first question, experts were questioned whether they agree that a CRUD matrix is suitable for defining data responsibilities on planning phase level. The second question was about to what extent they agree that such a matrix helps to identify the data owner. The results are shown in Figure 8-14.

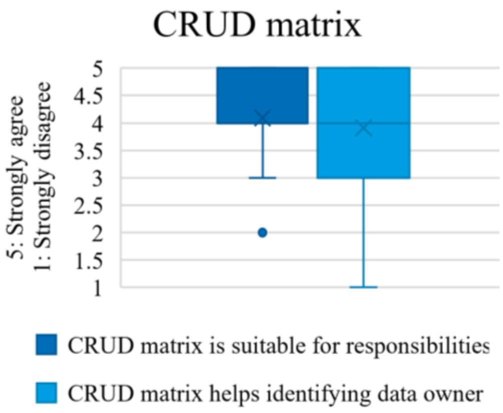

Figure 8-14 CRUD matrix validation

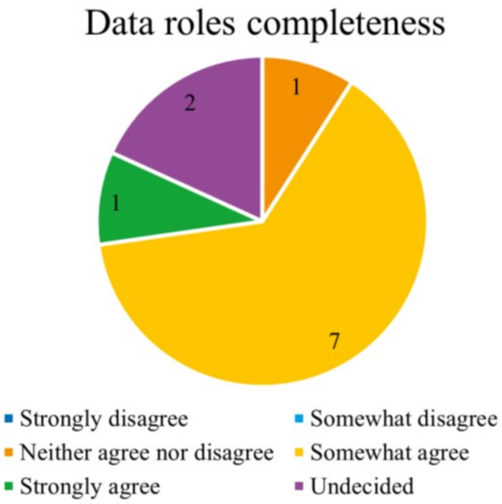

Figure 8-15 Data roles validation

According to the experts, a CRUD-matrix can be used to identify data responsibilities on planning phase level. However, for identifying the owner, a CRUD-matrix is not always useful. This is because "multiple planning phases might want to create or edit the data" and "an actor is not immediately the owner". Two experts said that it could help to identify the owner. One of them suggested using the ArchiMate business layer.

The data roles figure (Figure 7-8) was presented to the experts and they were asked about the completeness (A5 – A6 in Appendix G). The results are visualized in Figure 8-15. Nobody disagreed about the completeness of the data roles. Eight out of 11 agreed strongly or somewhat, which is 73%. One expert stated to neither agree nor disagree, while two experts stated to be undecided. Extra comments experts gave to this questions were: "the distinction between roles is not made so strict in practice. Ideally there is one officer for every role, in practice this is expensive.", "employees with knowledge of data management and material knowledge are scarce", "the process and roles are rather complex, especially considering that smaller PTOs run their business from data owner to data consumer" and "within the scheduling domain, most of the data is mandated (by EU law) to be open data".

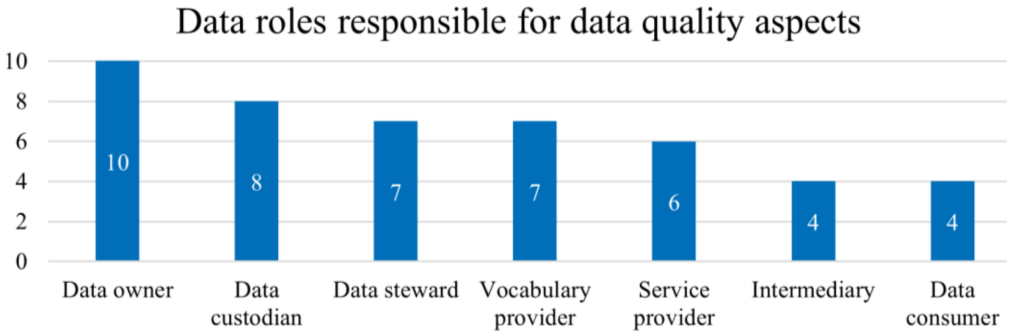

Figure 8-16 Data roles responsible for data quality aspects

The last question about data ownership and responsibility (A7 in Appendix G) covered the previously identified data roles. As 11 experts answered this question, the data owner was seen by almost every expert (10) as the role which – among others – is responsible for the data quality aspects. Data custodian was indicated as responsible relatively often (eight out of 11). The other roles also seem to be important according to the experts, however, their importance declines. Two experts answered that every role is responsible for the data quality, one of them added that "all roles have an impact on the overall data quality". The responsibilities of every role are provided in the proposed approach.

8.4 Artifact Goal Validation

To test whether the designed artifact would contribute to the goals (as listed in Section 7.1), the experts were asked to imagine the designed artifact within the context of a PTO's planning process (V1 – V28 in Appendix G). The goals can be divided into three parts, which are: data, approach and business process improvements. From the total of 15 validation results, 11 experts answered the questions in this part of the validation survey. The results are shown and discussed in the next subsections. If not all 11 experts answered on a particular goal, it is indicated in the text. This means that one or more experts answered 'Undecided', which are not taken into account in the figures.

8.4.1 Goals regarding Data

The scoring of the goals regarding the PTO's data situation is visualized in Figure 8-17. On average, the experts think that the artifact contributes positively to the goals regarding the data situation.

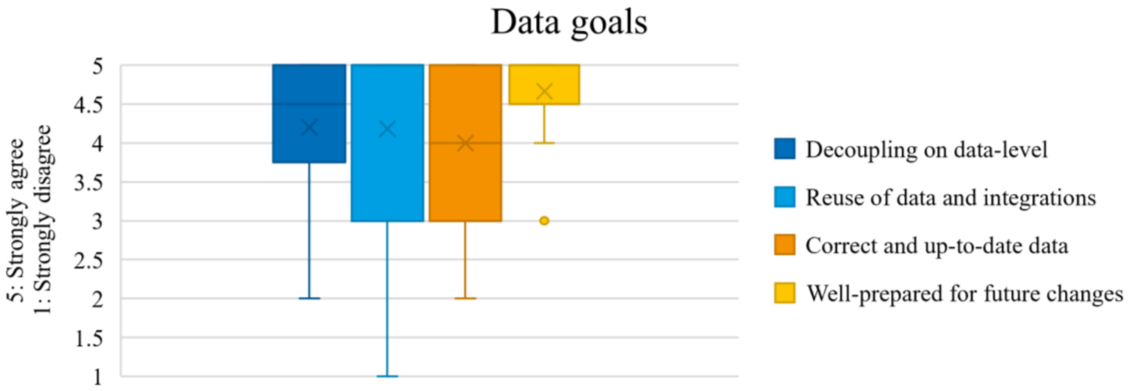

Figure 8-17 Goal validation regarding the data situation

8.4.1.1 Decoupling on Data-level

Ten experts answered this question. On average the result is quite positive. The expert who scored 'somewhat disagree' says that "data providing and data consuming happen in two separated universes, and are therefore already loosely coupled". On the contrary, the experts were rather positive about the extent to which the data integration design approach would contribute to the decoupling on data-level. Necessary are however: standardization, data ownership and enforcing the data transfer via the data access service. One expert said that for a brownfield situation the target architecture with decoupling is hard to reach. Another expert quested whether decoupling is always necessary, but often used data should become and be managed as master data for the entire enterprise.

As the research has shown, decoupling is beneficial for a more flexible IT landscape. Standardization and data ownership are accounted for in the data integration design approach, which satisfies the experts' considerations. Enforcing the data transfer through the data access service is inherent to the design, but indeed should be enforced during and after implementation.

8.4.1.2 Reuse of Data and Integrations

Eleven experts answered this question. Just like the previous goal, the average result is quite positive. One big outlier to 'strongly disagree' is present. This expert stated that: "reuse is driven by making data and half-fabrics available (...)". As this is true from a practical point of view, the proposed target data architecture shows how this can be made available. Other experts were either neutral or strongly agreed. They said that "planning would serve as a hub rather than information chains", "number of integrations can be reduced/created uniformly", "depending on the application, which is selected based on functional requirements and not on data reusability", "never-ending data copying should stop", "it facilitates a consistent definition of data(structures) and interpretation, thus easier to reuse", "reuse can be reached, but at the cost of lower flexibility".

It is acknowledged by experts that reuse of data and integration is made possible by the design, however, flexibility on the level of the data architecture indeed decreases (compared to separate application-specific data models).

8.4.1.3 Correct and up-to-date Data

Eleven experts answered this question. The goal to use more correct and up-to-date data is scored relatively positive. Experts state that with the proposed data integration design approach, it is likely that “data quality of the source can be guaranteed by the data access service”, “modeling the complete flow of data minimizes the risk of conflicting data” and “it would be at the cost of lower flexibility”. Again, this goal is acknowledged, but results in the same type of inflexibility as explained for the previous goal. However, the advantages probably even out the inflexibility.

8.4.1.4 Prepared for Future Changes

Nine experts answered this question. From all artifact goals, this goal is scored the highest by the experts. Also, no real outliers are present in the results, the minimum score is ‘neither agree nor disagree’ (see Figure 8-17). Experts state that “this is a huge benefit, since the company will have a portfolio of data to share with new applications. It will lower costs and effort for integration.”, “data management in general is an investment for future functionalities”, “by modeling the data flow and other processes it is easier to (consistently) insert new elements” and “preparing for the future is typically realized through open/modular design and standards-based interfaces”. Next to these positive findings, one expert was a bit more critical and stated that “it is nice that everyone could talk together, but this does not mean that new integrations would not use different newly invented terms for prepared views that are not foreseen at the moment”. This is correct, but we think (and experts agree) that the data integration design approach will facilitate this process.

8.4.2 Goals regarding the Approach

The scoring of the goals regarding the approach is visualized in Figure 8-18. On average, the experts think that the artifact contributes positively to the goals regarding the approach.

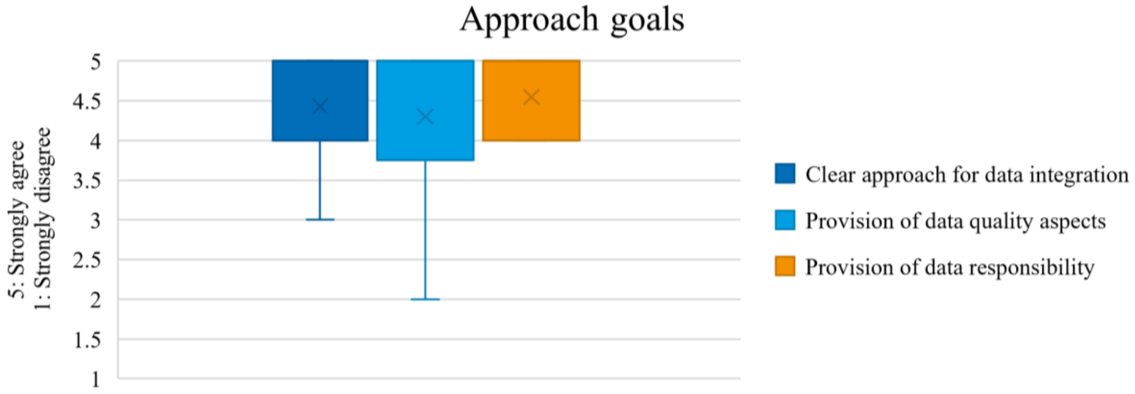

Figure 8-18 Goal validation regarding the approach

8.4.2.1 Clear Approach for Data Integration

Seven experts answered this question, the other experts commented that the question was not clear to them. The seven experts who answered the question provided comments on their answer as well, which showed that they understood the question the way we intended. The result is quite positive, however, some remarks were made by the experts. It could serve as a “template”, “starting point” and “guide” for improving the data integration situation. For brownfield situations, which are all PTOs, a “transitioning approach would be useful”. Another expert stated that “also monitoring and implementation (software and processes) should be easier”. Lastly, one expert said that “executing the design approach and creating the design itself will give new insights in how this can be accommodated by a somewhat altered approach”, which means that the approach could improve iteratively.

Since the data integration design approach consists of a target architecture, it is intended to be an end-state. The design can indeed be improved after implementation shows that this is required. However, since the implementation is out of scope for this research, we do not focus on that.

8.4.2.2 Provision of Data Quality Aspects

Ten experts answered this question. The mean is relatively high. Experts mostly agree and state that "having data at one place and reduce data duplication will lead to higher quality", "correct attention and focus on data management contributes to higher data quality" and "the data integration design approach provides consistency of data". "Getting the non-functional requirements on data quality correct at the beginning is crucial for benefits and costs", according to an expert. One expert, however, stated that they "do not see how quality reports flows back to data owners". This led to an improvement of the artifact (Section 7.3.3), in which data roles and their data quality responsibility became more clear.

8.4.2.3 Provision of Data Responsibilities

Eleven experts answered this question. Within the goal validation of the approach, this goal is scored the highest by the experts. Answers added by the experts are "only if everyone would agree on the roles", "data origin is known, which makes the responsibility clear", "the design makes the responsibilities explicit" and "being clear on responsibilities is the foundation for addressing concerns later on". This shows that the proposed artifact contains clear responsibility aspects on data entity level.

8.4.3 Goals regarding Business Process Improvements

The scoring of the goals regarding the PTO's improved business processes is visualized in Figure 8-19. On average, the experts think that the artifact contributes positively to the goals regarding business process improvements.

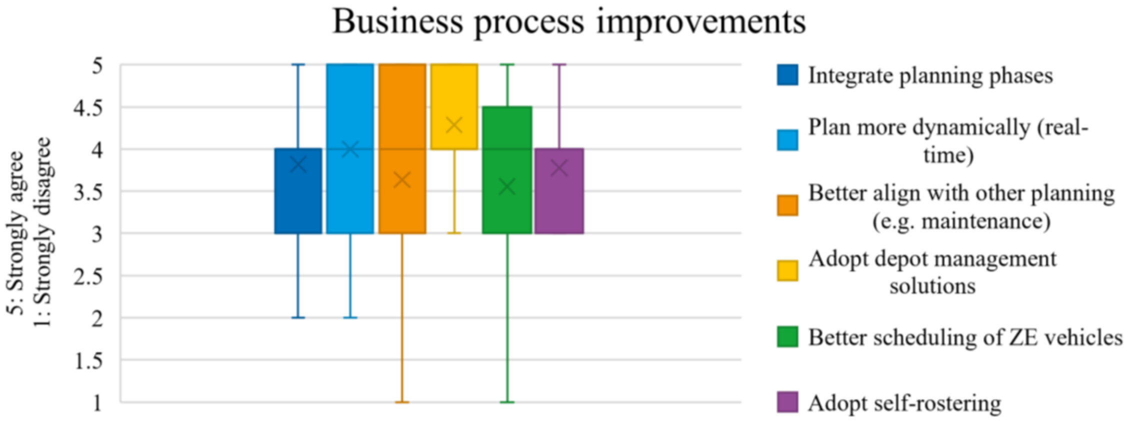

Figure 8-19 Goal validation regarding the business process improvements

8.4.3.1 Integration of Planning Phases

Eleven experts answered this question. Most of the experts 'somewhat agree' with this goal. They state that: "the data integration helps, but still would require the tool that can manage it", "it would be difficult to allow all kind of transactions during planning optimization" and "data dependencies and the overall picture have been made explicit". One expert ('somewhat agree') said that "In brown-field organizations optimization can only gradually be achieved as the costs and risks of total business process redesign is typically beyond what the organization can and will handle (…)". The data integration design approach can serve as a target to gradually move towards.

8.4.3.2 Dynamic (Real-time) Planning

Eleven experts answered this question. One of them said to 'somewhat disagree', because "this depends highly on the data provider's business processes". Two stated to 'neither agree nor disagree'. One of these two answered that "it allows more applications to be in sync near real-time, but it requires scalable solutions (e.g. for database access)". The agreeing experts said that "changes can be adopted more frequently" and "it is easier (safer) to make consistent (!) changes in the process". It is acknowledged that having the data in one place, would help planning more dynamically, with the precondition that technical issues are tackled during implementation (e.g. scalable data(base) access).

8.4.3.3 Align with Maintenance Planning

Eleven experts answered this question. The reactions were very different, therefore this goal is on average not scored positively but stays somewhat in the middle. Experts' reactions range from "vehicle management should use the timetable as input" (no alignment), "depends on the way vehicle maintenance is carried out" and "planning can already take into account mileage-based scheduling" to "data is available to more applications, thus maintenance planning can also use it". We believe that the general goal of providing data to other business areas that handle other planning tasks can be reached. This also depends on the implementation of the other planning tasks.

8.4.3.4 Adopt Depot Management Solutions

Seven experts answered this question. The result is quite positive, however, it is a limited number of responses. As experts indicated, standardization and real-time interfaces are of high importance to reach this goal, yet the proposed artifact contributes to this.

8.4.3.5 Better Scheduling of Zero-Emission Vehicles

Nine experts answered this question. The average answer among experts is neutral. The agreeing experts state that the provided data integration design approach is essential, whereas other experts note that charging ZE-vehicles is 'simply' a pre-condition for the planning. The added value of the data integration design approach is that it actually states that this data is necessary, but it does not improve the scheduling itself, according to one expert. Another expert says that this is a functional use case for the data integration design approach. This is true, since that use case is one of the enterprise's goals.

8.4.3.6 Adopt Self-rostering

Nine experts answered this question. On average, the result for contributing to this goal is positive. Experts state that with easier access to data and clear dependencies self-rostering can be implemented with less risk. One expert states that this is one of the functional use cases of having a proper data integration situation. Data provision is key in this process, which is included in the proposed design, as explained before for other goals.

8.4.4 Goals in General

The experts were asked whether they can imagine other positive and negative influences of the proposed data integration design approach. These are listed (in random order) below.

Positive influences

- General focus on data quality and quality of travel information.
- The data integration design approach makes it easier to explain to new people.
- It would bring public transport operators (and their IT departments) to a better information position, especially for organizations that do not have a clear architecture in place.
- Savings in employee costs for the departments responsible for integration.

Negative influences

- An approach like this could be nice for a starting company, but I do see not enough advantages to implement it in a complex running environment.
- In certain cases, especially with high performance and highly integrated applications, this will lead to overhead. It could even be unacceptable in certain cases. But I think in general this is a good approach.
- It encompasses a lot, so in practice probably a company will not implement/adhere to everything.
- Pessimistically, the data integration design approach leads to a monoculture of a single software vendor. This does not stimulate innovation within the organization.
- Requires a long implementation time. Needs commitment from management and a long breath. A risk that support will decline during the process because it takes too long to produce results.

8.5 Acceptance and Use Validation

For the validation of the acceptance and future usage of the data integration design approach, the Unified Theory for Acceptance and Use of Technology by Venkatesh et al. (2003) is used. This unified theory is a combination of several technology acceptance models in the field of IT acceptance research and is shown in Figure 8-20.

The model consists of four constructs that play a significant role as direct determinants of user acceptance and usage behavior: performance expectancy (PE), effort expectancy (EE), social influence

(SI) and facilitating conditions (FC) (Venkatesh et al., 2003). As can be seen in Figure 8-20, the moderating variables for the constructs are gender, age, experience and voluntariness of use (VU). The moderating variables are not analyzed in this research, since the number of responses is too low to investigate these relations.

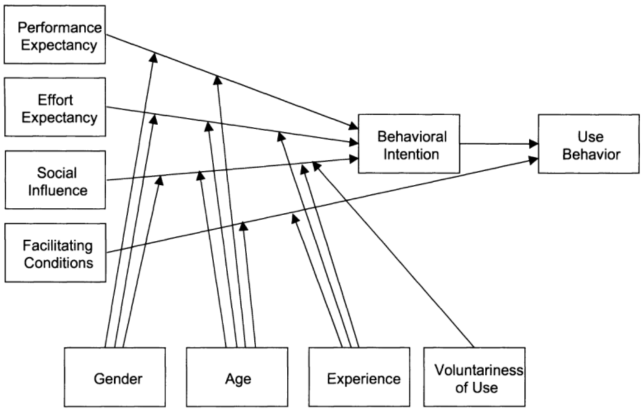

Figure 8-20 Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003)

Every construct and moderating variable is measured using several validated statements which are presented to the experts. They were asked to score the statements on a 5-point Likert scale, reaching from 'Strongly disagree' to 'Strongly agree'. The questions were slightly adapted to match the context and purpose of this research (e.g. replacing 'system' with 'data integration design approach'). In case the experts did not want or could not answer, they could answer 'Undecided'. Nine experts filled out the UTAUT validation.

The UTAUT validation can be found in Appendix G, while the results are shown in Appendix P. In the remainder of this section, every construct is shortly described, the statements are shown and the results are presented.

8.5.1 Performance Expectancy

The degree to which an individual believes that using the data integration design approach will help him or her to reach gains in job performance is called performance expectancy (PE) (Venkatesh et al., 2003). This construct indicates how useful the artifact would be for the respondent. Moderating variables for this construct are gender and age. Generally, the influence of performance expectancy is found to be higher for younger men (Venkatesh et al., 2003). The statements to measure this construct are listed in Table 8-8.

Table 8-8 Performance expectancy statements

ID	Statement
PE1	I would find the data integration design approach useful for my job.
PE2	Using the data integration design approach enables me to accomplish tasks more quickly.
PE3	Using the data integration design approach increases my productivity.

Results for performance expectancy are shown in Figure 8-21 and explained thereafter. PE1 is scored by nine experts and PE2 and PE3 are answered by eight experts.

As can be seen in Figure 8-21, it is found that the performance expectancy for the use of the data integration design approach provided in this research is relatively high. Experts see the added value of it. Extra comments made by the experts are: "integrated data design can facilitate clear agreements about ownership and responsibility of data, which is very important for an efficient organization and system landscape (...) but apart from having the overview, insight and understanding of orchestration is very important as well to support the business" and "the acceptance of the approach is the crux, it should be supported among users".

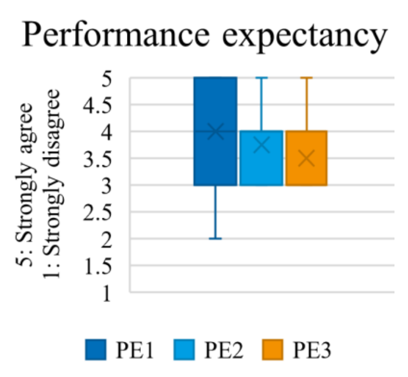

Figure 8-21 Performance expectancy results

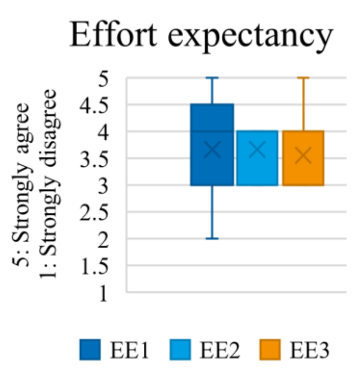

Figure 8-22 Effort expectancy results

8.5.2 Effort Expectancy

The ease associated with the use of the data integration design approach is defined as effort expectancy (EE) (Venkatesh et al., 2003). This construct indicates how much effort the respondent thinks it would cost to use the artifact. Moderating variables for this construct are gender, age and experience. Generally, the influence of effort expectancy is found to be higher for women, older workers and those with limited experience (Venkatesh et al., 2003). The statements to measure this construct are listed in Table 8-9.

Table 8-9 Effort expectancy statements

ID	Statement
EE1	My interaction with the data integration design approach would be clear and understandable.
EE2	It would be easy for me to become skillful at using the data integration design approach.
EE3	I would find the data integration design approach easy to use.

Results for effort expectancy are shown in Figure 8-22. The questions are answered by nine experts. As can be seen, the effort necessary to use the provided data integration design approach is scored a bit above average, but not far. This means that the experts were a bit more positive than neutral ('neither agree nor disagree' and 'somewhat agree') about the effort expectancy necessary to use the artifact. For the last question (EE3), one expert stated that it would only work with "more than one prophet within the organization".

8.5.3 Social Influence

Venkatesh et al. (2003) describe social influence (SI) as the degree to which an individual perceives that important others believe that they should use a technology artifact. Moderating variables for this construct are gender, age, experience and voluntariness of use. Generally, the effect of social influence is stronger for women, older workers, under conditions of mandatory use and with limited experience (Venkatesh et al., 2003). The statements to measure this construct are listed in Table 8-10 and the results shown in Figure 8-23. SI1 and SI2 are answered by six experts and SI3 and SI4 by eight experts.

Table 8-10 Social influence statements

ID	Statement
SI1	People who influence my behavior think that I should use the data integration design approach.
SI2	People who are important to me think that I should use the data integration design approach.
SI3	The senior management of the company I work for will be helpful in the use of the data integration design approach.
SI4	In general, the organization would support the use of the data integration design approach.

As can be seen in Figure 8-23, the questions regarding the social influence for the user of the data integration design approach are scored quite differently, especially SI2. The average of the questions rests in the middle of the 5-points Likert scale, which lets us conclude that the social influence is highly dependent on the expert, and thus the company and function/roles. No difference between male and female could be found, since no women answered this question. One expert added the following information: "within the PTOs I work with, I see a direct (negative) influence of IT departments wanting to 'own' the entire business, and not in a facilitating role as 'data custodian'". The data integration design approach could help to solve this problem.

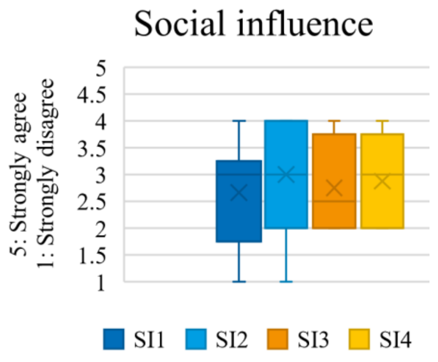

Figure 8-23 Social influence results

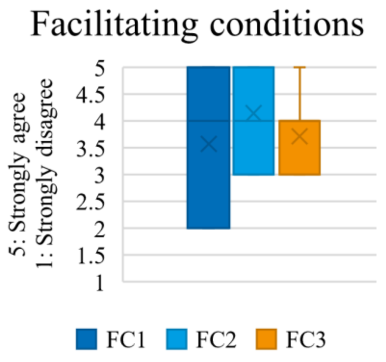

Figure 8-24 Facilitating conditions results

8.5.4 Facilitating Conditions

The belief of whether or not the organizational and technical infrastructure exists to support the use of the data integration design approach is called facilitating conditions (FC) (Venkatesh et al., 2003). Moderating variables for this construct are age and experience. Generally, the influence of facilitating conditions is higher for older workers with increasing experience (Venkatesh et al., 2003). Kleinreesink (2017) proved that facilitating conditions influence user adoption the most. The statements to measure this construct are listed in Table 8-11.

Table 8-11 Facilitating conditions statements

ID	Statement
FC1	I have the resources necessary to use the data integration design approach.
FC2	I have the knowledge necessary to use the data integration design approach.
FC3	The data integration design approach is compatible with other architectures, methods and/or frameworks I use.

Results for facilitating conditions are shown in Figure 8-24. All questions are answered by seven experts. As can be seen, except for FC1, the facilitating condition questions were not answered negatively. This means that – in general – the experts are either neutral or agree with the fact that the data integration design approach can be used within the companies they work for.

8.5.5 Behavioral Intention

Behavioral intention is the intention of a user to use an IT artifact. It significantly predicts the usage behavior of an IT artifact (Venkatesh et al., 2003). Behavioral intention is a dependent variable and depends on performance expectancy, effort expectancy and social influence (as can be seen in Figure 8-20). The statements to measure this construct are listed in Table 8-12. The results are shown in Figure 8-25 and explained thereafter. BI1 and BI3 are answered by six experts and BI2 is answered by seven experts.

Table 8-12 Behavioral intention statements

ID	Statement
BI1	I intend to use (part of) the data integration design approach within the next 2 years.
BI2	I predict I would use (part of) the data integration design approach in the next 2 years.
BI3	I plan to use (part of) the data integration design approach in the next 2 years.

As can be seen in Figure 8-25, the answers to the questions about behavioral intent are answered very differently. However, the biggest parts of the boxes are located on the positive side of the figure. This means that most of the experts think that they will use or plan to use the data integration design approach with the next two years.

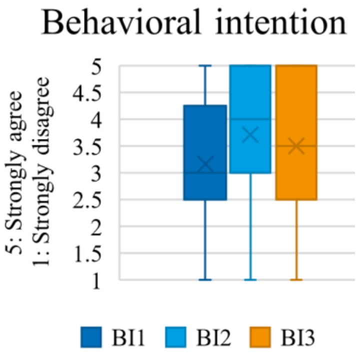

Figure 8-25 Behavioral intention results

8.6 Generalization

Generalizability can be seen as the application of findings to other contexts (Noble & Smith, 2015). The goal of this research was to design a data integration design approach for a PTO. The findings consist of the validated target data architecture and the approach, which includes important considerations for implementing the architecture. Generalization can be either case-based or sample-based (Wieringa, 2014). For this research, the generalization will be case-based, in line with the method explained in Section 2.6.2.

As turned out during the validation, within the railways' sector other standards next to NeTEx exist, which are the TAP TSI (passenger) and TAF TSI (freight) standards. These standards are different compared to NeTEx and the conceptual model Transmodel. However, similarities of the TSI standards and Transmodel and NeTEx were also found by a study of the European Union (Bourée et al., 2019). Experts working within the national railway sector stated that the proposed design is not entirely correct and complete for their sector.

Generalizing the data integration design approach is possible by defining a case-based generalization. After validation it turned out that the proposed artifact can be applied to every as long as the following conditions are met:

- The PTO offers public transport services according to the definition stated in Section 1.1.2.
- The PTO understands data integration the way it is formulated in Section 1.1.4.
- The PTO's planning process is similar to the proposed planning process in Section 4.3.
- The PTO's processes and data adhere to the conceptual model Transmodel.
- The PTO's data format adheres to the data standard NeTEx.

Applying these conditions to the PTOs of the validation is done by first grouping their contexts:

- Urban PTO (any modality except train-only; 6 experts from 2 enterprises)
- Urban and regional PTO (any modality except train-only; 2 experts from 2 enterprises)
- National PTO (any modality except train-only; 1 expert)
- National train-only PTO and enterprise dealing with train-only business (3 experts from 2 enterprises)

When matching the conditions to the context presented above, it becomes clear that the data integration design approach can be generalized to urban, urban and regional and national PTOs, for any modality (modality-wide), as long as they do not exclusively provide (national) train services. In this latter case, the processes are expected to be too different, the standard is not (entirely) based on Transmodel and NeTEx is likely not to be the preferred data standard. It is difficult to generalize the data integration design approach to a context outside the public transport business, since the processes and data requirements are very unique and the Transmodel conceptual model and NeTEx data standard are specifically designed for public transport purposes.

9 CONCLUSION

This chapter discusses the results and conclusions on the research questions. In Section 9.1, the findings of this research are highlighted and the relation to other research and developments is discussed. Subsequently, in Section 9.2, the conclusions of this research are presented based on the research questions. In the remainder of this chapter, the contributions to academic research, contributions to practice and GVB, limitations and future research recommendations are addressed.

9.1 Relation to other Research and Developments

The relation of this research to other research and developments is presented in this section. These topics have been divided into four sections.

9.1.1 Coverage of PTO's Planning Process

The data requirements provided in this research are based on both literature and interviews. Compared to the extensive public transport (planning) studies from Ceder (2016) and Scholz (2016), the current research provides a more holistic view of the planning process and its external providers and consumers. Furthermore, the step from operations research towards more tangible assets are made by the provisioning of the target data architecture and approach.

Although the proposed design does not take into consideration the operational phase *Control transport*, the data flows between this phase and the planning process were defined. Most of the real-time data comes from the operations, and planning needs to be adapted based on this input. The tasks for vehicle and crew disruptions (two of the three public transport triangle aspects of Figure 4-9) are delegated back to the planning process. The third aspect of the triangle, journeys, is handled by the control center directly. Concludingly, even though operations was placed out of scope, the integrations with the other planning phases are accounted for.

9.1.2 Artifact in Practice

The designed architecture is a target architecture and can therefore be considered an end-state of a PTO's data management improvement project. The proposed architecture cannot be put directly into practice, since a brownfield application landscape often needs to be accounted for. This is further discussed in Section 9.6.4.

Having proper data integration in place for the planning process does not mean that the entire PTO's data integration situation is improved. Problems are foreseen when data must be integrated with data from other sources than the planning domain, which is still based on the older data integration situation (point-to-point integrations, data chains, several truths for one data object, etc.). Therefore, the proposed architecture and approach lead to a partial data integration solution for PTOs. The proposed design, however, can help to improve the data situation for other business areas, as the process of designing a target data architecture can be used. Moreover, the approach also contains considerations for data management in general which can be used for other domains as well.

9.1.3 Data Quality and Reuse of Data

Data quality turned out to be very important and this research showed that it can directly help to mitigate data integration challenges identified in practice and literature. Having data of high quality will ensure that PTOs can make better decisions. This is very important within the PTO's planning process, as the development of the public transport network, timetables, vehicle blocks and crew duties are all based on (historic) data. Consequently, monitoring data quality is important to know whether the available data allows for good decision making. The proposed artifact does not include this monitoring aspect because of scoping reasons (Section 1.3.2), but it is highly necessary to monitor the quality after the design has been implemented by using the proposed approach.

Within academic research, FAIR data principles gain importance. This is an acronym for findable, accessible, interoperable and reusable data. According to Mons et al. (2017), it refers to "a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible". It seems applicable to the data used within public transportation, especially because public transportation data is often shared. FAIR was not part of this research, but the four aspects can be related to the data quality aspects from ISO 25012:2008 as used in this research.

9.1.4 Future of the Artifact

The future of the proposed artifact is discussed based on two aspects: future technology and future planning process. Both aspects combine the future usage of the artifact.

Regarding technology, it is possible and likely that PTOs use software applications from vendors specialized in the public transport planning process. An increasing number of applications are already or will probably become a Software-as-a-Service (SaaS) or Platform-as-a-Service (PaaS), and even iPaaS solutions are introduced. These services tend to offer, among other features, data management practices. The proposed artifact can also be used by these vendors. However, for a PTO it is important to stay the owner of the data and stay in control of what to do with your data. Furthermore, entrusting your data to a vendor implicitly means a huge vendor lock-in. To overcome this, it is recommended to remain in control of one's data and develop a proper data integration situation. The data integration design approach can be used for this. Furthermore, the proposed artifact can help to direct PTOs in the right direction regarding data management practices, especially for data integration and data quality.

Other technological changes such as robotic process automation, artificial intelligence (AI) and low-code are foreseen in the near future as well. All these technologies need and use data, which is easily accessible by employing the proposed target data architecture. These technologies can help reach business goals, and by doing so, they benefit from a proper data integration situation. Furthermore, AI might be used to map different data objects to the Transmodel ontology, which improves the usability and ease of use of integration techniques (see Section 9.6.3). As the importance of sharing data is likely to increase in the future, this is taken into account for the design of the artifact as well (explained in Section 9.1.3).

As explained at the beginning of this section, the other changing aspect is related to the developments of the PTO's planning process. Due to the ongoing improvements of the entire mobility sector, it is foreseen that the PTO's planning process will change accordingly. It is expected that more external data will be taken into account while defining public transport planning (this is also endorsed by interviews and expert validation). These external data sources can be more easily combined with the correct, up-to-date and real-time internal data when the proposed artifact is used (see Section 9.6.1). Furthermore, PTOs may further optimize their services on mutual routes (services offering (partly) the same routes and/or connections). This can be done more easily if every PTO has a data integration situation based on the best-practice planning process which complies with the European Transmodel standard.

9.2 Research Findings

This research aimed to answer the following main research question:

> What constitutes a good data integration design approach for
> the planning process of a Dutch urban public transport operator?

To answer this main research question, several sub research questions were defined. The conclusion of this research is provided per sub research question and presented next. The relations between the main research questions and the sub research questions are visualized in Figure 1-5.

RQ1 What is the current data integration situation within public transportation and what data standards, data challenges and data integration methods exist?

The data integration situation for a PTO is rather complex. Various data is used among the enterprise and shared externally. The planning process turned out to have multiple and more than average upstream and downstream data dependencies, both internally and externally. This makes the public transport planning process highly data-dependent. Requirements for the design in this research were defined by identified challenges: standardization for data exchange, one single source for every data object, reusable data integrations and clear ownership (see Section 3.1.2).

Data standards in the field of public transport planning include the conceptual model Transmodel, NeTEx for the data exchange about the public transport planning and SIRI for the real-time data exchange. These three standards are defined by the European Committee for Standardization.

In literature, data integration challenges have already been present for a long time. 51 challenges were identified and consolidated into 12 unique challenges (Table 3-7), which were categorized into: systems, logic, social and administrative. The following data integration methods were identified and compared to solve these challenges: global schema, canonical data model (CDM) and ontology matching. A global schema and ontologies are further used in this research, whereas a CDM is more an implementational choice. Yet, a CDM can be based on a global schema.

RQ2 What is the best-practice planning process of a public transport operator?

The best-practice planning process was derived from literature and practice and validated by experts. It consists of nine phases (Figure 4-8). These phases were categorized into planning (design network, plan lines, plan timetable and plan crew rosters), scheduling (schedule vehicle blocks and schedule crew duties), operational (assign vehicles and assign crew) and controlling (control transport). The design network and control transport phase were placed out of scope for this research, as they are not mainly focused on the actual public transport planning (as explained in Section 4.3 and discussed in Section 9.1.1). This resulted in the planning process as shown in Figure 5-2. Phases can trigger each other in sequential order, but can also trigger previous phases. The latter is required for optimization purposes.

RQ3 What changes in the planning process that are highly dependent on data (integration) are foreseen?

Several innovations in the planning process were identified and an overview is presented in Table 4-3. It became apparent that planning data is often necessary for these innovations. Furthermore, every project clearly showed the need for having correct and up-to-date data, which is easily accessible in a real-time manner. Also, extra newly defined data objects are necessary to support the innovations. Examples of these objects are the desired crew roster and crew requests. It is also expected that the innovations in this domain will continue, which increases the general need for an optimal data integration situation.

RQ4 What data requirements (both internally, to adjacent business areas, and externally) does each step in the planning process have?

From literature, interviews and previously conducted research, a list of 214 planning tasks was formed. This list was consolidated into 23 unique planning tasks mapped to the planning process as defined for RQ2 (Table 5-2). Subsequently, the data requirements were defined, based on these planning tasks. This led to a total of 41 data objects, including both upstream and downstream requirements, which are grouped per planning phase. The complete data catalog is presented in Table 5-3. Most of the data objects are used by more than one planning phase, i.e. the planning process is highly dependent on data objects generated in other planning process phases. Furthermore, many providers and consumers outside the scope of the planning process were identified, both from within the PTO as well as from external parties. An in-depth analysis of the planning tasks and their data requirements can be found in Appendix L.

RQ5 What data quality aspects are important to address when defining a data integration design approach for the planning process?

The data quality aspects for the data integration design approach are based on the ISO 25012:2008 standard (ISO, 2008). These aspects are mapped to the data integration challenges identified in practice and literature (Table 6-2). Furthermore, a selection of quality aspects is made for the design of the data integration design approach. This selection is based on a dynamic planning scenario and the applicability of the quality aspects to a data integration design approach and is shown in Table 6-4.

Data quality aspects are either included in the design (compliance, consistency, portability and understandability), enabled by the design (accessibility, availability, currentness, efficiency and precision), considered inherent data quality and therefore not directly included in the design (accuracy, completeness, credibility and traceability) or stated less important for dynamic planning purposes (confidentiality and recoverability). The application of these aspects to the data integration design approach is presented in Section 6.4. After validation, it turned out that the top 5 relevant data quality aspects for dynamic planning according to the experts (currentness, efficiency, portability, availability and precision) are included or enabled in the data integration design approach.

RQ6 How can all insights be combined into a data integration design approach in order to ensure cost savings, offer better service quality and be prepared for future integrations?

For combining the planning process phases, its data requirements, data quality aspects and other requirements (Table 7-2) for realizing the goals of this research (Table 7-1), the TOGAF ADM (The Open Group, 2018) was used. Phase C of this method focuses on the development of a data architecture and also includes considerations for the approach of such a data architecture.

To include every possible data requirement in the planning process, the ownership of data entities was defined firstly. This classification is presented in Table 7-3. Data entities owned by the planning process were used for the target data architecture: Figure 7-2. This resulted in a similar domain classification as was found during the interviews and literature review (Figure 4-9): journey, vehicle and crew. The architecture is based on Transmodel.

To embody the target architecture within a context in which it can be used to reach the goals of this research, a data access service was presented in Figure 7-6. Even though we refrain from the actual implementation, such a service is important to ensure data reuse and data quality, and is responsible for providing data entities in one standard, to authorized consumers.

An approach is presented to support PTOs toward properly working with the target data architecture and in reaching the PTO's goals. This consists of considerations for data architectures and was designed based on proposed aspects of TOGAF's ADM (The Open Group, 2018). Considerations regarding the implementation, data authorizations, data quality and data roles are included in the approach and are based on the ISO 25012:2008 standard (ISO, 2008) and IDSA's (2019b) reference architecture for data sharing.

RQ7 How well does the proposed data integration design approach contribute to the data integration situation of a public transport operator's planning process?

To validate the data integration design approach, 15 experts from within the field of public transportation, its planning process and IT took part in expert reviews (see Table 8-1 for the expert overview). Almost all experts qualified themselves as familiar with the PTO's planning process, data architecture and integration. All transport modalities are covered by the experts.

Validation of the business content (Section 8.3.1) was positive, as the entire proposed best-practice planning process was recognized by every expert. Furthermore, it was scored as complete, except for enterprises offering and dealing with train services, as some processes are not accounted for (see RQ8). Data ownership was not completely agreed upon at first (Section 8.3.2), but after considering the feedback, we assume that the ownership of data objects is accepted by most of the experts.

The data access service validation (Section 8.3.3) was positive and the responsibilities of each application function were clear. They originate from the best-practice planning process and are responsible for the data owned by them. The access service was seen as a good solution, however, it was not totally agreed upon that the service is not being split into smaller services. This is seen as an implementational choice as described in the main text and is used for future research recommendations (Section 9.6.2).

The top five data quality aspects which are the most important for real-time data exchange according to the experts (Section 8.3.4) are taken into account. This shows that the proposed design accounts for the most important aspects.

Experts scored the authorization part of the approach relatively high due to its clarity about authorizations on data-level (Section 8.3.5). Ownership, however, could not be based on this, which is therefore expected to be shown in a better way in Table 5-3. The data roles connected to the quality aspects are agreed upon by the experts for the greatest part. According to many experts, every role is important for the data quality (Figure 8-16). This shows the relevance of the inclusion of data roles in the approach.

Data-level goals can be realized to a large extent, as it was scored positively by experts (Figure 8-17). This means that loosely coupling, reuse, correctness, up-to-date data exchange and the degree of being prepared for the future are recognized and enabled by the design. The approach was also clear to experts and the quality and responsibility aspects were taken into account to a large extent (Figure 8-18). Lastly, most of the PTO's business goals can be enabled by the implementation of the data integration design approach (Figure 8-19). Especially the integration of planning processes, more real-time planning and the adoption of depot management solutions were scored positively.

Often it was stated by the experts that the proposed design is a greenfield approach. This is true, as it defines a target data architecture. The approach is provided to assist the transition from a current situation to the end state by providing important considerations for implementation.

Lastly, the acceptance and use validation turned out to be positive as well (Section 8.5). The performance and effort expectancy, on average, was scored close to 4. This means that the experts expect that the artifact would be beneficial to the PTO's performance and that it would be of low effort to use and understand the artifact. Social influence was considered average, which means that experts neither agree nor disagree about the social influence/oblige to use the artifact. Facilitating conditions for the artifact were scored positively. Lastly, the experts' behavioral intention was scored positively. This shows that experts foresee the use of (part of) the artifact in the near future.

RQ8 Is the proposed data integration design approach generalizable to other public transport operators and other industries?

Although the designed data integration design approach is based on practical input from one Dutch urban PTO, generalizability has been accounted for since the start of the research. This is done by including operations research literature through a systematic literature review. The combination of the qualitative sources set the basis for the data integration design approach.

The design was validated by 15 experts in the field of public transportation and its planning process. For generalization purposes, these experts were categorized into several types of PTOs. It came to light that the design was expected to be useful for Dutch urban, urban and regional and national PTOs that do not exclusively offer train services. However, for national train service PTOs, the process and its data elements seem to lack important process steps (such as advanced shunting planning) and data objects. Furthermore, within rail services, the NeTEx data standard is not the most important, whereas TAP TSI (and TAF TSI for freight) is used within Europe.

For generalizability purposes, a list of conditions is defined which should be met by PTOs for using the proposed data integration design approach. These conditions are presented in Section 8.6. In short, these conditions are that a PTO should have a planning process such as the proposed planning process (RQ4; Figure 4-8), should be aware and be able to understand the European Transmodel and should be using or willing to use the European NeTEx standard. Due to these last two conditions, it is difficult to generalize to a context outside the public transportation domain.

In conclusion, since the data integration design approach is based on the European conceptual model Transmodel and data standard NeTEx and is based on the expert validations, the proposed data integration design approach is generalizable towards European PTOs offering (demand responsive) bus, tram, light rail, metro and ferry services.

Concluding Remark

This concluding remark aims to answer the main research question. The research has shown that improving data integration within the planning process contributes to the stakeholder goals as presented in Section 1.3.3 and optimization of business and IT processes. The in-depth planning phase and task analysis, identification of consuming and providing (internal and external) data requirements and data quality aspects that solve challenges from practice and literature constitute the proposed data integration design approach. Together with the ISO 25012:2008 standard (ISO, 2008) for data quality and IDSA (2019b) reference architecture for data sharing, a target data architecture and design approach are provided. These were validated by PTO experts and turned out to improve the data integration situation for the planning process and reach the stakeholders' goals.

To conclude, the research findings of this research contribute to the research objective introduced in Section 1.3.1. It provides a data integration design approach that can be used by PTOs to improve their data integration situation. As such, their data management practice can be improved and higher-level stakeholder goals (Section 1.3.3) can be reached.

9.3 Contributions to Academic Research

The proposed target data architecture can be seen as a reference architecture for data integration in public transport planning to enhance the data management situation of a PTO. It extends the research carried out within this domain, e.g. by Ceder (2016) and Scholz (2016), as this research also accounts for data requirements from and to other PTO's business areas and external data consumers.

Furthermore, it includes the connection between the operations research perspective (the business architecture in EA) and the actual usage of IT artifacts. The proposed target data architecture can be seen as the operationalization of the data needs identified in practice and literature.

The inclusion of these requirements led to minor changes compared to the process as found in academic literature. Small adaptations have been made to formulate the best-practice planning process, based on the actual data needs for the different planning tasks. These adaptations were validated by experts in this research. This can be further used for research into improving the planning process.

In literature, it is often stated that planning process phases should be integrated for optimization purposes (Canca et al., 2016; David et al., 2018; Desaulniers & Hickman, 2007; Nagy & Tick, 2019; Scholz, 2016; Shen et al., 2016; Steinzen, 2007; Weider, 2007). Data is key to this optimization process. The presented research provides a holistic data integration architecture for the planning process, which allows to better integrate planning process phases, as the correct, up-to-date and high-quality data can always be accessed.

The combination of the approach steps proposed by TOGAF ADM (The Open Group, 2018) with the ISO 25012:2008 data quality standard (ISO, 2008) and a reference architecture for data roles (IDSA, 2019b) provides an operationalized TOGAF ADM application. This can be used for other design projects and to further improve enterprise and data architecture frameworks such as TOGAF. Furthermore, it provides the next step for improving the data management situation for a PTO.

Lastly, this research contributes to the field of data integration focused on the challenges. From literature, a total of 51 data integration challenges were identified and consolidated into a classification of 12 unique challenges. Some of these challenges were also found in GVB's practice. Furthermore, these academic and practical challenges were mapped to data quality aspects from the ISO 25012:2008 standard. This indicates the relevance of the data quality aspects and can be used for further improvements of the standard. Furthermore, it provides directions to solve the data integration challenges identified by Jarke et al. (2014), which were already present for many years.

9.4 Contributions to Practice

In this section, the contributions to practice are provided. General PTO contributions are presented first, after which specific contributions to GVB's practice are provided.

9.4.1 General PTO Contributions

European PTOs can use the proposed and validated data integration design approach to improve their data management level, to reach business goals concerning the usage of better, real-time and high-quality data. It is validated that the use of the proposed artifact will lead to more IT flexibility towards the business, cost savings due to fewer integrations and preparation for future innovations and changes in the PTO's business processes. It enables practitioners in the field of public transport planning process and IT to increase their data management level and to benefit from the improvements possible when having a proper data integration environment. Moreover, PTOs can also use the identified best-practice planning process and data dependencies to improve their planning process on business-level.

The quality aspects and considerations for the approach are of added value for improving data quality. Improving data quality helps the PTO to improve decision-making. Furthermore, as shown in this research (Table 6-2), accounting for data quality aspects mitigates problems around data management and data integration. This helps PTOs to understand the value of data quality.

Moving towards a target data architecture is not considered to be an easy step, since business processes, responsibilities and authorizations likely need to be changed. For this process, the approach as defined in this research includes important considerations that should be accounted for while implementing the target data architecture. This can help PTOs move towards the target data architecture.

In general, the artifact serves as a holistic view of what data and approach are necessary, what needs to be carried out and what the implications could be once a PTO decides to work in a more real-time manner. It serves as guidance for improving the PTO's data management level which in the end leads to:

- The usage of correct, up-to-date and high-quality data.
- The possibility to plan more dynamically.
- The possibility to optimize an integrated planning process instead of suboptimization.
- The possibility to faster and more flexibly fit the business' needs.
- A better preparation for future (data) integrations.
- A decrease in IT costs for data integration maintenance and development.

9.4.2 Contributions for GVB

This research started with the data integration problem investigation at GVB. Based on this, together with literature, the data integration design approach was designed. In this section, the research contributions are related to GVB's practice which aims to provide recommendations for improving their data management situation. It is explained using the identified data integration challenges in Section 3.1.2 and the validation presented in Chapter 8.

The first identified problem within GVB was the lack of data standardization. The proposed architecture is based on Transmodel and allows to use NeTEx and SIRI data standards. Data standardization is necessary for proper data integration. As such, the added value of the target data architecture and data access service can be reached. This increases the integration flexibility and decreases costs.

Another problem was found in the several sources of data objects. The proposed data access service should ensure that only one truth of every data object is present within the data service, based on the provided target data architecture. This data is always retrieved from the system of record and no duplicate storage takes place since always the original data is used for integrations.

Combining the abovementioned two solutions leads to the solution for the third identified challenge, which is the low level of data (integration) reuse. As explained before, reuse of data and data integrations is enhanced by the proposed design. This means that the number of integrations could be reduced drastically, which ensures a more stable, flexible, cheaper (by means of maintenance and development) and better monitored IT landscape.

Ownership should be managed carefully. The approach presented in this research helps to allocate this within the enterprise. Furthermore, it is explained which data quality aspects are most important. It is key to monitor data quality, as important decisions are made based upon data.

The validation results of experts working at GVB is presented in Appendix Q. The validation of the proposed artifact was positive among the total sample of experts, however, that does not necessarily mean that this is the case for the specific GVB situation. The appendix contains the same goal satisfaction and usage and acceptance graphs as presented in Chapter 8, but they only include the validation data of GVB experts. As can be seen in Appendix Q, the goal satisfaction validation (especially on data-level and regarding the approach) and the use and acceptance validation were also scored positively by GVB employees. This substantiates the applicability and contributions to GVB's practice.

9.5 Research Limitations

The first limitation of this research can be found in the problem statement and analysis of the current data integration situation for PTOs. GVB's situation is analyzed to define the problem and objectives for this research. However, the effects of this limitation are expected to be minimal, since the generalization of the design proved that the problem is also present at other PTOs. Yet, it is possible that not all practical data integration challenges were identified, as the input from other PTOs is not taken into account.

Next to this, the qualitative approach of the data used as input for the design is considered a limitation. Summarizing and coding the interviews and literature was subjective, since the research was carried out by one researcher only. To mitigate this limitation, the interviews were validated by the respondents and the literature was compared to this practical evidence. Finally, a validation study took place to refrain from issues such as incompleteness and incorrectness, due to the qualitative approach of this study.

Two other limitations can be found in the interview sessions. During the first interview session, 12 respondents took part. However, in the second interview session, only two respondents participated. Even though we have argued that this was a valid choice (Section 2.3.2), it does not mean that saturation is reached automatically. Comparing the findings to literature, however, increased the validity and completeness of the results. The other limitation regarding the interview sessions is subject to the backgrounds of the respondents. All respondents worked for GVB, which means that input for the design in this research merely originates from GVB's practice. This limitation is attempted to be mitigated by applying triangulation and performing validation with experts from other PTOs as well.

Other limitations are subject to the validation of this research. Not all data objects have been validated individually by the experts, since this would have resulted in a too complex validation. Instead of individual validation, the data objects were validated as a whole. Furthermore, three experts (GVB employees) in the validation study were also involved during the problem investigation and treatment design parts of this research. This might result in a validation bias. However, in the validation, also other (non-)GVB experts were involved. Another limitation is the group of involved experts. No experts from PTOs outside the Netherlands took part in the validation and the number of experts involved was not high enough to reason about the results quantitatively. This limitation affects the validation and generalization of the research.

The last limitation was identified after the final version of the data integration design approach was designed. This final design has not been validated, because some changes took place after the validation as explained in Chapter 8. Especially the data quality aspects identification and application are not validated as such. Due to the resource constraints of this research, the last version of the artifact presented in this research is not validated entirely. However, the changes were based on the validation by experts, which justifies the changes.

9.6 Future Research

Recommendations for future research in the field of data management, data integration and the PTO's planning process can be divided into four different topics, as explained hereafter.

9.6.1 Data Standardization

As explained in Section 7.2.2, the NeTEx data standard does not contain standardization for most of the data in the vehicle and crew domains of the proposed data architecture. Furthermore, it is also expected that not every single data object in the journey domain is covered. To better investigate the coverage of NeTEx to the proposed design, future research is necessary. The proposed architecture can likely be consistent with the architecture, but not entirely compliant. If this is the conclusion, it is recommended to conduct future research to broaden the scope of NeTEx, at least to a point that it completely satisfies Transmodel's planning domain (which contains conceptual information about crew and more, as explained in Section 7.2.2). SIRI and ITxPT standards should also be considered, since these might cover important data entities that are uncovered right now. If an even broader scope of Transmodel is necessary, it would be wise to broaden it as well. Improving Transmodel and NeTEx is expected to be a better solution than developing a new model or standard for the missing parts, otherwise overlap between standards may arise. Only with standardization of data, the data integration – and thus the data management situation – can be improved. It will also improve the possibilities and portability of data within and the vendor lock-in with software packages such as Hastus[6].

As explained during the generalization of the designed artifact, PTOs offering train services need to adhere to the European TAP TSI (passenger) and TAF TSI (freight) standards. Bourée et al. (2019)

found similarities with Transmodel, but the standards are not entirely compliant. It is also questionable whether or not TAF TSI can become compliant with Transmodel, since freight transportation is not included in Transmodel. The applicability of the designed artifact should be further investigated by conducting research to identify the applicability of Transmodel to train services, since no clear answer to that question is presented in this research.

External data sources were left out of scope, as they are not owned by the planning process. Examples of these external data sources are weather and congestion information, but also crew working rules (according to the law and company-specific). To use this information within the planning process, it is also important to standardize this data and to ensure that this data is always up-to-date. Research should focus on how to include these kinds of external information in the data integration architecture.

9.6.2 Data Access Service

The data access service proposed in this research is one service that offers all identified data in one single service. During the validation, some experts mentioned the disadvantage of having one service in place. For this reason, they recommended the division of services. Although this is possible, it should be investigated what this means for the data objects which are often changed by different planning phases, what it implies for transactions, authorizations and administration. Due to the complex nature of processes, data ownership and authorizations, this should be addressed carefully.

9.6.3 Data Sharing and AI

For data reuse and sharing purposes, the International Data Spaces Association's reference architecture for data sharing (IDSA, 2019b) was used. As the importance of data sharing will very likely increase in the future, such a reference architecture is accounted for. Further innovations in the development of this reference architecture can influence the proposed approach. No other data sharing reference architecture was accounted for and some more exists within the field of logistics (TKI Dinalog, 2020). It is recommended to align with these data-sharing initiatives to benefit from the latest research within the field.

Artificial intelligence (AI) can contribute to the transition of mobility to accessibility, sustainability and the reduction of accidents (TNO, 2020). Data is necessary to apply AI within the public transport sector. The proposed data integration design approach is expected to enable this, but it is recommended to further investigate the applicability of the artifact to such technological developments, as it can also help to improve the planning process of a PTO.

9.6.4 Using the Data Integration Design Approach

As noticed by many experts during the validation study, the proposed data integration design approach is the easiest to be applied to a greenfield situation. While a PTO often has a brownfield situation in place, it is not directly suitable for every PTO. Even though the approach provides the PTO with important considerations for the implementation, an exact method that prescribes the process step-by-step is not given. For this reason, it should be explored whether appropriate methods exist to move towards such a target data architecture. If such a method is not available, research can be carried out to find the optimal method for implementing the target data architecture and approach. It should be considered whether TOGAF ADM is the best method to be used for this further research.

Since the actual artifact implementation is not part of this research, the artifact has not been validated in a real problem context. For further improvements of the proposed artifact in this design, it should be used in practice within several PTOs. Conducting such research would potentially lead to an improved version of the data integration design approach, as the implementation evaluation can be used as the problem identification of a new research process (Wieringa, 2014).

BIBLIOGRAPHY

Adams, W. C. (2015). Conducting Semi-Structured Interviews. In *Handbook of Practical Program Evaluation* (Vol. 72, Issue 12, pp. 492–505). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119171386.ch19

Alexander, I. F. (2005). A Taxonomy of Stakeholders: Human Roles in System Development. *International Journal of Technology and Human Interaction (IJTHI)*, *1*(1), 23–59.

Arneodo, F. (2015). *Public Transport Network Timetable Exchange (NeTEx) Introduction.* http://netex-cen.eu/?page_id=14

Bahga, A., & Madisetti, V. K. (2015). Healthcare Data Integration and Informatics in the Cloud. *Computer*, *48*(2), 50–57. https://doi.org/10.1109/MC.2015.46

Bandara, W., Furtmueller, E., Gorbacheva, E., Miskon, S., & Beekhuyzen, J. (2015). Achieving Rigor in Literature Reviews: Insights from Qualitative Data Analysis and Tool-Support. *Communications of the Association for Information Systems*, *37*, 154–204. https://doi.org/10.17705/1CAIS.03708

Batini, C., Lenzerini, M., & Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys (CSUR)*, *18*(4), 323–364. https://doi.org/10.1145/27633.27634

Békési, J., Brodnik, A., Krész, M., & Pash, D. (2009). An Integrated Framework for Bus Logistics Management: Case Studies. In *Logistik Management* (pp. 389–411). Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2362-2_20

Bernstein, P. A., & Haas, L. M. (2008). Information integration in the enterprise. *Communications of the ACM*, *51*(9), 72–79. https://doi.org/10.1145/1378727.1378745

Bertossi, A. A., Carraresi, P., & Gallo, G. (1987). On some matching problems arising in vehicle scheduling models. *Networks*, *17*(3), 271–281. https://doi.org/10.1002/net.3230170303

BKK Budapest. (n.d.). *Information on public transport boats.* Retrieved July 2, 2020, from https://bkk.hu/en/boats/

Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic Approaches to a Successful Literature Review* (2nd ed.). SAGE Publications Ltd.

Bourée, K., De Vries, B., Duquesne, C., Dodson, C., Jugelt, S., Martirano, G., Minghini, M., & Pignatelli, F. (2019). *INSPIRE-MMTIS: overlap in standards related to the Delegated Regulation (EU) 2017/1926.* Publications Office of the European Union. https://doi.org/10.2760/404745

Busscher, R. (2020). Blog: de potentie van datagedreven werken. *OV Magazine.* https://www.ovmagazine.nl/2020/12/over-de-potentie-van-datagedreven-werken-1123/

Cambridge University Press. (n.d.-a). *Approach | Meaning in the Cambridge English Dictionary.* Retrieved November 10, 2020, from https://dictionary.cambridge.org/dictionary/english/approach

Cambridge University Press. (n.d.-b). *Cambridge Essential American English Dictionary.* Retrieved August 25, 2020, from https://dictionary.cambridge.org/dictionary/essential-american-english/

Cameron, B., & McMillan, E. (2013). Analyzing the current trends in enterprise architecture frameworks. *Journal of Enterprise Architecture, 9*(1), 60–71.

Canca, D., De-Los-Santos, A., Laporte, G., & Mesa, J. A. (2016). A general rapid network design, line planning and fleet investment integrated model. *Annals of Operations Research, 246*(1–2), 127–144. https://doi.org/10.1007/s10479-014-1725-0

CAO Multimodaal. (2019). https://www.fnv.nl/cao-sector/streekvervoer/cao-multimodaal

CAO Openbaar Vervoer. (2020). https://www.fnv.nl/cao-sector/streekvervoer/cao-openbaar-vervoer

Carey, B., & DeLayne Stroud, J. (n.d.). *SIPOC leads to process mapping and project selection.* Retrieved November 6, 2020, from https://www.isixsigma.com/implementation/project-selection-tracking/sipoc-leads-process-mapping-and-project-selection/

Ceder, A. (2016). *Public Transport Planning and Operation: Modeling, Practice and Behavior* (2nd ed.). CRC Press.

Ceder, A., & Wilson, N. H. M. (1986). Bus network design. *Transportation Research Part B: Methodological, 20*(4), 331–344. https://doi.org/10.1016/0191-2615(86)90047-0

CEN. (n.d.-a). *NeTEx - Standards Context.* Retrieved February 13, 2020, from http://netex-cen.eu/?page_id=58

CEN. (n.d.-b). *Standard Interface for Real-time Information.* Retrieved February 18, 2020, from http://www.transmodel-cen.eu/standards/siri/

CEN. (2005). *SIRI (Service Interface for Real-time Information) Management Overview - White Paper.* https://www.vdv.de/siri.aspx

CEN. (2019). *Transmodel at a glance.* http://www.transmodel-cen.eu/downloads/

Claus, S. (2018). De Noord-Zuidlijn gaat eindelijk rijden, maar niet ver genoeg. *Trouw.* https://www.trouw.nl/nieuws/de-noord-zuidlijn-gaat-eindelijk-rijden-maar-niet-ver-genoeg~bf3b0b77/

DAMA International. (2014). *DAMA-DMBOK2 Framework.* https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf

David, B., Hegyhati, M., & Kresz, M. (2018). Linearly priced timed automata for the bus schedule assignment problem. *2018 4th International Conference on Logistics Operations Management (GOL)*, 1–7. https://doi.org/10.1109/GOL.2018.8378104

Desaulniers, G., & Hickman, M. D. (2007). Chapter 2 Public Transit. In *Handbooks in Operations Research and Management Science* (Vol. 14, Issue C, pp. 69–127). https://doi.org/10.1016/S0927-0507(06)14002-5

Dingemans, B. (2019). *Data modelleren in de praktijk.* Brave New Books.

Doan, A., Ardalan, A., Ballard, J. R., Das, S., Govind, Y., Konda, P., Li, H., Paulson, E., C., P. S. G., & Zhang, H. (2017). *Toward a System Building Agenda for Data Integration.* http://arxiv.org/abs/1710.00027

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of Data Integration.* Elsevier Inc.

Doody, O., & Noonan, M. (2013). Preparing and conducting interviews to collect data. *Nurse Researcher, 20*(5), 28–32. https://doi.org/10.7748/nr2013.05.20.5.28.e327

Evgeniou, T. (2002). Information Integration and Information Strategies for Adaptive Enterprises. *European Management Journal, 20*(5), 486–494. https://doi.org/10.1016/S0263-2373(02)00092-0

Firican, G. (2018). What is data stewardship? *Lights On Data*. https://www.lightsondata.com/data-stewardship-definition/

Firican, G. (2019). *5 main data roles found in data governance programs*. *Lights On Data*. https://www.lightsondata.com/5-main-data-roles-data-governance/

Friedrich, M. (2011). Wie viele? Wohin? Womit? Was können uns Verkehrsnachfragemodelle wirklich sagen? *Tagungsbericht Heureka 11*.

Friedrich, M., Leurent, F., Jackiva, I., Fini, V., & Raveau, S. (2016). *From Transit Systems to Models: Purpose of Modelling* (pp. 131–234). Springer International Publishing. https://doi.org/10.1007/978-3-319-25082-3_4

Garde, A. H., Albertsen, K., Nabe-Nielsen, K., Carneiro, I. G., Skotte, J., Hansen, S. M., Lund, H., Hvid, H., & Hansen, Å. M. (2012). Implementation of self-rostering (the PRIO-project): effects on working hours, recovery, and health. *Scandinavian Journal of Work, Environment & Health*, *38*(4), 314–326. https://doi.org/10.5271/sjweh.3306

Giachetti, R. E. (2004). A framework to review the information integration of the enterprise. *International Journal of Production Research*, *42*(6), 1147–1166. https://doi.org/10.1080/00207540310001622430

GIRO. (n.d.). *HASTUS for operations managers*. Retrieved October 8, 2020, from https://www.giro.ca/en-ca/our-solutions/hastus-software/hastus-for-operations-managers/

Golshan, B., Halevy, A., Mihaila, G., & Tan, W.-C. (2017). Data Integration: After the teenage years. *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS '17*, 101–106. https://doi.org/10.1145/3034786.3056124

Guido, G., Rogano, D., Vitale, A., Astarita, V., & Festa, D. (2017). Big data for public transportation: A DSS framework. *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 872–877. https://doi.org/10.1109/MTITS.2017.8005635

GVB Activa BV. (2020). *Jaarverslag 2019*. https://over.gvb.nl/content/uploads/2020/05/GVB-Activa-BV-Jaarverslag-2019.pdf

GVB Holding NV. (2020a). *Jaarverslag 2019*. https://jaarverslag.gvb.nl/pdfondemand/printpdf?docId=249158

GVB Holding NV. (2020b). *Lijnenkaart 2020*. https://www.gvb.nl/sites/default/files/lijnenkaart2020.pdf

GVB Holding NV. (2020c). *Toelichting Planketen project*.

Halevy, A., Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A., & Sikka, V. (2005). Enterprise information integration: successes, challenges and controversies. *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data - SIGMOD '05*, 778–787. https://doi.org/10.1145/1066157.1066246

Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data integration: The teenage years. *32nd International Conference on Very Large Data Bases (VLDB)*, 9–16.

Harbering, J. (2017). Delay resistant line planning with a view towards passenger transfers. *TOP*, *25*(3), 467–496. https://doi.org/10.1007/s11750-017-0436-5

Hausladen, I., & Schosser, M. (2020). Towards a maturity model for big data analytics in airline network planning. *Journal of Air Transport Management*, *82*(September 2019). https://doi.org/10.1016/j.jairtraman.2019.101721

Huisman, D. (2016). *Goed, beter, best! - Over optimalisatie in het openbaar vervoer* [Erasmus

University Rotterdam]. https://personal.eur.nl/huisman/26994_ERIM_Oratie_Dennis_Huisman.pdf

IDSA. (n.d.). *The Association - International Data Spaces Association*. Retrieved December 3, 2020, from https://www.internationaldataspaces.org/the-association/#mitglieder

IDSA. (2019a). *International Data Spaces Fact Sheet and Core Statements*. https://www.internationaldataspaces.org/wp-content/uploads/2019/10/IDSA-Fact-sheet-and-core-statements-English.pdf

IDSA. (2019b). *Reference Architecture Model*. https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf

INIT. (n.d.). *Efficient management of an e-bus fleet*. Retrieved October 8, 2020, from https://www.initse.com/ende/news-resources/knowledge-database/articles/2018/initiative2-chargingmanagement/

ISO. (n.d.). *ISO/IEC 25012:2008*. Retrieved November 21, 2020, from https://www.iso.org/standard/35736.html

ISO. (2008). *ISO/IEC 25012:2008 Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model*. https://www.iso.org/standard/35736.html

iso25000.com. (n.d.). *ISO 25012*. Retrieved November 21, 2020, from https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

ITxPT. (n.d.). *ITxPT Specifications*. Retrieved December 4, 2020, from https://itxpt.org/technology/itxpt-specifications/

ITxPT. (2020). *ITxPT - Information Technology for Public Transport*. https://wiki.itxpt.org/images/e/ed/2020_ITxPT_Presentation.pdf

IVU. (n.d.). *Depot Management*. Retrieved October 8, 2020, from https://www.ivu.com/eready/vehicle-dispatch/depot-management.html

Jarke, M., Jeusfeld, M., & Quix, C. (2014). Data-centric intelligent information integration—from concepts to automation. *Journal of Intelligent Information Systems*, *43*(3), 437–462. https://doi.org/10.1007/s10844-014-0340-5

Kallio, H., Pietilä, A.-M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, *72*(12), 2954–2965. https://doi.org/10.1111/jan.13031

Kamargianni, M., & Matyas, M. (2017). The Business Ecosystem of Mobility-as-a-Service. *96th Transportation Research Board (TRB) Annual Meeting*, 1–14.

Katasonov, A., & Lattunen, A. (2014). A Semantic Approach to Enterprise Information Integration. *2014 IEEE International Conference on Semantic Computing*, 219–226. https://doi.org/10.1109/ICSC.2014.23

Kirkpatrick, D. (2011). *Forbes: Now Every Company Is A Software Company*. https://www.forbes.com/sites/techonomy/2011/11/30/now-every-company-is-a-software-company/

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*.

Kleinreesink, M. G. P. W. (2017). *Predicting The Acceptance of Model-Driven Software Applications in Organisations*. University of Twente.

Lampkin, W., & Saalmans, P. D. (1967). The Design of Routes, Service Frequencies, and Schedules for a Municipal Bus Undertaking: A Case Study. *Operational Research Quarterly*, *18*(4), 375.

https://doi.org/10.2307/3007688

Lemcke, J., Stuhec, G., & Dietrich, M. (2012). Computing a Canonical Hierarchical Schema. In R. Poler, G. Doumeingts, B. Katzy, & R. Chalmeta (Eds.), *Enterprise Interoperability V: Shaping Enterprise Interoperability in the Future Internet* (pp. 305–315). Springer. https://doi.org/10.1007/978-1-4471-2819-9

Lenstra, J. K., & Kan, A. H. G. R. (1981). Complexity of vehicle routing and scheduling problems. *Networks*, *11*(2), 221–227. https://doi.org/10.1002/net.3230110211

Lenzerini, M. (2002). Data integration: a theoretical perspective. *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - PODS '02*, 233–246. https://doi.org/10.1145/543613.543644

Machel, R. E. (1965). *System engineering handbook*. McGraw-Hill.

Metselaar, D. (2020). GVB en HTM breiden reisplanner uit met drukte-indicator. *OV Pro*. https://www.ovpro.nl/innovatie-2/2020/10/08/gvb-breidt-reisplanner-uit-met-drukte-indicator/

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Nadella, S. (2018). *Future Decoded event*. https://news.microsoft.com/en-gb/2018/11/07/microsoft-ceo-satya-nadella-on-fuelling-tech-intensity-in-the-uk/

Nagy, A., & Tick, J. (2019). Tasks of operative planning for transport management systems. *2019 IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 199–204. https://doi.org/10.1109/SAMI.2019.8782766

Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence Based Nursing*, *18*(2), 34–36. https://doi.org/10.1136/eb-2015-102054

Okoli, C. (2015a). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, *37*(1), 879–910. https://doi.org/10.17705/1cais.03743

Okoli, C. (2015b). The View from Giantss Shoulders: Developing Theory with Theory-Mining Systematic Literature Reviews. *SSRN Electronic Journal*, *16*(1), 24–25. https://doi.org/10.2139/ssrn.2699362

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, *34*(5 Pt 2), 1189–1208. http://www.ncbi.nlm.nih.gov/pubmed/10591279%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1089059

Peng, C., Goswami, P., & Bai, G. (2019). A literature review of current technologies on health data integration for patient-centered health management. *Health Informatics Journal*, 1–26. https://doi.org/10.1177/1460458219892387

Pratama, F. G., Astana, S., Yudhoatmojo, S. B., & Nizar Hidayanto, A. (2018). Master Data Management Maturity Assessment: A Case Study of Organization in Ministry of Education and Culture. *2018 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 1–6. https://doi.org/10.1109/IC3INA.2018.8629524

PSI. (n.d.). *Depot Management*. Retrieved October 8, 2020, from https://www.psitrans.de/en/solutions/depot-management/

Qiu, G., Song, R., He, S., & Song, Z. (2018). Diversified Bus Services and Enterprise Information System: An Example of Beijing. *2018 Sixth International Conference on Enterprise Systems*

(ES), 172–179. https://doi.org/10.1109/ES.2018.00034

RET NV. (n.d.). *Dienstregeling*. Retrieved December 13, 2020, from https://www.ret.nl/home/reizen/dienstregeling/fast-ferry.html

Reynolds, S. (2019). *NeTEx Workshop # 2 - Routes & Timetables*. http://netex.uk/farexchange/doc/2019.07.16_Workshops/DFT-Workshop2-Routes_and_Timetable-2019.07.16-v02.pdf

Rijksoverheid. (2020). *Beschikbaarheidsvergoeding OV goedgekeurd door Europese Commissie*. https://www.rijksoverheid.nl/actueel/nieuws/2020/11/03/beschikbaarheidsvergoeding-ov-goedgekeurd-door-europese-commissie

Rowley, J. (2012). Conducting research interviews. *Management Research Review*, *35*(3/4), 260–271. https://doi.org/10.1108/01409171211210154

Saint-Louis, P., Morency, M. C., & Lapalme, J. (2019). Examination of explicit definitions of enterprise architecture. *International Journal of Engineering Business Management*, *11*, 1–18. https://doi.org/10.1177/1847979019866337

Salguero, A., Araque, F., & Delgado, C. (2008). Ontology based framework for data integration. *WSEAS Transactions on Information Science & Applications*, *5*(6), 953–962. http://www.wseas.us/e-library/transactions/information/2008/27-358.pdf

Samonas, S., & Coss, D. (2014). The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security. *Journal of Information System Security*, *10*(3), 21–45.

Scholz, G. (2016). *IT Systems in Public Transport*. dpunkt.verlag.

Sezgin, M. S., Bayrak, A. T., & Yildiz, O. T. (2019). A Hybrid Approach to Dynamic Enterprise Data Platform. *2019 IEEE International Conference on Big Data (Big Data)*, 3486–3492. https://doi.org/10.1109/BigData47090.2019.9006043

Shen, Y., Xu, J., & Zeng, Z. (2016). Public transit planning and scheduling based on AVL data in China. *International Transactions in Operational Research*, *23*(6), 1089–1111. https://doi.org/10.1111/itor.12164

Shu, N. C., Housel, B. C., Taylor, R. W., Ghosh, S. P., & Lum, V. Y. (1977). EXPRESS: A Data EXtraction, Processing, and REStructuring System. *ACM Transactions on Database Systems*, *2*(2), 134–174. https://doi.org/10.1145/320544.320549

Singh, P. M., Van Sinderen, M., & Wieringa, R. (2017). Reference Architecture for Integration Platforms. *2017 IEEE 21st International Enterprise Distributed Object Computing Conference (EDOC)*, 113–122. https://doi.org/10.1109/EDOC.2017.24

Steinzen, I. (2007). *Topics in Integrated Vehicle and Crew Scheduling in Public Transport*. University of Paderborn.

TfL London. (n.d.). *River - Transport for London*. Retrieved July 2, 2020, from https://tfl.gov.uk/modes/river/

The Open Group. (n.d.). *ArchiMate® 3.1 Specification*. Retrieved September 1, 2020, from https://pubs.opengroup.org/architecture/archimate3-doc/toc.html

The Open Group. (2018). *The TOGAF standard - Version 9.2*. Van Haren Publishing.

TKI Dinalog. (2020). *The Logistics Data Sharing Infrastructure*. https://www.dinalog.nl/wp-content/uploads/2020/08/Dinalog_Whitepaper-Data-Infrastructure_DEF.pdf

TNO. (2020). *Artificiële intelligentie in mobiliteit en transport*. https://nlaic.com/wp-content/uploads/2020/07/Position-Paper-AI-in-Mobiliteit-en-Transport-1.pdf

Transmodel. (2019). *Transmodel V6.0 Definitions of concepts*. http://www.transmodel-cen.eu/wp-content/uploads/sites/2/2015/01/TRM6_DataDefinitions.pdf

UITP. (2017). *Stakeholder Cooperation on Data in Public Transport*. https://www.uitp.org/news/public-transport-data-driven-business

UITP. (2019a). *Public Transport Trends 2019*. http://www.uitp.org/public-transport-trends

UITP. (2019b). *Ready for MaaS? Easier Mobility for Citizens and better Data for Cities*. https://www.uitp.org/sites/default/files/cck-focus-papers-files/Policy Brief_MaaS_V3_final_web_0.pdf

UITP. (2020). *Data at the Heart of Public Transport*. https://www.uitp.org/sites/default/files/Data4PT_PR_.pdf

Van de Velde, F. (2019). *GVB Hybrid Integration Platform selection (20191118)*.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 425–478.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii.

Weider, S. (2007). *Integration of Vehicle and Duty Scheduling in Public Transport* [Technische Universität Berlin]. https://opus4.kobv.de/opus4-zib/files/1035/Diss_Weider_Druck.pdf

Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer. https://doi.org/10.1007/978-3-662-43839-8

Winter, T., & Zimmermann, U. T. (2000). Real-time dispatch of trams in storage yards. *Annals of Operations Research*, *96*, 287–315.

Wray, S. (2020). Autonomous public transport bus trials to launch in five European cities. *Cities Today*. https://cities-today.com/autonomous-public-transport-bus-trials-to-launch-in-five-european-cities/

Zhang, H., Guo, Y., Li, Q., George, T. J., Shenkman, E., Modave, F., & Bian, J. (2018). An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Medical Informatics and Decision Making*, *18*(2), 41. https://doi.org/10.1186/s12911-018-0636-4

Appendix A Stakeholder Taxonomy (Alexander, 2005)

The stakeholder taxonomy by Alexander (2005) as used within this research is presented in Table A-1.

Table A-1 Stakeholder taxonomy by Alexander (2005)

Group	**Stakeholder**	**Description** *(adapted from Wieringa (2014))*
Stakeholder interacting with the artifact	Normal operator	Gives routine commands to the artifact, sometimes called 'end users'
	Maintenance operator	Interacts with the artifact to keep it running
	Operational support	Supports normal operators in their use of the artifact and help to keep the artifact operational
Stakeholders in the immediate environment of the artifact	Functional beneficiary	Benefits from the output of the artifact, sometimes called 'users'
	Interfacing system	Has interest in the requirements and scope of the artifact
Stakeholders in the wider environment of the artifact	Financial beneficiary	Benefits from the artifact financially
	Political beneficiary	Benefits from the artifact in terms of status, power, influence, etc.
	Negative stakeholder	Would be worse off when the artifact is used in the problem context
	Threat agent	Wants to hurt the system
Stakeholders involved in the development of the artifact	Sponsor	Initiates the development and provides budget and is an important source of goals and requirements
	Purchaser	Is responsible for terminating the development successfully
	Developer	Builds the artifact, is not a normal operator
	Consultant	Supports the development of the artifact
	Supplier	Delivers components of the artifact

Appendix B TOGAF ADM Cycle Phases (The Open Group, 2018)

The TOGAF Architecture Development Method (ADM) Cycle is visualized in Figure 2-4 and its original explanation from The Open Group (2018, sec. 2.4) is given in Table A-2.

Table A-2 TOGAF ADM cycle phases objectives (The Open Group, 2018)

Phase	**Explanation** *(The Open Group, 2018, sec. 2.4)*
Preliminary	Describes the preparation and initiation activities required to create an Architecture Capability including customization of the TOGAF framework and definition of Architecture Principles.
Requirements Management	Examines the process of managing architecture requirements throughout the ADM.
A: Architecture Vision	Describes the initial phase of an architecture development cycle.
B: Business Architecture	Describes the development of a Business Architecture to support the agreed Architecture Vision.
C: Information Systems Architecture	Describes the development of Information Systems Architectures to support the agreed Architecture Vision.
D: Technology Architecture	Describes the development of the Technology Architecture to support the agreed Architecture Vision.
E: Opportunities and Solutions	Conducts initial implementation planning and the identification of delivery vehicles for the architecture defined in the previous phases.
F: Migration Planning	Addresses how to move from the Baseline to the Target Architecture by finalizing a detailed Implementation and Migration Plan.
G: Implementation Governance	Provides an architectural oversight of the implementation.
H: Architecture Change Management	Establishes procedures for managing change to the new architecture.

Appendix C SLR – Data Integration in Public Transport Planning

The systematic literature review (SLR) is based on Okoli's (2015a) standalone SLR guide, which is presented in the main text in Figure 2-2. Every phase is explained in this appendix.

Purpose

The purpose of this literature review is to get a clear overview of the conducted research and its conclusions in the field of data integration and data management, possibly in the business area of (public) transportation. The SLR tries to answer the following question:

RQ1 What is the data integration situation within public transportation and to which requirements does this lead?

Since this part of the research used to be part of a larger, standalone and unpublished research, the scope for this SLR used to be broader. Within the scope of this original SLR were also the following concepts: definition of data integration and data management, levels of data integration, challenges associated with data integration, data integration methods and data integration and data management methods available specifically designed for PTOs.

Protocol

A description of the consulted databases, search terms used by the execution of the static literature review and criteria for inclusion can be found in Table A-3, Table A-4 and Table A-5, respectively. The criteria for exclusion are the exact opposites of the criteria for inclusion and are therefore not stated separately.

Table A-3 Consulted databases

Google Scholar[8]
IEEE Xplore
Scopus
Transportation Research Information Services
Web of Science

Table A-4 Used search terms

"data integration framework"
"data integration method*"
"data integration" AND "literature review"
"data integration" AND "public transport*"
"data management" AND "public transport*"
"enterprise information integration"

Table A-5 Criteria for inclusion I

The paper matches the defined search terms
The paper was published in English
The paper was published after 2000
The paper originates from an academic journal or conference

Selection

Applying the protocol as defined above, the result set contained 1,438 matching papers. From this set, duplicate papers (because of the multiple sources) were removed. Furthermore, papers were filtered based on their titles. After this, a selection of papers took place based on the abstract, introduction and conclusion, full text and forward and backward reference search. This led to a total number of 26 papers. The process is illustrated in Figure A-1.

[8] Not a database, however, Google Search was used to ensure completeness of the SLR.

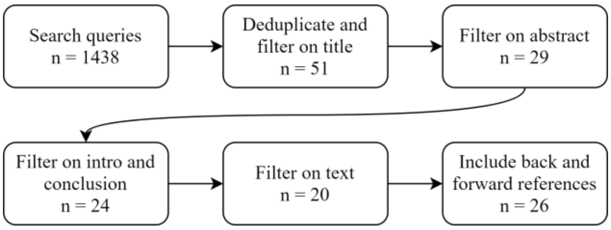

Figure A-1 Literature selection process I

Next to the papers found in the academic journals and conference proceedings, also one text-book is used for this literature review, which was found using Google Scholar. This leads to a total of 27 sources as listed in Table A-6, of which 4 are excluded from the current research (strikethrough).

Table A-6 Literature overview I

Author	Title
~~Arshah et al. (2008)~~	~~The need of Information Systems (IS) Integration Complexity Model for IS integration project~~
Bahga & Madisetti (2015)	Healthcare Data Integration and Informatics in the Cloud
Batini et al. (1986)	A comparative analysis of methodologies for database schema integration
Bernstein & Haas (2008)	Information integration in the enterprise
DAMA International (2014)	DAMA-DMBOK2 Framework
Doan et al. (2012)	Principles of Data Integration
~~Doan et al. (2017)~~	~~Toward a System Building Agenda for Data Integration~~
Evgeniou (2002)	Information Integration and Information Strategies for Adaptive Enterprises
Giachetti (2004)	A framework to review the information integration of the enterprise
~~Gold-Bernstein (1999)~~	~~EAI market segmentation~~
Golshan et al. (2017)	Data Integration: After the teenage years
Guido et al. (2017)	Big data for public transportation: A DSS framework
Halevy et al. (2005)	Enterprise information integration: successes, challenges and controversies
Halevy et al. (2006)	Data integration: The teenage years
Hausladen & Schosser (2020)	Towards a maturity model for big data analytics in airline network planning
Jarke et al. (2014)	Data-centric intelligent information integration – from concepts to automation
Katasonov & Lattunen (2014)	A Semantic Approach to Enterprise Information Integration
Lemcke et al. (2012)	Computing a Canonical Hierarchical Schema
Lenzerini (2002)	Data integration: a theoretical perspective
Peng et al. (2019)	A literature review of current technologies on health data integration for patient-centered health management
Pratama et al. (2018)	Master Data Management Maturity Assessment: A Case Study of Organization in Ministry of Education and Culture
Qiu et al. (2018)	Diversified Bus Services and Enterprise Information System: An Example of Beijing
Salguero et al. (2008)	Ontology based framework for data integration
Sezgin et al. (2019)	A Hybrid Approach to Dynamic Enterprise Data Platform
Singh et al. (2017)	Reference Architecture for Integration Platforms
~~Spruit & Pietzka (2015)~~	~~MD3M: The master data management maturity model~~
Zhang et al. (2018)	An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival

Extraction and Quality Appraisal

This phase consists of the quality appraisal and data extraction method. The first is explained hereafter, whereas the data extraction method is explained in the main text.

Several methods to assess quality are proposed by Okoli (2015a). Since this literature review is qualitative and the goal is "to get a clear overview of conducted research and its conclusions in the field of data integration and data management, possibly in the business area of (public) transportation", the relatively extensive methods are not considered to be useful. Another reason for this is that Okoli's (2015a) methodology is meant to be carried out within a team, which incorporates differences in the selection of papers between team members. This led to the application of the checklist for qualitative studies proposed by Kitchenham & Charters (2007). Since no source scored drastically low on the checklist, no exclusion of sources took place.

Appendix D Interviews – Data Integration in Public Transport Planning

Interview Design

The interview is designed based on Kallio et al.'s (2016) practical tool for developing an interview guide for semi-structured interviews, which is presented in the main text in Figure 2-3.

Identifying the Prerequisites for using Semi-structured Interviews

The goals and justification for choosing semi-structured interviews can be found in the main text and are therefore not repeated here.

Retrieving and using Previous Knowledge

For this research, the understanding of the subject originates from previous knowledge of the researcher based on their study and consulting experts within GVB's IT & Innovation department. These experts were IT architects and information managers. Combining the theoretical background from the study, expert stories and the introduction of this research led to a comprehensive and adequate understanding of the subject for the semi-structured interviews.

Formulating the Preliminary Semi-structured Interview Guide

The first version of the interview guide was made based on the theory mentioned above and previous knowledge gained by talking to experts (during the previous phase). The goal of the interview is to answer the research questions, which are therefore of high importance in the interview guide. The questions asked during the interview should provide answers to the following research questions (these are research questions from the original study and answer RQ1 of the current research):

1. Who are the stakeholders and what are their goals?
2. What are the current data integration and data management situation at GVB?
    a. What data dependencies are present between GVB's business functions?
    b. What data dependencies are present between GVB and external parties?
    c. What data dependencies are foreseen?
    d. Who is responsible for the quality, confidentiality, integrity and availability of the data identified in questions 2a, 2b and 2c?
    e. What opinion exists within GVB about the adoption of a HIP?
3. What common data and IT standards exist for European public transport operators?

The interview guide was improved iteratively between this phase and the next phase. The final semi-structured interview guide is presented in the last phase.

Pilot Testing of the Interview Guide

This phase took place iteratively with the previous phase. The interview guide was tested, adapted and tested again. This is done to confirm the coverage and relevance of the changes. For this research, expert assessments were used to pilot test the interview guide. Experts were asked to give their opinion about the preliminary interview guide. This led to some minor changes and improvements to the interview guide. Experts included in this pilot test were GVB's IT architects and information managers.

Presenting the complete Semi-structured Interview Guide

The semi-structured interview guide as used during the interviews in this research can be found below.

Interview Guide

Introduction

Do you allow the research to record the interview?
Can you please introduce yourself?
• Name, function, role, years of experience (at GVB and before) • Relation with GVB's planning process

IT Product Team

Can you please introduce your team?

- Name
- Working area
- Tasks
- Responsibilities

Applications

Which applications does your team manage?

- Custom software or off-the-shelf?
- Which business functions/areas are supported by this application?

Is the business satisfied with the services provided by your team?

- Is business & IT alignment monitored?
    - How?
- What can be improved?

Data

Which data do the applications which are managed by your team need to function properly?

- Internal or external sources?
- Who is the data owner?
- Does CIA (confidentiality, integrity, availability) rating exist for this data?
- Are data models available? [ask for access]

Which data is provided to other consumers by the applications which are managed by your team?

- Consumers
    - Internal or external?
- Who is the data owner?
- Does CIA (confidentiality, integrity, availability) rating exist for this data?
- Are data models available? [ask for access]

Are there problems to consume the right data?

- Integration
- Availability
- Quality
- Ownership

Are there problems to provide the right data?

- Integration
- Availability
- Quality
- Ownership

What is your vision for the future regarding data (integration)?

- Personal vision
- From within the team
- Applications; e.g. phasing out
- If known: does IT align with the strategic goals within GVB?

Ending

Are there any other topics that would be interesting for this research, but which have not been addressed?

Do you have any suggested colleagues whom I should also talk to in order not to miss any important information?

Can I contact you after this interview in case follow-up questions arise?

Do you want to receive the results of this study?

Appendix E SLR – Planning Process and Data

Interview Design

The systematic literature review (SLR) is based on Okoli's (2015a) standalone SLR guide, which is presented in the main text in Figure 2-2. Every phase is explained in this appendix.

Purpose

The purpose for performing the SLR is twofold, as it must provide answers for answering two research questions (RQs) of this research:

RQ2 What is the standard planning process of a public transport operator?
RQ4 What data requirements (both internally, to adjacent business areas, and externally) does each step in the planning process have?

RQ2 requires information from the field of operations research and can be answered through an SLR, assuming that research is carried out in this field of research. RQ4 is about data requirements for the planning process, which will be identified on the level of planning tasks. These tasks can be retrieved from literature. Furthermore, the answers on RQ2 helps by identifying the planning tasks, and, consequently, by identifying the data requirements.

Protocol

The consulted database for this systematic literature review is Scopus[9]. The query used for retrieving results is: *("public transport*") AND ("planning process").* The asterisk (*) after 'public transport' was used to also retrieve results that incorporate terms such as 'public transportation' or any other ending after 'transport'. The criteria for inclusion are presented in Table A-7.

Table A-7 Criteria for inclusion II

The literature matches the defined search terms
The literature was published in English
The literature was published in or after 2014
The literature is peer-reviewed
The literature originates from an academic journal, conference or book

Selection

Applying the search protocol as defined above resulted in a set of 94 literature results. This set is brought back to 6 results by filtering on title, abstract, introduction, conclusion and, finally reading the entire article. Soon it became clear that the planning process in the resulting literature was often based on the same sources, leading back to literature in the field of operations research published in 1967 and 1986. To offer a complete overview of literature, forward and backward referencing was therefore also included in the resulting literature set, which is also suggested by Webster & Watson (2002). This resulted in a total of 18 sources, including two textbooks: Ceder (2016) and Scholz (2016). The process is illustrated in Figure A-2 and the final literature list can be found in Table A-8.

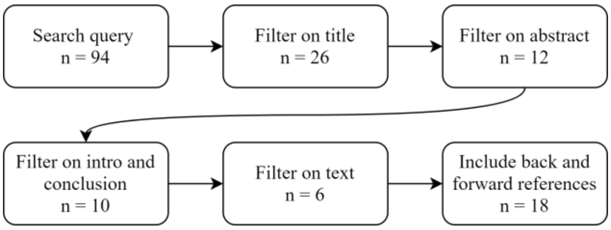

Figure A-2 Literature selection process II

[9] Scopus is an abstract and citation database of peer-reviewed literature: www.scopus.com

Table A-8 Literature overview II

Author	Title	In initial search or reference
Békési et al. (2009)	An Integrated Framework for Bus Logistics Management: Case Studies	Reference
Bertossi et al. (1987)	On some matching problems arising in vehicle scheduling models	Reference
Canca et al. (2016)	A general rapid network design, line planning and fleet investment integrated model	Initial search
Ceder (2016)	Public Transport Planning and Operation: Modeling, Practice and Behavior	Reference
Ceder & Wilson (1986)	Bus network design	Reference
David et al. (2018)	Linearly priced timed automata for the bus schedule assignment problem	Initial search
Desaulniers & Hickman (2007)	Chapter 2 Public Transit	Reference
Friedrich (2011)	Wie viele? Wohin? Womit? Was können uns Verkehrsnachfragemodelle wirklich sagen?	Reference
Friedrich et al. (2016)	From Transit Systems to Models: Purpose of Modelling	Initial search
Harbering (2017)	Delay resistant line planning with a view towards passenger transfers	Initial search
Lampkin & Saalmans (1967)	The Design of Routes, Service Frequencies, and Schedules for a Municipal Bus Undertaking: A Case Study	Reference
Lenstra & Kan (1981)	Complexity of vehicle routing and scheduling problems	Reference
Nagy & Tick (2019)	Tasks of operative planning for transport management systems	Initial search
Scholz (2016)	IT Systems in Public Transport	Reference
Shen et al. (2016)	Public transit planning and scheduling based on AVL data in China	Initial search
Steinzen (2007)	Topics in Integrated Vehicle and Crew Scheduling in Public Transport	Reference
Weider (2007)	Integration of Vehicle and Duty Scheduling in Public Transport	Reference
Winter & Zimmermann (2000)	Real-time dispatch of trams in storage yards	Reference

Extraction and Quality Appraisal

The previous phase resulted in a list of literature sources. These sources now need to be analyzed in more detail in order to extract data for the literature review in this research. Okoli (2015a) suggests extracting data by using a data extraction form and proposes the methodology of Bandara et al. (2015). An overview of the data extraction form applied to the literature review in this research can be found in Table A-9 and is based on Bandara et al. (2015) and Okoli (2015b).

Table A-9 Data extraction form (based on Bandara et al. (2015) and Okoli (2015b))

Source details	Source
	Year
General description	Study objectives
	Research questions
Research design	Research methods employed
	Empirical data characteristics
Elements of theory	Definitions of concepts
	Characteristics, dimensions, and level of analysis of the concepts
	Relationships between concepts
	Theoretical explanations for relationships
	Planning phases, planning tasks and data requirements
	Contexts of reported studies (boundary conditions)
Practical issues	Reported success factors
	Reported issues or failure factors
Future work	Future work suggested

The concepts in the data extraction form as shown in Table A-9 are important in the literature review process. However, the data extraction instrument does not need to be reported thoroughly, since it would be similar to the results shown in the next phase (Okoli, 2015b). For this research, the concepts as shown in Table A-9 are used to extract data from the literature sources. This is done by reading, analyzing and marking concepts and findings, by applying the data extraction form, in the literature sources in an application called Mendeley[10].

[10] Mendeley – Reference Management Software & Researcher Network: www.mendeley.com

Execution

The concept-centric approach of Webster & Watson (2002) as proposed by Okoli (2015a) is used for synthesizing and writing the review. The concept matrices for both RQ2 and RQ4 can be found in Table A-10 and Table A-11, respectively.

Table A-10 Concept matrix for planning topics

Topic **Literature**	Planning phases	Planning tasks	Required data
Békési et al. (2009)	✓	✓	
Bertossi et al. (1987)	✓		
Canca et al. (2016)	✓		
Ceder (2016)	✓	✓	✓
Ceder & Wilson (1986)	✓	✓	✓
David et al. (2018)	✓		
Desaulniers & Hickman (2007)	✓	✓	
Friedrich (2011)	✓	✓	
Friedrich et al. (2016)	✓	✓	✓
Harbering (2017)	✓	✓	
Lampkin & Saalmans (1967)	✓	✓	
Lenstra & Kan (1981)	✓	✓	
Nagy & Tick (2019)		✓	✓
Scholz (2016)	✓	✓	✓
Shen et al. (2016)	✓	✓	
Steinzen (2007)	✓	✓	
Weider (2007)	✓	✓	
Winter & Zimmermann (2000)	✓	✓	

Table A-11 Concept matrix for planning phases

Planning phase **Literature**	Design network	Plan lines	Develop timetable	Schedule vehicles	Schedule crew	Manage transport perturbations	Depot management
Békési et al. (2009)				✓	✓		
Bertossi et al. (1987)				✓			
Canca et al. (2016)	✓	✓		✓	✓		
Ceder (2016)	✓	✓	✓	✓	✓		
Ceder & Wilson (1986)	✓	✓	✓	✓	✓		
David et al. (2018)				✓			
Desaulniers & Hickman (2007)	✓	✓	✓	✓	✓	✓	✓
Friedrich (2011)	✓	✓	✓	✓	✓		
Friedrich et al. (2016)	✓	✓	✓	✓	✓		
Harbering (2017)		✓				✓	
Lampkin & Saalmans (1967)		✓	✓	✓	✓		
Lenstra & Kan (1981)				✓			
Nagy & Tick (2019)				✓	✓		
Scholz (2016)	✓	✓	✓	✓	✓	✓	
Shen et al. (2016)	✓	✓	✓	✓	✓	✓	
Steinzen (2007)				✓	✓		
Weider (2007)				✓	✓		
Winter & Zimmermann (2000)				✓			✓

Appendix F Interviews – Planning Process and Data

Interview Design

The interview is designed based on Kallio et al.'s (2016) practical tool for developing an interview guide for semi-structured interviews, which is presented in the main text in Figure 2-3.

Identifying the Prerequisites for using Semi-structured Interviews

The goals and justification for choosing semi-structured interviews can be found in the main text and are therefore not repeated here.

Retrieving and using Previous Knowledge

For this interview, the understanding of the subject originates from previous knowledge of the researcher based on their study, the experience gained during the internship and previously conducted research and consultation of experts within GVB's IT & Innovation department. These experts were IT architects and information managers. Combining the theoretical background from the study, practical experience and expert stories led to a comprehensive and adequate understanding of the subject for the semi-structured interviews.

Formulating the Preliminary Semi-structured Interview Guide

The first version of the interview guide was made based on the theory above and previous knowledge gained by talking to experts within GVB. The goal of the interview is to answer the research questions, which are therefore of high importance in the interview guide. The questions asked during the interview should enable answering the following research questions:

RQ2 What is the standard planning process of a public transport operator?
RQ3 What changes in the planning process which are highly dependent on data (integration) are foreseen?
RQ4 What data requirements (both internally, to adjacent business areas, and externally) does each step in the planning process have?

The interview guide was improved iteratively between this phase and the next phase. The final semi-structured interview guide is presented in the last phase.

Pilot Testing of the Interview Guide

This phase took place iteratively with the previous phase. The interview guide was tested, adapted and tested again. This is done to confirm the coverage and relevance of the changes. For this research, an expert assessment was used to pilot test the interview guide. An expert was asked to give his opinion about the preliminary interview guide. This led to some minor changes and improvements to the interview guide. The expert included in this pilot test was the enterprise architect of GVB.

Presenting the complete Semi-structured Interview Guide

The semi-structured interview guide as used during the interviews in this research can be found below.

Interview Guide

Introduction

Do you allow the research to record the interview?
Can you please introduce yourself?
• Name, function, role, years of experience (at GVB and before) • Relation with GVB's planning process

As-is Planning Process

Can you divide the planning process into several main phases?
• If yes: which phases are present at GVB? • If no: Read 'planning process' instead of 'phase(s)' for the next questions.
How do these phases look like?

Which business department is responsible for which phase?

Which decisions need to be made per phase? Which processes need to be carried out?

- Which information is necessary to make these decisions and carry out the processes?

Which phases are dynamically changed by influences such as the Covid-19 pandemic?

Which other business processes/departments are dependent on data generated within the planning process?

- For which purposes?

To-be Planning Process

Does the current planning process fall short?

- If yes: where? What? How? Etc.
- If no: next question.

GVB is planning to expand, improve and change the planning process. How does this look like?

- What needs to be done?

Which extra functionalities would the improved planning process facilitate?

Which other business processes/departments are (in the future) dependent on data generated within the planning process?

Ending

Are there any other topics that would be interesting for this research, but which have not been addressed?

Do you have any questions regarding this interview?

Do you have any suggested colleagues whom I should also talk to in order not to miss any important information?

Can I contact you after this interview in case follow-up questions arise?

Do you want to receive the results of this study?

Appendix G Validation Survey

The validation survey was presented to the respondents through a Google Forms form within the Google Workspace of the University of Twente.

Consent

C1 I have read and understood the study information. I have been able to ask questions about this study and my questions have been answered to my satisfaction.

Yes / No

C2 I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw at any time, without having to give a reason.

Yes / No

C3 I understand that taking part in this study involves answering questions that are provided in this validation survey.

Yes / No

C4 I understand that information I provide will be used for validating the artifact of this study. This might lead to improvements of the artifact.

Yes / No

C5 I understand that personal information collected about me that can identify me, such as my email address, will not be shared beyond the researcher.

Yes / No

C6 I understand that no (financial) compensation is provided to the participants of this study.

Yes / No

Best-practice Planning Process

Provided was Figure 4-8 and an explanation of every phase.

B1 – B9 How relevant are the planning process phases as presented in the figure above within the company you work for?

	Not relevant	Slightly relevant	Moderately relevant	Relevant	Very relevant	Undecided
Design network						
Plan lines						
Plan timetable						
Plan crew rosters						
Schedule vehicle blocks						
Schedule crew duties						
Assign vehicles						
Assign crew						
Control transport						

B10 Do the above-mentioned planning phases give a complete view of the planning process at the company you work for?

Yes / No / Undecided

B11 Please explain your answer to the previous question:

B12 General remark(s) about the questions in this section (optional):

Data Requirements

Provided was Table 7-3 and an explanation of every data object.

I1 The categorization in three categories as shown in the figure is useful.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

I2 Please explain your answer to the previous question:

I3 Do you think the data objects belong to the right categories?

Yes / No / Undecided

I4 Please explain your answer to the previous question:

I5 General remark(s) about the questions in this section (optional):

Data Access Service

Provided was Figure 7-6 and a brief explanation.

I6 For me it is clear what the application functions are based upon.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

I7 For me it is clear what the responsibility boundaries are of the application functions.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

I8 The data access service is a good solution when implemented correctly.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

I9 Please explain your answer to the previous questions:

I10 How many application components/software packages realize the application functions at the company you work for?

1 / 2-4 / 5-7 / >7 applications/software packages

I11 What is your general opinion/feedback/advice about the figure presented at the top of this page?

I12 General remark(s) about the questions in this section (optional):

Data Quality

Scenario 1 - Static planning: The public transport planning is made only several times per year (including detour planning) and the data is provided as static data to other processes, business areas and external consumers. It cannot perfectly respond to changes in the environment other than planned detours.

Scenario 2 - Dynamic planning: The public transport planning is made based on more real-time input and the data is provided to other processes, business areas and external consumers whenever it changes. It can perfectly respond to changes in the environment, since planning phases and tasks can be executed in a real-time manner.

An explanation of every quality aspect was provided to the respondents.

I13 – I27 Please score the relevancy of the following data quality aspects for 'Scenario 1 - Static planning' for data objects within the scope of the planning process of a public transport operator.

	Not relevant	Moderately relevant	Relevant	Undecided
Accessibility				
Accuracy				
Availability				
Completeness				
Compliance				
Confidentiality				
Consistency				
Credibility				
Currentness				
Efficiency				
Portability				
Precision				
Recoverability				
Traceability				
Understandability				

I28 Please explain your answers:

I29 – I43 Please score the relevancy of the following data quality aspects for 'Scenario 2 - Dynamic planning' for data objects within the scope of the planning process of a public transport operator.

	Not relevant	Moderately relevant	Relevant	Undecided
Accessibility				
Accuracy				
Availability				
Completeness				
Compliance				
Confidentiality				
Consistency				
Credibility				
Currentness				
Efficiency				
Portability				
Precision				
Recoverability				
Traceability				
Understandability				

I44 Please explain your answers:

I45 General remark(s) about the questions in this section (optional):

Data Quality – Compliance to Standards

Provided was Figure 7-2 and the following text:

One of the data quality aspects is 'Compliance' ("The degree to which data has attributes that adhere to standards (...)" (ISO, 2008)). Soon, Dutch public transport operators are obliged to provide data in the NeTEx standard, which covers most of the data objects in the schedule and vehicle domains.

I46 NeTEx covers most of the data in the 'journey domain' as presented in the figure above.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided / I don't know NeTEx (well enough)

I47 Why (not)?

I48 NeTEx covers most of the data in the 'vehicle domain' as presented in the figure above.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided / I don't know NeTEx (well enough)

I49 Why (not)?

I50 NeTEx covers most of the data in the 'crew domain' as presented in the figure above.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided / I don't know NeTEx (well enough)

I51 Why (not)?

I52 In case you think that the NeTEx standard cannot cover all three domains (entirely), what is your opinion and advice for data standardization within the domain(s) that are not (entirely) covered by the NeTEx standard?

I53 General remark(s) about the questions in this section (optional):

Approach

A1 For understanding how and where data entities are created, stored, transported and reported, a CRUD (Create, Read, Update, Delete) matrix which shows the relationship between data objects and planning phases within the planning process would be suitable.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

A2 Why (not)?

A3 A CRUD (Create, Read, Update, Delete) matrix helps to identify the data owner.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

A4 Why (not)?

Provided was Figure 7-8 and an explanation of every data role.

A5 The figure shows all important data roles around data exchange at a public transport operator.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

A6 Why (not)?

A7 Which data roles from the figure above are responsible for the data quality aspects as defined by the ISO 25012:2008 standard? More than one answer possible.

An explanation of every quality aspect was provided to the respondents.

Data owner / Data custodian / Data steward / Vocabulary provider / Service provider / Intermediary / Data consumer / Other ...

A8 General remark(s) about the questions in this section (optional):

Artifact Validation

V1 The data integration design approach would contribute to the goal of decoupling the providing and consuming applications on data-level.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V2 Please explain your answer:

V3 The data integration design approach would contribute to the goal of reaching a higher level of reuse of data and integrations.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V4 Please explain your answer:

V5 The data integration design approach would contribute to the goal of ensuring that every consumer uses the correct and most up-to-date data.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V6 Please explain your answer:

V7 The data integration design approach would contribute to the goal of offering the possibility to better integrate separate planning phases in order to reach overall optimization (instead of sub-optimization of the individual process phases).

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V8 Please explain your answer:

V9 The data integration design approach would contribute to the goal of offering the possibility to plan more dynamically (real-time).

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V10 Please explain your answer:

V11 The data integration design approach would contribute to the goal of offering the possibility to better align the public transport planning with other planning tasks such as vehicle maintenance planning.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V12 Please explain your answer:

V13 The data integration design approach would contribute to the goal of offering the possibility to adopt depot management solutions.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V14 Please explain your answer:

V15 The data integration design approach would contribute to the goal of better scheduling of (the charging of) battery-equipped (zero-emission) vehicles/vessels.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V16 Please explain your answer:

V17 The data integration design approach would contribute to the goal of offering the possibility to adopt self-rostering for crew members.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V18 Please explain your answer:

V19 The data integration design approach would contribute to the goal of being well-prepared for the future (not yet defined/known) integrations.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V20 Please explain your answer:

V21 The data integration design approach would contribute to the goal of providing a clear approach for the data integration design.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V22 Please explain your answer:

V23 The data integration design approach would contribute to the goal of providing data quality aspects for the planning process.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V24 Please explain your answer:

V25 The data integration design approach would contribute to the goal of providing data responsibility aspects for the planning process.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

V26 Please explain your answer:

V27 In general: would the artifact have other positive impact(s) than the goals mentioned before? Please elaborate.

V28 In general: would the artifact have negative impact(s)? Please elaborate.

V29 General remark(s) about the questions in this section (optional):

Acceptance of Use and Technology (UTAUT by Venkatesh et al. (2003))

Performance expectancy

U1 I would find the data integration design approach useful for my job.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U2 Using the data integration design approach enables me to accomplish tasks more quickly.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U3 Using the data integration design approach increases my productivity.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U4 General remark(s) about the questions in this section (optional):

Effort expectancy

U5 My interaction with the data integration design approach would be clear and understandable.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U6 It would be easy for me to become skillful at using the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U7 I would find the data integration design approach easy to use.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U8 General remark(s) about the questions in this section (optional):

Social influence

U9 People who influence my behavior think that I should use the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U10 People who are important to me think that I should use the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U11 The senior management of the company I work for will be helpful in the use of the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U12 In general, the organization would support the use of the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U13 General remark(s) about the questions in this section (optional):

Facilitating conditions

U14 I have the resources necessary to use the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U15 I have the knowledge necessary to use the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U16 The data integration design approach is compatible with other architectures, methods and/or frameworks I use.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U17 General remark(s) about the questions in this section (optional):

Behavioral intention

U18 I intend to use (part of) the data integration design approach within the next 2 years.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U19 I predict I would use (part of) the data integration design approach in the next 2 years.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U20 I plan to use (part of) the data integration design approach in the next 2 years.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U21 General remark(s) about the questions in this section (optional):

Voluntariness of use

U22 Although the data integration design approach might be helpful, using it is certainly not compulsory for my work and responsibilities.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U23 My work and responsibilities do not require me to use the data integration design approach.

Strongly disagree / Somewhat disagree / Neither agree or disagree / Somewhat agree / Strongly agree / Undecided

U24 General remark(s) about the questions in this section (optional):

Personal Information

P1 Your gender:

Female / Male/ Non-binary / Prefer not to say

P2 Your age:

< 30 / 30-40 / 41-50 / 51-60 / >60 years old / Prefer not to say

P3 How familiar are you with ‘data integration’?
Not required but kindly requested to be specified.

Unfamiliar	1	2	3	4	5	Expert

P4 How familiar are you with ‘data architecture’?
Not required but kindly requested to be specified.

Unfamiliar	1	2	3	4	5	Expert

P5 How familiar are you with ‘enterprise architecture’?
Not required but kindly requested to be specified.

Unfamiliar	1	2	3	4	5	Expert

P6 How familiar are you with ‘public transportation planning process’?
Not required but kindly requested to be specified.

Unfamiliar	1	2	3	4	5	Expert

P7 Company name:
Not required but kindly requested to be specified.

P8 I allow the researcher to mention the name of the company I work for as validation partner of this study in the final publication.

Yes / No / Undecided

P9 Company type:
Please select the option which suits the company you work for the best. If you do not work for a public transport operator, please specify at 'Other'.

Urban / Regional / Urban and regional / National / International public transport operator / Prefer not to say / Other ...

P10 Company’s modalities (more than one answer possible):

Bus / Tram / Metro / Light rail / Train / Ferry / Demand responsive bus / None of the above / Other ...

P11 Your function/role within the company:

P12 Working experience:

0-3 / 4-6 / 7-10 / 10-20 / >20 years / Prefer not to say

P13 General remark(s) about the questions in this section (optional):

P14 Do you allow the researcher to contact you in case any questions arise from your answers and/or do you want to be informed about the results of this study?

Yes, for both / Yes for questions / Yes for results / No

P15 Email address:
Only in case the answer on the previous question was ‘Yes (...)’

Appendix H GVB Projects as Data Integration Incentive

At the moment of conducting this research, some ongoing projects within GVB need data from several different sources. This stresses the importance of well-designed and well-executed data integration and can thus be considered as data integration incentives. In this section, a short overview of the ongoing digitalization of the PTO's business is provided. This listing is not limited to the planning process, but many projects require the planning data.

New Payment Methods

Within The Netherlands, the PTOs are introducing new payment methods for public transportation (nation-wide). Partly because of the disadvantages of the current system (OV-Chipkaart) and partly because of introducing Mobility-as-a-Service (explained hereafter). These new payment methods are barcodes and EMV (Europay, Mastercard and Visa) cards. The biggest difference between these two new methods and the current method is that the new methods require communication with a backend system, whereas the current method (OV-Chipkaart) requires local communication, since data can be written on the card. This means an increase in the number (checking backend for validity), security (bank account details) and performance (gate should open within *x* milli-seconds after scanning barcode/EMV card) of data integrations regarding payment.

Mobility-as-a-Service (MaaS)

The Mobility-as-a-Service (MaaS) model centralizes the transportation offers from all *MaaS operators* and their different modalities (public transport, bike and car-sharing, taxis, etc.) to let the *MaaS provider* offer a complete journey to the customer. This includes planning, ticketing and paying. Both the *MaaS operator* and *MaaS provider* are roles within the MaaS ecosystem. See Figure A-3 for the difference between the current situation and the situation when having a MaaS ecosystem in place.

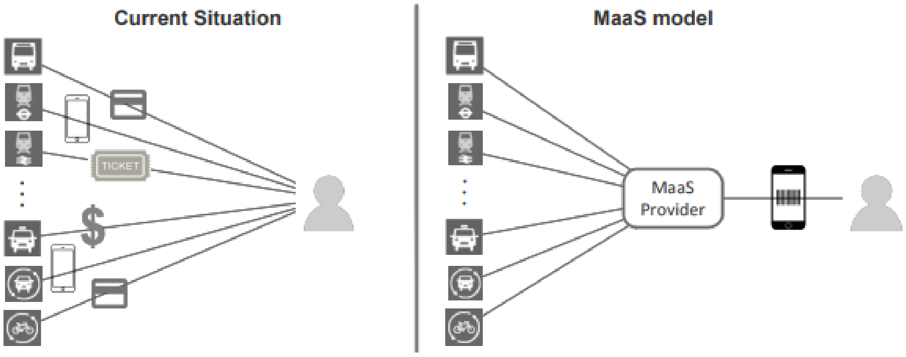

Figure A-3 Mobility-as-a-Service (Kamargianni & Matyas, 2017)

MaaS will, very likely, become the PTO's business model of the future (UITP, 2019b). This is also endorsed by the municipality of Amsterdam, which wants to reduce the number of cars in the city and thus wants to offer alternative modes of transportation to citizens and tourists.

It is expected that GVB will perform the two aforementioned roles within the MaaS ecosystem in the future. Firstly, GVB will be the *MaaS operator* offering public transportation to travelers. Secondly, GVB will – to fulfill the demands of the municipality of Amsterdam – be a *MaaS provider* as well. GVB will then offer both their own (*MaaS operator*'s) services as well as other *MaaS operator*'s services (which are likely to be GVB's competitors).

Scheduling, pricing and real-time information need to be integrated within a MaaS ecosystem. Currently, these three aspects are already integrated between PTOs in the Netherlands (see Section 3.2). However, a new prerequisite for MaaS is that the *MaaS provider* should be able to offer one single ticket for different transport modalities to travelers. Hence, *MaaS operators* should be able to offer their tickets to *MaaS providers*, and validate issued *MaaS provider*'s tickets for their own *MaaS operator*'s network.

My GVB App

While writing this report, GVB is running a project for the design and development of a new 'My GVB' application for smartphones. This app is being developed to offer travelers a journey planning for traveling within Amsterdam and the possibility to buy tickets through the application. For tourists, the app offers functionality to set up your hotel address and check points of interest on the map. For commuters, the daily ride from home to work can be set.

The next development will be incorporating data from the traveler's My GVB account, such as opt-ins, subscriptions and monthly plans. Furthermore, as soon as GVB is ready to offer services as a *MaaS provider*, this should be incorporated in the same application, which means that the journey planner also takes into account offers from other *MaaS operators*.

Project EFO

EFO is an abbreviation for Exploitatie Flexibeler Organiseren which is Dutch for 'organize operations more flexibly'. This large project is divided into two projects: speed up the planning process (Dutch: versnellen planketen) and depot management (Dutch: depotmanagement).

Speed up Planning Process

As the name says, GVB wants to speed up the planning process. This process starts with GVB's ideas and obligations set by the Vervoerregio Amsterdam (VRA), but can also be initiated due to special circumstances like road works or the spread of a global virus like COVID-19 which started in March 2020. In short, these ideas and obligations are translated into a transport schedule. Before this schedule can be executed on the day of operations, many more steps have to take place. Some examples are vehicle scheduling (how many vehicles you need for which line), the development of duties for drivers, allocation of drivers to duties and re-allocation of drivers and vehicles in case of sickness and malfunction, respectively. This process is described in detail in Chapter 4 and Chapter 5.

Right now, this process is supported by many different applications. The data integration between these applications is not real-time, but often uses file transfer or batch processing which runs only once a day or once a week.

The project aims at implementing a total planning, scheduling and operations application (from one vendor). This should ensure higher flexibility in the schedules, faster and more accurate data transfers and less redundant integrations. However, for this to succeed, many data sources from different business functions need to be integrated within the new application, because the business function *Transport planning* (see Table 3-2) has a vast number of upstream data dependencies.

Depot Management

The project 'depot management' should improve the depot management activities within the bus and tram depots. Right now, buses and trams are manually planned (by using an Excel-file), taking into account malfunction and maintenance schedules from several systems. The new depot management solution should register the sign-in of drivers and let them know which vehicle they have to take from where, taking into account the registrations of malfunctions, maintenance and the vehicle type constraints. When vehicles return to the depot, the correct parking position to park the vehicle should be communicated to the driver (preferably via the board computer in the vehicle).

As can be derived from the above, data is needed from several business functions and applications. Examples are HR, security, vehicle management, planning and operations.

Appendix I Data Integration Challenges in Literature

Table A-12 List of data integration challenges in literature

Study	Challenge
Bahga & Madisetti (2015)	Different semantical standards
Bahga & Madisetti (2015)	Different technical standards
Batini et al. (1986)	Common concepts with different meanings (they can be identical, equivalent, compatible, incompatible)
Batini et al. (1986)	Conflicting inter-schema properties
Batini et al. (1986)	Different perspectives
Batini et al. (1986)	Equivalence among constructs of the model
Batini et al. (1986)	Incompatible design specifications (cardinalities)
Bernstein & Haas (2008)	Changing the structure of just one data source can force an integration redesign
Bernstein & Haas (2008)	Inconsistencies and incompleteness of different sources.
Bernstein & Haas (2008)	Iterative process where data should be understood, prepared for integration by cleansing and standardization, specification about what data should be integrated and how they are related, finally integration program is generated. The results are examined, and any anomalies must be resolved, which often requires returning to step one and studying the data.
Doan et al. (2012)	Data owner is reluctant to share data because of competition, mission-critical data (higher load is not desired) and an integration system might not enforce restrictions regarding safety and privacy
Doan et al. (2012)	Data fiefdoms
Doan et al. (2012)	Different query processing powers between systems
Doan et al. (2012)	Efficiently querying over multiple system
Doan et al. (2012)	Finding or having the right data source electronically
Doan et al. (2012)	Legitimate legal reasons (e.g. medical records), anonymization is necessary
Doan et al. (2012)	Reconcile technical differences between systems
Doan et al. (2012)	Semantic heterogeneity because of different context
Doan et al. (2012)	Semantic heterogeneity because of human nature
Evgeniou (2002)	Different information is represented by different data in different sections across different business units and functions
Evgeniou (2002)	Information is duplicated in many areas, and errors are lurking everywhere
Giachetti (2004)	Different perspectives: local definitions of the same or similar concepts
Giachetti (2004)	Equivalence among constructs: more than one method for modeling a domain
Giachetti (2004)	Inter-schema properties
Giachetti (2004)	Systems were designed, built and optimized for local (domain/department) needs and not for the entire enterprise. Therefore various data representation formats, data semantics, programming languages, hardware platforms
Giachetti (2004)	Weak semantics: often semantics are informally defined or not defined at all, leading to difficulty in determining the precise meanings
Golshan et al. (2017)	It is often difficult for data owners to agree on how a global schema should be represented and which sources are more authoritative
Golshan et al. (2017)	Vast number of data sources
Halevy et al. (2005)	One of the key issues faced in data integration projects is locating and understanding the data to be integrated.
Halevy et al. (2005)	Adapters need to disappear, replaced by sufficiently standard interfaces
Halevy et al. (2005)	To couple data loosely, and for isolated data to become information, formal semantics must be added
Halevy et al. (2005)	Our customers have learned from painful experience that integration progress happens incrementally by developing and exploiting data standards within focused communities of interest.
Halevy et al. (2005)	Performance for real-time access to distributed information
Halevy et al. (2005)	Security: only authorized users should get access
Halevy et al. (2005)	The rate at which new sources of data are appearing in an enterprise also increasing rapidly, as is the different types of information that need to be integrated.
Halevy et al. (2005)	Time invested in schema management per new source integrated increases costs (schema-centric approach)
Halevy et al. (2005)	Users need to be able to find data that they need (search and navigation)
Halevy et al. (2006)	By nature, data coming from multiple sources will be uncertain and even inconsistent with each other
Halevy et al. (2006)	Convincing data owners that concerns will be addressed (e.g. privacy, effects on performance)
Halevy et al. (2006)	Convincing people to share data and offer an incentive
Halevy et al. (2006)	Finding appropriate data
Halevy et al. (2006)	Meaning of data in a different context (actually a business question)
Jarke et al. (2014)	Increased complexity in terms of data volume, heterogeneity, and especially size and number of data models
Jarke et al. (2014)	Semantics of the data, in practitioner terms: the data quality
Lemcke et al. (2012)	60% of the structures are semantically the same, however, only 5% was syntactically similar
Lemcke et al. (2012)	Different structures of the schemas
Lenzerini (2002)	Mutually inconsistent data sources; transformation and cleaning procedures are therefore necessary
Peng et al. (2019)	Heterogeneity of data models makes integration difficult, it is not just aggregating data, it needs to be organized and identifiable, the data consumer needs to understand the data
Salguero et al. (2008)	Multiple sources for the same data element
Sezgin et al. (2019)	Different data definitions for 'the same data'
Zhang et al. (2018)	Data integration is a daunting task because data from different sources can be heterogeneous in syntax (e.g., file for- mats, access protocols), schema (e.g., data structures), and semantics (e.g., meanings or interpretations).
Zhang et al. (2018)	The effort required to connect different sources is substantial due to the lack of clear definitions (i.e., data semantics) of variables, measures, and constructs.

Appendix J Interview Summaries – Planning Process and Data

Note: the interview summaries in this appendix are translated from the original Dutch versions.

Business Analyst (R1)

Function	Business Analyst (R1)
Date(s)	21 July 2020 and 29 July 2020
Duration(s) (mm:ss)	61:07 and 61:07
Setting	Microsoft Teams audio call

Introduction

The respondent (R1) is a business analyst (or business consultant) at GVB within the department IY & Innovation in the group Information Management and Architecture. It is R1's responsibility to formulate business' demands and secure them during projects. R1 also helps the business to define requirements of IT systems to reach higher-level goals, such as market conformity and shorter throughput times. R1 does not specify the functionality of the IT systems, but specifies what the business expects from them. R1 is working for GVB for 5 years now. Before that, R1 gained experience as a manager within a healthcare institution and as a manager of business development for (e-)marketing within the commercial sector. 100% of R1's working time is allocated to the project responsible for the implementation of a new IT system for planning. R1 is the only business analyst within GVB.

Planning as-is

According to R1, the planning is divided into a conceptual phase, timetable phase, crew rostering phase, operational (crew and vehicles) phases and operations and controlling phases. Detour management is also part of the scope, as well as the provision of data to adjacent business areas. The process according to R1 is visualized in Figure A-4.

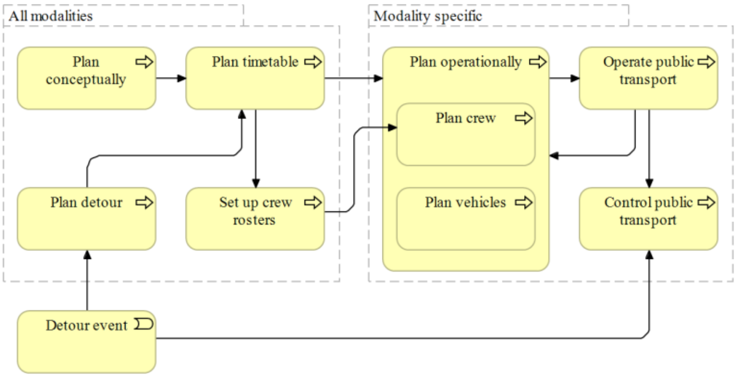

Figure A-4 Planning process according to R1

Plan conceptually

R1 described the first phase as the conceptual phase. This phase is responsible for the development of the public transport plan. This plan has the validity of one year, which makes this a yearly activity. It contains information such as lines, routes, frequencies, first/last journeys, headways, priorities at traffic lights, transfer possibilities, accessibility of vehicles, stops and stations, sustainability and social safety. The plan is mostly based on a cost/benefit analysis of GVB. This analysis is based on expected frequencies and minimal stop-distance for citizens as prescribed by the VRA. Historical journey (OV Chipkaart) data, passenger counting data and the number of crew members and vehicles are also taken into account. It can also be the case that a change in the number of crew members and vehicles. The plan is also based on the previous annual plan and can therefore be seen as a yearly improvement.

The public transport plan needs to be approved by the VRA and most often many negotiations take place. Not only between GVB and VRA, but also with passenger organizations, elderly organizations and more. If the plan gets approval from the VRA, it becomes the contract. This implies that the public transport plan becomes the demand and that the operations are compared to this plan. Canceled journeys need to be reported and reasoning should be provided to the VRA.

The public transport plan as a value when it is made available, however, soon after that the values decreases since the outside world changes day by day.

The responsibility of the conceptual phase is located at GVB's department called traffic direction (verkeersregie) and the plan is designed GVB-wide (all modalities, except for the ferries).

Plan timetable

The design of the timetable is the last step of the planning process which takes place GVB-wide but in cooperation with the different modalities. This is to secure the connections and other mutual dependencies between the modalities.

In this phase, the public transport plan is defined on journey-level. This is called the baseline timetable and is completely based on the plan including the requirements set in there. In the case of GVB, the winter timetable is the baseline and the summer timetable an adaptation on this (fewer journeys).

The public transport plan contains all parameters for the transport services, and they know what are the constraints regarding crew working rules and the availability of them and the vehicles. Historic data is of great value for defining the timetable. The plan, however, can be seen as a reference and also contains information such as important stops for connection and mandatory transfers. Historical data is based on OV Chipkaart and counting systems in the vehicles. Like the public transport plan, also the timetable is a copy of the previous year including some adaptations (most often improvements for the passengers). Except for bigger changes, such as the introduction of the new metro line in Amsterdam in 2018. Back then, the entire public transport plan and timetable were designed from scratch.

Set up crew rosters

This phase does not always start after the timetable was designed, but it can take place simultaneously. This is because the crew roster needs to be known more or less one year before. This means that the rosters are not based on the actual public transport plan (since it cannot be ready yet), but on the previous plan, and adhere to crew working rules and law. It is possible that rosters need to change due to a changed service level. The rosters contain indicative time frames between which the actual duty (explained next) will take place.

Plan operationally – crew

The next step is to define crew shifts based on the timetable. These shifts are most often based on the vehicle blocks (explained next), but not necessarily. Vehicle workings may change (in case of a detour), while shifts remain the same. The shifts need to adhere to crew working rules and law. Duties are connected to the rosters' indicative time frames (explained before), which means that duties and vehicle workings can change without changing the roster. Within GVB, the crew knows 4 days beforehand which exact duty they will have to work (within their roster). From this moment on, GVB is not allowed to change the duty any longer. In some cases this is however still necessary. This does not lead to any financial compensation for the crew member, but GVB has to put maximal effort to refrain from these last-minute changes.

When assigning duties to crew (rosters), GVB should take into account the granted leaves. Leave is approved or declined by the crew planning department of the modality and the crew member's manager. Next to granted leaves, the planning also should take into account education and training for the crew members. Annual training sessions are planned far ahead, since it is about a large number of crew members. However, for example, Covid-19 training is planned within a shorter period: around the 4 days deadline.

It is possible to grant leave for 2 hours, for example for a doctor's appointment. In these cases, a shift can be changed for one particular day, such that the crew member can visit the doctor. The journeys which are taken away from that shift, are assigned to another crew member.

While planning crew members, GVB also accounts for any crew disruptions, such as sickness. These accidents are handled within this phase and not in operations. Whenever the crew member gets sick during operation, this is handled by the control center firstly, but communicated to the crew planning.

Plan operationally – vehicles

Vehicle blocks are made with the timetable as input. Other inputs for this phase are the vehicle (type) requirements and constraints and constraints in the network. Vehicle workings are highly influenced by detours, both scheduled and unscheduled, and any other incidents during operations. Schedules detours are accounted for in this phase, whereas unscheduled detours and incidents are handled by the control center.

Depot management does also partly belong to this phase. It facilitates a part of the operations at the metro, tram and bus modalities. Depot management can be seen as managing vehicles around operations (before and after). Within this process, it is decided which vehicle will operate which vehicle block, but also alignment with the maintenance planning takes place.

Allocating vehicles for operations is mostly done by the maintenance departments of the modalities, as they account for maintenance, service per x kilometers, annual checks, and so on. These departments know how many vehicles are necessary for operations and make these available for the operations department.

Assigning vehicles to vehicle blocks is done by the depot coordinator for tram and the vehicle manager for metro. At the bus modality, specific operations employees account for maintenance and assign vehicle types (not actual vehicles) to a vehicle block. Crew members can choose a vehicle of the predefined vehicle type.

Detour management

Detours take place continuously and are actually always small public transport plans which are developed. On average, GVB deals with scheduled detours per year, which means that 500 multiple-days changes of lines/routes/timetable/service have to be taken into account. These are out of GVB's control, except for GVB Rail Services which can initiate rail works. Reasons for detours can be road/rail works, Covid-19 pandemic, heavy storms, defective bridges, events, etc.

Because of detours, GVB costs can increase (driving more) or decrease (driving less). Passengers should never pay more than they do according to the baseline timetable, therefore often negotiations with the VRA take place because of lower income or higher operational costs. The planning of detours is the responsibility of traffic direction (verkeersregie) and is transferred to operations after planning. The process takes place pro-actively. Reactive detour management takes place during operations and is handled by the control center.

Operations

Operations is seen as the moment from which the vehicle leaves the depot to start its journey. From that moment on, a combination of a vehicle, crew member(s) and journey is made.

During operations, unforeseen circumstances can lead to incidents, such as accidents, sickness, malfunction, etc. In these cases, every three operations aspects (vehicle, crew member and journey) need to be managed. An integrated solution is necessary for this and is managed by the control center. Ferries are not managed by the control center of GVB.

Next to depot management taking place right before operations, it also takes place during operations. It ensures that buses, trams and metros enter the depot in the way which suits the next operational activities the best. Right now, this process is carried out manually with Excel files and, for bus, manually communicated to crew members through LED displays.

Business processes dependent on planning data

Next to the obvious business processes such as payroll and reporting, also travel information for passengers is a large consumer of planning data. Furthermore, BI uses a vast amount of data for providing insight into passengers, income, ticket inspections, etc.

Planning to-be

The process of the to-be situation is similar to the as-is process (Figure A-4). This means that the phases are still identical. However, the big difference is that the new planning process will be supported by one application. This application ensures – among other things – that changes in one process phase are immediately transferred and communicated to other adjacent phases. This is called an integrated planning solution. In the current situation, every phase is an island and hardly any changes can be made as soon as the planning is transferred to the next phase.

Such an integrated solution can be explained in two different aspects. The first being time. As soon as an early phase changes something that influences the following phases, the application will show which other phases also need to change. An example is a bigger detour for which some lines need to change. It is very likely that also vehicle blocks and crew duties need to change, because they do not align with the new lines anymore. The second aspect is the combination of vehicle, crew member and journey. Whenever one of these three aspects changes or is missing, then also the other two aspects need to be managed somehow.

When having an integrated planning process, it is possible to:

- Optimize the planning on every aspect (lines, timetable, frequencies, stops, vehicle blocks, crew, operations, incidents, emergencies).
- Predict how many crew members and vehicles are necessary for the future. For this, data about the vehicles, crew pool, expected number of passengers, maintenance planning and detours are necessary.
- Predict how many crew members and vehicles are necessary for detours and how much this will cost. For this, data about the vehicles, crew pool, expected number of passengers, lines, detours, crew costs and vehicle costs are necessary.
- Let the application show what the best alternative is given an incident. It tries to stay as close to the original timetable as possible. This can become very expensive, which means that more alternatives should be given after which one can be selected. For this, data about the incident, vehicles, routes, stops, duties, etc. is necessary. After an incident, the application

will propose how to get back to the normal timetable as easily as possible. These decisions take place based on experience, for which rulesets should be implemented in the application such that it can decide by itself.

- Request crew members to work extra. The application shows which crew members can be contacted (accounting for crew work rules, duties, rest, preferences, etc.)
- See which crew member is on which vehicle. Right now, this is a separate process and the control center needs to request at the planning who is actually on a vehicle.
- Introduce depot management solutions. This will help by planning the vehicle assigning and helps to manage disruptions. It also provides solutions when not enough vehicles are available: then it will be suggested which vehicle block can be canceled the best.
- Offer a self-service portal for crew members. Within this portal, crew members can see their roster, shifts, ask for leave, request shift, request shift changes with colleagues.

A lot of crew data is necessary within the planning process. HR's role includes the storage and administration of the following data: employees, contract, contract hours, manager and amount of leave. Granting leave is done within the planning application, since HR cannot do so. However, this data needs to be communicated with HR, to have the correct data about working hours and leave. Next to leave, sickness also needs to be registered at HR, since within GVB one hour of the sickness is always unpaid. Assumption: this data is sent to and registered by HR annually.

Within an integrated planning solution, it is important to have an employee who oversees all planning phases and their integrations. This employee should ensure that processes are aligned perfectly and that the planning personnel responds to each other work the way it is intended. When conflicts arise, the employee has to decide who's decision is the most important and which planning personnel needs to change their work. A change is foreseen to centralize the planning, instead of decentral planning per modality. This ensures that situations such as having a giant detour for metro or tram do not take place together with major maintenance tasks for buses.

Cost savings can also be found in self-rostering for crew members. The crew can (partly) decide which days and shift-types they want to work. The expectation is that this will save costs since the crew members will have higher job satisfaction and a lower sickness rate. GVB decided not to implement this right now, but will be implemented in the (near) future.

Since the introduction of the new planning application will lead to changes in the business processes, change management is very important. Employees see their work change and disappear, however, their experience and knowledge are very important to be included in the application as rulesets. Furthermore, it is expected that the type of planning personnel changes from having developing/changing capabilities to personnel who is capable of analyzing outcomes and changing ruleset parameters.

Product Owner Team Planning (R2)

Function	Product owner team planning (R2)
Date	22 July 2020
Duration (mm:ss)	74:59
Setting	Microsoft Teams video call

Introduction

The respondent (R2) is product owner of team planning for half a year. Before, since 2008, R2 has been in the roles of functional application manager, application coordinator and information manager. In the role of information manager, R2 was responsible for the domains of planning and BI in the period 2014 – 2020.

Planning as-is

Globally, the planning process can be divided into three main steps: define lines and routes, create timetable and planning vehicles and personnel. The planning of vehicles and crew are done separately, as well as the creation of crew rosters. After the planning is made, public transport is operated accordingly. After operations, the actual journeys are compared with the planned journeys, which can lead to input for the planning. The planning process is visualized in Figure A-5.

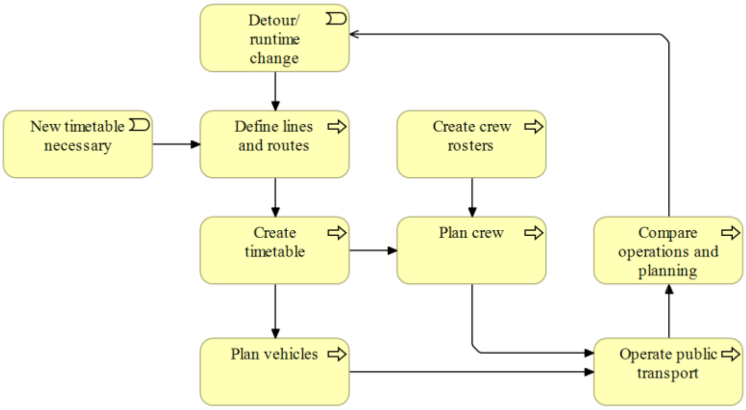

Figure A-5 Planning process according to R2

Define lines and routes

Lines and routes need to comply with the VRA's requirements for public transportation. Next to these, requirements for the maximum distance between stops and distance between stops and citizens are set. Every citizen needs to be able to reach a stop within X meters of their home address. Furthermore, the VRA comes with requirements regarding vehicle types, crew, etc. These can be seen as requirements regarding vehicles and personnel.

On strategical level, it is considered how much personnel is necessary and how many of them need to be flexible. This is relevant for contract management regarding own personnel but also contracts with employment agencies. The crew planning department also needs this information to know whether they need to hire more personnel, start campaigns, etc.

Further upfront the actual operations it is necessary to forecast the vehicle demand, this is necessary because the throughput time for acquiring new vehicles is longer than for personnel. Important factors for this process are mobility developments, demographical growth, tourist expectations, etc. This is outside the scope of the planning process but is considered an important input for the network design.

Create timetable

Based on the previous step, the timetable is made. This happens normally one year in advance and consists of two different timetables: winter and summer. The first being the 'normal' timetable, the latter the timetable with less journey, most often less rush-hour journeys.

Create crew rosters

The timetable defines the work to be done, however, the crew rosters are not based on the actual timetable. This is impossible, since the crew rosters need to be defined at least nine months in advance, whereas the exact timetable is not always known and is also prone to change. This latter is the case due to detours, events, etc., which are not accounted for in the normal timetable.

Detours

Detours are an important aspect in the planning process, since they often require changes in the network, plans and timetable. Sometimes these changes require adaptations in the timetable, vehicle blocks and crew duties. Because sometimes lines are connected and that detours can add up to each other, the process can be difficult.

It can also be the case that detours imply changes in fares. Normally the passenger pays for journeys, this can be more or less compared to the normal timetable. In some exceptional cases, this is not the case, e.g. when journeys become way more expensive. In such a case, GVB tries to get compensated by the VRA, since the costs increase while the income stays equal. This information comes from traffic direction (verkeersregie) and is not seen as part of the planning process.

Plan vehicles

The timetable is known and based on this, the vehicle blocks are planned. The goal is to plan as efficiently as possible, which means that it is tried to drive as many journeys as possible with the lowest number of vehicles. This will lead to lower costs. It can never be the case that more vehicles are planned than we actually have. For this reason, it is important to monitor this process. Bigger detours (for tram and metro lines) can be an exception to this rule.

Assigning the vehicles to the vehicle blocks does not really happen right now. During operations, consultations between operations and vehicle management take place to define which vehicles can be assigned for operations and which are planned for maintenance. Right now this happens with Excel lists and PowerBI reports, for which information from the maintenance department comes in via phone/email. The lists of actual vehicles assigned to vehicle blocks are made by hand every day and printed for the drivers.

Plan crew

To create crew duties, data from the timetable, vehicle blocks, workers' council, labor agreements and company-specific rules are taken into account. The crew duties need to be approved by the workers' council.

Many duties are exchanged between drivers. In case this is not conflicting with regulations and law, this is possible. Changes that happen often are for example between Christian and non-Christian holidays/crew members.

Leave and sickness registration is necessary for the planning process as well, as the planning personnel needs to know whether or not a crewmember can work or not. This data is also necessary for HR.

Operations

Operations is closely connected with planning and maintenance. However, operations does not always be the same as planning. This is the case when incidents happen, unscheduled detours are necessary, vehicle malfunctions take place and when crewmembers are called sick.

Comparing operations and planning

During and after operations, it is consulted to which degree the operations took place compared to the planning. It is verified whether the operations were according to the plan. This information goes via the reporting process to the VRA. In these reports, punctuality and information about canceled journeys are accounted for. Explanations are given by the responsible departments. Furthermore, the comparison is used for the runtime check, which can be used to improve the timetable, vehicle blocks and crew duties.

Business processes dependent on planning data

Travel information is very dependent on planning data. This is the static data, since the real-time data originates from the operations department. The information is necessary to inform passengers about the services, and thus, in the end, earn money. The fare data for passengers also needs to come from the planning department. The decision about prices is done by the VRA and Metropoolregio Amsterdam. GVB has hardly any influence on this.

Continuous planning

As a PTO, you are continuously adapting the planning to the current circumstances. This is reactively changing the planning. Proactive changes also take place, e.g. detour planning. Every minute a PTO changes their services based on (earlier) finished road/rail works, a bad performance due to congestion, events which demand more vehicles/higher frequencies, bad state of bridges, etc. This planning is necessary, since the annual planning cannot account for these circumstances. As the planning can change every day, R2 calls this 'continuous planning'. This means that every aforementioned step needs to be carried out every time, since a change in the timetable often also means a change in vehicle blocks, crew duties, etc. Another example is the reaction on pandemics such as Covid-19, but also to account for bigger vehicle maintenance projects, such as the introduction of a new board computer in vehicles.

Continuous planning is relatively easy for the bus department, since a bus is way more flexible compared to trams, metros, ferries, etc. Metro is even more difficult, since a temporary stop cannot be made (which can be done for bus and tram). This flexibility however most often results in more work for bus, since more possibilities can be offered to passengers (other routes, stops, etc.). These changes need to be visible everywhere (travel information, but also information for drivers and planning).

Planning to-be

An improved version of the planning process consists of the same steps as for the as-is planning process (Figure A-5). Some advantages of a new planning process are the flexibility of planning (software suites which help planning personnel better). Furthermore, every individual step can be integrated in the future, which allows changes in one step to be immediately visible in the following steps. The GVB to-be planning process will use a software suite to cover all planning steps and has the following advantages:

- The new software suite will allow creating scenarios after which the best can be chosen. This also shows forecasts for personnel and vehicles, also for detour planning this is useful. This forecast ensures that the enterprise can react on time. Necessary data for this process is data about routes, lines, stops, vehicles, crew, working times, regulations, costs.
- Planning electrical vehicles requires another approach than diesel vehicles. A vehicle running on diesel can run the entire day, whereas electrical vehicles need to be charged in between. In the best scenario, this happens not too often, since this will result in higher vehicle demand, but also because the physical places for charging are not unlimited. The range of electrical vehicles, and thus their batteries, is not unlimited either. This is highly influenced by weather conditions (air-conditioning and heating), driving style of crewmembers, number of passengers, detours, age of batteries, etc. These parameters are only known very shortly before the operations. This means that in the best scenario, the vehicle blocks are planned only short before operations. The smarter planning of battery-equipped vehicles is therefore necessary and facilitated by the new software suite.
- Better alignment with the maintenance departments is possible because the vehicle forecast is always calculated. This means that the maintenance department exactly knows how many vehicles should be available at which day/time and, thus, on how

many vehicles they can carry out maintenance. This ensures that situations such as bigger maintenance plans will not take place together with bigger detours for which more vehicles are necessary.

- Right now, timetables and crew are planned, but assigning vehicles to vehicle blocks is not really. In the future, also vehicles need to be planned carefully (just like crewmembers), which also considers aspects such as predictive maintenance and mileage of the vehicles. It is always necessary to react to unforeseen circumstances (malfunction), but the number of ad-hoc decisions can be decreased, which can improve efficiency.
- Self-rostering for crewmembers is an important aspect of the field of HR. Crew has more control over their working times and duties, which can lead to operational benefits.
- In the current process, every crew duty needs to be approved by the workers' council. The parameters on which these duties are approved/rejected should be an important input for the system. In the best situation, the system only generates crew duties that always comply with the rules, regulations and wishes from the workers' council and law.
- Scenario planning is important and is facilitated by the software suite, as explained before. Parameters can be changed by planning personnel to find a more optimized planning.
- The approval/rejection of leave requests could take place automatically, as long as data of forecasts, actual duties, crew members, etc. is available.
- Real-time access to any data about the public transport, vehicles and crew members.

Appendix K Planning Tasks within the Planning Process

Table A-13 Total list of planning tasks

ID	Phase	Planning task	Source
1	[PL]	Create route network	Scholz (2016)
2	[PL]	Define stops	Scholz (2016)
3	[PL]	Define link courses	Scholz (2016)
4	[PL]	Define routes	Scholz (2016)
5	[PL]	Define runtimes	Scholz (2016)
6	[PL]	Define runtime profiles	Scholz (2016)
7	[PL]	Add fare information (tickets, products, properties)	Scholz (2016)
8	[PL]	Specify relief points at locations that are services by many vehicles on different routes, i.e. at main traffic points (in this way, the driver has to opportunity to return to the wheel of a vehicle after the break as quickly as possible)	Scholz (2016)
9	[PT]	Define trips for route (headway and exact times)	Scholz (2016)
10	[PL]	Define connections/transfers	Scholz (2016)
11	[PL]	Define vehicle type to be deployed	Scholz (2016)
12	[PL]	Define restrictions that may result from the routing (e.g. vehicle height)	Scholz (2016)
13	[PT]	Specify necessary special equipment (such as the accessible entrance to vehicle)	Scholz (2016)
14	[PT]	Adjust trips/connections based on historical data (based on trip analysis and frequent deviations)	Scholz (2016)
15	[PC]	Create roster lay-outs	Scholz (2016)
16	[PC]	Assign people to roster lay-outs	Scholz (2016)
17	[PC]	Schedule absence (holidays, training)	Scholz (2016)
18	[PC]	Monitor qualifications of employees and their renewal	Scholz (2016)
19	[SV]	Determine layover times to ensure the stability of the schedule	Scholz (2016)
20	[SV]	Create service packages for vehicles of a particular type (trips in sequential order)	Scholz (2016)
21	[SV]	Plan deadhead trips (non-revenue trips, pull-in (to depot), pull-out (from depot), turning, shunting, service blocks, building formations)	Scholz (2016)
22	[SV]	Specify whether a vehicle should remain occupied during vehicle breaks (because of danger/vandalism at a particular location)	Scholz (2016)
23	[SC]	Plan the work to be done by personnel (right balance between efficient and stable duty schedules)	Scholz (2016)
24	[AV]	Assign specific vehicles to scheduled vehicle blocks	Scholz (2016)
25	[AV]	Informs personnel dispatch about the location of a vehicle or tells the driver directly	Scholz (2016)
26	[AV]	Monitors inspection dates (time-based or mileage-based)	Scholz (2016)
27	[AV]	Plan maintenance workshop visits (monitor maintenance)	Scholz (2016)
28	[AV]	Plan parking of vehicles (depot management)	Scholz (2016)
29	[AV]	Perform capacity review	Scholz (2016)
30	[AV]	Ensure uniform mileage on long-term	Scholz (2016)
31	[AV]	Reserve replacement vehicles	Scholz (2016)
32	[AV]	Local depot balance to avoid trips with empty vehicles	Scholz (2016)
33	[AV]	Send technicians to vehicle	Scholz (2016)
34	[AV]	Request vehicles to a depot	Scholz (2016)
35	[AV]	Assess damages and agrees on deployment restrictions with the control center	Scholz (2016)
36	[AV]	Future: vehicles requestions vehicle workings if they know they are damaged for example	Scholz (2016)
37	[AV]	Shortly said three main areas for vehicle dispatching: vehicle allocation, maintenance management and performance record.	Scholz (2016)
38	[AC]	Monitors duty sign-ons	Scholz (2016)
39	[AC]	Ensures that other employees can jump in in case of last-minute absence (standby duties, 'spares')	Scholz (2016)
40	[AC]	Schedule absence (doctor's appointments, work council activities)	Scholz (2016)
41	[AC]	Solves conflicts regarding personnel assigning	Scholz (2016)
42	[AC]	Records planned-actual deviations as the dispatch data form the basis for the payroll	Scholz (2016)
43	[AC]	Inform personnel of any changes during duty sign-on	Scholz (2016)
44	[AC]	Match duties with shift class/roster	Scholz (2016)
45	[AC]	Arrange spare duties	Scholz (2016)
46	[AC]	Arrange standby duties	Scholz (2016)
47	[AC]	Arrange duty swapping	Scholz (2016)
48	[AC]	Taking requests into consideration (certain calendar days, shift class, shift length, specific duty, etc.)	Scholz (2016)
49	[AC]	Depot-exit check (most often done by control center)	Scholz (2016)
50	[AC]	Last-minute dispatch (sickness)	Scholz (2016)
51	[AC]	Manage disruptions	Scholz (2016)
52	[AC]	Record actual data	Scholz (2016)
53	[AC]	Levels of planning: long-term (shift class schedule, personnel roster); planned level (dispatch level, more specifically), actual level (on day X and afterward)	Scholz (2016)
54	[AV]	Depot-exit check for employees and vehicles	Scholz (2016)
55	[CT]	Keep the transport operations as close to the schedule as possible, comparing planned and actual values, check the location in relation to the transport network and the timetable	Scholz (2016)
56	[CT]	Recognize disruptions as quickly as possible	Scholz (2016)

57	[CT]	React on disruptions	Scholz (2016)
58	[CT]	Manage larger deviations (smaller are managed by drivers)	Scholz (2016)
59	[CT]	Change routes	Scholz (2016)
60	[CT]	Distribute vehicles evenly	Scholz (2016)
61	[CT]	Monitor connections (keep or break connection)	Scholz (2016)
62	[CT]	Managing accidents and missed connections	Scholz (2016)
63	[CT]	React to events which results in disruptions	Scholz (2016)
64	[CT]	Predefine actions for common events	Scholz (2016)
65	[CT]	Improvise action if predefined action is non-existent	Scholz (2016)
66	[CT]	Analyze deviations in real-time and try to get back to schedule	Scholz (2016)
67	[CT]	Strictly speaking, transport control also extends to the execution of trips. The monitoring tasks undertaken by control center employees are often extended to all installations and systems that can show disruptions: passenger information systems, ticket machines, turnstiles, lifts, escalators, cameras, as well as IT components such as web servers or network connections, and even software interfaces to other systems (e. g. for timetable information). General technical disruption management is added to the operational disruption management.	Scholz (2016)
68	[CT]	Informing passengers about actions done by the control center	Scholz (2016)
69	[DN]	Developing new service (routes and terminals)	Ceder (2016)
70	[PL]	Developing new routes, changing route structure, realignment of terminals and stops	Ceder (2016)
71	[PL]	Developing new (changes in) express and local routes, short-turn and zonal services	Ceder (2016)
72	[PL]	Updating vehicle size, frequency, and departure, running, layover and recovery times	Ceder (2016)
73	[PT]	Updating vehicle size, frequency, and departure times	Ceder (2016)
74	[PT]	Updating the design of vehicle scheduling, short-turn, and zonal services	Ceder (2016)
75	[PL]	Determine interchanges and terminals	Ceder (2016)
76	[PL]	Design of network of routes and stops	Ceder (2016)
77	[PT]	Analysis of frequencies and headways	Ceder (2016)
78	[PT]	Analysis and construction of alternative timetables	Ceder (2016)
79	[SV]	Fleet size analysis with and without interlinings (determine lower bound, reduction via deadheading, reduction via shifting departures times (can be again input for the timetable))	Ceder (2016)
80	[SV]	Analysis and construction of blocks	Ceder (2016)
81	[SC]	Analysis and construction of crew assignments	Ceder (2016)
82	[PC]	Establishment of crew rosters	Ceder (2016)
83	[PL]	Frequency and headway determination	Ceder (2016)
84	[PT]	Evaluate optional timetables in terms of required resources	Ceder (2016)
85	[PT]	Improve the correspondence of vehicle departure times with passenger demand while minimizing resources	Ceder (2016)
86	[PT]	Permit, in the timetable construction procedure, direct vehicle-frequency changes for possible exceptions (known to the planner/scheduler) that do not rely on passenger-demand data	Ceder (2016)
87	[PT]	Allow the construction of timetables with headway-smoothing techniques (similar to those performed manually) in the transition segments between adjacent time periods	Ceder (2016)
88	[PT]	Integrate different frequency-setting methods and different timetable construction procedures	Ceder (2016)
89	[SV]	Maintaining a balance in the number of vehicles starting and ending at a depot	Ceder (2016)
90	[SV]	Establish chains of daily trips or vehicle blocks	Ceder (2016)
91	[SC]	Assign drivers according to the outcome of vehicle scheduling	Ceder (2016)
92	[SC]	Group duties into rosters	Ceder (2016)
93	[PL]	Create transport plan including information about lines, routes, frequencies, times of first/last rides, headways, priorities at traffic lights, connections, transfers, accessibility (vehicles and stations/stops), sustainability and social safety	Interview R1
94	[PL]	Define contract with the transport authority	Interview R1
95	[CT]	Reporting about canceled rides	Interview R1
96	[PT]	Create timetable (last modality-wide step at GVB)	Interview R1
97	[PT]	Transform the transport contract with the transport authority into individual rides	Interview R1
98	[PC]	Create crew roster layouts with indicational working times, 9 months ahead	Interview R1
99	[SC]	Create crew duties based on the timetable	Interview R1
100	[AC]	Assign crew duties to roster layouts and crew to duties	Interview R1
101	[AC]	Accept/reject leave requests	Interview R1
102	[PC]	Planning of trainings	Interview R1
103	[AC]	Manage unplanned leave right before the operation	Interview R1
104	[CT]	Manage unplanned leave during operations	Interview R1
105	[AC]	Change crew duties for one day (e.g. doctor's appointment, first 2 rides other crew member)	Interview R1
106	[SV]	Create vehicle blocks	Interview R1
107	[SV]	Adapt vehicle blocks due to detour (proactive)	Interview R1
108	[CT]	Adapt vehicle blocks due to unexpected circumstances during operations	Interview R1
109	[AV]	Depot management right before leaving the depot, during depot activities and while entering the depot	Interview R1
110	[AV]	Provide vehicles for operation (done by maintenance department)	Interview R1
111	[AV]	Assign vehicle to vehicle block	Interview R1
112	[PL]	Plan detours (going through all phases)	Interview R1

113	[CT]	React on unforeseen circumstances during operations and manage ride, crew member and vehicle	Interview R1
114	*	Payroll	Interview R1
115	*	Reporting about travelers, sales, controls, and more	Interview R1
116	*	Travel information	Interview R1
117	ALL	Changes in planning phases urge changes in other adjacent planning phases	Interview R1
118	ALL	If something changes in a planning phases, this should also be changed in other phases	Interview R1
119	ALL	Always try to find the best combination between vehicle, crew and ride. If one of these is missing, the other 2 should be taken care of too.	Interview R1
120	ALL	Optimize on every level (lines, timetable, frequencies, stops, vehicle blocks, crew, operations, disruptions, emergencies)	Interview R1
121	[PL]	Predict how many vehicles/crew members are necessary	Interview R1
122	[PL]	Predict vehicle/crew for detour planning	Interview R1
123	[CT]	Give alternatives when disruption happens	Interview R1
124	[CT]	Providing reserve crew members, also when the first group is used (standby, free, etc.)	Interview R1
125	[CT]	Being able to see which driver is on which vehicle on which ride	Interview R1
126	[AV]	Depot management for leaving the depot. If vehicle X can't drive, vehicle Y and Z might be stuck in the depot	Interview R1
127	*	Self-service portal for crew members (check roster, duties, leave, requests, swap, etc.)	Interview R1
128	[PC]	Take into account crew members contract hours (also in SC and AC)	Interview R1
129	[SC]	Take into account crew members contract hours (also in PC and AC)	Interview R1
130	[AC]	Take into account crew members contract hours (also in SC and PC)	Interview R1
131	*	HR needs data about worked hours, leave, sickness, etc.	Interview R1
132	ALL	Ensure every phases is perfectly matching other phases, ensure no errors in the planning	Interview R1
133	[PC]	Self-rostering possibilities	Interview R1
134	[PL]	Define lines and routes	Interview R2
135	[PT]	Design timetable based on line plan and frequencies	Interview R2
136	[SV]	Create vehicle blocks	Interview R2
137	[AV]	Assign vehicles to vehicle blocks	Interview R2
138	[SC]	Create crew duties	Interview R2
139	[CT]	Operate according to the planning (as much as possible)	Interview R2
140	[CT]	Compare operations and planning	Interview R2
141	[PL]	Create lines and routes	Interview R2
142	[PL]	Crew prediction (relevant for contracts, forecast for external agencies, planning, campaigns, hire new personnel, etc.)	Interview R2
143	[PL]	Calculate vehicle prognosis for purchasing new vehicles	Interview R2
144	[PC]	Create crew roster layouts with indicative working hours and days-off	Interview R2
145	[PL]	Plan detours, due to which lies, routes, vehicle blocks, crew duties might need to change too	Interview R2
146	[PL]	Change prices in case of detours	Interview R2
147	[SV]	Create vehicle blocks	Interview R2
148	[SV]	Create as efficient as possible vehicle blocks, which means less vehicles and thus less costs	Interview R2
149	[AV]	Plan vehicles on vehicle blocks for operation	Interview R2
150	[AV]	Create pull-out lists for vehicles (depot management)	Interview R2
151	[SC]	Create crew duties	Interview R2
152	[SC]	Approve crew duties by workers council	Interview R2
153	[AC]	Swap crew duties between crew members (might be requested by crew members)	Interview R2
154	[AC]	Leave and sickness registration	Interview R2
155	[CT]	Operate public transport according to planning	Interview R2
156	[CT]	Adapt planning due to unexpected circumstances, emergencies, malfunctions, sick crew members, etc.	Interview R2
157	[CT]	Compare planning and operations	Interview R2
158	*	Data about punctuality and numbers about canceled rides for reporting to transport authority	Interview R2
159	[CT]	Compare driving times (might be input for timetable)	Interview R2
160	*	Travel information	Interview R2
161	*	OV Chipkaart system needs information about lines, prices and plans	Interview R2
162	[CT]	Reactively adapting planning based on circumstances	Interview R2
163	[PL]	Proactive adapting planning (for detours)	Interview R2
164	ALL	Continuous planning. Every day the planning is adapted to match the reality as well as possible. Roadworks, events, traffic jams, bridges, etc. influence this	Interview R2
165	ALL	Faster and more flexible planning when phases are integrated, changes can be rolled out more easily	Interview R2
166	[PL]	Create forecasts and scenarios	Interview R2
167	[SV]	Smarter planning of battery-equipped vehicles (zero-emission)	Interview R2
168	[AV]	Alignment with the planning of maintenance	Interview R2
169	[AV]	Planning of vehicles (taking into account maintenance planning)	Interview R2
170	[PC]	Self-rostering	Interview R2
171	[SC]	Approval from workers council for crew rosters and shifts	Interview R2
172	[AC]	Approval for leave requests	Interview R2
173	[AC]	Real-time possibility to see shifts, leave possibilities, vehicles for a shift, everything accessible via a self-service portal for crew members	Interview R2
174	[PL]	Model travel demand	Friedrich et al. (2016)

175	ALL	Developing a scenario consists of a line network, timetabling, vehicle deployment plan and driver deployment plan	Friedrich et al. (2016)
176	*	Dynamic passenger information	Scholz (2016)
177	*	Quality and contract manager information	Scholz (2016)
178	*	Accounting, customer service, sales, vehicle management	Scholz (2016)
179	*	Timetable display and information	Scholz (2016)
180	*	Sales and distribution	Scholz (2016)
181	*	Settlement, performance analysis and quality management	Scholz (2016)
182	[PT]	Timetable development	Ceder (2016)
183	[SV]	Vehicle scheduling	Ceder (2016)
184	[SC]	Crew scheduling	Ceder (2016)
185	[PL]	Network route design	Ceder (2016)
186	[PT]	Timetable development	Ceder (2016)
187	[SV]	Vehicle scheduling	Ceder (2016)
188	[SC]	Crew scheduling	Ceder (2016)
189	[CT]	Controlling events during operation	Scholz (2016)
190	[SV]	Schedule vehicles, vehicle shift starting with outbound and ending with inbound trip, minimizing costs thus minimizing the number of vehicles	Nagy & Tick (2019)
191	[SC]	Staff scheduling, minimizing costs is the goal	Nagy & Tick (2019)
192	ALL	Integrated vehicle and crew scheduling problem	Nagy & Tick (2019)
193	[PT]	Plan transport services based on historical data and requirements from the transport authority	RQ1
194	[SC]	Create crew duties based on timetable and vehicle blocks	RQ1
195	[AC]	Assign crew members to crew duties	RQ1
196	*	Operations of transport services	RQ1
197	*	Sales	RQ1
198	*	Travel information (online and offline)	RQ1
199	*	Social safety	RQ1
200	*	Reporting and business intelligence	RQ1
201	*	Operations	RQ1
202	*	Real-time travel information (online)	RQ1
203	[PL]	Plan detours	RQ1
204	[SC]	Staff scheduling	RQ1
205	[SV]	Vehicle scheduling	RQ1
206	[AC]	Staff scheduling (assign)	RQ1
207	[AV]	Vehicle scheduling (assign)	RQ1
208	[CT]	Transport operations	RQ1
209	*	Business intelligence	RQ1
210	*	Vehicle and vessel management	RQ1
211	*	Social safety	RQ1
212	*	Travel information	RQ1
213	*	Sales and collection	RQ1
214	*	HR management	RQ1

Legend:
[DN] = Design network; [PL] = Plan lines; [PT] = Plan timetable; [PC] = Plan crew roster; [SV] = Schedule vehicle bocks; [SC] = Schedule crew duties; [AV] = Assign vehicles; [AC] = Assign crew; [CT] = Control transport; ALL = All planning phases; * = task outside planning which consumes planning data

Appendix L In-depth Planning Process Analysis and its Data Requirements

Plan Lines [PL]

[PL1] Create Route Network

[PL1] Create route network can be seen as the main planning task of the planning phase *Plan lines [PL]*. It is about the actual development of the route network. It contains the actual topology, routes and lines of the public transport network. Moreover, runtimes, headways, frequencies, first and last journeys, transfers, traffic light priority and vehicle type (restrictions) are also set. The outcome of this task can be seen as the contract between the PTO and the transport authority (TA). This task normally takes place one or several times per year and is adapted afterward by *[PL5] Plan detours* and during operations by the control center. An overview of upstream and downstream data dependencies is given in Table A-14.

Table A-14 SIPOC [PL1] Create route network

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
TA	Transport authority requirements	[PL1] Create route network	Infrastructure	[PL], [PT], [SV]
TA	Utilities and land-usage			Compliance
PTO	Infrastructure			Employees
TA				Operations
PTO	Mobility behavior			Sales
TA				Safety & enforcement
PTO	Transport supply			Travel information
[PL]	Detour planning		Routes	[PL], [PT], [SV]
				Compliance
				Employees
				Operations
				Sales
				Safety & enforcement
				Travel information
			Line plan	[PL], [PT], [SV]
				Compliance
				Employees
				Operations
				Sales
				Safety & enforcement
				Travel information
			Vehicle forecast	[SV]
				Vehicle management & maintenance
			Crew forecast	[PC], [AC]
				HR

Upstream Data Dependency

A lot of different vast amounts of data is necessary to plan the PTO's route network including the main characteristics of every line as explained above. In most cases, a transport authority orders a PTO to offer public transport services. This order contains a lot of information about the public transport demanded by the TA. Also data about utilities (Friedrich et al., 2016) and land-use (Ceder, 2016) are necessary to correctly create the route network. Based on this more or less static data, also information about the mobility behavior (historic data about journeys, both from PTO as well as from TA), transport supply (how many available vehicles and depot/station/siding capacity) and, in some cases, data from *[PL5] Plan detours*.

Downstream Data Dependency

The data created within this planning task is the basis of the public transport operation. For this reason, the infrastructure (including information such as locations of bus stops, stations, railways, etc.), routes and line plan (including all kinds of information as explained before), are shared with many internal and external (in the form of travel information) stakeholders. As can be deducted from Table A-14, many other planning phases are dependent on the routes and line plans generated in this task as well. Next to this public transport basis data, also vehicle and crew forecasts are made, to ensure that the created route network can be operated (resources such as crew and vehicles can be scaled down or up).

[PL2] Determine Relief Points

According to Transmodel (2019), a relief point is a location "where a relief is possible, i.e. a driver may take on or hand over a vehicle". These locations are of high importance for the planning process, since a crew change (necessary due to crew work rules and optimizing the planning) can only take place at these locations. At these locations, the crew can have their break. This planning task most often takes place together with the design of new routes and line plans in *[PL1] Create route network*. However, with minor changes in the routes and line plans, the relief points do not always need to be reconsidered. An overview of upstream and downstream data dependencies is given in Table A-15.

Table A-15 SIPOC [PL2] Determine relief points

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PL]	Infrastructure	[PL2] Determine relief points	Relief points	[PT], [PC], [SV], [SC]
	Routes			
	Line plan			
PTO	Facilities			
External	Crew work rules			
PTO				
PTO	Actual driving time			

Upstream Data Dependency

To define the best locations for relief points, the infrastructure, route network and lines need to be known. The most optimal relief points are located on locations at which many vehicles on different routes pass by, since then the crew member can return to their job after their break as soon as possible. To determine these locations, data about canteen facilities, crew working rules and the actual driving times need to be known.

Downstream Data Dependency

The relief point locations are an important input for the next phases in the planning process. The planning process phases *Plan timetable [PT]* and *Schedule vehicle blocks [SV]* depend on relief points in the way that they have allowed time for crew changes at these specific relief points. Furthermore, *Plan crew roster [PC]* might use information about relief points for start and end locations for duties (depends on the PTO), whereas *Schedule crew duties [SC]* depends on relief points in the sense that breaks and driver changes need to take place at these locations.

[PL3] Provide Fare Information

The provision of fare information is seen as the responsibility of the planning process by literature (Scholz, 2016) and both interviews. Fare information consists of information about prices, tickets, subscriptions and any other information which has to do with the pricing of public transport services for customers. This planning task most often takes place once a year, or in case new tickets and/or subscriptions are introduced. An overview of upstream and downstream data dependencies is given in Table A-16.

Table A-16 SIPOC [PL3] Provide fare information

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
TA	Transport authority requirements	[PL3] Provide fare information	Pricing information	Operations
[PL]	Infrastructure			Sales
	Routes			Travel information
	Line plan			
PTO	Fare information			

Upstream Data Dependency

The fare information (prices, tickets, subscriptions, etc.) depends on requirements from the transport authority and decisions made at the PTO. This fare information needs to be mapped to infrastructure, routes and line plans, to calculate fares for different journeys.

Downstream Data Dependency

Fare information has only downstream dependencies towards other business areas of the PTO. Operations, sales and the provision of travel information are dependent on fare information provided by the planning process.

[PL4] Develop Scenarios

For common unscheduled events (disruptions during operations), scenarios are developed beforehand to easier the work during operations. One can think of predefined detours and the planning of replacement buses in case a metro, tram or ferry cannot operate. This planning task most often takes place together with the design of new routes and line plans in *[PL1] Create route network*. An overview of upstream and downstream data dependencies is given in Table A-17.

Table A-17 SIPOC [PL4] Develop scenarios

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PL]	Infrastructure	[PL4] Develop scenarios	Scenarios for transport control	[PT]
	Routes			Operations
	Line plan			
PTO	Crew			
PTO	Vehicle			

Upstream Data Dependency

Scenarios for transport control are dependent on the infrastructure and the already planned routes and line planning (in [PL1]). Based on this planning, common scenarios are developed. In these scenarios, crew and vehicle information is taken into account for calculating extra costs. The goal is to minimize the deviation from the planning together with decreasing the extra costs.

Downstream Data Dependency

The scenarios need to be further defined into a timetable, which results in a downstream dependency towards *Plan timetable [PL]*. The other downstream dependency is towards operations (scenarios).

[PL5] Plan Detours

Detours can be planned if they are known by or communicated to the PTO in advance. This detour planning includes every step which is normally also taken when planning the entire public transport plan. One can think of roadworks, events and any other reason why the normal public transport plan cannot be operated accordingly. According to one interview respondent, for GVB, there are more or less 500 detours per year. The difference with *[PL4] Develop scenarios* is that the planning of detours can be done way more detailed and is planned, since the exact detour information is known beforehand. An overview of upstream and downstream data dependencies is given in Table A-18.

Table A-18 SIPOC [PL5] Plan detours

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PL]	Infrastructure	[PL5] Plan detours	Detour planning	[PL]
	Routes			Employees
	Line plan			Travel information
[PT]	Timetable		Crew forecast	[PC], [AC]
PTO	Crew			HR
PTO	Vehicle		Vehicle forecast	[SV]
External	Detour information			Vehicle management & maintenance

Upstream Data Dependency

As this planning task follows on *[PL1] Plan lines*, the most important input are the routes and line planning defined in that task. Furthermore, information such as infrastructure, timetable data from the next phase and detour information is also necessary to plan detours correctly. Also, crew and vehicle data is necessary to see which resources are available and how much the detour will cost. In case of (way) higher costs, the PTO can request compensation at the transport authority.

Downstream Data Dependency

The main downstream dependency is the detour planning, which is input for *[PL1] Create route network*. Via this planning task, the planning is adapted in all the following phases of the planning process. Since the detour information is known beforehand, employees and customers can be informed accordingly. Furthermore, crew and vehicle forecasts can be made for the corresponding departments. This allows, for example, to distribute the information towards HR for extra crew hiring and maintenance planning for the adjustment with the maintenance department.

Plan Timetable [PT]

[PT1] Determine Timetables

This planning task is responsible for the determination and provision of all timetables of the PTO. Journeys, headways and connections are defined on individual journey-level. Timetables are most often made several times per year and can be adapted based on external factors such as events and detours. A better forecast for crew and vehicles can also be made after the timetables are defined. An overview of upstream and downstream data dependencies is given in Table A-19.

Table A-19 SIPOC [PT1] Determine timetables

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
TA	Transport authority requirements	[PT1] Determine timetables	Timetable	[PL], [PT], [PC], [SV], [SC], [AV]
[PL]	Infrastructure			Employees
	Routes			Operations
	Line plan			Sales
	Relief points			Safety & enforcement
PTO	Mobility behavior			Travel information
TA			Crew forecast	[PC], [AC]
PTO	Actual driving time			HR
PTO	Transport supply		Vehicle forecast	[SV]
				Vehicle management & maintenance

Upstream Data Dependency

To determine timetables, most of the data generated in *Plan lines [PL]* is necessary. Also, requirements from the TA are taken into account. When routes and lines are not new, one can use actual driving times from the operations department to better calculate the driving times and thus the timetable. Furthermore, in the case of events and detours, timetables might be changed. This information comes from *[PL5] Plan detours*.

Downstream Data Dependency

The largest and most important downstream dependency from this planning task is the provision of the timetable to the further planning process, other business areas within the PTO and the external consumers of the data (think of travel information). Furthermore, this timetable data is the basis of reporting towards the TA (real-time data is compared with timetable data). As explained before, after the timetables are set, a better forecast for crew and vehicles can be made, which is demanded by other planning phases but also business areas such as HR and Vehicle management & maintenance.

[PT2] Adjust/improve Timetables

Whenever timetables are determined, it happens sometimes that runtimes on specific lines are not feasible. In this case, timetables should be adjusted slightly. Since this planning task does not need all information as *[PL1] Determine timetable* needs, it is presented separately. This task takes place when the actual driving times from the operations department differ a lot from the runtimes, which might happen regularly. Furthermore, in the future, this planning task might be extended with external real-time information to improve the timetables. An overview of upstream and downstream data dependencies is given in Table A-20.

Upstream Data Dependency

Upstream dependencies for this task are the timetable as determined in *[PT1] Determine timetable*, mobility behavior from the TA and PTO (such as passenger counting in vehicles or ticket-data) and actual driving times from the operations department. Based on this input, the timetable can be further optimized, and, eventually, be dynamically optimized when connected to more (external) real-time data (e.g. weather forecasts).

Table A-20 SIPOC [PT2] Adjust/improve timetables

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PT]	Timetable	[PT2] Adjust/ improve timetables	Timetable	[PL], [PT], [PC], [SV], [SC], [AV]
PTO	Mobility behavior			Employees
TA				Operations
PTO	Runtime analysis			Sales
				Safety & enforcement
				Travel information
			Crew forecast	[PC], [AC]
				HR
			Vehicle forecast	[SV]
				Vehicle management & maintenance

Downstream Data Dependency

This task has the same downstream data dependencies as *[PT1] Determine timetable*, since it is actually part of this higher-level process. The same consumers need to be provided with timetable and forecasting information.

[PT3] Design Frequency Changes

Whereas in *Plan lines [PL]* scenarios are developed which can be used during the operation whenever unexpected events occur, these scenarios need to be extended with timetables. An example is a timetable for a metro replacement bus. These scenarios and timetables are necessary for a PTO to quickly react to unforeseen circumstances during the operation. This planning task most often takes place several times per year, as it is triggered by *[PL4] Develop scenarios*. An overview of upstream and downstream data dependencies is given in Table A-21.

Table A-21 SIPOC [PT3] Design frequency changes

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PT]	Timetable	[PT3] Design frequency changes	Timetable	[PL], [PT], [PC], [SV], [SC], [AV]
[PT]	Scenarios for transport control			Employees
				Operations
PTO	Mobility behavior			Sales
TA				Safety & enforcement
PTO	Transport supply			Travel information
			Crew forecast	[PC], [AC]
				HR
			Vehicle forecast	[SV]
				Vehicle management & maintenance

Upstream Data Dependency

As indicated above, this planning task is triggered by *[PL4] Develop scenarios*, of which the scenarios are an important input for this task. Timetables developed within this task are made for the scenarios. Furthermore, during the development of this scenario-timetables, the normal timetable, mobility behavior and transport supply are taken into account (to offer the most optimal timetable).

Downstream Data Dependency

The timetables developed in this planning task are not part of the normal public transport plan. However, they should be communicated to every planning phase, business area and external consumer as soon as they are used (whenever a scenario is activated during operations). For this reason, this task has the same downstream dependencies as *[PT1] Determine timetable* has.

Plan Crew Rosters [PC]

[PC1] Create Roster Lay-outs

Most often, crew members need to know in advance when they have to work. This roster does not need to contain exact working times, but at least some level of deviation between duties should be known (think of an early, day, late, night duty), and are therefore called roster lay-outs. They often need to be accepted by the workers' council. How far in advance these rosters need to be communicated to the crew members, depending on the national and PTO-specific rules. Since the roster layouts are planned before the actual transport supply is known, the roster cannot be based on the actual work. For this reason, and among other reasons such as sickness, spare and stand-by duties are also taken into account while creating roster lay-outs. An overview of upstream and downstream data dependencies is given in Table A-22.

Table A-22 SIPOC [PC1] Create roster lay-outs

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PT]	Timetable	[PC1] Create roster lay-outs	Crew roster lay-outs	[PC], [SC], [AC]
[PL]	Relief points			Employees
[PL], [PT], [PC]	Crew forecast			Operations
PTO	Mobility behavior			
TA				
PTO	Crew			
External	Crew work rules			
PTO				

Upstream Data Dependency

Since roster layouts cannot be based on the actual workload, the rosters should be created based on an estimate. This estimate is often made using information such as crew forecast and mobility behavior (historic figures). Furthermore, the layouts are based on the (historic) timetable and relief points. Also information from the crew and crew work rules is necessary to know how many crew members should be provided with a roster and which rules should be accounted for (e.g. daily and weekly rest periods). Most often, roster layouts are developed in such a way that these recur every x weeks.

Downstream Data Dependency

Once the roster layouts are made, these are distributed to other planning tasks in the current planning phase, but also to other phases such as *Schedule crew duties [SC]* and *Assign crew [AC]*. Furthermore, roster layouts are communicated with employees and operations.

[PC2] Assign Crew to Roster Lay-outs

Once the roster layouts are ready, the crew should be assigned to these rosters. This process highly depends on the implementation per PTO, the available crew roster layouts and the preferences of the crew members (one prefers late-night duties, whereas others only prefer the early morning duties). After assigning crew members to the roster layouts, a better forecast for the crew can be given. An overview of upstream and downstream data dependencies is given in Table A-23.

Table A-23 SIPOC [PC2] Assign crew to roster layout

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PC]	Crew roster lay-outs	[PC2] Assign crew to roster layout	Crew in crew roster	[PC], [AC]
PTO	Crew			Employees
				Operations
			Crew forecast	[PC], [AC]
				HR

Upstream Data Dependency

This planning task connects the crew members to the different available rosters, taking into account aspects such as contract hours, preferences, co-working with partners, etc. For this, information about the roster layouts and crew members is necessary.

Downstream Data Dependency

As explained above, this planning task connects the crew members to the available crew roster lay-outs. Once this is done, these combinations are further processed into the planning process, but also communicated with employees and operations. Furthermore, the crew forecast can be indicated more carefully and is therefore adapted, which is used by other planning phases and the HR department.

[PC3] Plan Training and Holidays

As the planning tasks within *Plan crew roster [PC]* are about the long-term crew planning, also training and holidays can be very well planned in this phase. Training is most often necessary to keep a driving or any other working license valid and holidays are one of the important secondary employment conditions. Training often consists of several days a year and holidays of several consecutive weeks. For planning these, it is important to spread the days and weeks off among the employees, to prevent a huge crew shortage. An overview of upstream and downstream data dependencies is given in Table A-24.

Table A-24 SIPOC [PC3] Plan training and holidays

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PC]	Crew in crew roster	[PC3] Plan training and holidays	Crew	HR
PTO	Crew		Crew in crew roster	[SC], [AC]
[PL], [PT], [PC]	Crew forecast			Employees
[PC]	Holiday/leave request			Operations
External	Crew work rules		Crew forecast	[PC], [AC]
PTO				HR
External	Training information			
PTO				

Upstream Data Dependency

For planning training and holidays for employees, all information about the crew members, their rosters and valid licenses is necessary. Furthermore, the training possibilities and holiday requests should be known. To perfectly plan and take into consideration the even distribution among crew members, crew forecasts and crew work rules are necessary.

Downstream Data Dependency

As soon as the training and holidays are planned, the crew member's roster should be changed. Furthermore, HR should be informed about training and/or holiday hours. Also, the crew forecast can be adapted based on planned training and holidays.

[PC4] Arrange Self-rostering

As we explained earlier, in the future more and more companies likely offer self-rostering to crew members (see Section 4.4.4). This means that crew members can propose their own desired roster to the PTO, after which the PTO combines all the crew rosters and tries to provide everyone with a roster as close as possible to their desired roster. An overview of upstream and downstream data dependencies is given in Table A-25.

Table A-25 SIPOC [PC4] Arrange self-rostering

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
HR	Crew	[PC4] Arrange self-rostering	Crew in crew roster	[SC], [AC]
[PL], [PT], [PC]	Crew forecast			Employees
[PC]	Desired crew roster			Operations
External	Crew work rules		Crew forecast	[PC], [AC]
PTO				HR

Upstream Data Dependency

The most important upstream data dependency is the desired crew roster. This originates from the crew members and is possibly based on the crew roster layouts created in *[PC1] Create roster lay-outs*. Other inputs for this task are information about the crew members, the forecast and the working rules.

Downstream Data Dependency

As downstream data dependency, this planning task leads to the crew in crew roster information and an improved forecast on crew level, just like the other cases in this planning phase.

Schedule Vehicle Blocks [SV]

[SV1] Determine Layover Times

Layover time is a time allowance at the end of every vehicle journey. This extra time ensures the stability of the network, since a delay in journey x is not automatically taken to journey x + 1. The layover times are most often some minutes and are dependent on the timetable, passenger demand, transport supply and vehicle types. This planning task most often takes place when bigger changes in the planning are implemented (e.g. new lines or routes). In case the previously mentioned objects remain the same, the layover times are not expected to change. An overview of upstream and downstream data dependencies is given in Table A-26.

Table A-26 SIPOC [SV1] Determine layover times

Upstream data dependency			*Downstream data dependency*	
Supplier	Input	Process	Output	Consumer
[PT]	Timetable	[SV1] Determine layover times	Layover times	[SV]
PTO	Mobility behavior			
TA				
PTO	Transport supply			

Upstream Data Dependency

The most important input for this planning task is the timetable. Decisions in this task might also trigger a change in the timetable to reach a higher level of overall optimization. Another important input is the passenger demand (mobility behavior) and the supply of public transport. Based on these inputs, layover times are determined.

Downstream Data Dependency

The layover times need to be taken into account while scheduling vehicle blocks. For this reason, the output is the layover times, which might differ per line, day, time of the day, vehicle type, etc.

[SV2] Plan Dead Runs

Dead runs are non-service journeys of a vehicle. These journeys are necessary to reach a starting point of a line, or reach the depot from an ending point. Dead runs are also possible in between the end and start points of two lines, from station to charging point, gas station, car wash, and any other journey which is necessary to realize the public transport service, as long as passengers are not allowed on the journey. For rail services, this is often called shunting. This planning task most often takes place whenever bigger changes such as the introduction of new lines, different start/endpoints of lines, etc. Only then, the dead runs need to be planned, in the other cases the dead runs are already present. In some cases, dead runs are always planned newly, however, as long as the locations do not change, the dead runs stay the same. An overview of upstream and downstream data dependencies is given in Table A-27.

Table A-27 SIPOC [SV2] Plan dead runs

Upstream data dependency			*Downstream data dependency*	
Supplier	Input	Process	Output	Consumer
[PL]	Infrastructure	[SV2] Plan dead runs	Dead runs	[SV]
	Routes			
	Line plan			
[PT]	Timetable			
PTO	Facilities			
TA	Mobility behavior			
PTO				
PTO	Transport supply			

Upstream Data Dependency

To plan the dead runs, the lines, routes and infrastructure need to be known. The timetable is also assessed, since dead runs might differ per day and time of the day (e.g. during rush hour). Furthermore, information about facilities (depots, canteens, etc.), mobility behavior (passenger demand, history, etc.) and transport supply are needed. Part of the last object is also the costs of dead runs. Since these journeys are not creating any revenue, the dead runs are most often minimized as much as possible.

In some scenarios, dead runs might lead to a change of the timetable (e.g. when an empty vehicle needs to drive along a route of a particular line).

Downstream Data Dependency

The downstream data objects are the dead runs. These dead runs are necessary for the next task *[SV3] Create vehicle blocks*.

[SV3] Create Vehicle Blocks

This is the main planning task in the planning phase *Schedule vehicle blocks [SV]*, since the actual vehicle blocks are created. The balance of vehicles at depots, journeys, dead runs, layover times, maintenance planning and more are taken into account while creating these vehicle blocks. Also, the vehicle forecast is continuously adapted based on the creation of these blocks. While creating vehicle blocks for battery-equipped vehicles, also charging needs to be taken into account. The vehicle blocks should always match the timetable and always strive for a minimum of vehicles to be operated, since this saves costs. Other cost-saving methods such as interlining (a vehicle that covers more lines) are also taken into account in this stage of the planning. This planning task takes place every time a new timetable is created. An overview of upstream and downstream data dependencies is given in Table A-28.

Table A-28 SIPOC [SV3] Create vehicle blocks

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PL]	Infrastructure	[SV3] Create vehicle blocks	Vehicle blocks	[SC], [AV]
	Routes			Employees
	Line plan			Operations
	Relief points			Travel information
[PT]	Timetable		Vehicle forecast	[SV]
[SV]	Layover times			Vehicle management & maintenance
	Dead runs			
[PL], [PT], [SV]	Vehicle forecast			
PTO	Mobility behavior			
TA				
PTO	Transport supply			
PTO	Maintenance planning			
PTO	Facilities			

Upstream Data Dependency

For creating vehicle blocks, basic network information such as the infrastructure, routes, lines and the timetable are necessary. Moreover, relief points, layover times and dead runs are taken into account as well. For vehicle block scheduling, also maintenance planning is considered. Most often, between the morning and afternoon rush hour, vehicles can undergo a maintenance visit. This is only possible when the vehicle blocks allow this. For battery-equipped vehicles, also the charging locations are considered (facilities). The mobility behavior and transport supply contain information about the passenger demand and vehicle type (restrictions) per line. Not every line can be and should be deployed by the same vehicle. Hence, creating vehicle blocks is done based on vehicle type as well.

Downstream Data Dependency

The main downstream data dependency from this planning phase is the vehicle blocks data object. This data is used further in the planning process, but also during operations. Based on these vehicle blocks, the vehicle forecast can also be adapted.

Schedule Crew Duties [SC]

[SC1] Create Crew Duties

Crew duties are created to let the crew members work and operate public transport most optimally (the less personnel you need, the cheaper the operation will be). These duties contain journeys and/or vehicle blocks, dead runs, personal care time, break time and any other specified work such as refueling, washing, etc. (most often based on local work (council) rules). Next to duties which contain working activities, also spare duties (being on-site and stand-in when necessary) and stand-by duties (being home and available to work) are created. At many Dutch PTOs, the workers' council needs to accept the created crew duties. In this best scenario, all their requirements/constraints are loaded as a ruleset into an application, which generates efficient crew duties based on this input. In this way, the created crew duties always comply with the workers' council rules and requirements. An overview of upstream and downstream data dependencies is given in Table A-29.

Table A-29 SIPOC [SC1] Create crew duties

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PL]	Relief points	[SC1] Create crew duties	Crew duties	[SC], [AC]
[PT]	Timetable			Employees
[PC]	Crew roster layout			Operations
[SV]	Vehicle blocks		Crew forecast	[PC], [AC]
PTO	Crew			HR
External	Crew work rules			
PTO				

Upstream Data Dependency

For creating crew duties, operational information about relief points and the timetable is necessary. Crew changes and breaks can only take place at these specified relief points. Furthermore, crew duties are often based on vehicle blocks, which are therefore also important input. To match the crew roster layout created in *Plan crew roster [PC]*, these rosters are also used as input. Moreover, information about the crew (such as size, contract hours, etc.) and crew work rules are also taken into account, to create enough valid duties, of necessary duty lengths. As mentioned above, the ruleset for accepting duties by the workers' council can be seen as part of crew work rules.

Downstream Data Dependency

The main downstream data dependency from this planning phase is the crew duties data object. This data is used further in the planning process, but also during operations. Based on these crew duties, the crew forecast can also be adapted.

[SC2] Assign Crew Duties to Roster

Whenever the duties are created, they need to be assigned to a roster lay-out. That is also the point in time from which the crew exactly knows what duty they have to work (except for last-minute detours and other unknown disruptions). This assigning of duties to roster layouts depends on many (most often company-specific) rules and strives for the highest efficiency. An overview of upstream and downstream data dependencies is given in Table A-30.

Table A-30 SIPOC [SC2] Assign crew duties to roster

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PC]	Crew roster layout	[SC2] Assign crew duties to roster	Crew roster with duties	[AC]
[SC]	Crew duties			Employees
External	Crew work rules			Operations
PTO				

Upstream Data Dependency

For assigning the duties to the roster layouts, the crew duties, crew roster layouts and the crew work rules are necessary input data objects.

Downstream Data Dependency

The only downstream data dependency is the mapping of duties to roster layouts: crew roster with duties.

Assign Vehicles [AV]

[AV1] Assign Vehicle to Block

The responsibility of this planning task is the actual realization of the relationship between a specific vehicle and vehicle block. By doing this, the vehicle is also automatically assigned for routes and lines (vehicle type restrictions) and maintenance possibilities (including predictive maintenance). To facilitate this process, the decision is based on planning, network and maintenance information. By taking into account the maintenance planning, this planning task should also account for a uniform mileage of the vehicles. In the future, vehicles might be able to request vehicle blocks themselves, as they know 'their' condition and maintenance necessities. This planning task most often takes place daily. An overview of upstream and downstream data dependencies is given in Table A-31.

Table A-31 SIPOC [AV1] Assign vehicle to block

Upstream data dependency	
Supplier	**Input**
[PT]	Timetable
[SV]	Vehicle blocks
[AV]	Vehicle assign plan
PTO	Vehicle
PTO	Maintenance planning

Process
[AV1] Assign vehicle to block

Downstream data dependency	
Output	**Consumer**
Vehicle on vehicle block	[AV]
	Employees
	Operations
	Vehicle management & maintenance

Upstream Data Dependency

As the planning task is to match vehicles with vehicle blocks, these are the most important input for this task. The vehicle assign plan as determined in *[AV3] Plan vehicle parking* is also input, as the integration of these two tasks leads to a better solution. Regarding maintenance, the maintenance planning is necessary (in which the maintenance department requests vehicles for maintenance, based on milage, runtime and/or predictive maintenance).

Downstream Data Dependency

The information of which vehicle will be operating which vehicle block is shared among the planning department, employees (such that they know which vehicle to take from the depot), operations and vehicle management.

[AV2] Manage Vehicle Disruptions

Last-minute changes happen daily. This planning task handles these unexpected disruptions regarding vehicles. Most often, this task is triggered by the control center (operations), since they receive information that a vehicle is missing, defect or has any other problem. In this case, this task has the responsibility to ensure that the originally planned vehicle (in *[AV1] Assign vehicle to block*) is replaced by another suitable vehicle. The trigger might also take place a bit earlier, e.g. right before leaving the depot. In case no solution can be found and the journey cannot be driven by a vehicle, this planning task has the right to cancel particular journeys from the timetable. An overview of upstream and downstream data dependencies is given in Table A-32.

Note: at some PTOs, this planning task is housed at and carried out under the responsibility of the control center.

Table A-32 SIPOC [AV2] Manage vehicle disruptions

Upstream data dependency	
Supplier	**Input**
[PT]	Timetable
[SV]	Vehicle blocks
[AV]	Vehicle on vehicle blocks
PTO	Real-time vehicle information
PTO	Vehicle/vessel
PTO	Maintenance planning

Process
[AV2] Manage vehicle disruptions

Downstream data dependency	
Output	**Consumer**
Vehicle on vehicle block	[AV]
	Operations
	Employees
	Vehicle management & maintenance
Vehicle incident	Vehicle management & maintenance
Timetable	Employees
	Operations
	Safety & enforcement
	Travel information

Upstream Data Dependency

To manage the vehicle disruptions, quite some data is necessary. From the operations, the real-time vehicle information (such as location) is necessary, basic vehicle information and its maintenance planning (should still be taken into account as much as possible). The planned timetable, vehicle blocks and connected vehicles as assigned by *[AV1] Assign vehicle to block* are necessary to analyze the current situation and find the solution.

Downstream Data Dependency

In case a disruption happens, most often the vehicle needs to be replaced. In that case, the data object 'vehicle on vehicle block' is changed from that particular moment on. This is communicated to all necessary consumers. Furthermore, the so-called vehicle incident which is owned by the operations, is edited and sent to the vehicle management department to log what happened and which solution was provided. In case no solution can be found, journeys need to be canceled. This can be done by editing the timetable. The key is to communicate these cancellations to the business, passengers and other external consumers.

[AV3] Plan Vehicle Parking

This planning task contains the depot management functionality regarding the parking of vehicles. As explained in Section 4.2, depot management has to do with every task regarding vehicles that need to be carried out at the depot. For this planning task, parking planning is the most important. Vehicles need to be parked in the way they will have to leave the depot on the next operational day (as defined by *[AV1] Assign vehicle to block*. This should be communicated to the drivers of the vehicles, to let them park correctly. This planning task most often takes place together with *[AV1]*, thus daily. An overview of upstream and downstream data dependencies is given in Table A-33.

Table A-33 SIPOC [AV3] Plan vehicle parking

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[SV]	Vehicle blocks	[AV3] Plan vehicle parking	Vehicle assign plan	[AV]
PTO	Vehicle			Employees
PTO	Maintenance planning			Vehicle management & maintenance
PTO	Parking			
PTO	Real-time vehicle information			

Upstream Data Dependency

For planning the vehicle parking, the vehicle blocks are necessary for the current day (which vehicles will return at what time) and the next day (for planning the parking). As well as planning task *[AV1] Assign vehicle to block*, this task also accounts for the maintenance planning. To see whenever vehicles will return with a delay that changes the planning, the current position of the vehicles is necessary. The data object called parking should contain facility data such as parking possibilities, depots, refueling/charging locations and washing installations.

Downstream Data Dependency

The only downstream data dependency is the vehicle assign plan. This data is shared among the *Assign vehicle [AV]* phase, the employees (to let them know where to park) and the vehicle management department.

Assign Crew [AC]

[AC1] Assign Crew to Duties

Within the *[AC1] Assign crew to duties* planning task, the crew members are assigned to the crew duties which are scheduled in *Schedule crew duties [SC]*. Most of the crew members are automatically assigned to a crew duty, since the roster layout specifies which crew duty belongs to which crew member (after *[SC2] Assign duties to roster* is carried out). However, as explained before, the crew duties and rosters also contain spare duties. These spare duties can be changed into a 'normal' duty in case of e.g. sickness and holidays. When assigning crew to duties, qualifications need to be (automatically) checked, since some work requires specific qualifications or certification. This planning task takes place daily. An overview of upstream and downstream data dependencies is given in Table A-34.

Upstream Data Dependency

As the goal is to assign every crew member to a crew duty, this task should have all roster layouts, crew members, crew duties, contract hours and crew work rules as input. Furthermore, specific preferences of crew members are available through the crew-object.

Table A-34 SIPOC [AC1] Assign crew to duties

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PC]	Crew roster lay-outs	[AC1] Assign crew to duties	Crew assigned to duty	[AC]
	Crew in crew roster			Employees
[SC]	Crew duties			HR
	Crew roster with duties			Operations
PTO	Crew			
External	Crew work rules			
PTO				

Downstream Data Dependency

For the planning process, operations (control center), HR and the employees themselves, it is important to provide the data of crewmembers assigned to duties.

Note: often employment agencies work for PTOs to provide a flex pool of crew members. In this case, data is provided to this employment agency about crew duties that are not covered by the PTO and need to be assigned by the employment agency. This downstream data dependency is not shown in the table and figures.

[AC2] Manage Crew Disruptions

Last-minute disruptions occur daily which means that this planning task also takes place daily. Most often, the task is triggered by operational information from the PTP's control center. In case a driver calls sick, is absent or not available for work for any reason, this disruption needs to be handled. It can be solved by calling the particular crew member (maybe they are running late) or ask the spare-duty crew member to stand-in. Whenever no spares are available anymore, the crew members with stand-by duties are asked to work. For several reasons, it might be the case that some journeys cannot be driven. In that case, this planning task is also allowed to change the timetable (e.g. by canceling journeys). An overview of upstream and downstream data dependencies is given in Table A-35.

Note: at some PTOs, this planning task is housed at and carried out under the responsibility of the control center.

Table A-35 SIPOC [AC2] Manage crew disruptions

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[PT]	Timetable	[AC2] Manage crew disruptions	Crew assigned to duty	[AC]
[PC]	Crew in crew roster			Employees
	Crew roster lay-outs			HR
[SC]	Crew duties			Operations
	Crew roster with duties		Crew incident	HR
[AC]	Crew assigned to duty		Timetable	Employees
PTO	Crew			Operations
External	Crew work rules			Safety & enforcement
PTO				Travel information
PTO	Real-time crew information			

Upstream Data Dependency

For handling disruption in real-time, a lot of information is necessary. All information regarding crew members, their duties and the timetable is necessary to assess what can go wrong or is going wrong. In such a situation, it is also important to know which crew members are on spare and stand-by duties, because they could immediately help to solve the disruption. Real-time information most often comes from the control center, but can also be retrieved directly from the crew member to the *Assign crew [AC]* workers. When managing the disruption, the crew work rules always have to be taken into account, and are therefore also an input for this planning task.

Downstream Data Dependency

In case a disruption takes place, most often crew members need to be changed, or journeys need to be canceled. For this reason, both assigned crew members and the timetable are important downstream data dependencies. These need to be provided to employees, operations, safety & enforcement and also travel information. Next to these operational data, also a crew incident data object needs to be generated in which the problem and solution of a specific crew member or situation are logged. This information is also important for reporting purposes about possibly cancelled journeys.

[AC3] Manage Leave, Duty Swap, Requests

Short-term leave requests and duty (swap) requests are also handled within this planning task. Long-term leave requests (e.g. yearly holiday planning) are the responsibility of *Plan crew rosters [PC]*. The closer to the day of operation, the better a crew forecast can be made. Furthermore, requests could be processed automatically by an application, when enough data about demand and supply is available. Personal requests such as a crew member who prefers late-night duties over early morning duties are also processed within this planning task. This planning task takes place daily, whenever a crewmember files a request. An overview of upstream and downstream data dependencies is given in Table A-36.

Table A-36 SIPOC [AC3] Manage leave, duty swap, requests

Upstream data dependency			*Downstream data dependency*	
Supplier	**Input**	**Process**	**Output**	**Consumer**
[AC]	Crew assigned to duty	[AC3] Manage leave, duty swap, requests	Crew	HR
[PC]	Crew roster lay-outs		Crew assigned to duty	[AC]
PTO	Crew			Employees
[PL], [PT], [PC]	Crew forecast			HR
[AC]	Holiday/leave request			Operations
[AC]	Duty swap or request			

Upstream Data Dependency

For managing crew requests, it is important to know who is requesting (contract hours, leave balance, etc.) and for which duty they are actually scheduled for. This latter can be based on the duty planning, but also on the roster layouts (whenever duties are not yet defined). The crew forecast is important to take into account when assessing requests, since this forecast determines whether or not a crew member's request can be accepted.

Downstream Data Dependency

This particular planning task is changing personal crew rosters and crew information. Whenever requests are accepted, the duties to which the crew member is assigned, might need to be changed. Furthermore, the crew's leave balance needs to be communicated to HR.

Appendix M Data Object Definitions

Table A-37 Data object definitions

Data object	Definition
Actual driving time	Actual public transport journeys, times, connections, passengers, etc. Used for checking runtimes.
Crew	Crew data including details such as name, ID, contract hours, leave, email address, phone number(s), qualification(s), licenses.
Crew assigned to duty	The relation between a *crew* member and *crew duty.*
Crew duties	Duties for *crew* members, exact information about what is expected to be done by the *crew* member.
Crew forecast	Forecast about how much personnel is expected to be necessary on a particular date.
Crew in crew roster	The relation between a *crew* member and *crew roster lay-out.*
Crew incident	Incident data generated within the real-time operation (transport control) which triggers the *Assign crew* planning phase in case of a disruption and provides with necessary information.
Crew roster lay-outs	Rosters for *crew* members which include duty types (e.g. early/day/late/night/spare/stand-by) and days-off for the upcoming period (PTO-specific, often around 1 year).
Crew roster with duties	The relation between a *crew roster layout* and *crew duties.*
Crew work rules	Regulations (EU and national), collective labor agreement and rules (company-specific) related to *crew* working days, hours and activities.
Dead runs	Any *vehicle* journey without passengers that is necessary for the realization of the *timetable.*
Desired crew roster	Input from a *crew* member about their desired *crew roster lay-out.*
Detour information	Information about a detour in the future. Input for the planning process and provided by e.g. a province, municipality, road authority, transport authority, event organizer.
Detour planning	Adapted *timetable* based on the *detour information.* This might also have an impact on any other planning-related data object (e.g. *routes, line plan, crew duty, vehicle block*).
Duty swap or request	Input from a *crew* member about a duty swap with a colleague or request for a *crew duty.*
Facilities	PTO's facilities such as depots, canteens, refueling points, charging points, etc.
Fare information	Fare, ticket, pricing and subscription data based on the transport network intended to be provided together with the public transport planning data (i.e. for passenger information).
Holiday/leave request	Input from a *crew* member about their holiday/leave request.
Infrastructure	Infrastructural (topology) data about stops and stations.
Layover times	The extra time between journeys to ensure the stability of the *timetable* (such that a delay in journey A does not affect journey B).
Line plan	The plan of a particular line, including frequencies, expected runtimes, vehicle types, constraints, information about first/last journeys, etc.
Maintenance planning	Maintenance planning of the *vehicles.*
Mobility behavior	Data about the mobility behavior of passengers, including origin/destination matrices and historical data (based on ticket sales, OV Chipkaart and counters in vehicles).
Parking	Parking information for depots and relief points, including information about charging possibilities.
Pricing information	Information as input for determination of *fare information.* Can originate from the PTO, but also from a transport authority.
Real-time crew information	Real-time information about crew members, think about presence, but also last-minute sickness. This might result in a *crew incident* which might have consequences for the planning.
Real-time vehicle information	Real-time vehicle information, think about location, defects and state of charge. This might result in a *vehicle incident* which might have consequences for the planning.
Relief points	A location where a relief is possible, i.e. a driver may take on or hand over a vehicle (Transmodel, 2019)
Routes	All routes for particular public transport lines in a *line plan.*
Scenarios for transport control	The scenario developed to be implemented easily during operations. Designed based on situations that are expected to happen (e.g. tram or metro replacement buses).
Timetable	The exact service journeys per public transport line in *line plan.*
Training information	Information about training for crew members. Think about subject, hours, category, license, etc.
Transport authority requirements	Requirements from the transport authority about the public transport network. One can think of compulsory lines, connections and vehicle types, but also about prices and minimum distance for passengers to a stop or station.
Transport supply	Supply in terms of vehicles and crew, based on both PTO's as well as transport authority's data. Consists – among other data – of *crew* and *vehicle* data.
Utilities and land-usage	Data about utilities and land-usage for the determination of the public transport route network (*routes* and *line plan*).
Vehicle assign plan	The assignment of vehicle blocks to specific vehicles, most often for the next operational day, which includes the order of parking. Based on the *transport supply* for the next operational day and the *maintenance planning.*
Vehicle blocks	The actual work that has to be carried out by a vehicle. This includes public transport journeys, dead runs and any time-allocation for charging, refueling, etc.
Vehicle forecast	Forecast about how many vehicles are expected to be necessary on a particular date.
Vehicle incident	Incident data generated within the real-time operation (transport control) which triggers the *Assign vehicle* planning phase in case of a disruption and provides with necessary information.
Vehicle on vehicle block	The relation between a specific *vehicle* and *vehicle block.*
Vehicle/vessel	Vehicle data including details such as type, number, age, necessities, maintenance data.

Appendix N Target Data Architecture including the Data Requirements

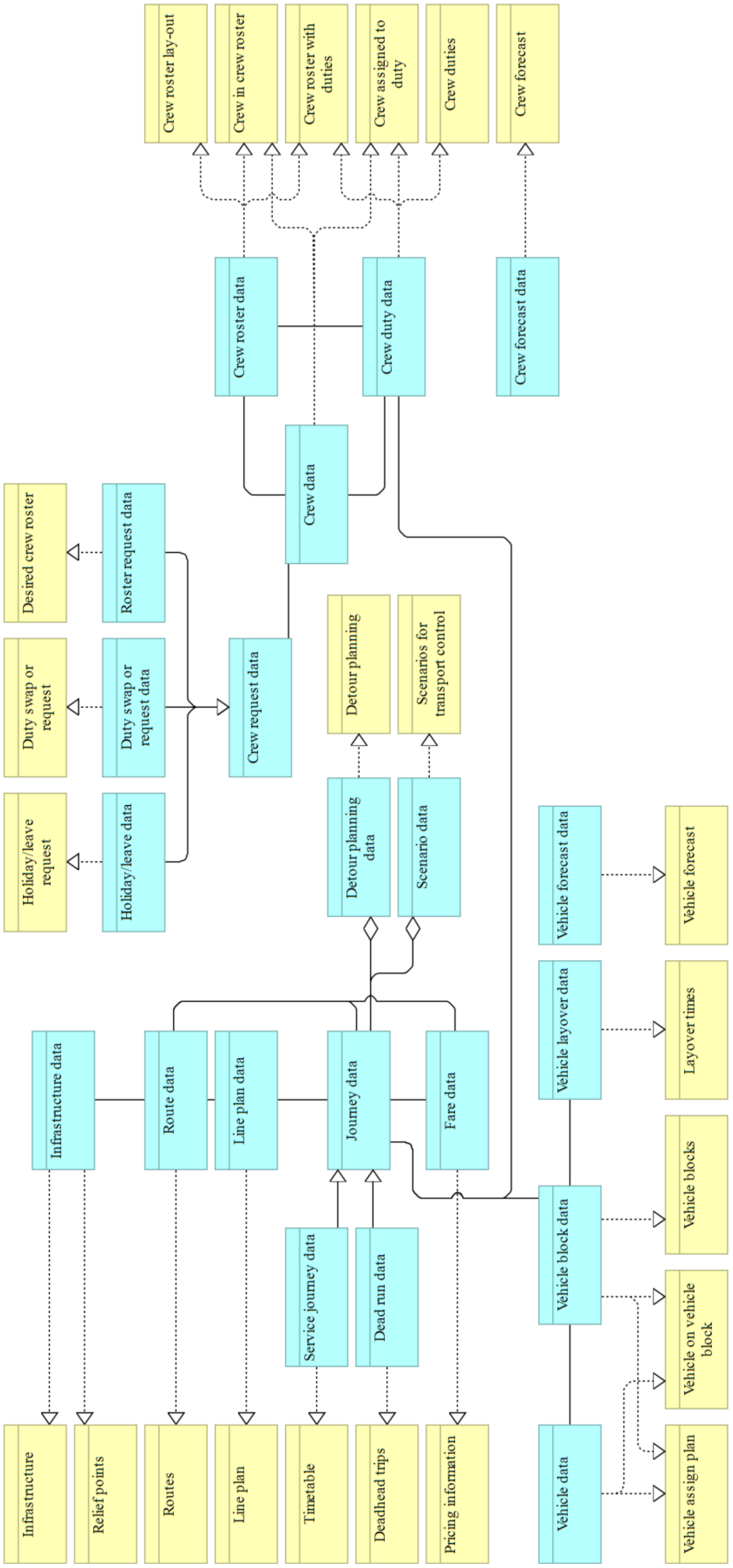

Figure A-6 Target data architecture and its data requirements

Appendix O Data Roles and Interactions (IDSA, 2019b)

This appendix contains data roles and interaction information set by the IDSA's (2019b) reference architecture.

Table A-38 Data roles by IDSA (2019b)

Role category	Role	Description
Core participant	Data owner	Has the legal right for creating data and/or executing control over it. Defines usage contracts and policies, provides access to data and defines payment models.
	Data provider	Makes data available for being exchanged between a Data owner and Data consumer.
	Data consumer	Receives data from a Data provider and use Data Apps (see App provider) to enrich and transform the data for the Data user.
	Data user	Has the legal right to use the data of a Data owner as specified by the usage policy.
	App provider	Develops so-called Data Apps to be used in the International Data Spaces which can be accessed and used by Data consumers and Data providers.
Intermediary	Broker service provider	Intermediary that stores and manages information about the data sources available in the International Data Spaces. Focuses on receiving (from Data providers) and providing (to Data consumers) metadata.
	Clearing house	Logs all activities performed in the course of a data exchange. Both the Data provider and Data consumer confirm the data transfer by logging at the clearing house. Facilitates in resolving conflicts, reporting and billing.
	Identity provider	Offers a service to create, maintain, manage, monitor and validate identity information of and for participants in the International Data Spaces.
	App store provider	Responsible for managing information about Data Apps offered by App providers. Provides interfaces for publishing and retrieving Data Apps plus corresponding metadata.
	Vocabulary provider	Manages and offers vocabularies (i.e. ontologies, reference data models or metadata elements) that can be used to annotate and describe datasets. Provides the information model of the International Data Spaces and other domain-specific vocabularies.
Software/Service provider	Service provider	Hosts the required infrastructure to make data available. Receives data from the Data provider and offers the data in the International Data Spaces.
	Software provider	Provides software for implementing the functionality required by the International Data Spaces. Delivers software over its distribution channels to the Data consumer, Data provider, Broker service provider, etc.
Governance Body	Certification body	Makes sure that only compliant organizations are granted access to the trusted business eco-system. The certification body supervises the actions and decisions of the Evaluation facility.
	Evaluation facility	
	IDSA	Develops the Reference Architecture Model continuously.

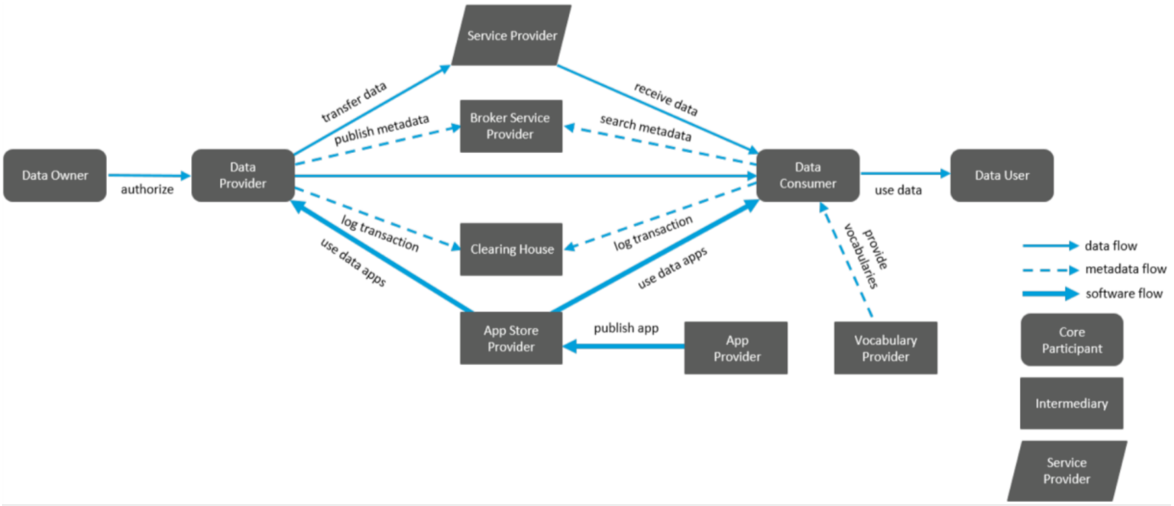

Figure A-7 Data roles and interactions by IDSA (2019b)

Appendix P Validation Survey Results

This appendix contains the answers to the closed question in the validation survey. The numbers and letters given as answers match the answer possibilities of the validation survey in Appendix G. In case answers are not answered, the respondent chose not to answer that particular part of the validation survey. For anonymity reasons, the answers to the open questions are not provided here.

Table A-39 Validation survey results

	E01	E02	E03	E04	E05	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15
C1	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
C2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
C3	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
C4	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
C5	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
C6	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
B1	N/A	5	5	1	4	5	5	5	5	5	3	5	5	N/A	3
B2	N/A	5	5	1	4	5	5	5	5	5	3	5	5	N/A	5
B3	N/A	5	5	1	5	5	5	5	4	5	5	5	5	N/A	5
B4	N/A	5	5	1	1	1	5	5	4	5	5	5	5	N/A	5
B5	N/A	5	5	1	4	3	5	5	5	5	5	5	5	N/A	5
B6	N/A	5	5	1	1	1	5	5	4	5	5	5	5	N/A	5
B7	N/A	5	5	1	4	2	5	5	4	5	5	5	5	N/A	5
B8	N/A	5	5	1	1	1	5	5	4	5	5	5	5	N/A	5
B9	N/A	5	5	1	5	5	5	5	4	5	5	5	5	N/A	5
B10	N/A	Y	Y	N	Y	N	Y	N	Y	N	Y	Y	Y	N/A	U
I1	3	3	4	2	3	3	5	2	2	2	5	5	5	5	5
I3	U	U	U	N	N	U	Y	N	N	N	Y	N	N	N	U
I6	5	5	5	5	4	5	5	4	5	1	4	4	5	5	5
I7	4	5	4	5	4	4	5	4	4	6	4	6	5	4	6
I8	4	4	4	2	4	2	5	2	6	6	3	6	4	5	6
I10	2	2	5	5	5	3	4	4	2	5	2	4	4	5	1
I13	2	3	3	3	3	3	3	3	3	3	2	3	3	4	3
I14	3	3	3	2	2	3	3	3	3	2	3	2	3	2	3
I15	2	3	3	3	3	3	2	2	2	2	2	3	3	4	3
I16	3	3	3	2	2	3	2	3	3	2	3	2	3	2	3
I17	3	3	3	1	2	3	2	2	3	2	3	4	3	4	3
I18	2	2	3	2	2	3	2	1	3	3	2	1	3	2	2
I19	3	3	3	3	3	3	2	3	3	2	3	3	3	3	3
I20	3	2	3	3	2	3	2	3	3	1	3	3	2	3	3
I21	3	2	3	2	3	3	2	1	3	1	1	2	3	2	3
I22	3	2	3	3	3	3	2	1	3	2	4	2	2	2	3
I23	2	2	3	2	1	3	1	1	2	1	4	3	2	1	3
I24	3	2	3	2	2	3	2	2	3	1	3	2	3	2	3
I25	2	2	3	2	2	3	2	2	3	2	3	3	3	3	3
I26	3	2	3	2	2	3	2	3	3	2	2	3	2	4	4
I27	2	3	3	2	2	3	2	3	3	2	3	3	3	3	3
I29	2	3	3	3	3	3	3	3	3	3	3	3	3	4	3
I30	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3
I31	3	3	3	3	3	3	3	3	3	3	3	3	3	4	3
I32	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3
I33	2	3	3	3	2	3	2	3	3	2	3	4	3	4	3
I34	2	2	3	1	1	3	2	1	3	3	3	1	3	4	2
I35	3	3	3	3	3	3	3	3	3	2	3	2	3	3	3
I36	2	3	3	3	3	3	3	3	3	2	3	3	2	3	3
I37	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
I38	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3
I39	2	3	3	2	3	3	1	3	3	3	3	3	2	3	3
I40	3	2	3	3	3	3	3	3	3	2	3	2	3	3	3
I41	2	2	3	3	3	3	3	3	3	2	3	3	3	1	3
I42	2	2	3	3	3	3	3	3	3	2	2	3	3	2	3
I43	3	2	3	1	3	3	3	3	3	3	3	3	3	3	3
I46	4	5	7	5	5	7	3	7	5	7	4	7	4	5	7
I48	4	4	7	1	3	7	4	7	2	7	4	7	2	5	7
I50	1	2	3	1	6	7	2	7	1	7	1	7	2	7	7
A1	4	4	N/A	2	4	5	5	3	5	N/A	4	N/A	4	5	N/A
A3	5	4	N/A	1	5	5	2	3	5	N/A	4	N/A	4	5	N/A
A5	4	6	N/A	4	4	4	3	4	5	N/A	6	N/A	4	4	N/A

A7	1;2;3;4	1	N/A	1;2;5;7	1;2;3;4;5;6	1;3;4;8	1;2;3;4	1;2;3;4;5;6;7	1;2;3;4;5;6;7	N/A	1;2;3	N/A	1;5;7	2;4;5;6	N/A
V1	2	6	N/A	4	5	4	5	3	5	N/A	4	N/A	5	5	N/A
V3	5	3	N/A	1	3	5	5	5	5	N/A	4	N/A	5	5	N/A
V5	3	4	N/A	2	3	5	5	5	4	N/A	4	N/A	4	5	N/A
V7	4	4	N/A	4	3	4	5	4	5	N/A	2	N/A	3	4	N/A
V9	2	4	N/A	4	3	4	5	5	5	N/A	3	N/A	5	4	N/A
V11	1	4	N/A	2	3	4	5	3	5	N/A	5	N/A	4	4	N/A
V13	4	6	N/A	3	6	6	5	4	5	N/A	4	N/A	5	6	N/A
V15	4	4	N/A	4	1	6	5	3	5	N/A	6	N/A	3	3	N/A
V17	6	4	N/A	3	6	4	5	3	4	N/A	4	N/A	3	4	N/A
V19	6	6	N/A	3	5	5	5	4	5	N/A	5	N/A	5	5	N/A
V21	5	3	N/A	5	6	6	6	4	6	N/A	4	N/A	5	5	N/A
V23	5	3	N/A	2	5	5	5	5	6	N/A	4	N/A	4	5	N/A
V25	4	4	N/A	4	5	5	5	5	5	N/A	4	N/A	5	4	N/A
U1	N/A	3	N/A	5	2	4	5	N/A	5	N/A	4	N/A	3	N/A	5
U2	N/A	3	N/A	4	3	4	6	N/A	4	N/A	4	N/A	3	N/A	5
U3	N/A	3	N/A	4	3	3	6	N/A	4	N/A	3	N/A	3	N/A	5
U5	N/A	2	N/A	5	3	4	4	N/A	5	N/A	3	N/A	3	N/A	4
U6	N/A	4	N/A	4	3	4	4	N/A	4	N/A	3	N/A	3	N/A	4
U7	N/A	4	N/A	3	3	3	5	N/A	4	N/A	3	N/A	3	N/A	4
U9	N/A	6	N/A	3	3	6	1	N/A	6	N/A	4	N/A	2	N/A	3
U10	N/A	6	N/A	N/A	3	6	1	N/A	6	N/A	4	N/A	4	N/A	3
U11	N/A	2	N/A	2	2	6	3	N/A	4	N/A	2	N/A	3	N/A	4
U12	N/A	2	N/A	3	2	6	2	N/A	4	N/A	3	N/A	3	N/A	4
U14	N/A	2	N/A	5	5	4	4	N/A	6	N/A	3	N/A	2	N/A	6
U15	N/A	4	N/A	5	5	4	5	N/A	6	N/A	3	N/A	3	N/A	6
U16	N/A	3	N/A	4	3	4	5	N/A	6	N/A	4	N/A	3	N/A	6
U18	N/A	1	N/A	3	4	6	5	N/A	6	N/A	3	N/A	3	N/A	6
U19	N/A	1	N/A	5	4	6	5	N/A	6	N/A	4	N/A	3	N/A	4
U20	N/A	1	N/A	5	4	6	5	N/A	6	N/A	3	N/A	3	N/A	6
U22	N/A	5	N/A	2	4	3	5	N/A	5	N/A	4	N/A	4	N/A	6
U23	N/A	5	N/A	2	4	2	5	N/A	5	N/A	4	N/A	4	N/A	3
P1	**M**	**M**	**F**	**M**	**M**	**M**	**M**	**M**	**M**	**M**	**M**	**F**	**M**	**M**	**P**
P2	5	4	3	2	4	4	3	3	3	3	2	3	4	4	1
P3	4	4	3	5	5	4	5	5	4	2	4	4	3	4	3
P4	4	4	3	5	5	4	4	4	5	2	4	3	3	4	2
P5	4	4	4	5	5	4	5	4	4	3	3	3	3	3	2
P6	3	4	5	5	5	5	3	3	4	5	4	4	4	4	4
P9	1	4	1	7	7	7	1	7	3	4	1	1	1	7	3
P10	1;2;3;4;6;7	1;5;7	1;2;3;6;7	8	8	5	1;2;3;6;7	5	1;2;5	5	1;2;3;6	1;2;3;6	1;2;3;6	8	1;7
P12	5	5	5	4	5	5	5	5	5	5	4	4	5	5	1

Appendix Q GVB-only Validation Results for Goals, Usage and Acceptance

This appendix shows the validation results filtered on experts working for GVB. It is used for substantiating the recommendations for GVB provided in Section 9.4.2. In total, five GVB experts took part in the validation. However, not every expert answered every validation question.

Figure A-8 shows the goals related to the data-level and is answered by 3 GVB experts.

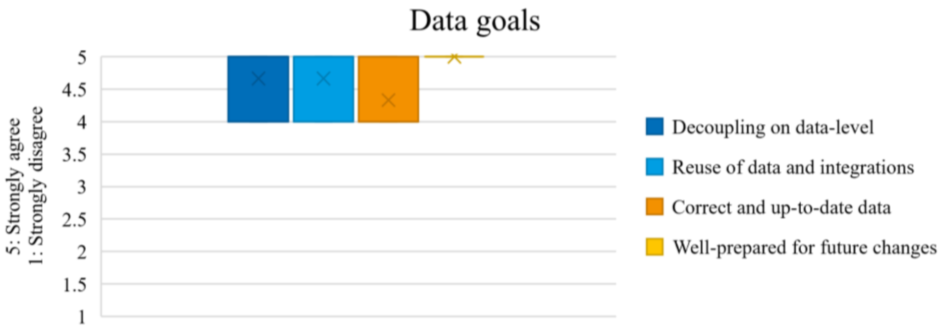

Figure A-8 Goal validation regarding the data situation (GVB-only)

Figure A-9 shows the goals related to the approach and is answered by 3 GVB experts.

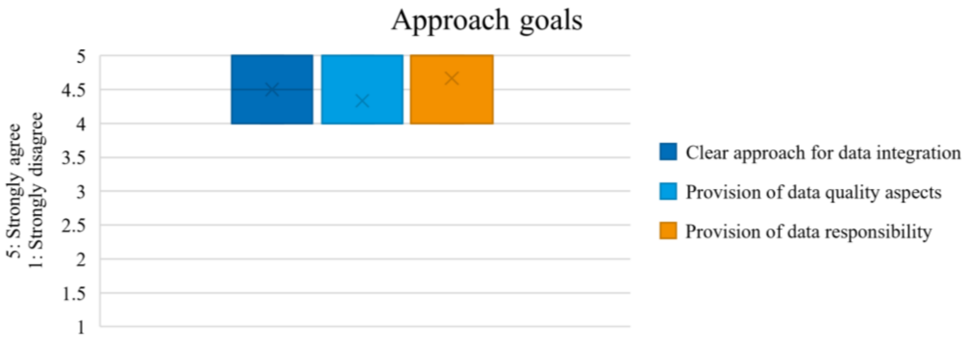

Figure A-9 Goal validation regarding the approach (GVB-only)

Figure A-10 shows the goals related to the business process improvements and is answered by 3 GVB experts.

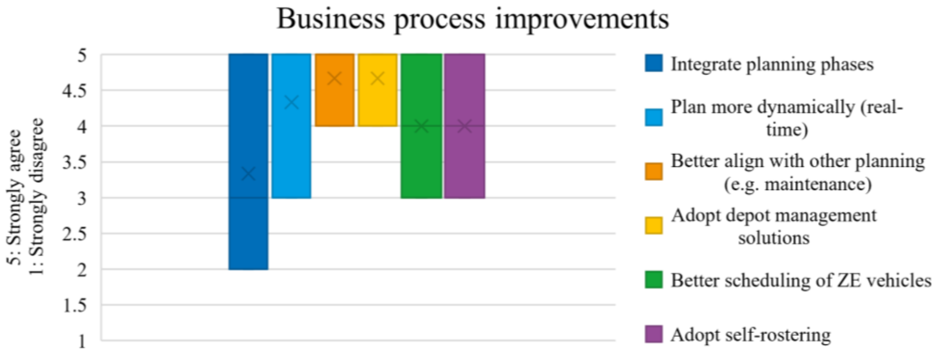

Figure A-10 Goal validation regarding the business process improvements (GVB-only)

Figure A-11 shows the performance expectancy and is answered by 2 GVB experts (PE1 by 3).

Figure A-12 shows the effort expectancy and is answered by 3 GVB experts.

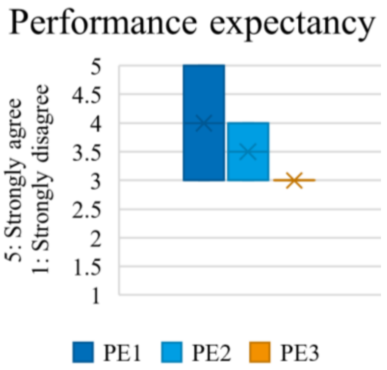

Figure A-11 Performance expectancy results (GVB-only)

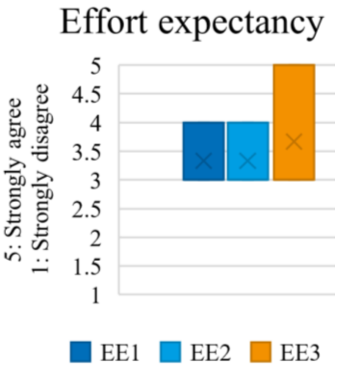

Figure A-12 Effort expectancy results (GVB-only)

Figure A-13 shows the social influence and is answered by 3 GVB experts.

Figure A-14 shows the facilitating conditions and is answered by 3 GVB experts.

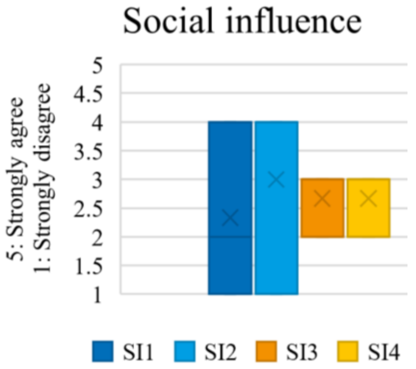

Figure A-13 Social influence results (GVB-only)

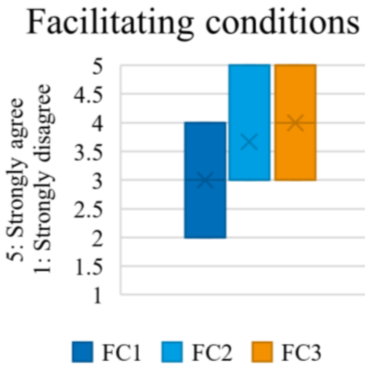

Figure A-14 Facilitating conditions results (GVB-only)

Figure A-15 shows the behavioral intention and is answered by 3 GVB experts.

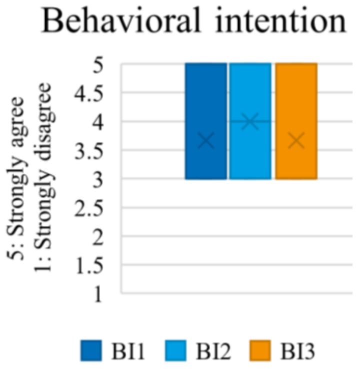

Figure A-15 Behavioral intention results (GVB-only)