UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

A context-driven approach to uncover root causes of data quality issues in intermodal planning: a case study

Ruben Dhaenens

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :
Prof. dr. Mieke JANS

CO-SUPERVISOR :
Prof. dr. Kris BRAEKERS

2020
2021

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculty of Business Economics

Master of Management

Master's thesis

A context-driven approach to uncover root causes of data quality issues in intermodal planning: a case study

Ruben Dhaenens

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :	**CO-SUPERVISOR :**
Prof. dr. Mieke JANS	Prof. dr. Kris BRAEKERS

This master thesis was written during the COVID-19 crisis in 2020-2021. This global health crisis might have had an impact on the (writing) process, the research activities and the research results that are at the basis of this thesis.

A context-driven approach to uncover root causes of data quality issues in intermodal planning: a case study

Ruben Dhaenens*, Mieke Jans[1], and Kris Braekers[1]

[1]Department of Business Economics, University of Hasselt,
Agoralaan Gebouw D, BE 3590 Diepenbeek

* E-mail: ruben.dhaenens@student.uhasselt.be.

Transportation portfolios of logistics companies are more diversified than ever. Moreover, they are increasingly reliant upon data and IT systems to optimally plan and execute transports. In order to adapt to this changing business reality, a new intermodal planning concept has been introduced, i.e. the leg planning system. A leg can be defined as 'a journey between 2 stops'. An optimal combination of legs can be combined into a load, which subsequently receives capacity. In this case study, the ODIGOS framework is applied in order to uncover root causes of data quality issues related to this planning concept. ODIGOS states that data quality issues are caused by interactions of different actors on processes, data, and event logs. Its objective is to reveal underlying relations within the planning processes (i.e. the planning of legs, loads, and capacity). All identified root causes are ultimately categorized into four groups: all issues relating to the receipt of erroneous information from external stakeholders, issues caused by (internal) communication (whether or not supported by business roles or software), data and process-related issues caused by individual mistakes, and issues caused by the use of sub-optimal IT systems. Further research possibilities lie in the analysis of data to better understand the frequencies and relevancies of data quality issues, as well as the management of underlying root causes.

Keywords *– Data quality, Odigos Framework, Intermodal Planning, Supply Chain*

1. Introduction

Supply chains are not static – they evolve and change in size, shape, and configuration, and in how they are coordinated, controlled, and managed (MacCarthy, Blome, Olhager, Srai, & Zhao, 2016). They are continuously faced with (technological) challenges and are becoming increasingly data-dependent, as data-derived insights are seen as an instrument to improve supply chain performance (Brinch, 2018). The concepts of big data, data mining, and business analytics offer decision-making practices to improve firm-level and process-level performance (Lee, Ooi, Chong, & Seow, 2014). Supply chain managers are, for example, increasingly reliant upon data to gain visibility into expenditures, identify trends in costs and performances, and support inventory monitoring, production and planning optimization, and process improvement efforts (Hazen, Boone, Ezell, & Jones-Farmer, 2014). However, the degree to which data can be used is largely determined by their quality (O'Reilly, 1982). Following the "garbage-in garbage-out" principle, the usage of low-quality data has a negative impact on business decisions, leading to both significant tangible (e.g. economical) and intangible effects (e.g. related to job satisfaction) (Hazen, Boone et al. 2014). For example, poor information quality results in 8% to 12% loss of revenue in a typical enterprise and has been estimated to be responsible for 40% to 60% of expenses in service organizations (Ge & Helfert, 2013).

Data can be defined as streams of raw facts representing events occurring in organizations or the physical environment, before they have been organized and arranged into a form that people can understand and use (Salem Al-Mamary, Shamsuddin, & Aziati, 2014). In general, data is collected as a by-product of the operation of systems that support process execution, such as Enterprise Resource Planning (ERP) systems. Data can be used for process mining (PM) purposes, which requires significant manipulation in order to be converted (and cleaned) to event logs suitable for PM techniques (Andrews, Emamjome, ter Hofstede, & Reijers, 2020). However, data is also heavily used for the day-to-day execution of processes (process data). Nonetheless, research shows no increase in attention to the quality of data (both raw data, process-related data, and data for event logs), despite the increase of importance and relevancy (Emamjome, Andrews, & ter Hofstede, 2019).

It’s clear that the importance of high-quality data, which is characterized by their completeness, consistency, validity, conformity, accuracy, and integrity, provides a strong impetus for organizations to actively improve all aspects of data gathering, processing, and analytics activities (Hazen, Weigel, Ezell, Boehmke, & Bradley, 2017; Ranjit & Kawaljeet, 2010). While multiple methods are used for ad-hoc data improvements, and some frameworks were developed (Mans, van der Aalst, Vanwersch, and Moleman (2013) and Bose, Mans, and Aalst (2013)), neither helps with identifying the causes of data quality issues nor with recommending possible remedies (Andrews et al., 2020). Going beyond mere reactive action (e.g. identification and correction of data defects), data quality management works as a preventive concept (M. Hüner, Schierning, Otto, & Österle, 2011). In this respect, the ODIGOS framework was developed (Emamjome, Andrews, ter Hofstede, & Reijers, 2020). The ODIGOS framework facilitates an informed way of dealing with data quality issues - in both raw data, process-related data and event logs - through supporting both prognostic (foreshadowing potential quality issues) and diagnostic (identifying root causes of discovered quality issues) approaches. It aims at capturing the context of a (process mining) project or process, by mapping all actors and influences that potentially affect the quality of data. By understanding the context of a project or process, it is possible to understand/explain the causes of data quality issues (Emamjome et al., 2020).

The goal of this paper is to uncover data quality issues in the supply chain context and to understand their root causes, through the application of ODIGOS. This is done by means of a qualitative study in cooperation with a leading Belgian logistics service provider. Accordingly, the following research question will be answered: “Which are the root causes of data quality issues in intermodal planning-related data, event logs, and processes?”. This paper fills a void in the literature by focusing on data quality in supply chain management.

2. Background

General data quality frameworks aid in the classification of specific types of quality issues (Bose et al. (2013); Kim, Choi, Hong, Kim, and Lee (2003) and Rahm and Do (2000)). Examples of data quality issues include but are not limited to: missing values, spelling errors, duplicated records, values outside domain ranges, etc. These classifications are most useful for future data cleaning purposes, as more targeted techniques could be used to fix said issues. Other frameworks define broad problem categories, like the accuracy, completeness, believability, timeliness, etc. of data (Wang & Strong, 1996). Winkler (2004) proposed a statistician perspective on methods for data editing and data cleaning to remove duplicates.

Recently, research has been focused on different business sectors to better converge focus and expertise. Numerous researchers have used data from electronic health records (EHR) for Operational Research purposes (e.g. the construction of emergency department simulations) in the healthcare sector. As the output of complex hospital simulation models is highly dependent on the input parameters, several frameworks aid in the detection of EHR-specific data quality issues (Orfanidis, Bamidis, and Eaglestone (2004); Vanbrabant, Martin, Ramaekers, and Braekers (2018) and Weiskopf and Weng (2013)). Likewise, the quality of EHR is improved for diagnostic purposes, for example, to provide better cardiac catheterization practices (Byrd et al., 2013) and kidney transplantation procedures (Srinivas et al., 2017). Choudhary, Harding, and Tiwari (2008) provide an overview of the PM applications in the manufacturing industry, including automatic defect detection and job shop scheduling. Bokrantz, Skoogh, Lämkull, Hanna, and Perera (2017) investigate data quality in the automotive manufacturing industry and relate 11 simulation data quality dimensions (e.g. accuracy, relevance, reputation) to four generic categories of data quality (accessibility, intrinsic, contextual, representational). By doing so, they propose a set of practical guidelines that can support manufacturing companies in improving data quality. K. Wang, Tong, Lionel, and Eynard (2008) consider 5M1E (Man, Machine, Material, Methods, Measurement, and Environment) the major influential factors of data quality problems.

This review of existing studies on data quality shows that, although recognized as a critical success factor, a more systematic and generalizable approach for dealing with data quality issues (and their root causes) is needed. Aimed at bridging this gap, the ODIGOS framework (chapter 3.4) was introduced.

3. Research Methodology

In this study, we aim to introduce the concept of preventive data quality management in the logistics sector. The goal is to map potential root causes of data and process-related quality issues in intermodal planning. This detection is a two-step process. In a first step, a business case is defined, for which practical expertise is congregated. Then, a root-cause identification methodology (i.e. the ODIGOS framework) is applied.

To do so effectively, we perform a case study in cooperation with a logistics service provider. The company is one of the most important players in the field of transport and logistics throughout Europe. The company follows an asset-based strategy, owning the entire fleet, all warehouses, and IT systems, making it both representative of the sector and interesting data-wise.

3.1 Case study set-up

As previously mentioned, the case study consists out of two main phases. In the first phase, an introductory session was given, i.e. a presentation in which we introduced ourselves and our research, the importance of data quality management, and the procedure of this research, including the working of root-cause detection framework ODIGOS. This was followed by the first of two semi-structured interviews. Its goal was to gather information such as the areas in which data is used within the company and to get an overview of which data-related issues are currently encountered. Based on the results of the interview, a relevant business problem (case) was identified.

In a second phase (i.e. a second interview session), the business case is taken under a magnifying glass. The main goal was to identify the context of the business case. This was done specifically according to the principles of ODIOGOS, i.e. by identifying all actors and influences that (potentially) impact the quality of data, event logs and processes.

3.2 Interview process

Interviews are one of the main data collection methods in qualitative research; they provide a means to gather rich data. A semi-structured interview ensures all questions are covered during the interview. However, digressions to related topics of discussion were permitted to increase the richness of the information captured (Andrews et al., 2020). In this case study, two interview sessions were organized. Both Business and Functional Analysts, along with the Business Unit Manager of the Business Intelligence (BI) department, were interviewed. All interviewees have first-hand experience with intermodal planning processes and are to some extent involved with the development of the intermodal leg planning concept.

The first interview initially aimed at understanding the scope and operations of the company. It then progressed to how and where data is used. The final set of questions related to data quality and current issues. Furthermore, a business case was identified, which is outlined in the next section. The second interview combined experience from all interviewees related to the business case, its actors, and their respective influence on data and processes. Using a digital whiteboard and post-its, actors and relations were mapped according to the ODIGOS' principles, further outlined in chapter 3.4. Iteratively, (potential) causes of data and process-related quality issues caused by these relations were captured.

3.3 Business case

To better understand which actors have (which) impacts on data, event logs, and actual processes, a specific business case was used as a starting point. That is, by demarcating the scope of a Supply Chain related project, processes, data, and event logs specifically needed for the project can be taken under a magnifying glass; in this case to identify actors and influences on data and process quality. This narrowing down was an essential part of this research, as a clear "focus" was needed to ensure detailed root causes were uncovered, rather than a summation of general potential sources.

Business cases originate from a need or a problem within a company and are tied together with a research question the company aims to solve. While a relevant and representative business case was identified in this research, answering the underlying research question(s) would deviate from this research's purpose.

3.3.1 The intermodal leg planning system

The transportation business has evolved significantly over the last decade (Hoppe, Christ, Castro, Winter, & Seppänen, 2014). Consequently, the transportation portfolios of Supply Chain companies are more diversified than ever. Ongoing evolutions such as the globalization, an increase in transported volumes and the continuous opening of logistics nodes/hubs throughout the world add additional complexity to contemporary transport management systems. A transportation management system (TMS) is a logistics platform that uses technology to help businesses plan, execute, and optimize the physical movement of goods, both incoming and outgoing, and is used to make sure the shipment is compliant with all regulations (Oracle, 2021). To adapt the current TMS system to the changing business reality, a new planning concept is introduced. Using what will be called a "leg-planning" system, the company featured in this research aims to better optimize its resources, costs, and other time-related KPI's (such as the lead time) of its intermodal transports.

The main idea is that transport orders are split into the most optimal combination of "legs" (figure 1) given the available transport network. A leg can be defined as 'a journey between 2 stops'. Each leg gets one of the following transport modes: either Road, Rail, Water or Air. However, a leg cannot have multiple transport modes, and switching between separate legs (transport modes) can only be done by using specific multimodal terminals.

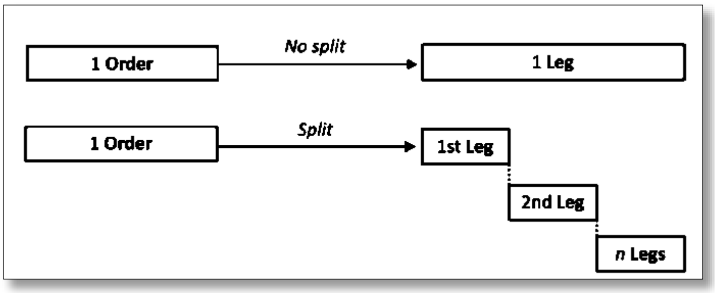


Figure 1: Splitting of orders into legs

Within this new planning concept, the created legs can then be combined into the most optimal loads. A load can be defined as "a leg or a combination of legs that will be transported together with the same transport mode, with a clear planned sequence of the activities and by taking into account all the limits of the actual transport (such as weight, volume, timing,…)". In other words, loads receive road or intermodal capacity and need to be organized with a correct sequence of activities and planned timings. Once all details of a load are known, the load can become a trip, which can be executed according to the existing functionalities. Figure 2 represents a simplified overview of the planning concept.

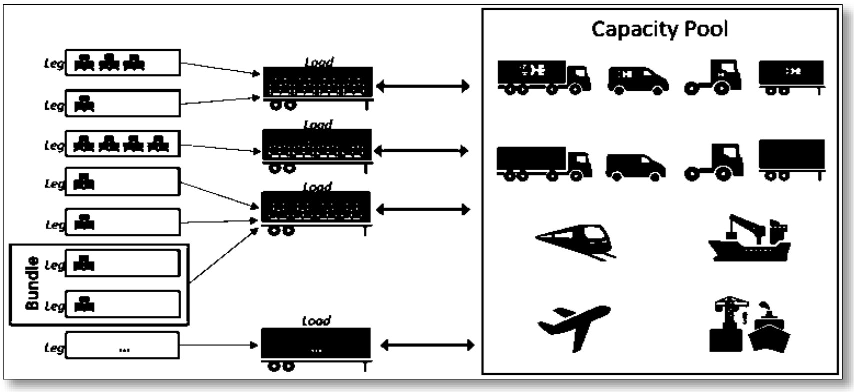


Figure 2: Legs into loads, a visual representation

The creation of legs and the following transformation into loads entail many different individuals, systems, and social structures, all intertwined with each other. This interconnectivity of process actors is further complexified by the international character of the concept, which in effect amplifies the risk on issues regarding data or process quality (e.g. caused by communication errors).

By contextualizing this planning process (i.e. by mapping all actors involved in the system and their interactions/influences on respective data, event logs, or processes), potential causes of data quality issues can emerge..

3.4. ODIGOS-framework to contextualize projects

Process or event data, usually considered as the starting point for any (PM) project, is generated as a result of interactions between different actors, i.e. process participants, automation pieces (e.g. bots), data curators (people in charge of deriving process data from information systems) and information systems, all influenced by the organizational rules, procedures, norms, and culture. This understanding of data implies that quality issues are caused as a result of interactions between these different actors (process participants, bots, data curators, etc.), systems, and the context (Emamjome et al., 2020). The ODIGOS framework, based on the semiotics framework of Mingers and Willcocks (2014) and adapted to the context of process mining by Emamjome, Andrews, ter Hofstede, and Reijers (2020) (figure 1), allows for a systematic mapping of these interactions. In other words, the framework allows for contextualizing influences on data, and thus all potential causes of quality issues. To do so, Emamjome et al. (2020) categorized the actors into three main “worlds” i.e. the social, personal and material worlds.

The social world is an ensemble of social structures, culture and norms, practices, and conventions realized in the form of role positions and social practices. The social world includes but is not limited to laws, regulations, or agreements within or between companies. Emamjome et al. (2020) further specified situational (relating to the firm itself), and macro-social structures. The latter includes the broader social context: the economy, history, culture, gender, etc. that influences the behavior of actors.

The personal world refers to the actors who are involved in the process of creating content (i.e. data, actual processes, and event logs) as well as their beliefs, values, motivations, and expectations. In relation to the concept of data quality, two main actors are recognized: process participants and data curators. The role of process participants is to perform the processes and to create event data, such as the registrations of data by white-collar employees. The data curator's role is to create event logs from event data for the purpose of analysis. This job is usually reserved for business analysts, planners, etc.

The material world relates to the physical structure of communication media, whether it's technological or not. Three layers are specified:

● The presentation layer, which includes all physical structures (such as personal computers or other devices) and interfaces (such as forms, query interfaces, and reports) used for communication purposes;

● The application layer, which consists of program instructions that support business rules and transactions;

● The data layer, which consists of data warehouse technologies that support intensive data analysis, the hard coding of software, etc.

As visually represented in figure 3, the three worlds within the ODIGOS framework aim to capture and categorize all actors that (potentially) influence the "semiotic content", i.e. raw data, process related data, and data for and from event logs.

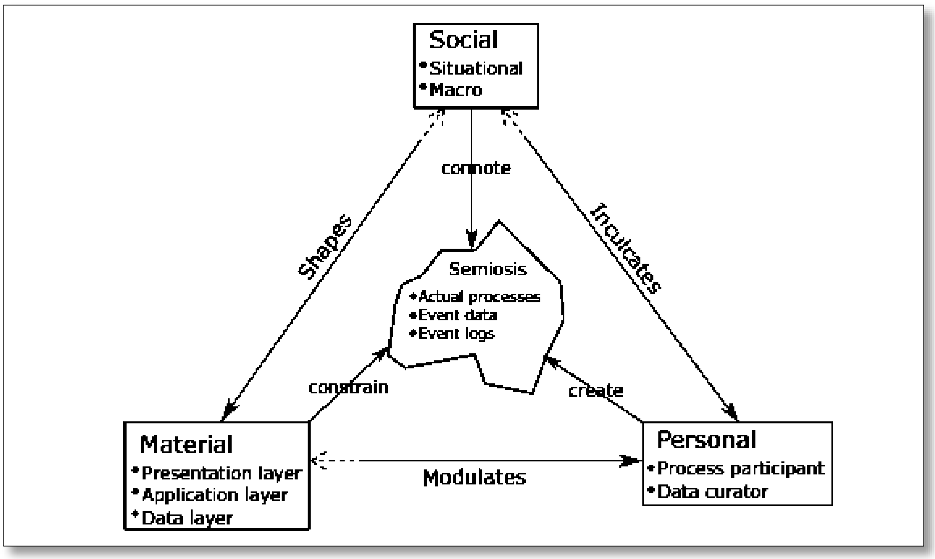


Figure 3: The ODIGOS framework visually represented

Social structures can influence Process Participants' intentions, attitudes, and behaviors relating to how they perform their tasks. This interaction is denoted "incalculate" in the ODIGOS' visual representation. Moreover, social structures can influence ("shape") the use and development of IT and communication systems. A practical example would include IT-related features that are developed and embedded into ERP systems, relating to the segregation of different roles in a company. That is, inconsistencies of assigned roles and consequently, the order how of tasks can be executed, can have a severe effect on data and/or event log quality.

Data quality issues can also emerge from the interaction between the Material World (the IT systems) and the Personal World (the Process Participants and Data Curators). A common example includes data entry errors, i.e. when a Process Participant (such as an order entry employee) wrongly records certain information using an interface. Another example, and one uncovered during this research, relates to the joining of information from multiple incompatible IT systems by the Data Curator (e.g. the transport planner).

4. Research findings

Data quality issues often come to light when a (process mining) project, analysis, or process is well underway. However, root causes can originate from early on in the project and from a wide range of causes. In order to uncover root causes related to the leg planning system, different actors involved in the underlying processes are mapped. This was done chronologically, i.e. starting from a pre-planning phase. Figure 4 visualizes all phases within the leg planning concept.

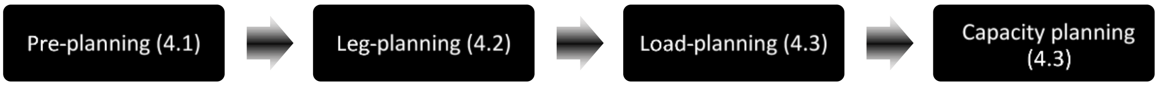


Figure 4: Phases within the leg planning concept

4.1. Pre-planning phase

The pre-planning phase consists out of both the receipt and the checking of information needed to execute all downstream planning processes. Information such as the pick-up and delivery location, the amounts, weights, and volumes of the goods, an indication about the stackability, a signal if the cargo is dangerous or not, etc. are communicated by the clients to the company's order registration team. According to experts involved in this case study, this is the first root cause of data quality issues, as wrong (input) data is frequently sent by customers.

Which data, its form, and how it's communicated is established in agreements with individual clients. This results in a plethora of different forms being used (e.g. phone calls, e-mails, excel files, etc.). The order information is extracted from the forms and is reviewed, in order to ensure logical, plausible, and complete data will be used in the planning processes. The TMS system has features implemented to minimize risks related to order entry issues. For example, a problem solver, a tool concerning geo-mapping (in case of wrong or unclear addresses), and others are embedded in the system. Within the new leg planning system, a stricter approach is maintained, as it is currently possible to create transports with critical information missing. The reviewing employee and the assisting IT systems are identified as main actors causing data quality and process-related issues. That is, the manual execution of the reviewing process and labour intensive order entry tasks are prone to mistakes. Moreover, the supporting software is found to heavily influence ("modulate") the employees' work. Following the visual representation of the ODIGOS framework, the reviewing process and its underlying actors and influences can be mapped accordingly (figure 5).

An additional root cause of data quality issues was acknowledged when zooming in on the distribution of orders to planning pools. When orders are received by the order registration team, orders are segregated and dispersed conform to underlying business rules. This is done for safety's sake, i.e. not everyone is allowed to plan every order. However, experts mentioned that wrongful distributions have occurred, which had an impact on downstream (planning) processes as well as the quality of data (i.e. databases were polluted with wrong orders).

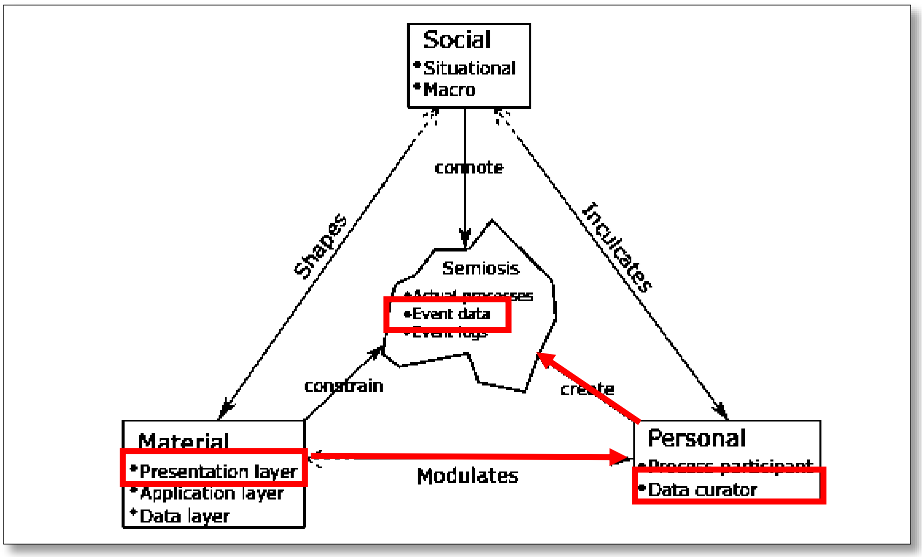
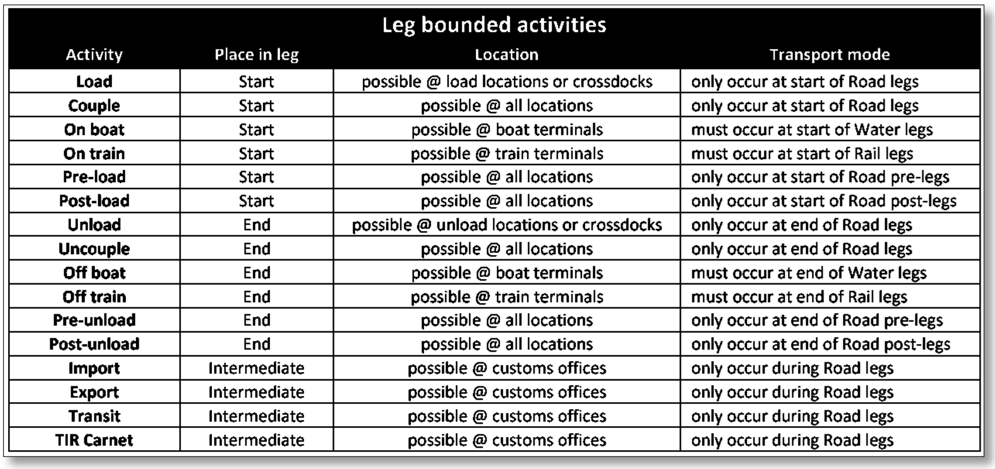


Figure 5: Actors, influences, and root-causes of data quality issues in the reviewing process

4.2. Leg-planning phase

Additionally, causes of data quality issues were identified during the transmission of information from the pre-planning environment towards the leg-planning environment. That is, the information registered in the TMS by the order registration employees needs to be translated into "events" or "activities", in order to be useful later on. Figure 6 identifies all of those possible "activities".

A practical example of such data quality issues observed includes the translation of time zones from the TMS (which utilizes the Central European Time system, CET) to the time zones utilized within the leg planning environment (which uses the Coordinated Universal Time format, UTC). Both the risk on the front-end regarding a wrong input/entry of the time format, as well as the risk on the back-end regarding the conversion of time zone formats have emerged as root causes.



Leg bounded activities			
Activity	Place in leg	Location	Transport mode
Load	Start	possible @ load locations or crossdocks	only occur at start of Road legs
Couple	Start	possible @ all locations	only occur at start of Road legs
On boat	Start	possible @ boat terminals	must occur at start of Water legs
On train	Start	possible @ train terminals	must occur at start of Rail legs
Pre-load	Start	possible @ all locations	only occur at start of Road pre-legs
Post-load	Start	possible @ all locations	only occur at start of Road post-legs
Unload	End	possible @ unload locations or crossdocks	only occur at end of Road legs
Uncouple	End	possible @ all locations	only occur at end of Road legs
Off boat	End	possible @ boat terminals	must occur at end of Water legs
Off train	End	possible @ train terminals	must occur at end of Rail legs
Pre-unload	End	possible @ all locations	only occur at end of Road pre-legs
Post-unload	End	possible @ all locations	only occur at end of Road post-legs
Import	Intermediate	possible @ customs offices	only occur during Road legs
Export	Intermediate	possible @ customs offices	only occur during Road legs
Transit	Intermediate	possible @ customs offices	only occur during Road legs
TIR Carnet	Intermediate	possible @ customs offices	only occur during Road legs

Figure 6: Leg bounded activities

When order information is successfully entered, "leg templates" can be utilized. Leg templates could be described as a pre-defined combination of leg-splits, to assist in the planning on regularly exploited routes. Leg templates can be configured in advance by the planning employees. However, with both the reliance on the human cognitive capacity for creating the templates (cfr. "personal world"), as well as the reliance on underlying coding (cfr. "material world"), risks related to the quality of data arise. Leg planners, who in this case function as data curators, can split legs sub-optimally and can, for example, split legs into two practically illogical sequences. According to experts in the field, when the set of split legs is further complexified, for example when revisions or changes are made, the number of human errors increases heavily. An example of such revisions could entail the removal of an air freight leg split and replacing it with a train transport split. Or, the change of dates assigned to the "activities" overviewed in figure 6.

To assist the planners with planning legs optimally, the leg planning system employs a state-of-the-art algorithm. The idea is that the algorithm is fed with order data and additionally required input data about the transport network (e.g. costs, cross docks, train and boat departures, etc.). The algorithm then processes the data and returns the best possible leg split (i.e. with a minimized cost to execute). The integration of the algorithm with other components of the TMS is recognized as a challenge. That is, similar to the compatibility issues between the order entry and leg planning environment, the compatibility between the algorithm and the TMS may be a potential root cause of (data) quality issues down the line.

Once the leg planners have designated all information as described in figure 7, legs can be put on a load.

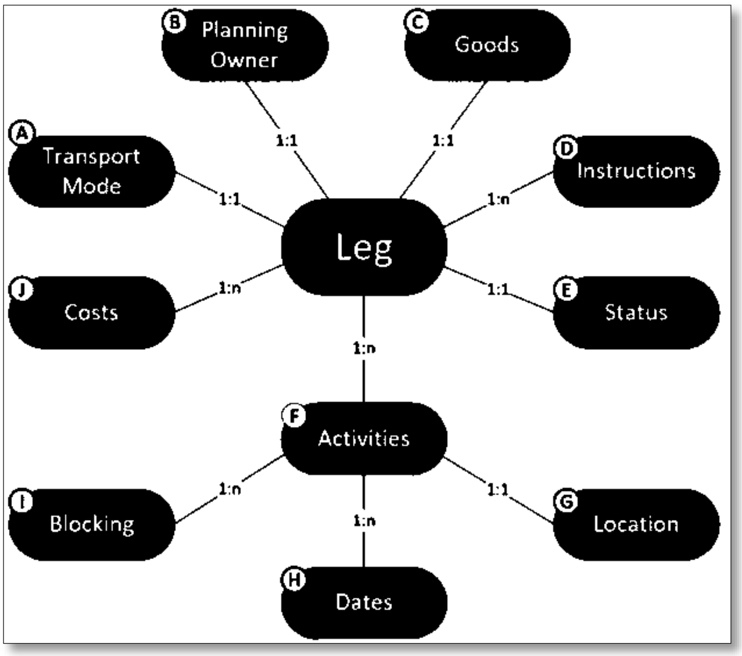


Figure 7: Information required to plan a leg

4.3. Load-planning phase

The following phase in the planning process is the combination of legs into loads (figure 2). This is where load planners come into play: they compile soon-to-be physical transports out of an optimal combination of legs. Both full truckloads (i.e. loads consisting out of one leg) and less than container loads (i.e. loads consisting out of multiple legs) are possible. Load planning is a complex process and relies heavily on the cognitive capacity of the planners. This characteristic, again, inevitably results in the process being prone to errors – and process/data-related quality issues.

Moreover, external laws and regulations, such as protocols regarding the combination of dangerous goods, rules about the temperature of pharma products, laws concerning the total axle weights, etc. have to be taken into account when planning loads. Framing in the "social world" of the ODIGOS framework, these external factors heavily influence both the data curators (the planners) and the development of IT-related support systems (for example, the inclusion of certain features, limitations, and checks) both potentially influencing the quality of the processes, data and/or event logs. These risks are visually mapped with the aid of ODIGOS in figure 8.

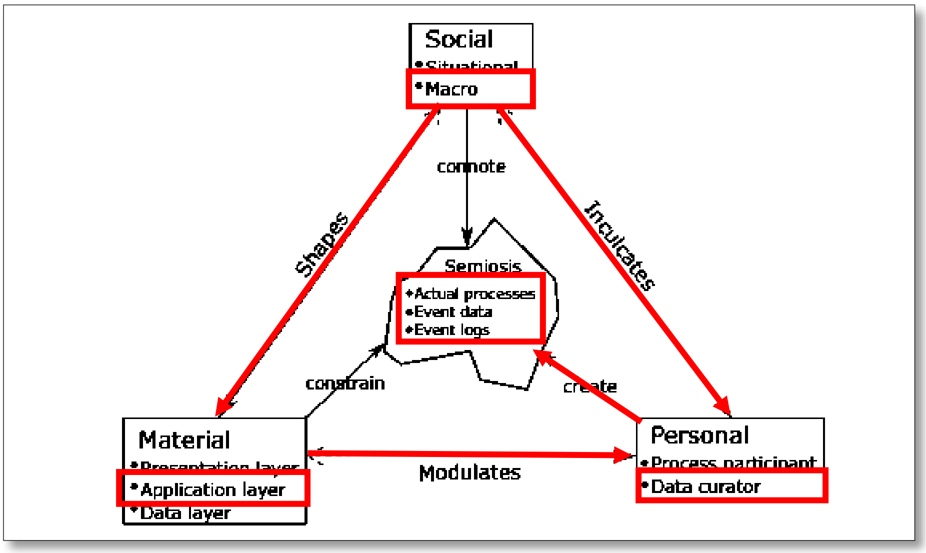


Figure 8: Actors and influences during the load planning process

4.4. Capacity-planning phase

The final phase within the leg planning concept is the assignment of loads to capacity. This is the responsibility of capacity planners, who aim to optimally allocate loads to physical resources. A clear distinction is made between road capacity (e.g. trucks, trailers, and drivers) and intermodal capacity (trains, boats, containers, etc.). While the available capacity is already taken into consideration when assigning legs to loads, individual resources such as the specific drivers, are now appointed. During the allocation process, many parameters and constraints are considered. Some examples include:

- the availability of spots on boats, trains, and airplanes
- the required container type (i.e. reefers, insulated containers, etc.)
- the regulations regarding driving/resting times
- the certification of drivers (i.e. who is allowed to transport dangerous goods)

From an ODIGOS point of view, actors categorized in both the social and personal worlds are identified to have influences on the execution of this process, as well as its underlying data. The ‘macro’ component of the social world covers all regulations the capacity planners have to take into account.

The personal world entails risks on data and process quality originating from the precarious characteristic of external planning schedules, such as the departure schedules of trains and airplanes.

Once capacity has been assigned, formal and informal agreements are communicated from the planning application to internal and external staff as well as their interfaces (such as drivers' onboard computers). This communication happens automatically and is fully digital, but transpires international and organizational borders. It is identified as a root cause of data/process-related issues as well, as many details get "lost in translation".

It is of note that, during the interviews, no root causes of data quality issues were identified regarding the flow of documents. All processes related to (physical) documents are handled outside of the leg planning concept and are outside of this research's scope.

5. Results and conclusions

During this research, a leg planning concept for intermodal transport is taken under a magnifying glass. The goal of the concept is to better accommodate the TMS to intermodal transports, by introducing the concepts of "legs" and "loads". The planning concept is identified as a current and relevant business case and is utilized to concentrate attention, expertise, and data in the search for causes of data (and process-related) quality issues. The uncovering of such causes is guided by the ODIGOS framework. The framework aims at capturing the context of a project (in this case the leg planning), by identifying actors and influences on underlying data, processes, and event logs.

With the entire leg planning concept comprising out of a multitude of actors, (IT) systems, business rules, and procedures, all intertwined with each other, multiple root causes of data and process-related issues were identified. They can be grouped into the following four categories:

1. Erroneous information received from external stakeholders

This category includes the receipt of wrong, incomplete, or unreliable information from, among others, customers (i.e. orders containing wrong information) or external stakeholders, such as shipping companies (i.e. unreliable departure schedules and time slots) or governmental institutions (i.e. relating to changes in laws and regulations).

2. Issues caused by (internal) communication, whether or not supported by business roles or software

This classification includes all root causes stemming from the risks of the communication and transmission of information internally. Both the communication relying on potentially imperfectly constructed business rules (e.g. the distribution of transport orders to wrong planning pools), as well as miscommunications occurring due to software-related issues, fall under this category. Examples of the latter include errors in the translation of time-zone formats from the order entry to load-planning environments, as well errors in the communication of (in)formal agreements from capacity planners to drivers.

3. Data and process-related issues caused by individual mistakes

This category contains all issues caused by the incorrect registration, planning, or execution of tasks by employees. During the interviews, these issues were often brought up, as the planning processes heavily rely on the capacities of individual planners. Examples include but are not limited to wrong order entry, sub-optimal leg splitting, and wrong leg template configurations.

4. Issues caused by the use of sub-optimal IT systems

This classification umbrellas all root causes uncovered relating to the reliance on potentially sub-optimal software (such as the algorithm for automatic leg splitting proposals) and the development of software-related features, limitations, and tools (such as the inclusion of problem-solving tools, limitations regarding when orders can (not) be planned, etc.)

The following paragraphs dive deeper into the root causes of revealed data quality issues, following the chronological phases within the leg-planning concept, i.e. the pre-planning phase, the planning of legs, the planning of loads, and the assignment of loads to capacity.

During the pre-planning phase, the order information is collected, reviewed, and registered by order entry employees. A first root cause of data quality issues emerges immediately, as experts detected wrong or incomplete order data sent in by customers (e.g. wrong addresses, weights, and volumes, etc.). The reviewing of order information, as well as the registration in the TMS, were pinpointed additionally, as the labour-intensive and concentration-demanding nature of the tasks often lead to order entry mistakes. While supporting procedures and mechanisms are in place, for example, geo-mapping tools to solve issues regarding addresses, such IT tools may still produce sub-optimal results. Additionally, the reliance on business rules concerning the distribution of orders to planning pools was recognized.

Throughout the leg planning phase, which can be defined as the phase in which orders are split into an optimal combination of legs, three causes of data quality issues emerged. During the transmission of (order) information from the pre-planning to the leg-planning environment, this research uncovered issues regarding the formatting of time zones. Furthermore, a risk of incorrectly or illogically constructed "leg templates" was recognized. Leg templates are defined as a combination of pre-defined leg splits, to assist in the planning of regular transportation itineraries. However, similar to the entry of order information, the construction of leg templates is a human task that entirely relies on the input of the employees (leg planners). Finally, with the usage and heavy reliance on the black-box algorithm for automatic leg-splitting, questions regarding its accuracy and compatibility with other parts of the TMS (and consecutively its potential impact on data quality) were raised.

The load planning process comprises of the combination of legs into loads. The process of merging legs is prone to human mistakes as well and is thus identified as a potential root cause of process-related issues. Additionally, external laws and regulations are identified as main actors and influences on both the development of software and the quality of work delivered by planners. Following the categorization of actors within the ODIGOS framework (i.e. the material and personal world both being influenced by the social world), the underlying relations and effects on data and processes were revealed.

Within the final planning phase, i.e. the allocation of capacity to loads, the volatile characteristic of external information (such as the rapidly changing availability of capacity on trains, airplanes, and boats as well as their schedules), was identified as the main cause of data quality issues. Furthermore, the reliance on the cognitive capacity of capacity planners (and, again, in specific their proneness to mistakes), as well as communication errors between planners and downstream employees, are distinguished as important root causes.

6. Limitations and research opportunities

While multiple root causes of data and process-related quality issues were successfully unearthed, the scope of the investigated project was limited to the planning phases (i.e. from pre-planning to capacity planning). However, during the execution of the intermodal transports, certain information is injected back into the planning process to support posterior (planning) processes. The communication back to the TMS as well as all actors, structures, and tools that come into play, potentially influence the integrity of downstream data and processes. However, they were not included in this study. Furthermore, the leg planning concept is uniquely customized to the organization. While similar planning concepts may exist, actors, influences, and underlying relations affecting data, processes, and (the art of creating) event logs may be drastically different. Due to the limited sample size (e.g. few employees from only one company), the research findings may be prone to sampling bias.

From investigations regarding the improvement of communication technologies to management practices regarding the handling of incoming “dirty” data, research opportunities are plenty. More specifically, researchers could zoom in to the proposed four categories of uncovered root causes, with the goal of minimizing their related risks to data quality. That is, practical approaches on how to tackle and solve root causes of data quality-related issues (instead of just the symptoms, i.e. data cleaning) are currently few. Additional data analyses may provide insights into the severity and frequency of quality issues, which in effect may help in the prioritization for tackling their root causes. That said, during this research, no actual data was analyzed, allowing for future research opportunities.

References

Aalst, W., Guenther, C., Recker, J., & Reichert, M. (2006). Using Process Mining to Analyze and Improve Process Flexibility - Position Paper.

Alshawi, S., Missi, F., & Irani, Z. (2011). Organisational, technical and data quality factors in CRM adoption — SMEs perspective. *Industrial Marketing Management, 40*(3), 376-383. doi:https://doi.org/10.1016/j.indmarman.2010.08.006

Andrews, R., Emamjome, F., ter Hofstede, A., & Reijers, H. (2020). An expert lens on data quality in process mining. In *2nd International Conference on Process Mining*: IEEE.

Bokrantz, J., Skoogh, A., Lämkull, D., Hanna, A., & Perera, T. (2017). *Data Quality Problems in Discrete Event Simulation of Manufacturing Operations*.

Bose, R. P. J. C., Mans, R. S., & Aalst, W. M. P. v. d. (2013, 16-19 April 2013). *Wanna improve process mining results?* Paper presented at the 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM).

Brinch, M. (2018). Understanding the value of big data in supply chain management and its business processes: Towards a conceptual framework. *International Journal of Operations & Production Management*. doi:10.1108/IJOPM-05-2017-0268

Byrd, J. B., Vigen, R., Plomondon, M. E., Rumsfeld, J. S., Box, T. L., Fihn, S. D., & Maddox, T. M. (2013). Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: The VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. *The American Heart Journal, 165*(3), 434-440. doi:http://dx.doi.org/10.1016/j.ahj.2012.12.009

Choudhary, A. K., Harding, J. A., & Tiwari, M. K. (2008). Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing, 20*(5), 501. doi:10.1007/s10845-008-0145-x

Emamjome, F., Andrews, R., & ter Hofstede, A. H. M. (2019, 2019//). *A Case Study Lens on Process Mining in Practice.* Paper presented at the On the Move to Meaningful Internet Systems: OTM 2019 Conferences, Cham.

Emamjome, F., Andrews, R., ter Hofstede, A., & Reijers, H. (2020). Alohomora: Unlocking data quality causes through event log context. In *Proceedings of the 28th European Conference on Information Systems (ECIS2020)* (pp. 1-16). United States of America: Association for Information Systems.

Emamjome, F., Andrews, R., ter Hofstede, A., & Reijers, H. (2020). Alohomora: Unlocking data quality causes through event log context. In *Proceedings of the 28th European Conference on Information Systems (ECIS2020)* (pp. 1-16). United States of America: Association for Information Systems.

Even, A., Shankaranarayanan, G., & Berger, P. D. (2010). Evaluating a model for cost-effective data quality management in a real-world CRM setting. *Decision Support Systems, 50*(1), 152-163. doi:https://doi.org/10.1016/j.dss.2010.07.011

Ge, M., & Helfert, M. (2013). IMPACT OF INFORMATION QUALITY ON SUPPLY CHAIN DECISIONS. *The Journal of Computer Information Systems, 53*(4), 59-67. Retrieved from https://search.proquest.com/docview/1429691450?accountid=27889

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics, 154*, 72-80. doi:10.1016/j.ijpe.2014.04.018

Hazen, B. T., Weigel, F. K., Ezell, J. D., Boehmke, B. C., & Bradley, R. V. (2017). Toward understanding outcomes associated with data quality improvement. *International Journal of Production Economics, 193*, 737-747. doi:https://doi.org/10.1016/j.ijpe.2017.08.027

Hoppe, M., Christ, A., Castro, A., Winter, M., & Seppänen, T.-M. (2014). Transformation in transportation? *European Journal of Futures Research, 2*. doi:10.1007/s40309-014-0045-6

Kim, W., Choi, B.-J., Hong, E., Kim, S.-K., & Lee, D. (2003). A Taxonomy of Dirty Data. *Data Min. Knowl. Discov., 7*, 81-99. doi:10.1023/A:1021564703268

Lee, V.-H., Ooi, K.-B., Chong, A. Y.-L., & Seow, C. (2014). Creating technological innovation via green supply chain management: An empirical analysis. *Expert Systems with Applications, 41*(16), 6983-6994. doi:https://doi.org/10.1016/j.eswa.2014.05.022

M. Hüner, K., Schierning, A., Otto, B., & Österle, H. (2011). Product data quality in supply chains: the case of Beiersdorf. *Electronic Markets, 21*(2), 141-154. doi:10.1007/s12525-011-0059-x

MacCarthy, B. L., Blome, C., Olhager, J., Srai, J. S., & Zhao, X. (2016). Supply chain evolution – theory, concepts and science. *International Journal of Operations & Production Management, 36*(12), 1696-1718. doi:10.1108/ijopm-02-2016-0080

Mans, R. S., van der Aalst, W. M. P., Vanwersch, R. J. B., & Moleman, A. J. (2013, 2013//). *Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions.* Paper presented at the Process Support and Knowledge Representation in Health Care, Berlin, Heidelberg.

Mingers, J., & Willcocks, L. (2014). An integrative semiotic framework for information systems: The social, personal and material worlds. *Information and Organization, 24*, 48–70. doi:10.1016/j.infoandorg.2014.01.002

O'Reilly, C. A. (1982). Variations in Decision Makers' Use of Information Sources: The Impact of Quality and Accessibility of Information. *Academy of Management journal, 25*(4), 756-771. doi:10.2307/256097

Orfanidis, L., Bamidis, P. D., & Eaglestone, B. (2004). Data Quality Issues in Electronic Health Records: An Adaptation Framework for the Greek Health System. *Health Informatics Journal, 10*(1), 23-36. doi:10.1177/1460458204040665

Petrovic, M. (2020). Data quality in customer relationship management (CRM): Literature review. *Strategic Management, 25*, 40-47. doi:10.5937/StraMan2002040P

Price, R., & Shanks, G. (2011). The Impact of Data Quality Tags on Decision-Making Outcomes and Process. *Journal of the Association for Information Systems, 12*(4), 323-346. Retrieved from https://search.proquest.com/docview/866307276?accountid=27889

Rahm, E., & Do, H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull., 23*, 3-13.

Ranjit, S., & Kawaljeet, S. (2010). A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. *International Journal of Computer Science Issues, 7*.

Reid, A., & Catterall, M. (2005). Invisible data quality issues in a CRM implementation. *Journal of Database Marketing & Customer Strategy Management, 12*(4), 305-314. doi:10.1057/palgrave.dbm.3240267

Srinivas, T. R., Taber, D. J., Su, Z., Zhang, J., Mour, G., Northrup, D., . . . Mauldin, P. D. (2017). Big Data, Predictive Analytics, and Quality Improvement in Kidney Transplantation: A Proof of Concept. *American journal of transplantation, 17*(3), 671-681. doi:10.1111/ajt.14099

van der Aalst, W. M. P. (2016). *Process Mining : Data Science in Action*. Berlin, Heidelberg, GERMANY: Springer Berlin / Heidelberg.

Van Der Aalst, W. M., Becker, J., Bichler, M., Buhl, H. U., Dibbern, J., Frank, U., . . . Zdravkovic, J. (2018). Views on the Past, Present, and Future of Business and Information Systems Engineering. *Business and Information Systems Engineering*, 60(6), 443–477. doi: 10.1007/s12599-018-0561-1

Vanbrabant, L., Martin, N., Ramaekers, K., & Braekers, K. (2018). Quality of input data in emergency department simulations: Framework and assessment techniques. *Simulation Modelling Practice and Theory, 91*. doi:10.1016/j.simpat.2018.12.002

Wang, K., Tong, S., Lionel, R., & Eynard, B. (2008). Analysis of Data Quality and Information Quality Problems in Digital Manufacturing. *Proceedings of the 4th IEEE International Conference on Management of Innovation and Technology, ICMIT*. doi:10.1109/ICMIT.2008.4654405

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst., 12*(4), 5–33. doi:10.1080/07421222.1996.11518099

Weinhardt, C., Kloker, S., Hinz, O., & van der Aalst, W. M. P. (2020). Citizen Science in Information Systems Research. *Business & Information Systems Engineering*, 61(4). doi: 10.1007/s12599-020-00663-y

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association, 20*(1), 144-151. doi:10.1136/amiajnl-2011-000681

Winkler, W. (2004). Methods for evaluating and creating data quality. *Information Systems, 29*, 531-550. doi:10.1016/j.is.2003.12.003

Xu, H., Nord, J. H., Brown, N., & Nord, G. (2002). Data quality issues in implementing an ERP. *Ind. Manag. Data Syst.*, *102*, 47-58.