1 Introduction

Since the transfer of the logistics approach from the military to the private sector in the 1950s (Morgenstern 1955), logistics has developed into one of the most important economic sectors in Germany and is a decisive factor in global competition (Kübler et al. 2015; Backhaus et al. 2019). As an independent scientific discipline, logistics is responsible for the design of the flow of materials, information, finance, and energy by means of logistics processes (Fleischmann 2008a; Göpfert 2013; Koether 2018b). The task of logistics is to process logistics objects, such as material goods, people, and information according to customer demand (Huber and Laverentz 2018). The goal of logistics can be summarized with the "eight R's of logistics" (Hausladen 2016), which, especially in German-speaking countries, is characterized by the initial definition from Jünemann and Pfohl (1989). Hausladen (2016) describes the goal of logistics as providing the right product, at the right time, in the right quantity, at the right place, with the right quality, to the right customer, at the right cost, and with the right information. Logistics systems of industrial companies are typically complex global networks that are subject to uncertainty (Arnold et al. 2008). Such a network is known as a Supply Chain (SC) (Bichler et al. 2017). Within an SC, a multitude of heterogeneous actors are involved in different locations of the network (Werner 2017). A part of an SC and a prominent example for such complex networks involving different actors is a materials trading network, through which a wide variety of products are distributed from producers to customers (Barth et al. 2015; Bretzke 2020). SCs in general and materials trading networks in particular have been experiencing an increase in complexity in recent years (ten Hompel et al. 2014). The induced uncertainty is confirmed by studies that cite endogenous and exogenous examples for the complexity (cf. Kersten et al. 2017). In particular, the studies point out that digital transformation is driving these trends, and has a decisive influence on the complexity of a materials trading network (Kersten et al. 2017).

Mastering the complexity of a materials trading network is a challenging task for the Supply Chain Management (SCM), which is responsible for the cross-company design of planning, controlling, and monitoring the logistics processes within the network (ten Hompel and Heidenblut 2011). Thus, SCM is confronted with a multitude of logistics tasks, which are typically subject to evaluating decision options and making decisions (Eapen 2009). In order to maintain a materials trading network in a competitive state, a human actor coping with a logistics task would have to evaluate a very large number of scenarios that each potential set of changes might induce on the network. Due to the complexity of such networks, a manual approach of decision-making does not appear to be feasible.

Therefore, decision-making in the context of logistics tasks within a materials trading network is supported by Information Technology (IT) (Hausladen 2016). An example of such an IT system is a Logistics Assistance System (LAS), which

supports decision-makers from the SCM in solving logistics tasks. In this context, the quality of the underlying data is of essential importance. LAS rely on vast amounts of data generated by the logistics processes and allow the application of data analytics methods. Popular methods originate from fields like data mining and machine learning. However, the general availability of data poses further challenges for the application of such methods. A commonly encountered challenge and a major cost driver in companies is a lack of sufficient data quality (Otto et al. 2011; Hazen et al. 2014). Another drawback of data from logistics processes is that they limit the types of insights possible, since methods of data mining and machine learning may only discover correlative, or even spurious, patterns in the data (Sanchez and Sanchez 2017; Sanchez 2018).

A frequently-used method in this context is simulation (Gutenschwager et al. 2017; Pfohl 2022). Based on an experimentable model derived from the data of a system, simulation experiments can be used to investigate the model behavior (Rabe et al. 2008b). Within a simulation study, this can contribute to cope with complexity and, thus, is beneficial for decision support. With increasing digitization and technological progress, the general availability of vast storage and computing power allows for the application of computationally expensive simulation studies for purposeful data generation (Sanchez 2018). The method of using modeling and simulation, Design of Experiments (DOE), and analysis of purposeful generated simulation result data is called Data Farming (DF) (Horne and Meyer 2016). Here, the metaphor of "farming" or "growing" data refers to the purposeful manipulation of a simulation model using experiments to generate data (Sanchez 2018). As a result, it is possible to represent a comprehensive model behavior through the resulting data basis and, thus, gain an accurate understanding of the model. The origin of DF comes from a military application area ("Project Albert") of the United States of America (Brandstein and Horne 1998; Forsyth et al. 2005). Thus, previous research contributions have focused mainly on tasks and use cases in this area (Horne and Meyer 2016). These include issues related to defense organization, troop movements, or joint and coordinated operational planning. Horne and Meyer (2010) report about 150 research teams worldwide whose focus is in this topic area. To conduct a DF study in a purposeful manner, procedure models have been developed. Examples are the "Data Farming Loop" by Horne and Meyer (2005) or the "Framework for Systematic Data Farming" by Choo et al. (2008). From the aforementioned explanations, a challenge for data generation using DF in the context of materials trading networks arises, since established approaches seem to provide an excellent basis, but are so far not directly transferable to this problem domain. This concerns in particular the suitability of the generated databases in order to be able to answer logistics tasks of decision-makers from SCM in a targeted and efficient way. Furthermore, Verification and Validation (V&V) is not yet an explicit part of the DF procedure models, but it is referred to in a number of the papers (cf. Sanchez 2021).

Before analyzing the result data, defining a suitable structure for the simulation result data is an important intermediate step. This has implications for both the process of modeling to create a simulation model and data modeling. For answering logistics tasks from materials trading networks, the complex relations between data are subjects of interest. The network structure of actors and their relationships can be interpreted as a graph. From a mathematical point of view, a graph consists of a finite set of nodes and a finite set of edges as well as a mapping rule from two nodes to one edge (Aigner 2015). This poses another challenge for the persistent storage of data generated by DF as a basis for decision support in SCM. Relational databases, which express data and their relationships in the form of tables, have dominated worldwide since their development in the 1970s. This type of database is increasingly reaching conceptual limits due to increasing volume and interconnections of data (Moniruzzaman and Hossain 2013; Störl 2015; Jose and Abraham 2017; Meier 2018). Evidence suggests that the mapping of relationships between data is complex and error-prone in relational databases (Fernandes and Bernardino 2018). This complicates the use of decision-support methods to analyze the database of a materials trading network. For example, a direct representation of relations based on data (e.g., clustering of sinks on transport routes between locations within a materials trading network) is not possible, so that often a procedure-specific data management or specific tools are necessary. To address this issue, non-relational databases, such as graph databases, offer a promising approach. In a graph database, the focus is on the relationships between data. Therefore, data and their relationships are stored in a graph database as nodes as well as edges and are, from a mathematical point of view, supported on a formal level by graph theory, in particular by the concept of labeled property graphs (Meier and Kaufmann 2019).

Using a graph database for storing simulation result data in the context of materials trading networks poses another challenge. For the evaluation of decision options and supporting decisions, providing knowledge to decision-makers of SCM is of major importance. An established process of discovering knowledge in databases in the context of SCs is Knowledge Discovery in Databases (KDD), which contains data mining as a central phase (Fayyad et al. 1996). Data mining describes the step in which previously unknown patterns can be extracted from the underlying data using specific algorithms (Fayyad et al. 1996). Today, although promising, the combination of DF and data mining has received less attention in the problem domain described. Furthermore, the structure and storage of the generated data have a direct impact on the analysis. A graph database enables the direct use of supporting methods from graph theory and algorithms on the database, such as performing a breadth-first or depth-first search (Robinson et al. 2015). A subset of data mining that seems promising in the domain of materials trading networks is the application of graph mining on simulation result data. Graph mining is a specific type of data mining on data stored in the form of a graph (Rehman et al. 2012).

To address these challenges, the primary goal of this dissertation is to develop a method for combining DF and data mining in an LAS for materials trading networks based on graph databases. Since the method is used by human deciders of a materials trading company, a software prototype for an LAS is presented to validate the method in a practical use case. The prototype is kept modular, enabling a flexible modification and exchange of the software components. The core elements of the method include a simulation-based approach for data generation as well as a subsequent process of knowledge discovery in simulation result data, allowing users to reap the benefits from both research domains. Furthermore, an approach for an accompanying modeling concept is needed. This modeling concept uses labeled property graphs, combining DF and KDD based on a consistent modeling concept and data storage for materials trading networks.

The primary goal of this dissertation is structured into secondary goals, which in turn are divided into several sub-goals. The first of the secondary goals is to develop a DF method element. Essential for the method element is a procedure model, which includes steps for conducting a DF study in materials trading networks in a structured manner. In order to develop a suitable DF procedure model, a requirements catalog is necessary, which must be available as the first sub-goal. Based on the requirements, established DF and simulation procedure models can be evaluated and two representatives are set as a reference as further sub-goals. On this basis, an adapted procedure model for DF in materials trading networks can be presented as a further sub-goal, which combines the advantages of the selected reference models. These advantages include suitable experiment designs for large-scale experiments as well as structured and efficient model development to create a simulation model. Another sub-goal includes the conceptual design of a result space for simulation result data, since these data form the basis for a subsequent analysis. A further sub-goal includes the integration of an accompanying V&V into the procedure model to ensure the necessary credibility of the achieved results. For the usability of the method element in practice, the last sub-goal is to provide a suitable simulation tool.

The second of the secondary goals is to integrate a method element for knowledge discovery in simulation result data. As with the DF method element, a KDD procedure model is relevant to structure the process of knowledge discovery. Therefore, the sub-goals from developing a DF method element can be transferred to the development of a knowledge discovery method element. These sub-goals include a requirements catalog, assessment of established KDD procedure models based on these requirements, and a selection of a model as a reference. Based on the selected KDD reference model, an adapted procedure model for knowledge discovery in simulation result data can be presented. Essential is that extensive data preprocessing for the application of mining methods is omitted, since the corresponding DF method element generates a suitable result space as an input for the knowledge discovery process. In addition, as in the case of the first goal,

further sub-goals are the integration of a target-oriented V&V and the selection of a suitable knowledge discovery tool.

The main basis for achieving the first two goals is the development of an accompanying graph-based concept for modeling and persistent data storage as the third of the secondary goals. To develop a concept, requirements from the domains of materials trading network, DF, and KDD are required, which have to be systematized as the first sub-goal. Based on this, a concept for modeling and persistent data storage based on labeled property graphs is presented as the following sub-goal. Thereupon, the effects of the modeling concept on the DF and KDD method element have to be outlined in particular, since a graph-based modeling concept ensures a consistent transition from DF to KDD, and specific graph mining methods can be applied directly to the simulation result data. Finally, the modeling concept and a suitable tool for persistent data storage are integrated as final sub-goals.

In order to achieve the complex interdependent cross-domain goals, it is necessary to define the research theory in order to support the present work in an engineering context. Decisive for this is the research area of information systems. According to March and Smith (1995) and Hevner and March (2003), this research area is divided into two essential paradigms: behavioral science and the design science approach. The former incorporates a social science view and the latter an engineering, science-technical view. The differentiation of the paradigms is still a subject of debate in the established literature (cf. Baskerville et al. 2011; Österle et al. 2011). A more in-depth discussion in this context can be found in Stahl (2009). Especially in German-speaking and Northern European countries, the term "Business Informatics" has become established in this context, which is located in the design science paradigm (Frank 2006). Therefore, the present work follows the established and widely used research methodology of design science research according to Peffers et al. (2007), whose structured process is given in Figure 1.1. The essential characteristic of the research method is the creation of so-called artifacts. These are artificially created and are differentiated into four elementary artifacts: constructs, models, methods, and implementation.

The research process according to Peffers et al. (2007) describes the process of artifact creation in a structured way via six steps and four possible entry points for research. This dissertation follows the problem-centered approach in design science research and, therefore, starts in Chapter 2 by examining the problem domain of materials trading networks. Hence, an introduction to logistics in general and logistics networks in particular is given, which enables the subsequent introduction of trading networks. In order to keep a complex socio-technical system such as a materials trading network in an efficient and competitive state, the concept of Key Performance Indicator (KPI) as part of a performance management is introduced. This is the basis for a discussion of different logistics tasks in materials trading. Since solving logistics tasks is a challenge for decision-makers of materials trading networks, data as an essential basis are needed. Subsequently, the terms

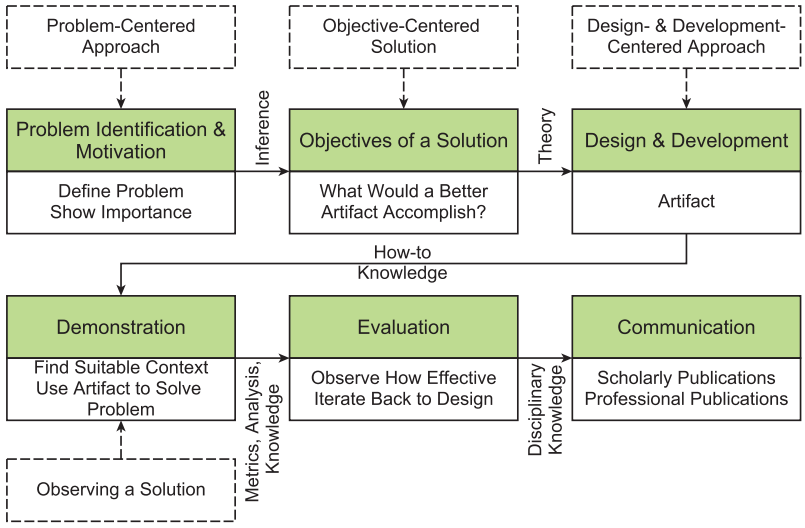

Figure 1.1: Design science research according to Peffers et al. (2007, p. 54)

data, information, and knowledge are classified and distinguished from each other. Based on this, specific information systems in materials trading networks, LAS, are introduced. This concludes in an analysis of data management in materials trading networks, since persistent data storage is an important part of an LAS. The discussion starts by introducing a Database System (DBS) in general and examines both relational and non-relational concepts of persistent data storage. The focus of the discussion is on graph DBS with an emphasis on DBS implementing a labeled property graph. Chapter 3 links the individual areas of the problem domain introduced in Chapter 2. The focus is on the examination of DF and KDD, which are vital parts of the farming-for-mining method developed in this dissertation. Therefore, an introduction to simulation-based data generation is given. A systematic examination of the fundamentals of simulation is followed by a review of the particularities of DF. An emphasis is on procedure models, both from the perspective of structure and descriptive means as well as the analysis of established procedure models in simulation and DF. Based on this, KDD and associated procedure models are introduced to analyze simulation result data. A focus is on discussing the central phase of data mining. Since data in a labeled property graph pose a challenge for data mining algorithms, specific graph mining techniques are introduced and subsequently linked to logistics tasks in materials trading networks. Concluding, a summary of previous research activities in the different domains is given, emphasizing limitations and research gaps. This is the basis to motivate the research approach. Lastly, a primary research question is

formulated to guide the course of research: How should a suitable method be designed that combines simulation-based data generation and knowledge discovery to purposeful support decision-makers in materials trading networks in solving logistics tasks? The primary research question is, furthermore, detailed by three consecutive research questions.

1. How can DF be used to generate data for solving logistics tasks in a materials trading network?
2. Which data structure is suitable to support DF and knowledge discovery to provide an appropriate basis for decision-making in answering logistics tasks from materials trading networks?
3. How can simulation result data be used as a basis for a knowledge discovery process in materials trading networks to provide decision-makers with purposeful knowledge?

To achieve the aforementioned goals and answer the research questions, Chapter 4 and Chapter 5 systematically describe the steps to develop a method to combine DF and KDD based on graph databases in materials trading networks. First, an overview of the methodical basis is given. This includes a specification of the selected research method and an overview of the developed main artifacts, which are the method and the LAS software prototype. Subsequently, core components of the method are introduced to elaborate the link between DF and KDD on the basis of graph DBS. In order to systematize the interrelationships methodically, a method element of initialization is first introduced. This is the basis to develop a specific procedure model for DF in materials trading networks. To this end, the requirements for a procedure model must first be derived so that a suitable model can be selected. Depending on the selected references, a specific procedure model for the method element of DF is developed and described in detail using suitable descriptive means. Furthermore, a suitable graph-based modeling concept is integrated, allowing end-to-end, consistent, and break-free model development and data modeling. Relevant are the integration of a suitable persistent storage of the input and result data in graph databases as well as a suitable simulation tool to avoid a break in the descriptive means. Here, the interactions of the core components of the developed LAS are considered and integrated. As a result of the developed DF procedure model, a result space is presented that contains a selected subset of simulation result data. The result space is used subsequently as an input for the method element of knowledge discovery. Thereupon, the process of knowledge discovery is first placed in the methodological context and connected to DF. The focus is on expanding the graph-based modeling concept to KDD and to describe the impact of purposeful generated input data. This has significant implications for the following adaptation of a suitable KDD procedure model. In order to integrate an adapted model, requirements must first be derived to be able to select an existing procedure model as a reference. To this end, the combination of DF and KDD allows for omitting the extensive data preprocessing. Thus, the

use of suitable mining techniques on the one hand and a suitable presentation of discovered patterns on the other hand have to be taken into account. To ensure the reliability of the results obtained, an accompanying V&V is integrated for both DF and KDD. At the end of the method development in this dissertation, the implemented software prototype is described.

To demonstrate the feasibility of the developed method in the industry, the prototypical implementation is used in a real-world case study in Chapter 6. After describing the general evaluation process and briefly introducing the use case, the method elements are structured into application fields. On this basis, the initialization, the DF, as well as the KDD are systematically applied using the software prototype in a farming-for-mining study. Finally, the findings are summarized and critically discussed.

2 Decision Support in Materials Trading Networks

Chapter 2 presents the state of the art of decision support in materials trading networks and the underlying IT-related technologies. First, the characteristics of such networks are analyzed (Section 2.1) with the indicators that measure their efficiency and the required tasks to operate them for providing customer satisfaction at reasonable costs. These tasks require adequate IT support, and especially effective storage and processing of the huge amounts of data that are generated for and by these tasks. Therefore, these data receive a classification (Section 2.2).

The application context of this dissertation includes an LAS that supports the involved managers' decisions by providing data, indicators, and action proposals. In order to be aware of the related work, decision support systems are discussed and existing LAS are presented with their specific architectures and major features (Section 2.3). Following the demand for the processing of big amounts of data as motivated above, the methods for managing these data in the specific context of materials trading networks are elaborated (Section 2.4). Here, a focus is set on the database structures to persistently store the data with low effort for their storage and retrieval. For this purpose, relational databases – which are by far the most-applied database structures today – are investigated, but also graph databases that can be expected to model the graph structures contained in trading networks in a more elegant and value-adding manner.

2.1 Fundamentals of Materials Trading Networks

This section introduces the terminology relevant to the problem domain. The terms of logistics, SC, and SCM are discussed in Section 2.1.1. Building on this, Section 2.1.2 classifies and delineates materials trading networks. Based on these two sections, Section 2.1.3 focuses on elaborating relevant key figures. The section concludes with a presentation and classification of logistics tasks from SCM in the context of materials trading networks.

2.1.1 Logistics und Supply Chains

After its original development for military operations in the 1950s, the definition of the term *logistics* is now largely determined by the practice of companies (Heiserich et al. 2011; Muchna et al. 2018). An overview of the literature shows that the concept of logistics is undergoing a strong change due to globalization and digitization. This has contributed to the development of a multitude of differ-

ent approaches to logistics and definitions of the term (Fleischmann 2008a; Pfohl 2016). The three primary approaches are referred to as flow-oriented, life-cycle-oriented, and service-oriented (Pfohl 2022). Predominant in science and practice is the flow-oriented approach, which focuses on a consideration of logistics based on flows (Pfohl 2016; Pfohl 2022). This understanding of the different flows is reflected in the "eight R's of logistics" introduced in Chapter 1. Table 2.1 provides a selection of existing definitions of the term.

Table 2.1: Definitions of the term logistics

Reference	Definition
Morgenstern (1955, p. 130)	"A logistic operation consists in the supply of definite quantities of physical means and services for activities that according to their missions consume these means and services in order that the activities be maintained at particular present or expected future rates."
Ghiani et al. (2013, p. 1)	"Logistics [...] is the discipline that studies the functional activities determining the flow of materials (and of the relative information) in a company, from their origin at the suppliers up to delivery of the finished products to the customers and to the post-sales service."
Göpfert (2013)	Logistics is a modern management concept for the development, design, control, and realization of effective and efficient flows of objects (goods, information, money, and financial flows) in company-wide and cross-company value creation systems.
Christopher (2016, p. 2)	"Logistics is the process of strategically managing the procurement, movement and storage of materials, parts and finished inventory (and the related information flows) through the organisation and its marketing channels in such a way that current and future profitability are maximised through cost-effective fulfillment of orders."
Zsifkovits (2018)	Logistics is the integrated planning, design, handling, and control of the entire flow of materials from the supplier into the company, within the company, from the company to the customer, the return of goods in a cycle, and the information flows required to control the flow of materials.
Visser and van Goor (2019, p. 22)	"Logistics entails the organization, planning, control and execution of the flow of goods from development and purchasing, through manufacturing and distribution to the consumer (end user), up to and including the reverse flow. The aim is to meet market demand at lowest cost and best use of capital, and build long-term relationships with customers."