

Using Natural Language Processing for Supply Chain Mapping: A Systematic Review of Current Approaches

Henning Schöpper and Wolfgang Kersten

Hamburg University of Technology, Am Schwarzenberg-Campus 4, Hamburg, Germany

Abstract

Purpose: The COVID-19 crisis has shown that the global supply chains are not as resilient as expected. First investigations indicate that the main contributing factor is a lack of visibility into the supply chain's lower tiers. Simultaneously, the willingness to share data in the supply chain is low as companies mainly consider their data as proprietary. However, large amounts of data are available on the internet. The amount of this data is steadily increasing; however, the problem remains, that this data is hardly structured. Therefore, this paper investigates current approaches to use this data for supply chain transparency and derives further research directions.

Methodology: The paper uses a systematic review of the literature followed by content analysis. The research process further follows established frameworks in the literature and is subdivided into distinct stages.

Findings: Descriptive and clustering results show a fragmented research field, where current approaches disconnect from prior research. We classify the methods using a simple taxonomy and show developments from rule-based to supervised techniques and horizontal to vertical mining approaches. The techniques with rule-based-matching procedures mainly suffer from low recall. The current approaches do not satisfy yet essential requirements on supply chain mapping based on natural language.

Originality: To the best of the authors' knowledge, no prior research has been attempted to review textual data usage for supply chain mapping. Therefore, this paper's main contribution is to fill this gap and add further evidence to the use of data-driven supply chain management methods.

Keywords 1

Natural Language Processing, Supply Chain Mapping, Systematic Literature Review, Supply Chain Management

1. Introduction

The COVID 19 pandemic has been a great challenge for supply chains. This situation can be described aptly by using one product: Toilet paper. According to a recent study, many customers regularly faced empty shelves during the pandemic. The central issue [1] identified was not hoarding purchases but rather a lack of responsiveness in institutional supply chains, which proved incompatible with private value chains. However, the disruptions caused by the pandemic were not limited to individual products but severely disrupted whole industries such as the food supply chain [2, 3]. Even many large and globally-oriented companies ran into trouble. Fiat, for example, had to shut down temporarily one plant in Serbia at the beginning of February because procurement could not get essential parts from China. Simultaneously, many other original equipment manufacturers (OEM) in the automotive area reported dangerously low inventory levels from different production sites [4].

Global supply chains did not respond well to the pandemic's effects, but the question arises why the effects were so drastic. To take appropriate and effective measures against supply risks, information

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine

EMAIL: henning.schoepper@tuhh.de (H. Schöpper); logu@tuhh.de (W. Kersten)

ORCID: 0000-0002-3629-3648 (H. Schöpper); 0000-0003-0894-4121 (W. Kersten)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

about the supply chain needs to be present. Studies indicate, however, that companies often have a severe lack of knowledge about their value chains' deeper tiers. According to a *Resilinc* field study, 70 % of companies were still manually investigating whether they had indirect suppliers in China's affected region at the end of January 2020 [5], and even multinationals are often not aware of their dependency on inputs from Asia [6]. A very recent study about the pandemic's impact showed that about half of German logistics service providers have poor or very poor knowledge about their supplier's supplier [7]. A 2019 survey demonstrated that only 8 % of companies in the apparel industry could trace their products back to their origin [8], and *The Sustainability Consortium* argues that more than 80 % of consumer goods manufacturers have no or only minimal information about their suppliers' sustainability activities [9].

With the lack of supply chain information, further threats for companies arise besides the management of challenges caused by the pandemic. An increasing number of end-customers want to know where their products come from and value insights into the value chain's social practices [10]. Violations of social or environmental standards increase reputational risks as customers hold companies responsible also for their supply chains [11]. Following this development, companies face growing regulatory pressure to act strictly according to corresponding standards [12], of which the planned *Supply Chain Act* in Germany is one recent example [13].

Although drivers towards supply chain information are strongly present, the gap to industry practice remains. The issue stems from multiple barriers: On the one hand, companies mainly consider information about their supply chain as proprietary and are therefore cautious about sharing it with others [6]. On the other hand, gaining and exploring the supply chain data can also be quite challenging because value chains have become increasingly global and complex. Collecting the data is also not sufficient alone by itself. Studies generally recommend an approach known as *supply chain mapping* to aggregate the supply chain data and inform strategic decision-making [14]. For example, one recent study suggested using mapping to prepare for future pandemics [1]. Following this thesis and for the same reasons mentioned above, obtaining the mapping procedure data remains the most pressing issue [15].

While direct supply chain data from companies are scarce, vast amounts of data are available on the internet and the world wide web today. A significant part of this data contains essential information about the supply chains and competitive position of companies. However, this information is primarily present in sources, such as web pages, news articles, or social media posts, written in natural language. The extraction of supply chain information from these sources is challenging; however, recent advances in natural language processing (NLP) and machine learning appear to create causes for hope. To the best of the author's knowledge, no attempt is present, systematically identifying and analyzing the approaches for extracting supply chain information from natural language texts. Therefore, the article's target is to determine appropriate methods, compare them accordingly, and provide further research suggestions. Thus, to address the problem, we formulate the following three research questions, which we address in this publication:

- **RQ1:** What approaches exist for extracting supply chain information based on natural language text?
- **RQ2:** How can the identified approaches be compared and evaluated?
- **RQ3:** What are the limitations of the identified approaches, and what further research is needed?

2. Related work

We very briefly introduce and further elaborate on essential concepts in the context of supply chains. These concepts will help the reader as a groundwork when reading the later stages of the review.

2.1. Supply Chains

Supply chains consist of companies and the linkages between them [16]. From a focal company's perspective, these relations can be either *upstream* on the supply side or *downstream* on the demand side [17]. From the members and links between them emerges a supply network, which has two

structural dimensions. The *horizontal* dimension refers to the number of tiers across the whole supply chain, whereas the *vertical* dimension refers to the number of suppliers/customers at each level [16]. *Supply Chain Management* (SCM) can be defined as the “integration of key business processes from end user through original suppliers that provides products, services, and information that add value for customers and other stakeholders” [16].

The global supply chains were subject to significant structural changes in the past centuries. Cost savings potentials of 50 % and more, accompanied by digital innovations, have led to intensive offshoring of production capacities to low-wage countries [18, 19]. This led to globally distributed value chains. Concurrent a continuing trend toward greater modularization of product structures served to manage increasing product complexity and realize further cost-saving potential [20, 21]. As a result, many Original Equipment Manufacturers (OEMs) significantly reduced the number of direct suppliers, which outsourced value creation by themselves, making the supply chains more complicated in the process [22, 23].

The pandemic impressively questioned the *resilience* of global supply chains. [24] define *Supply Chain Resilience* as the capability to “react to and recover from a disruptive event, and to regain performance by absorbing negative impacts”. Regarding the pandemic, a recent study found that risk management in supply chains (SCRM) significantly improved supply chain resilience [25] as appropriate SCRM increases responsiveness by identifying risks on time, assessing and controlling them on an ongoing basis [26, 27]. However, the lack of transparency in supply chains is one primary barrier to the successful implementation of SCRM [28].

2.2. Supply Chain Visibility

In the academic literature, sharing and access to information in supply chains are primarily discussed under two related terms: *visibility* and *transparency*. Having *visibility* into the supply chain means for companies having “access to timely, high-quality, and useful supply chain information” [29]. However, the term is not consistently defined. In contrast, *supply chain transparency* is commonly referred to as disclosing information to others (i. e., suppliers' names) [30]. Visibility and transparency are often used interchangeably. By taking into account the above definitions, we offer a simple distinction: one company's transparency is another company's visibility and represent two sides of the same coin. However, transparency and visibility levels are generally low in today's supply chains. One important aspect is that companies see their supply chain information as proprietary and are unwilling to share it. Mainly this is because companies fear losing bargaining power or the danger of being cut out [31].

One solution is using track and trace technologies, sometimes with advanced concepts, such as the blockchain, to address trust issues [32]. However, this approach is often limited to supply chains with a substantial power gap between one actor and the upstream supply chain, through which this actor can force the supply chain to participate [33]. Furthermore, track and tracing products through material transformation remains challenging. Additionally, track and trace technologies are limited to specific parts of the supply chain companies already know, but [34] proved that critical *nexus suppliers* could appear anywhere in a supply network, potentially leading to severe effects [35].

2.3. Natural Language Processing

The research field of NLP has been a very active subfield of artificial intelligence (AI) for the past years. The role of NLP within AI-systems is often to derive structured information from semi-structured sources. With new deep neural network architectures, many simple NLP tasks significantly improved in performance. This development will further enable the technical systems to reach the ultimate goal of fully *understanding* human language. [36] consequently define NLP as “computer systems that analyze, attempt to understand, or produce one or more human languages”. To explain what happens in an NLP-system, [37] suggested taking a bottom-up approach to the different human language levels (c. p. Figure 1). *Morphology* is the first stage in textual analysis and refers to the study of various word forms. One common application in NLP is to normalize the text's words to their stem or root forms. *Syntax*, on the other hand, looks at the structure of sentences and how these are formed. One regular use would be to identify the subject, verb, or objects in a given sentence. Lastly, *Semantics* is concerned

with the meaning of words and sentences. A typical application is to identify specific noun phrases, such as Persons, Organizations, which is also referred to as Named-Entity-Recognition (NER) and is a subtask of *information extraction* (IE). Lastly, identifying semantic relations between words, such as supplier relations between companies, is a subtask of *relation extraction* (RE).

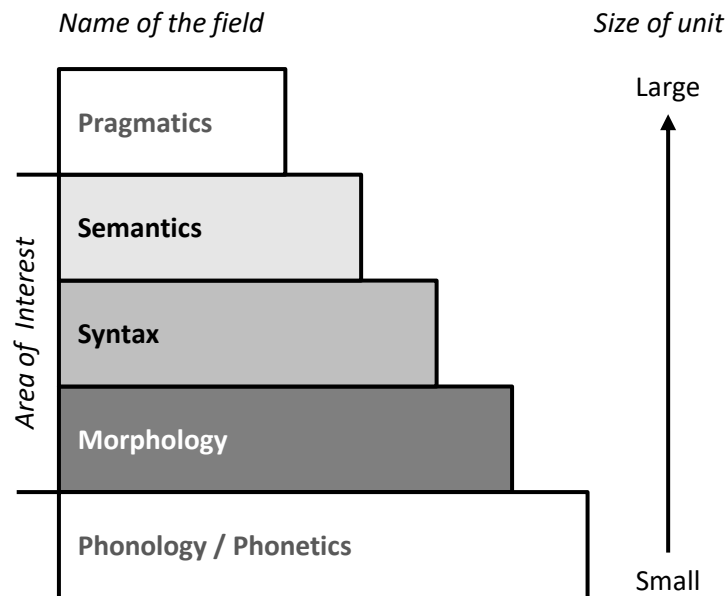


Figure 1: Bottom-up approach of linguistic analysis (Adapted from [37])

2.4. Supply Chain Mapping

Supply Chain Mapping “focuses on how goods, information, and money flow in both the upstream and downstream directions and through a firm” [14]. While the purpose of supply chain maps is strategic, they differentiate from other process mapping approaches by a higher level of detail and the extension from an intra-company to an inter-company perspective. There is no accepted mapping convention; however, [14] described feasible characteristics of a supply chain map, such as the number of tiers, the direction, length of the supply chain, and spatial representation. [15] further elaborates the concept and suggests that mapping needs to pragmatically balance complexity and information needs, and conclude that the supply chain map should focus on tier one to three. Additionally, [15] introduces the idea that a supply chain map should include information besides the network’s suppliers and customers. They highlight the identification of suppliers’ customers, which could be potential competitors or the customer’s supplier, leading to potential alliance opportunities. Following up this line of thought [38] suggests a new structural supply chain mapping model. This model distinguishes between a *vertical* mapping dimension, which contains supplier-customer relations, and a *horizontal* mapping dimension, including competitors and complementors. These mapping dimensions should not be confused with the classical supply chain dimensions (c. p. chapter 2.1). In this article, we adopt the model from [38] to compare the identified mapping approaches using NLP. Figure 2 displays a simplified version of this structural model (limited to one tier).

[39] suggest *essential* requirements and challenges for an NLP-based supply chain mapping solution (c. p. Table 1). They recommend an approach to cover a supply relationship’s respective direction. As a company usually supplies another company only for particular products or services, specific relations need to be included. Supply relations change in the course of time, which means that companies are forming new links and dropping others. The transitivity requirement reflects whether the method can draw valid inferences over multiple supply relationships for one specific end-product. Moreover, it is important to take into account companies having multiple roles in the network simultaneously. Finally, the approach should provide the results on different aggregation levels (i. e., company, product- or industry-specific).

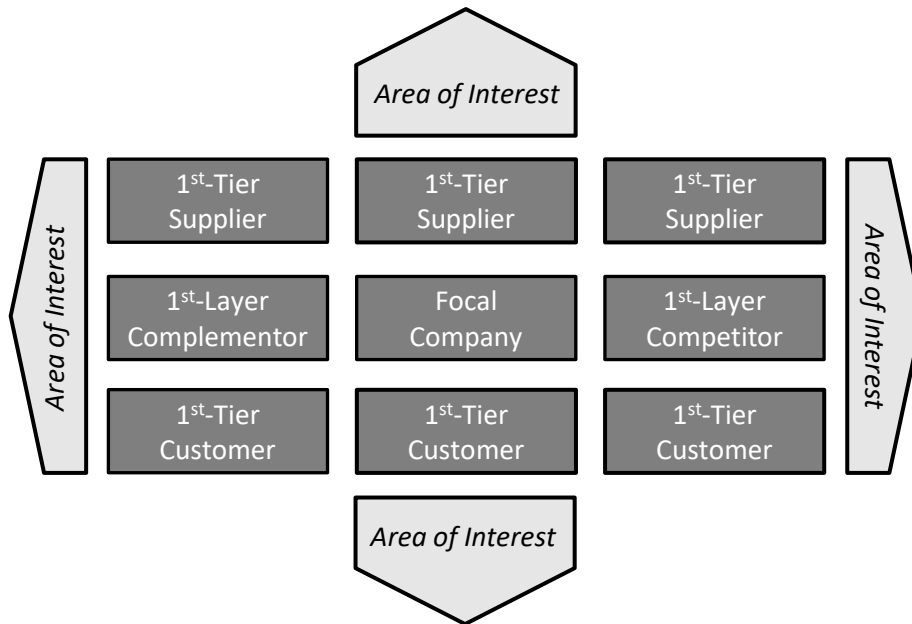


Figure 2: Structural Supply Chain Mapping model (adapted from [38])

Furthermore, [39] propose four primary challenges related to the data input. First, the relevant supply chain information is most likely available in different languages. Information written in natural language will almost certainly also contain wrong or ambiguous information. Therefore, an approach needs to take into account instruments for ensuring and accessing information quality. Data is also always limited in two important ways: On the one hand, the data containing relevant information might be scarce, depending on the data type. On the other hand, the amount of positive data available for generating either heuristics or training classifiers is also limited due to limited manual capacity. Lastly, the low-recall-problem is particularly challenging for information extraction in NLP-solutions. Mainly, this is because natural language is complex and multi-faceted. Overlooking some potentially critical supply relations, especially if this information presents itself in the data, is strongly undesirable from a risk perspective.

Table 1

Requirements and challenges for an NLP-based Mapping solution [39]

Supply Chain Requirements	Data Challenges
Directed relations	Multiple languages
Product/ service specific relations	Imperfect / Ambiguous information
Dynamics of relations	Limited data availability
Transitivity of relations	Low Recall
Multiple company roles	
Location	
Different aggregation levels	

3. Methodology

The present study adopts a systematic literature review approach to address the research questions. We used the methodological framework based on the concept initially outlined by [40] for our review and considered the recommendations for further improvement of the methodology made by [41]. Lastly, we took into account practical considerations as suggested by [42].

The study followed five steps:

1. Identification of the Research Questions / Scope of the Inquiry
2. Identification of Relevant Studies
3. Selection of Studies
4. Charting the Data
5. Collating, Summarizing, and Reporting the Results

First, we assembled a study team with context expertise, including participants with experience in SCM and NLP and performing systematic literature reviews. We discussed the research questions and scope of the inquiry and removed potential ambiguities. We found no single database covering all existing relevant scientific literature, so we combined different sources to form a more robust approach. We chose the two largest meta-databases, SCOPUS and Web of Science (WoS). These cover in conjunction a wide range of scientific literature. Nevertheless, some limitations in terms of over-and under-representation of countries and languages still exist [43]. We added a search strategy via Google Scholar to address limitations because it provides an alternative page rank retrieval strategy to identify relevant studies. Additionally, we implemented a forward- and backward search strategy for all relevant publications later in the review. Figure 3 shows the utilized databases and the search terms with the logical link between them.

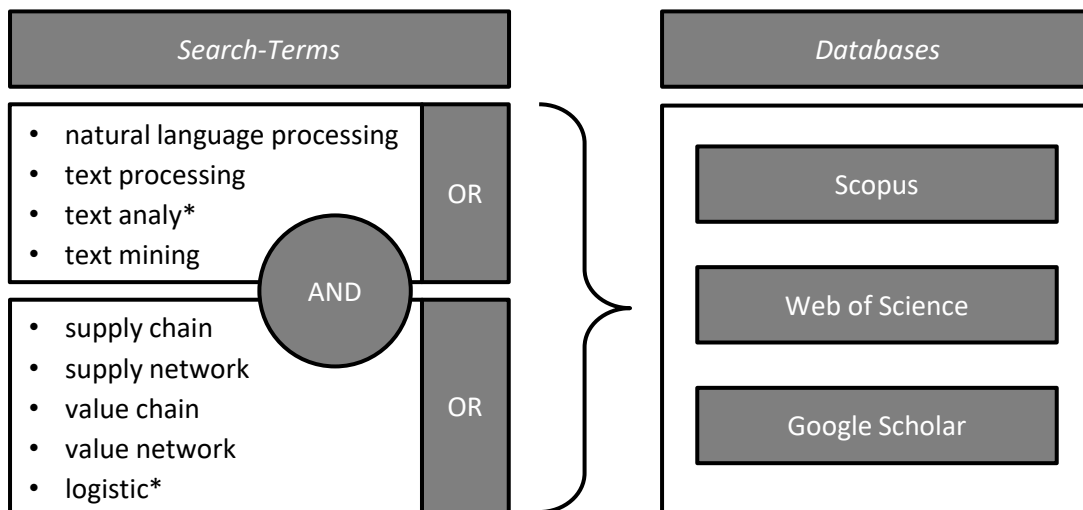


Figure 3: Database choices and logical link between search terms

We excluded the term *logistic regression* from the review because it caused many false-positive results in the search query. We carried out the initial search on August 05, 2020, with no further constraints on the time horizon. We searched Google Scholar using only the first 100 hits, which were prior sorted by relevance and imported all results into Citavi 6 for further analysis. We checked for duplications based on Title, Year, and Authors and removed entries if all three criteria matched.

Following the duplication check, we performed a two-stage screening process beginning with a Titel-Abstract-Screening. We excluded studies with no apparent connection to SCM or NLP. Also, we excluded studies in languages other than English due to limited translation capacity. We removed all literature reviews, which used NLP as a clustering method when these studies did not focus on the extraction of supply relations. In vague cases, we included the publication for further analysis. We computed inter-rater reliability and found Cohens Cappa to be 0.7, which we interpreted as substantial agreement [44]. Figure 4 displays detailed information on the selection and screening process. We coded all records in Citavi 6 and exported the data to Microsoft Excel for further analysis.

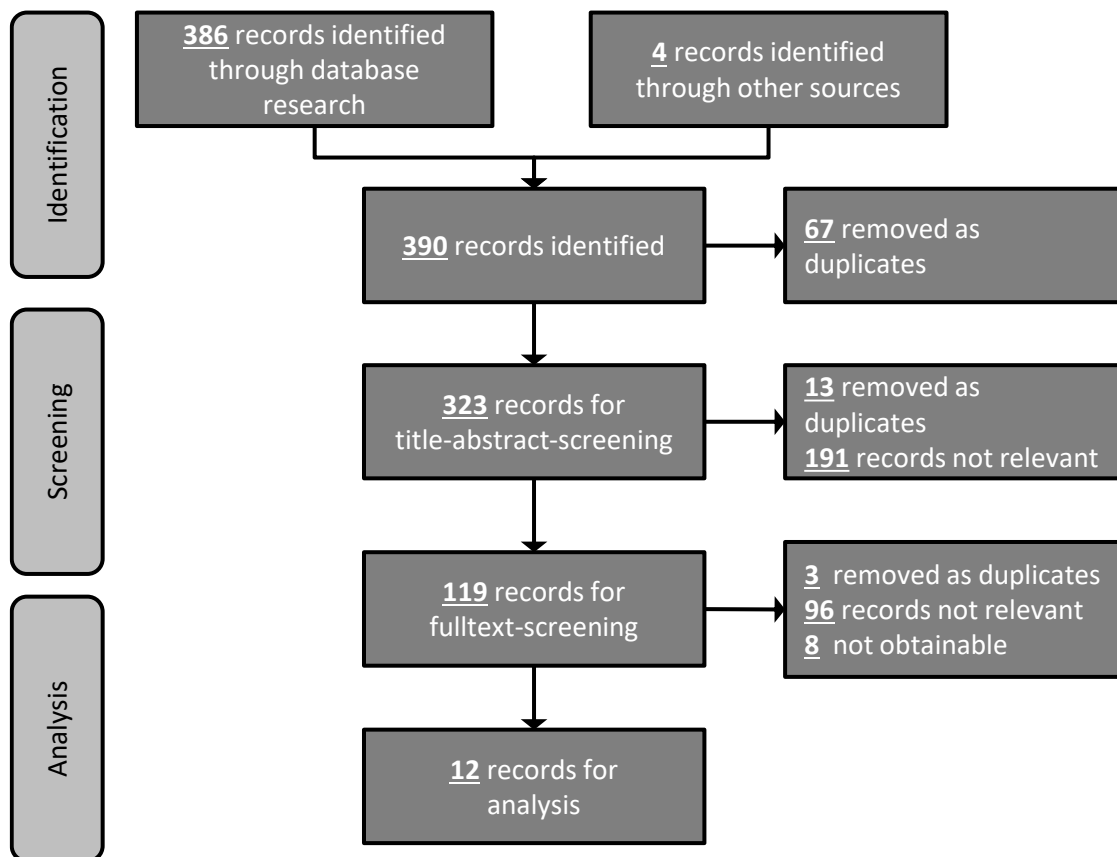


Figure 4: PRISMA Flowchart of the study selection process [45]

4. Results

Table 2 shows the list of the twelve relevant identified publications. Descriptive analysis reveals that two-thirds of the publications are articles published at conferences, emphasizing the field's prematurity. On the timescale, it appears that two periods of primary research interest exist, one around 2010, and another started a few years later.

4.1. Cluster Analysis

We decided to perform a deeper cluster analysis from the observation of two visually identifiable clusters on the timescale. We evaluated the bibliographic coupling [57] and obtained the cluster by visually analyzing the resulting network. This method treats two documents as much as linked as both cite the same publications. Accordingly, this method is backward-looking and therefore fixed because future publications do not change the results. We visualized the resulting network using Gephi, an open-source graph visualization tool [58]. Figure 5 shows the results of the cluster analysis. The coupling analysis reveals and confirms the research field's strong fragmentation: First, it shows the publications separate into two clusters, in which they refer to different literature as knowledge bases. Second, the clusters are only weakly coupled themselves except for publications from the same authors. The publications in *cluster A* center around 2010 and are somewhat connected. Two articles that appear central in this cluster have the most interconnections with other publications [51, 53]. Thematically, publications in cluster A are mainly concerned with the extraction of horizontal supply chain mapping items, specifically competitors. With one exception [46], the publications in cluster B are more recent. In contrast, this newer research appears only very weakly coupled around two publications [39, 56]. Publications in cluster B contrast to A, are more concerned with the extraction of vertical supply chain mapping items.

Table 2

List of publications for NLP-based supply chain mapping

Source	Authors	Year	Title	Type
[46]	Bao et al.	2008	<i>Competitor Mining with the Web</i>	Journal
[47]	Pant & Sheng	2009	<i>Avoiding Blind Spots: Competitor Identification Using Web Text and Linkage Structure</i>	Conference
[48]	Ma et al.	2009	<i>Discovering Company Revenue Relations from News: A Network Approach</i>	Journal
[49]	Xia et al.	2010	<i>Mining Commercial Networks from Online Financial News</i>	Conference
[50]	Zhao et al.	2010	<i>Business Relations in the Web: Semantics and a Case Study</i>	Conference
[51]	Ma et al.	2011	<i>Mining Competitor Relationships from Online News: A Network-Based Approach</i>	Journal
[52]	Lau & Zhang	2011	<i>Semi-supervised Statistical Inference for Business Entities Extraction and Business Relations Discovery</i>	Conference
[53]	Wei et al.	2013	<i>Mining Suppliers from Online News Documents</i>	Conference
[54]	Yamamoto et al.	2017	<i>Company relation Extraction from Web News Articles for Analyzing Industry Structure</i>	Conference
[55]	Carstens et al.	2017	<i>Modeling Company Risk and Importance in Supply Graphs</i>	Conference
[39]	Wichmann et al.	2018	<i>Towards Automatically Generating Supply Chain Maps from Natural Language Text</i>	Conference
[56]	Wichmann et al.	2020	<i>Extracting Supply Chain Maps from News Articles Using Deep Neural Networks</i>	Journal

4.2. Content Analysis

As mentioned above, [38] suggested differentiating between horizontal and vertical mapping. Because the approaches can be broadly assigned to either of these classes, we use this distinction to structure the following section. All identified works are conceptual and have an additional case study part, which verifies the concepts. Therefore, we briefly summarize the appropriate methods for each mapping dimension and recognize the work's limitations for each publication.

4.2.1. Horizontal Mining Approaches

[46] provide an approach called "CoMiner". Originating from a focal company, the method extracts and ranks potential competitors. As the first step, the algorithm uses pre-defined linguistic patterns to construct queries sent to a search engine (in this case: Google API) and then extracts competitor names from the received candidate webpages using the same linguistic patterns mentioned above. The CoMiner utilizes metrics as the competitor names quantity of appearances on the webpages to rank the identified competitors. In addition to competitor names, the CoMiner extracts competitor domains by querying the search engine with the focal company's name and potential competitors and searching for common noun phrases as domain candidates. The authors validate the algorithm with a manual annotated gold standard resulting in a precision of 83.3 % and a recall of 54.6 %. The significant limitations of CoMiner are twofold: First, using pre-defined rules for candidate identification is generally knowledge-intensive. However, human language is complex and fixed rules can hardly cover all cases, leading to a low recall. Second, the algorithm largely relies on third-party search engines and the retrieved webpages' quality. If the web pages do not contain competitor relationships, the method introduces noisy data into the results.

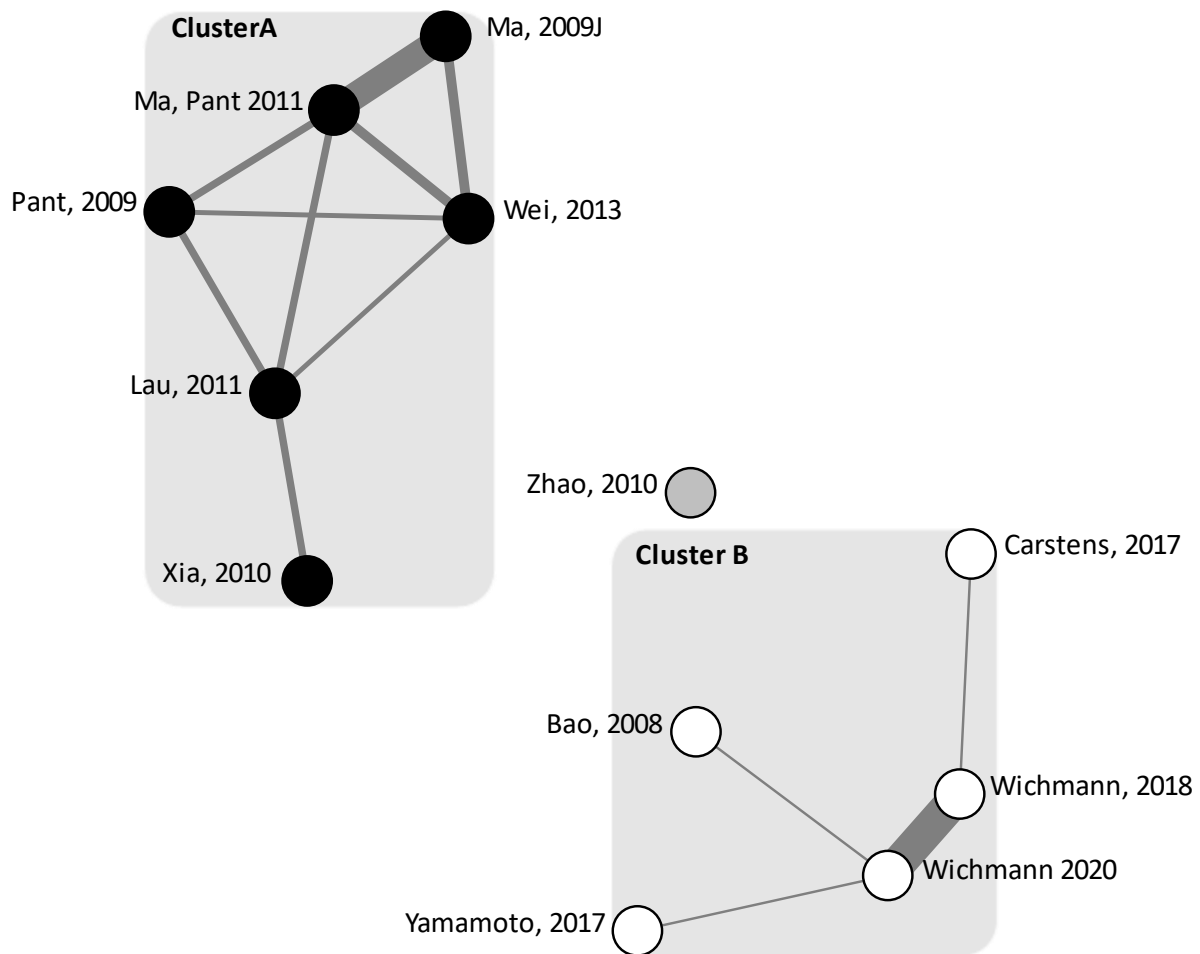


Figure 5: Literature clustering based on bibliographic coupling

In contrast to the content-related approach mentioned above, [47] develop a link-based competitor identification approach, which follows the assumption, that companies and their competitors are more likely to be linked in their web sources. The authors use companies from the Russel 3000 index and extract the companies' webpages from Yahoo Finance. The method first extracts three metrics of similarity based on link structure. Links to (In-links) the identified companies on another page may indicate demand-side substitutability and, therefore, potentially a competitive relation. The authors obtain this metric utilizing a third-party API. Links from (Out-links) different of the companies' webpages to a third company's webpage may indicate that both firms provide similar products or services and therefore indicate demand-side substitutability. The authors get the metric by utilizing a web crawler for each of the identified companies' websites. Last, the authors use the similarity between companies websites' content as a metric because the company's website offers a description of the products and services and can again indicate demand-side substitutability. Also, the authors use an external API to obtain a gold standard for the identified companies from the Russel 3000 index. The authors create different supervised learning models from this training data for predicting competitor relationships between the recognized companies. Evaluation of the best performing algorithm (Decision Tree) shows Precision of 0.79 % and Recall of 0.71 %. Although this approach uses a supervised learning approach to circumvent one major limitation of [46] in the form of rule-based identification, this approach suffers from several shortcomings. First, the method is limited to public-traded companies in the Russel 3000 index and can only detect competitors in the same index. Next, the gold standard (competitor companies) used for training rests solemnly on an external API. Consequently, the authors provide no indicator of the internal validity of the training data. Finally, the chosen metric may contain noisy data, as links between websites may also have other than competitive relations such as supplier-buyer relationships.

Another competitor mining approach is presented by [51]. This work supersedes and sums up some of the authors' prior work [48, 59]. The authors base the method on assuming that parallel citations of

companies in news documents indicate business relationships and that structural information in the resulting network can conclude a competitive relationship. The authors retrieve business news articles from Yahoo! Finance and rely on the fact that Yahoo ordered each news document to a specific company. Also, the authors use the fact that Yahoo called each mentioned company to its stock ticker, which simplifies the identification of companies in the text. The algorithm then leverages a third-party API to generate a gold standard of competitors for each identified company and uses an artificial neural network for training and evaluation. The reported precision of the method is 26.8 %, and recall is 22 %. The work uses news documents combined with a supervised learning approach; however, this work suffers from significant limitations. First, the proposed method is heavily dependent on the quality and availability of third-party APIs such as Yahoo Finance, which may change at any time, leaving the technique ineffective. Second, the approach does not extend to other news documents from different sources, which do not tag the companies explicitly in some form. Third and most substantial, the structural network attributed seems to have low predictive power for competitive relationships. The work's basic assumption that parallel mentions in news documents necessarily imply business relationships is very strong and might, in some cases, introduce noisy data.

[49] introduce the “CoNet” system, which employs simple lexical rules to identify business entities and their relations entirely based on the news articles' content. The authors collect news articles from Google Finance and utilize NLP techniques to identify “commercial entities” and their competitive and cooperative relationships. The authors evaluated the method by a subsample of 600 news articles, which they annotated manually. Entity tagging reaches a precision of 81 % and a recall of 61 %, while relationship extraction has a precision of 92 % and a recall of 67 %. [49] present the earliest work based entirely on news articles' content for entity recognition and relation extraction. Furthermore, the horizontal business relation is extended from competitive business relations to cooperative relations also. However, the work suffers primarily from low recall, which is a general issue of rule-based approaches, as described above. A similar approach is presented by [52], which extracts business entities and their respective competitive and cooperative relationships entirely content-based. The authors extend the basic NLP techniques by statistical considerations, which significantly improves the Business Entity Identification's performance. A more recent paper by [54] also attempts to extract company-to-company relations from the content of text focusing on cooperative and competitive business relations. In contrast to the approaches above, the authors use distant supervision for relation extraction, a combination of supervised learning with manual labels. The authors only report precision, which is 67 % for cooperative and 81 % for competitive relations. Although the authors use a supervised technique to address the restriction of rule-based approaches, the work suffers from one main limitation. Using manually crafted training sentences for training (distant training) is knowledge-intensive in the same way rule-based systems lead to a potentially low recall.

Summary:

The analysis of horizontal mining approaches shows two broad categories: content-based and link-based procedures. Link-based methods have solely utilized supervised learning techniques. Most content-based systems focused on rule-based approaches; however, the study shows a tendency towards data-driven strategies in more recent years and a more detailed breakdown into further categories other than pure competitive relations.

4.2.2. Vertical Mining Approaches

[53] present the first known work that solely focuses on the extraction of supply (and customer) relationships. The method rests on the basic assumption that news documents contain essential information about supply and customer relations between companies. The authors collect the source documents from Reuters and restrict company names to the Financial Times Global 500 in 2011. The researcher subdivides their proposed method into three steps. First, the algorithm divides the documents into sentences by using shallow NLP techniques and then classified the sentences into either containing a supply relationship or not. In the following classification step, the approach classifies the direction of the known relation as either supply or customer side. Both classification tasks employ a supervised learning approach. The authors generate this training data by manual annotation. In this regard,

annotating only those sentences, which contain at least two or more companies, provides a notable reduction of possible candidate sentences. Finally, 93 tagged sentences served as positive training data and as the gold standard. Afterward, the method uses an additional classification step, whether the relationship is true or not. The evaluation shows precision to be 46 %, recall 66 % for the first and precision of 84 %, and recall 56 % for the second classification step. The first substantial limitation in this work is the relatively low amount of training data, potentially leading to low precision or recall. Second and even more significant, the authors omit a company identification step in their work. Consequential company names need to be known beforehand, which is knowledge-intensive. Furthermore, this limits the method's explorative potential as it cannot detect previously unknown companies. Third, the last classification step's role appears cosmetic and unneeded if the first two steps' accuracy would already deliver acceptable performance.

Similarly, [55] employs a content-based supervised learning technique to extract a supply chain graph from news articles or filings from the Security and Exchange Commission (SEC). In contrast to [53], the method incorporates a company identification step provided by a third-party API. The algorithm then identifies candidate sentences with two or more companies; however, the technique includes additional linguistic rules to reduce candidate sentences further. The authors use *Mechanical Turk* to generate training data for a logistic regression model. The evaluation shows a precision of 76 % and recall of 46 %. Although the approach addresses some of the limitations of [53], the solutions offer restrictions themselves. The authors rely on the performance of a third-party API to identify companies. A poor-performing API can potentially increase the workload for annotation to reach a critical amount of training data and lead to a systematic failure to recognize individual companies (low recall). The integration of linguistic rules in the prefiltering of candidate sentences for annotation is knowledge-intensive and makes the method even more vulnerable to low-recall. Lastly, Mechanical Turk's use is potentially costly and requires additional verification steps, which the authors do not propose.

[39, 56] presents a substantial work referring to vertical supply chain mining. In a first and more conceptual work [39], derive requirements that a vertical mining approach should satisfy. The authors chose Toyotas supply chain as a case study and extracted the business entities and their relations based on simple pre-defined lexico-syntactic rules. The researcher used a private automotive industry database as the gold standard. The evaluation did not report a recall but showed a precision of 67 %.

In [56], the authors present a more sophisticated approach addressing many of the limitations above. The authors formalize the problem of detecting supply chain relations in a textual document via a simple two-step approach: First, the method detects business entities in the text if three well-reported open source libraries for NER-tagging show agreement. Second, the system classifies the relationship between these entities via a multi-class supervised learning approach. The researcher generates a corpus by randomly drawing from several publicly available news corpora to train the learning algorithm. The random drawing has the advantage that it potentially increases the method's generalizability to previously unseen data. Seven independent human annotators labeled the sentences, and subsequent, the authors report promising validity results on the inter-and intra-rater agreement. However, because the number of labeled sentences from annotation is still low, the authors incorporate additional labeled data from other sources (distant supervision). The authors used a BiLSTM deep neural network as a state-of-the-art algorithm with word embeddings obtained from the GloVe dataset. The authors used the ground truth of the annotated corpus for evaluation. The best performing algorithm's precision ranged between 33 % and 85 %, the recall between 22 % and 85 %.

Finally, the authors propose a simple method to integrate results into a “supply chain map” by merging the relationships and company names from different sentences. Although the approach presented by [56] addresses many of the limitations of prior works, especially regarding the corpus generation, the work suffers from substantial limitations. First, the introduction of additional labeled data (distant supervision) may lead to “so called ‘overfitting’ and result in false positives if the classifier is applied to previously unseen data” [56, p. 8]. Second, the combination of different NER techniques to detect company names may introduce additional noise in the model, which the current evaluation does not (even) cover. Third, the authors use a complex single-step but six-class classification schema instead of multi-step but less complicated classification steps. Using many classes can be problematic when the amount of training data is low. However, multi-class classification generally requires an even more tremendous amount of labeled data to be of similar accuracy. Moreover, [53] contrasts this choice because they show that a multi-step but binary classification procedure delivers adequate performance.

Last [50] presents the only work known to the authors that address horizontal and vertical relationships in one approach. Founding on prior work [60], the researcher developed an ontology that contains competitive, cooperative, supply, and sale relationships; however, the researcher presents no case study of a resulting relationship mining approach.

Summary:

The scientific work on the mining of vertical business relationships is of a more recent nature. All approaches we observe in the literature are content-based, and the majority use supervised learning techniques to identify business relations. When compared with the essential Supply Chain Mapping requirements (c. p. section 2.4), no approaches fulfill even the basic requirements. Furthermore, the techniques show a surprising homogeneity among each other (c. p. Table 3). However, we observe significant differences in the methods and showed that they contain considerable limitations.

Table 3
Match with supply chain mapping requirements

Requirements	[53]	[55]	[39]	[56]
Directed relations	yes	yes	yes	yes
Product/service-specific relations	no	no	no	no
Dynamic relations	partly*	partly*	partly*	partly*
End-product / Transitivity of relations	no	no	no	no
Multiple company roles	no	no	no	no
Location	no	no	no	no
Different aggregation levels	one	one	one	one

*no dropout of outdated relations

Table 4
Taxonomy of Supply Chain Mining Approaches

Source	Year	Supply Chain Mapping		Relation reference		Relation identification	
		Horizontal	Vertical	Content	Link	Rule	Supervised
[46]	2008	x		x		x	
[47]	2009	x			x		x
[48]	2009				x		
[49]	2010	x		x		x	
[50]	2010	x	x				
[51]	2011	x			x		x
[52]	2011	x		x		x	
[53]	2013		x	x			x
[54]	2017	x		x			x
[55]	2017		x	x			x
[39]	2018		x	x		x	
[56]	2020		x	x			x
Σ		7	5	8	3	4	6

4.3. Taxonomy

From the content analysis, we finally propose a simple taxonomy for NLP-based supply chain mapping approaches, which may be either horizontal, vertical, or both, base the relation on content or links, and identify the relation by either rule-based or supervised classification. Table 4 displays the identified literature with the classification. The majority of the approaches address horizontal supply chain mapping based on content rather than links and supervised learning techniques for classification.

However, visually emerges a development from horizontal to vertical approaches and from link to content-based methods. The classification techniques appear to be supervised rather than rule-based in recent years.

5. Conclusion and Outlook

We presented the first article to identify and analyze NLP-based supply chain mapping approaches with a systematic literature review to the best of our knowledge. We identified twelve supply chain mapping techniques in the literature (RQ 1) and showed that research is in an early stage, as only a few publications in established journals could be found. Moreover, our results highlight a fragmented research field on the timescale and concerning the conceptual knowledge bases. The recent advances in machine learning and NLP may explain the new emergence of a research cluster because they triggered new research interests. However, the scientific basis fragmentation limits the more recent approaches' ability to "learn" from prior research's strengths and pitfalls. Simultaneously there seems to be ambiguity about the scientific foundation for supply chain mapping approaches. Nevertheless, we observed strong conceptual similarities in the approaches and proposed a simple taxonomy from our analysis (RQ 2). Our results highlight a development from link- and rule-based horizontal mapping methods to content-based and supervised vertical mapping techniques. Additionally, this scheme can serve as a template for identifying, describing, and structuring future works.

In our content analysis, we identified and highlighted significant methodical and conceptual limitations. Significantly, the advanced supervised learning approaches did not show substantial performance improvements in precision and recall over rule-based systems. These results seem counter-intuitive, as we would expect at least significant advances in recall. We suggest future research to elaborate on criteria for good data quality and the annotation phase's performance. Furthermore, the approaches do not comply with essential supply chain mapping requirements (RQ 3) from a conceptual perspective. We suggest that further research considers additional requirements such as product or service-specific relations or the supply chain dynamics. Also, we observed no integration between vertical and horizontal mapping techniques. We propose that future methods integrate supply relations with other business relations into a more holistic framework for supply chain mapping. Lastly, the case studies in the identified publication focused solely on large companies. However, companies get smaller in the upstream direction of the supply chain. We suggest future research to validate whether and to what extent NLP-based supply chain mapping approaches can detect small and medium-sized companies across different industries.

6. References

- [1] F. B. Norwood, and D. Peel, „Supply Chain Mapping to Prepare for Future Pandemics“, *Applied Economic Perspectives and Policy*, 2020, doi: 10.1002/aepp.13125.
- [2] S. Singh, R. Kumar, R. Panchal, and M. K. Tiwari, „Impact of COVID-19 on Logistics Systems and Disruptions in Food Supply Chain“, *International Journal of Production Research*, pp. 1–16, 2020, doi: 10.1080/00207543.2020.1792000.
- [3] J. E. Hobbs, „Food Supply Chains During the COVID-19 Pandemic“, *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, volume 68, issue 2, pp. 171–176, 2020, doi: 10.1111/cjag.12237.
- [4] B. Foldy, and E. Sylvers, „Coronavirus Creates Domino Effect in Global Automotive Supply Chain“, *The Wall Street Journal*, 2020, 2020. URL: <https://www.wsj.com/articles/coronavirus-outbreak-could-affect-production-at-2-gm-plants-union-officials-say-11581697246>. Accessed: 16. December 2020.
- [5] T. Y. Choi, D. Rogers, and B. Vakil, „Coronavirus Is a Wake-Up Call for Supply Chain Management“, *Harvard Business Review*, 27. März 2020, 2020. URL: <https://hbr.org/2020/03/coronavirus-is-a-wake-up-call-for-supply-chain-management>. Accessed: 16. December 2020.

- [6] P. Haren, and D. Simchi-Levi, „How Coronavirus Could Impact the Global Supply Chain by Mid-March“, *Harvard Business Review*, 28. Feb. 2020, 2020. URL: <https://hbr.org/2020/02/how-coronavirus-could-impact-the-global-supply-chain-by-mid-march>. Accessed: 15. December 2020.
- [7] J. Schnelle, H. Schöpfer, and W. Kersten, „Corona: Katalysator für Digitalisierung und Transparenz?“, *Industrie 4.0 Management*, volume 37, issue 1, pp. 27–31, 2021, doi: 10.30844/i40m_21-1_s27-31.
- [8] J. Tatzenko, C. Hart, J. Hollister, M. Ryland, L. Medic und E. Taylor, *The 2019 Ethical Fashion Report: The Truth Behind the Barcode*. URL: <https://apo.org.au/sites/default/files/resource-files/2019-04/apo-nid229606.pdf>.
- [9] The Sustainability Consortium, *Greening Global Supply Chains: From Blind Spots To Hotspots To Action. Impact Report*. URL: <https://www.sustainabilityconsortium.org/wp-content/themes/enfold-child/assets/pdf/2016-impact-report.pdf>.
- [10] T. Kraft, L. Valdés, and Y. Zheng, „Supply Chain Visibility and Social Responsibility: Investigating Consumers’ Behaviors and Motives“, *Journal of Manufacturing & Service Operations Management*, volume 20, issue 4, pp. 617–636, 2018, doi: 10.1287/msom.2017.0685.
- [11] J. Hartmann, and S. Moeller, „Chain Liability in Multitier Supply Chains? Responsibility Attributions for Unsustainable Supplier Behavior“, *Journal of Operations Management*, volume 32, issue 5, pp. 281–294, 2014, doi: 10.1016/j.jom.2014.01.005.
- [12] The California Transparency in Supply Chains Act of 2010: SB 657.
- [13] Initiative Lieferkettengesetz, „Anforderungen an ein wirksames Lieferkettengesetz“. URL: https://lieferkettengesetz.de/wp-content/uploads/2019/09/Anforderungen-an-ein-wirksames-Lieferkettengesetz_Februar-2020.pdf.
- [14] J. T. Gardner, and M. C. Cooper, „Strategic Supply Chain Mapping Approaches“, *Journal of Business Logistics*, volume 24, issue 2, pp. 37–64, 2003, doi: 10.1002/j.2158-1592.2003.tb00045.x.
- [15] M. Theodore Farris, „Solutions to Strategic Supply Chain Mapping Issues“, *Int Jnl Phys Dist & Log Manage*, volume 40, issue 3, pp. 164–180, 2010, doi: 10.1108/09600031011035074.
- [16] D. M. Lambert, and M. C. Cooper, „Issues in Supply Chain Management“, *Industrial Marketing Management*, volume 29, issue 1, pp. 65–83, 2000, doi: 10.1016/S0019-8501(99)00113-3.
- [17] J. T. Mentzer et al., „Defining Supply Chain Management“, *Journal of Business Logistics*, volume 22, issue 2, pp. 1–25, 2001, doi: 10.1002/j.2158-1592.2001.tb00001.x.
- [18] D. Farrell, „Offshoring: Value Creation through Economic Change“, *Journal of Management Studies*, volume 42, issue 3, pp. 675–683, 2005, doi: 10.1111/J.1467-6486.2005.00513.X.
- [19] R. E. Baldwin, „Global Supply Chains: Why They Emerged, Why They Matter, and Where They are Going“, *CEPR Discussion Paper*, 2012.
- [20] C. Y. Baldwin, and K. B. Clark, „Managing in an Age of Modularity“, *Harvard Business Review*, pp. 84–93, 1997.
- [21] M. Brylowski, H. Schöpfer, and M. Krull, „Produktmodularisierung entlang der Supply Chain“, *Industrie 4.0 Management*, volume 35, issue 5, pp. 50–54, 2019, doi: 10.30844/I40M_19-5_S50-54.
- [22] J. Hsuan, „Impacts of Supplier–Buyer Relationships on Modularization in New Product Development“, *European Journal of Purchasing & Supply Management*, volume 5, 3-4, pp. 197–209, 1999, doi: 10.1016/S0969-7012(99)00026-X.
- [23] J. H. Mikkola, „Modularity, Component Outsourcing, and Inter-Firm Learning“, *R&D Management*, volume 33, issue 4, pp. 439–454, 2003, doi: 10.1111/1467-9310.00309.
- [24] Y. Yao, and N. Fabbe-Costes, „Can you Measure Resilience if you are Unable to Define it? The Analysis of Supply Network Resilience (SNRES)“, *Supply Chain Forum: An International Journal*, volume 19, issue 4, pp. 255–265, 2018, doi: 10.1080/16258312.2018.1540248.
- [25] J. El Baz, and S. Ruel, „Can Supply Chain Risk Management Practices Mitigate the Disruption Impacts on Supply Chains’ Resilience and Robustness? Evidence from an Empirical Survey in a COVID-19 Outbreak Era“, *Int J Prod Econ*, pp. 107972, 2020, doi: 10.1016/j.ijpe.2020.107972.
- [26] W. Kersten, P. Hohrath, M. Böger, and C. Singer, „A Supply Chain Risk Management Process“, *International Journal of Logistics Systems and Management*, volume 9, issue 2, pp. 152–166, 2011.

- [27] M. Schröder, „Strukturierte Verbesserung des Supply Chain Risikomanagements“ in *Supply Chain Management - Beiträge zu Beschaffung und Logistik*, M. EBig, W. Stölzle und W. Kersten, Hg., Wiesbaden: Springer Gabler, 2019, doi: 10.1007/978-3-658-26773-5.
- [28] W. Kersten, P. Hohrath, and Böger Mareike, „An Empirical Approach to Supply Chain Risk Management: Development of a Strategic Framework“ in *Proceedings of the POMS Annual Conference*, 2007.
- [29] B. D. Williams, J. Roh, T. Tokar, and M. Swink, „Leveraging Supply Chain Visibility for Responsiveness: The Moderating Role of Internal Integration“, *Journal of Operations Management*, volume 31, 7-8, pp. 543–554, 2013, doi: 10.1016/j.jom.2013.09.003.
- [30] N. Egels-Zandén, K. Hulthén, and G. Wulff, „Trade-offs in Supply Chain Transparency: The Case of Nudie Jeans Co“, *Journal of Cleaner Production*, volume 107, pp. 95–104, 2015, doi: 10.1016/j.jclepro.2014.04.074.
- [31] Y. Sheffi, and J. B. Rice Jr., „A Supply Chain View of the Resilient Enterprise“, *MIT Sloan Management Review*, volume 47, issue 1, pp. 41–48, 2005.
- [32] A. Kassahun et al., „Enabling Chain-Wide Transparency in Meat Supply Chains Based on the EPCIS Global Standard and Cloud-Based Services“, *Computers and Electronics in Agriculture*, volume 109, pp. 179–190, 2014, doi: 10.1016/j.compag.2014.10.002.
- [33] D. DiMase, Z. A. Collier, J. Carlson, R. B. Gray, and I. Linkov, „Traceability and Risk Analysis Strategies for Addressing Counterfeit Electronics in Supply Chains for Complex Systems“, *Risk analysis : an official publication of the Society for Risk Analysis*, volume 36, issue 10, pp. 1834–1843, 2016, doi: 10.1111/risa.12536.
- [34] T. Yan, T. Y. Choi, Y. Kim, and Y. Yang, „A Theory of the Nexus Supplier: A Critical Supplier From A Network Perspective“, *J Supply Chain Management*, volume 51, issue 1, pp. 52–66, 2015, doi: 10.1111/jscm.12070.
- [35] D. Ivanov, B. Sokolov, and A. Dolgui, „The Ripple Effect in Supply Chains: Trade-Off ‘Efficiency-Flexibility-Resilience’ in Disruption management“, *Int J Prod Res*, volume 52, issue 7, pp. 2154–2172, 2014, doi: 10.1080/00207543.2013.858836.
- [36] J. F. Allen, „Natural Language Processing“ in *Encyclopedia of Computer Science*, A. Ralston und E. D. Reilly, Hg., Wiley and Sons Ltd., 2003, pp. 1218–1222.
- [37] A. Reshamwala, P. Pawar, and D. Mishra, „Review on Natural Language Processing“, *IRACST - Engineering Science and Technology: An International Journal*, volume 3, issue 1, 2013.
- [38] A. Kappel, „Increasing Supply Network Transparency: Structure Model, Mapping Procedure and Performance Impact“. Dissertation, University of Twente, Twente, 2019.
- [39] P. Wichmann, A. Brintrup, S. Baker, P. Woodall, and D. McFarlane, „Towards Automatically Generating Supply Chain Maps From Natural Language Text“ in *Proceedings of the 16th Symposium in Information Control Problems in Manufacturing (INCOM)*, Bergamo, Italy, 2018, pp. 1726–1731, doi: 10.1016/j.ifacol.2018.08.207.
- [40] H. Arksey, and L. O'Malley, „Scoping Studies: Towards a Methodological Framework“, *International Journal of Social Research Methodology*, volume 8, issue 1, pp. 19–32, 2005, doi: 10.1080/1364557032000119616.
- [41] D. Levac, H. Colquhoun, and K. K. O'Brien, „Scoping Studies: Advancing the Methodology“, *Implementation science : IS*, volume 5, pp. 69, 2010, doi: 10.1186/1748-5908-5-69.
- [42] M. T. Pham, A. Rajić, J. D. Greig, J. M. Sargeant, A. Papadopoulos, and S. A. McEwen, „A Scoping Review of Scoping Reviews: Advancing the Approach and Enhancing the Consistency“, *Research synthesis methods*, volume 5, issue 4, pp. 371–385, 2014, doi: 10.1002/jrsm.1123.
- [43] P. Mongeon, and A. Paul-Hus, „The Journal Coverage of Web of Science and Scopus: A Comparative Analysis“, *Scientometrics*, volume 106, issue 1, pp. 213–228, 2016, doi: 10.1007/s11192-015-1765-5.
- [44] J. R. Landis, and G. G. Koch, „The Measurement of Observer Agreement for Categorical Data“, *Biometrics*, volume 33, issue 1, pp. 159, 1977, doi: 10.2307/2529310.
- [45] M. J. Page et al., „The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews“, *PLoS medicine*, volume 18, issue 3, e1003583, 2021, doi: 10.1371/journal.pmed.1003583.
- [46] S. Bao, R. Li, Y. Yu, and Y. Cao, „Competitor Mining with the Web“, *IEEE Trans. Knowl. Data Eng.*, volume 20, issue 10, pp. 1297–1310, 2008, doi: 10.1109/TKDE.2008.98.

- [47] G. Pant, and O. R. Sheng, „Avoiding the Blind Spots: Competitor Identification Using Web Text and Linkage Structure“ in Proceedings of the 30th International Conference on Information Systems (ICIS), 2009.
- [48] Z. Ma, O. R. Sheng, and G. Pant, „Discovering Company Revenue Relations from News: A Network Approach“, Decision Support Systems, volume 47, issue 4, pp. 408–414, 2009, doi: 10.1016/j.dss.2009.04.007.
- [49] Y. Xia, N. Ji, W. Su, and Y. Liu, „Mining Commercial Networks from Online Financial News“ in 2010 IEEE 7th International Conference on e-Business Engineering (ICEBE), Shanghai, China, 2010, pp. 17–23, doi: 10.1109/ICEBE.2010.54.
- [50] J. Zhao, P. Jin, and Y. Liu, „Business Relations in the Web: Semantics and a Case Study“, Journal of Software, volume 5, issue 8, pp. 826–833, 2010.
- [51] Z. Ma, G. Pant, and O. R. Sheng, „Mining Competitor Relationships from Online News: A Network-Based Approach“, Electronic Commerce Research and Applications, volume 10, issue 4, pp. 418–427, 2011, doi: 10.1016/j.elerap.2010.11.006.
- [52] R. Y. Lau, and W. Zhang, „Semi-supervised Statistical Inference for Business Entities Extraction and Business Relations Discovery“ in Proceedings of the 1st International Workshop on Entity-Oriented Search (EOS), Beijing, China, 2011, pp. 41–46.
- [53] C.-P. Wei, L.-C. Chen, H.-Y. Chen, and C.-S. Yang, „Mining Suppliers from Online News Documents“ in Proceedings of the 17th Pacific Asia Conference on Information Systems (PACIS), Jeju Island, Korea, 2013.
- [54] A. Yamamoto, Y. Miyamura, K. Nakata, and M. Okamoto, „Company Relation Extraction from Web News Articles for Analyzing Industry Structure“ in Proceedings of the 11th International Conference on Semantic Computing (ICSC), San Diego, California, US, 2017, doi: 10.1109/ICSC.2017.25.
- [55] L. Carstens, J. L. Leidner, K. Szymanski, and B. Howald, „Modeling Company Risk and Importance in Supply Graphs“ in Proceedings of the 14th Extended Semantic Web Conference (ESWC), Portorož, Slovenia, 2017, doi: 10.1007/978-3-319-58451-5_2.
- [56] P. Wichmann, A. Brintrup, S. Baker, P. Woodall, and D. McFarlane, „Extracting Supply Chain Maps from News Articles using Deep Neural Networks“, Int J Prod Res, 2020, doi: 10.1080/00207543.2020.1720925.
- [57] M. M. Kessler, „Bibliographic coupling between scientific papers“, Amer. Doc., volume 14, issue 1, pp. 10–25, 1963, doi: 10.1002/asi.5090140103.
- [58] M. Bastian, S. Heymann, and M. Jacomy, „Gephi: an Open Source Software for Exploring and Manipulating Networks“ in Proceedings of International AAAI Conference on Weblogs and Social Media, 2009.
- [59] Z. Ma, G. Pant, and O. R. Sheng, „A Network-based Approach to Mining Competitor Relationships from Online News“ in Proceedings of the 30th International Conference on Information Systems (ICIS), 2009.
- [60] J. Zhao, and P. Jin, „Ontological Foundation for Enterprise Competitive Intelligence in the Web“ in Proceedings of the International Symposium on Information Processing, F. Yu, J. Shu und G. Yue, Hg., 2009, pp. 445–448.