UDC: 001.82:159.953.5:334.716
Original Scientific Paper
Accepted March, 22, 2018
Corresponding author Wanja Wellbrock,
wanja.wellbrock@hs-heilbronn.de

CROSS-COMPANY DATA MANAGEMENT AS A STRATEGIC ADVANTAGE – AN EXPLORATORY STUDY

Wanja Wellbrock[1], Christoph Hein[2]

[1]Departemnt for Management and Sales / Heilbronn University
[2]HENDRICKS, ROST & CIE GmbH

Abstract: *The perfect technical collection and storage of data does not guarantee companies better information. The focus must be on a pro-active and strategic oriented data management within and between cooperating companies. An information supply chain can ensure a real utility value as long as the semantic limitations are considered. Based on the conceptual development of an corresponding information supply chain model, the evidence and status quo in business practice is assessed by a large-volume empirical study within German companies.*

Keywords: *data management, information supply chain, business analytics, cross-company, supply chain management, semantic trap*

WORLDWIDE INCREASING DATA VOLUME

The amount of electronically available data is increasing exponentially every year. This includes an enormous impact on storage media and the storage density of hard drives or SSDs is already on the verge of the physically possible. The Internet of things continues to generate demand, with its sensors penetrating every single part of life, boosting the tide of data. However, does every measurement also contain information and is this information important and therefore worth saving?

As an example, take a wear sensor for a brake pad. It continuously reports the strength of the surface. This data is certainly useful for an engineer who wants to calculate the wear and thus the expected changeover time. For the driver, they are only conditionally. It is sufficient for him to know when the rubber is to be changed. Thus, the highest benefit can only be achieved by a targeted pro-active data management

The aim of the paper is, to develop a conceptual model for an information supply chain as a basis for pro-active cross-company data management. In addition, the evidence and status quo of data management within in German companies is assessed by a large-volume empirical study.

METHODOLOGY OF THE EMPIRICAL STUDY

Data collection

To choose a sample that suits our research purpose companies from different industries and different size were considered. In all cases, the questionnaire was sent to experts and executives in digitalization, business intelligence or information technology on corporate level. The direct contact was either determined by internet research or requested by phone calls. The survey itself was carried out in three stages. At first, the questionnaire was created, validated in pretests with selected companies and further adjusted. The final questionnaire was sent to the corresponding partners with the request to participate online via a survey platform. Overall, after exclusion of incomplete answers, 228 companies could be integrated in the sample.

Table 1 shows the corresponding distribution of participants. Companies from 13 different industries are involved with a majority in information technology and consulting. In terms of company size, the participating companies partially differ greatly. More than 50% of all companies have less than 500 employees, whereas almost 30% have more than 5,000 employees. Overall, the analyzed companies represent a meaningful sample for the selected industries.

Table 1. Sample characteristics

Characteristics of the sample *(n=228; relative frequency)*						
Business sectors of participating companies						
Transport & logistics	Metal production and processing	Telecomm., technology & electronics	Food & consumer goods	Health care, chemistry & pharma	Public sector & education	Plant & mechanical engineering
2.2%	2.2%	2.6%	4.9%	5.3%	6.2%	6.6%
Finance	Other services	Automotive	Trade	Consulting	Information technology	Others
6.6%	6.6%	7.1%	8.8%	12.8%	19.8%	8.4%
Company size (employees)						
< 201	201-500	501-1,000	1,001-5,000	5,001-10,000	>10,000	
38.6%	11.8%	8.3%	11.8%	8.3%	21.1%	

Method

The questionnaire combines (quasi)-interval scaled (Gregoire and Driver, 1987; Jaccard and Wan, 1996) and dichotomous indicators (Sheskin, 2011). To determine significant results for quasi-interval scaled indicators the single sample t test was selected with whom significant mean differences from a given value of the underlying rating scale can be determined. The rating scale for all (quasi)-interval scaled indicators is aligned from one (very low) to five (very high), wherein the values three (moderate) and four (high) are used as relevant test values for the determination of significance. The focus is on directed hypotheses, which lead in response to the sample means to the following null and alternative hypothesis (Cooper and Schindler, 2011; Sekaran and Bougie, 2013):

- Analysis regarding scale value $\mu_0 = 3$ (moderate): H_0: $\mu_i \geq 3$ and H_1: $\mu_i < 3$ (for $\mu_i < 3$) or H_0: $\mu_i \leq 3$ and H_1: $\mu_i > 3$ (for $\mu_i > 3$); i = quasi-interval scaled indicators.
- Analysis regarding scale value $\mu_0 = 4$ (high): H_0: $\mu_i \geq 4$ and H_1: $\mu_i < 4$ (for $\mu_i < 4$) or H_0: $\mu_i \leq 4$ and H_1: $\mu_i > 4$ (for $\mu_i > 4$); i = quasi-interval scaled indicators.

For significance investigation at dichotomous indicators, the binomial test was selected. This test examines, if a certain frequency p_0 of a characteristic is present in the population. The question, if a characteristic in the population has at least or at most a specific frequency p_0 can be converted in the following statistical test problem (Sheskin, 2011):

- H_0: $p_j \geq p_0$ and H_1: $p_j < p_0$ (for $p_j < p_0$) or H_0: $p_j \leq p_0$ and H_1: $p_j > p_0$ (for $p_j > p_0$); j = dichotomous indicators.

Therefore, directed hypothesis are considered in which the frequencies 0.25, 0.50 and 0.75 are used as relevant test values p_o. Regarding the significance level the stages *($\alpha = 10\%$), **($\alpha = 5\%$) and ***($\alpha = 1\%$) are taken into account (Cooper and Schindler, 2011; Sekaran and Bougie, 2013). All statistical tests were carried out by using the software SPSS (Field, 2013). To ensure the data quality of the sample possible distortions are to be avoided. Statistical tests for distortions caused by non-response-bias did not lead to any significant results.

EXPECTATION DETERMINES THE PRICE

The purpose decides the meaningfulness of stored data and its transformation into information. In addition, much of today's data is redundant. For example, a video bought and downloaded to one`s laptop doubles the amount of overall data because a copy is also stored on the server; but it does not increase the information. In any case, information is smaller than the corresponding data volume. Moreover, what else determines the value of information? Answer: only its purpose. In terms of market economy, the expectations regarding the value of information and the related demand determine the actual price. For example, a pirated copy of the seventh season of "Game of Thrones" could have ruined HBO. Thus, even redundant information has a potential value. Nevertheless, how to ensure it (Burgin, 2010)?

How is it possible to find the valuable nuggets in a bunch of river sand? The clear answer is not at all. Likewise, Michelangelo could have been accused of having to knock out every-

thing of the block of marble that did not look like the statue of David. Therefore, information gathering, in economic terms, is not a primary sector, but a manufacturing industry. The benefit lies in the systematic collection of data, the transformation into information and the delivery to end-users. From this perspective, the term "data mining", commonly used for gathering information, is actually the wrong term and one should speak better of "information production". Gaining information from existing data is a creative process that creates something new from existing resources (Han *et al.*, 2012).

How can companies set up such an information value chain? Every business needs to collect, analyze and enrich their own and external data along business process chains with additional information. In addition, the task of companies is to ensure that the data can be stored safely and sustainably and can be made available to other processes or users with the required quality and at the right time (Edmunds and Morris, 2000).

RESTRICTIONS OF CURRENT DATA MANAGEMENT

According to the empirical results, the participating companies rank the present importance (strategic and operative) of data management on a scale from one (very low) to five (very high) with a value of 3.74, which exceeds a moderate level significantly (see table 2). In the future, the expected importance is even higher with up to 4.47 points for the strategic relevance (significantly above a high level).

Table 2. Importance of data management

Importance of data management (n=228; scale: 1 (very low) to 5 (very high))	Descriptive statistics		Significance	
	Average	σ	Test value 3	Test value 4
Present operative importance	3.74	1.02	>***	<***
Present strategic importance	3.74	1.04	>***	<***
Future operative importance	4.40	0.77	>***	>***
Future strategic importance	4.47	0.71	>***	>***

While companies see a strategic relevance for data-driven decision-making, only 55% of the surveyed companies use data for strategic decision making (no significant result for test value 0.50). On the operative level, the percentage is even lower with 29%. An automatic decision-making based on data is even completely absent with only 5% (see table 3).

A similar situation is visible for the organizational structure of data management. Only in 20% of all companies, a specialized department is responsible for data management within the organization. For each 31% data management is connected to the central it department or handled decentralized within single business departments. 14% even say that data management is done only informal and uncoordinated (see table 3).

Table 2. Status quo of data-driven decision-making

Status quo of data-driven decision-making (n=228)	Descriptive statistics		Significance		
	Absolute frequency	Relative frequency	$p_0 = 0.25$	$p_0 = 0.50$	$p_0 = 0.75$
Basically, decisions are not made on the basis of data	8	0.04	<***	<***	<***
Data-driven decision-making only takes place unstructured	28	0.12	<***	<***	<***
Data is used for structured decision-making on operative level	66	0.29	>*	<***	<***
Data is used for structured decision-making on strategic level	126	0.55	>***	>*	<***
Data is used for automatic decision-making on operative level	12	0.05	<***	<***	<***
Data is used for automatic decision-making on strategic level	11	0.05	<***	<***	<***

Table 3. Organizational structure of data management

Organizational structure of data management (n=228)	Descriptive statistics		Significance		
	Absolute frequency	Relative frequency	$p_0 = 0.25$	$p_0 = 0.50$	$p_0 = 0.75$
Informal and uncoordinated	31	0.14	<***	<***	<***
Decentralized within single business departments	71	0.31	>**	<***	<***
Central by the the IT department	70	0.31	>**	<***	<***
Central by a specialized rod department	45	0.20	<**	<***	<***

The results show that some of the companies use data for strategic decisions, but the organization structure is not implemented to achieve meaningful results. Most of them are still a long way from becoming a real "data-driven company". To make matters worse, they also have to deal with additional objectives in the future like increasing demand for innovation or the tapping of new business areas. Not to mention the promises of artificial intelligence with the methods of machine learning and deep learning.

INFORMATION SUPPLY CHAIN

However, even a structured data management alone does not lead to a knowledge gain. The surrounding processes must be considered. The best data will not help, if no information is obtained from it. Futurologist John Naisbitt explains it with the following sentence: "We drown in data and thirst for information" (Naisbitt, 1982).

What is the solution for this problem? The answer is very pragmatic: analogous to the logistics supply chain, the information supply chain enables a holistic view of information processes. For this purpose, the internal and external flow of information between different departments or even different companies must be identified and later systematically controlled. Important in this process is, not to focus on the method but always to maximize the benefits to the recipients of information: an objective that information supply chain shares with the traditional supply chain in logistics.

The focus of an information supply chain is on the overall view, which ranges from data creation to decision-making (see figure 1). The latter is made based on the information obtained, and includes tasks from different business areas. Within the information supply chain data is seen as a product, which needs to be refined.

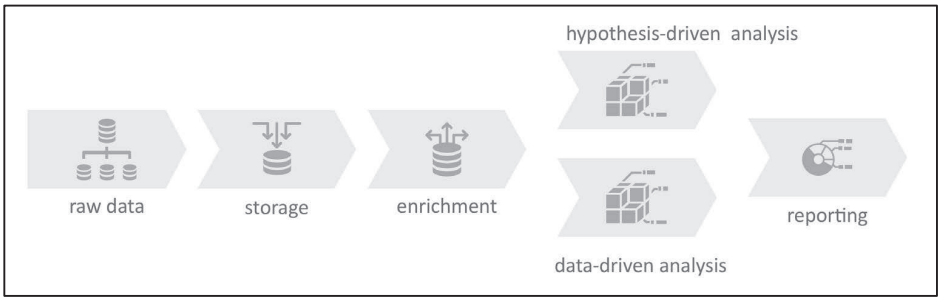

Figure 1. Information supply chain

The beginning of the information supply chain is raw data, collected for example by transactional systems or sensors. This data is mostly large in volume and low in structure. Therefore, an analysis is not advised (Roggen *et al.*, 2010).

In first step, the raw data has to be stored appropriately to secure the future use of the data sets. Persistence of the data and access to the data by all relevant persons, organizations or processes have to be ensured (Strong *et al.*, 1997). The importance of continuously maintaining the quality of stored data increased over the time in accordance to the growing use of automated analytical methods (e.g. machine learning or predictive analytics). Furthermore, a loss of data can lead to biased results in automated system (Batista *et al.*, 2003).

After the secured storage of the data, the data enrichment process starts. The data sets will be cleaned, transformed and if necessary consolidated to reduce the amount of data and speed up the following analysis process. Standardized metadata can be added to provide further possibilities to analyze the data (Conn, 2005). Those first three steps lay the groundwork for an effective and efficient data analysis.

According the final analysis, a distinction has to be made between hypothesis-driven and data-driven analysis. Data-driven analytics provide insights utilizing methods like data mining or predictive analytics. It is based on automated unsupervised analyses and therefore demands a high quality of the data sets with a minimal margin for errors. A hypothesis-driven analysis requires a manual definition of the data structure with the risk of ignoring specific areas of the data set. Afterwards there is a testing of the hypotheses with methods like visual or pivotal analytics. Because of the user-based approach errors within the data sets can be corrected through manual interpretation (Smalheiser, 2002; Kell, 2003).

The final step of the information supply chain is the reporting of findings gained via both analytical ways. This is the crucial step to provide the decision makers with all necessary information (Beattie *et al.*, 2003).

The production of information begins with the specification of a structure in which the data is to be collected already at the beginning of the information supply chain. Similar to a bill of material, it describes how data has to be made up for later use. The bill of material can be, for example, a form or mask for account assignment in financial accounting. The recording of an expense receipt can be used as an example: the entries required for the transaction, such as date, amount, currency, account, cost center or cost unit, are values that prepare a transaction for later analyzes (Chaudhuri *et al.*, 1997). The definition of the structure is based on a hypothesis about the desired result. In most cases, these are comparisons, time series, budget deviations and contribution margins. The deviation from the defined hypothesis within the data sets is subsequently analyzed more detailed. For example, all modern OLAP systems for multidimensional databases work this way (Sarawagi *et al.*, 1998).

SEMANTIC TRAP

Another problem with the described hypothesis-driven analysis is the possible limitation of knowledge gain. The data can only provide the information, which is created in the structure. This problem is called the "semantic trap". The mentioned example of the account assignment of a document can explain this danger: some information is not recorded because it was not needed for a specific purpose. The meaning of the data set is fixed rigidly; therefore, another interpretation is no longer technically possible. Such a limitation can be fatal. The document has much more information to offer, for example time, individual positions with their prices or the location. Additional interesting questions within this example can be: does the number of dishes match the number of people served, are children's menus included or did the date fall on a weekend? These are all information that can prove an expense scam that has already cost some employees the job. In most cases, the companies do not record the compliance-relevant information systematically (Zhu *et al.*, 2007). Therefore, especially compliance requirements are seen as a high potential of data management in the next years.

It gets even more complicated when supplier and customer data has to be integrated into the internal data management process. If a company requests data from a supplier or a customer, the semantic trap can have a major impact on the following analytics. If the supplier has a different data management process, the data can be collected in a different structure and maybe on a different level of detail. Sometimes, the comparison of those datasets is even impossible (Doan *et al.*, 2012).

INCREASING IMPORTANCE OF SUPPLIER DATA WITHIN THE INFORMATION SUPPLY CHAIN

Holistic data management in the sense of an information supply chain increases fundamentally the importance of supplier data. An outsourcing rate of up to 80% means that an isolated view of a company is no longer expedient. Interfaces with the value-adding partners (upstream and downstream) are becoming one of the most important competitive factors.

The vision of a digitized end-to-end supply chain is characterized by full real time visibility along the entire value chain. The central collection and location-independent availability of all supply chain relevant information is an elementary prerequisite for this. The goals are to identify potential for optimization by big data applications and to reduce control complexity through decentralized, autonomous decisions (Wagner and Kontny, 2017; Xue *et al.*, 2014). The adaptation to production, quantity and layout changes due to flexible, scalable and lean material flow structures are mentioned as additional potentials (Bierwith and Schocke, 2017; Gallay *et al.*, 2017; Rai, A. *et al.*, 2006; Schauer *et al.*, 2017). The fundamental basis for this vision is a cross-company, strategic oriented data management, which does not stop at the artificial boarder of a single company.

Although the need for a cross-company perspective is well known in practice, the empirical data of the underlying study show that almost half of the surveyed companies (45%) still focus only on their own data (see table 4). While 51% of the participants integrate at least customer data into their decision making process, the value is significantly lower for supplier data with 27%. Looking at the data strategy, the integration level for external data (2.85 and 2.89) is also still significantly lower than for internal data (3.22 and 3.28). A moderate level of three points is at least weakly significant rejected for the strategic integration of external data (see table 5).

Table 4. Range of data management

Range of data management (n=228)	Descriptive statistics		Significance		
	Absolute frequency	Relative frequency	$p_0 = 0.25$	$p_0 = 0.50$	$p_0 = 0.75$
Data management includes only company-internal data	103	0.45	>***	n.s.	<***
Vertical integration of customer data	116	0.51	>***	n.s.	<***
Vertical integration of supplier data	62	0.27	n.s.	<***	<***
Horizontal integration of competitor data (same industry)	54	0.24	n.s.	<***	<***
Horizontal integration of competitor data (different industry)	12	0.05	<***	<***	<***

Another interesting aspect is the low percentage of horizontal integration of competitor data. 27% of all surveyed companies integrate competitor data from the same industry,

whereas only 5% include competitor data from other sectors into their decision-making (see table 4). Especially when providing industry-wide data, suppliers become an important role. The example of self-driving cars shows that original equipment manufacturer can only collect data from their own products, while suppliers can perform cross-brand analyzes. In the case of the navigation system "Here", suppliers shall control the storage, processing and real time provision of sensor data in order to provide all original equipment manufacturer with an adequate cross-brand information base. The provided data from the cars is used to keep the maps of "Here" current, so that, for example, information on construction sites or accidents are immediately available to other motorists (Taub, 2018).

Table 5. Implementation level of a data management strategy

Implementation level of a data management strategy (n=228; scale: 1 (very low) to 5 (very high))	Descriptive statistics		Significance	
	Average	Σ	Test value 3	Test value 4
We have a clear strategy for collection and storage of company internal data	3.22	1.15	>***	<***
We have a clear strategy for analysis company internal data	3.28	1.13	>***	<***
We have a clear strategy for collection and storage of company internal and external data	2.89	1.16	<*	<***
We have a clear strategy for analyzing company internal and external data	2.85	1.19	<**	<***

Overall, the results show that the potential of an intensive IT-related link with suppliers has not yet been sufficiently exploited, which represents a clear obstacle on the way to a holistic information supply chain.

CONCLUSION

Companies are able to gain an added value from a cross-company information supply chain. There is a variety of deployment options in all sectors, depending on the availability of data and the willingness to restructure the flow of information. The possibilities are manifold; the decisive factor is the creativity in dealing with the data. A high quality information supply chain forms the basis for reliable added value. The semantic trap has to be considered. Afterwards companies have to start with a pro-active approach on data management and instead of working with the existing data, companies should look for new data sets within and especially outside of their own organization.

LITERATURE

1. Batista, G.; Monard, M. (2003). An analysis of four missing data treatment methods for supervised learning. In: *Applied Artificial Intelligence.* Vol. 17 Nr. 5/6, pp. 519-533.
2. Beattie, V.; Pratt, K. (2003). Issues concerning web-based business reporting: an analysis of the views of interested parties. In: *The British Accounting Review.* Vol. 35 Nr. 2, pp. 155-187.
3. Bethea, T.J., Krishna, V. and Zhu, G. (2007). Extracting relevant named entities for automated expense reimbursement. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, pp. 1004-1012.
4. Bierwirth, B. and Schocke, K.-O. (2017). Lead-time optimization potential of digitization in Air Cargo. In: Kersten, W.; Blecker, T. and Ringle, C.M. (Eds.). *Digitalization in supply chain management and logistics. Smart and digital solutions for an industry 4.0 environment*. Berlin, pp. 75-98.
5. Burgin, M. (2010). *Theory of information: fundamentality, diversity and unification*, Los Angeles.
6. Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and OLAP technology. In: *ACM Sigmod Record*. Vol. 26 Nr. 1, pp. 65-74.
7. Conn, S. (2005). OLTP and OLAP data integration: a review of feasible implementation methods and architectures for real time data analysis. In: *Proceedings. IEEE SoutheastCon 2005,* pp. 515-520.
8. Cooper, D.R. and Schindler, P.S. (2011). *Business research methods*. 11th ed. New York.
9. Dennert, A.; Gossling, A.; Krause, J.; Wollschlaeger, M. and Montoya, A. (2012). Vertical data integration in automation based on IEC 61499. In: *9th IEEE International Workshop on FactoryCommunication Systems,* pp. 99-102.
10. Doan, A.; Halevy, A. and Ives, Z. (2012). *Principals of data integration.* Waltham.
11. Edmunds, A. and Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. In: *International Journal of Information Management.* Vol. 20 Nr. 1, pp.17-28.
12. Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. 4th ed. Los Angeles.
13. Gallay, O.; Korpela, K.; Tapio, N. and Nurminen, J.K. (2017). A peer-to-peer platform for decentralized logistics. In: Kersten, W.; Blecker, T. and Ringle, C.M. (Eds.). *Digitalization in supply chain management and logistics. Smart and digital solutions for an industry 4.0 environment*. Berlin, pp. 405-426.
14. Gregoire, T.G. and Driver, B.L. (1987). Analysis of ordinal data to detect population differences. In: *Psychological Bulletin*. Vol. 101 No. 1, pp. 159-165.

15. Han, J.; Kamber, M. and Pei, J. (2012). *Data mining concepts and techniques*. 3rd ed. Amsterdam.

16. Jaccard, J. and Wan, C.K. (1996). *LISREL approaches to interaction effects in multiple regressions*. Thousand Oaks.

17. Kell, D. and Oliver, S. (2003). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. In: *Bioessays*, Vol. 26 Nr. 1, pp. 99-105.

18. Naisbitt, J. (1982). *Megatrends: ten new directions transforming our lives*, New York.

19. Rai, A.; Patnayakuni, R. and Seth, N. (2006). Firm performance impacts of digitally enabled supply chain integration capabilities. In: *MIS Quarterly*. Vol. 30 No. 2, pp. 225-246.

20. Roggen, D.; Calatroni, A.; Rossi, M.; Holleczek, T.; Förster, K.; Tröster, G.; Lukowicz, P.; Bannach, D.; Pirkl, G.; Ferscha, A.; Doppler, J.; Holzmann, C.; Kurz, M.; Holl, G.; Chavarriaga, R.; Sagha, H.; Bayati, H.; Creatura, M. and Millan, J.R. (2010). Collecting complex activity datasets in highly rich networked sensor environments. In: *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, pp. 233-240.

21. Sarawagi, S.; Agrawal, R. and Megiddo, N. (1998). Discovery-driven exploration of OLAP data cubes. In: Schek, H.J.; Alonso, G.; Saltor, F. and Ramos, I. (Eds). *Advances in Database Technology – EDBT'98*. Berlin, Heidelberg, pp. 168-182.

22. Schauer, S.; Stamer, M.; Bosse, C.; Pavlidis, M.; Mouratidis, H.; König, S. and Papastergiou, S. (2017). An adaptive supply chain cyber risk management methodology. In: Kersten, W.; Blecker, T. and Ringle, C.M. (Eds.). *Digitalization in supply chain management and logistics. Smart and digital solutions for an industry 4.0 environment*. Berlin, pp. 255-273.

23. Sekaran, U. and Bougie, R. (2013). *Research methods for business. A skill-building approach*. West Sussex.

24. Sheskin, D.J. (2011). *Handbook of parametric and nonparametric statistical procedures*. London.

25. Smalheiser, N. (2002). Informatics and hypothesis-driven research. In: *Embo reports*, Vol. 3 Nr. 8, pp. 701-803.

26. Strong, D.; Lee, Y. and Wang, R. (1997). Data quality in context. In: *Communications of the ACM*, Vol. 40 Nr. 5, pp. 103-110.

27. Taub, E.A. (2018). Car navigation system splot a course forward against phone apps, in: *New York Times*, February 1st, 2018.

28. Umanath, N.S. and Scamell, R.W. (2014). *Data modeling and database design*. 2nd ed. New York.

29. Wagner, J. and Kontny, H. (2017). Use-case of self-organizing adaptive supply chain. In: Kersten, W.; Blecker, T. and Ringle, C.M. (Eds.). *Digitalization in supply*

chain management and logistics. Smart and digital solutions for an industry 4.0 environment. Berlin, pp. 255-273.

30. Xue, L.; Zhang, C.; Ling, H. and Zhao, X. (2014). Risk mitigation in supply chain digitization: system modularity and information technology governance. In: *Journal of Management Information System*. Vol. 30 No.1, pp. 325-352.