Towards a Reference Model for Knowledge Driven Data Provision Processes

Wei Min Wang[1], Maurice Preidel[1], Bernd Fachbach[2], Rainer Stark[1]

[1] Chair of Industrial Information Technology, Technical University of Berlin, Pascalstr. 8-9, 10587 Berlin, Germany
[2] Virtual Vehicle Research Center GmbH, Inffeldgasse 21a. 8010 Graz, Austria
{w.wang, m.preidel, r.stark}@tu-berlin.de
Bernd.Fachbach@v2c2.at

Abstract. Value creation in most business areas takes place in networks that involve a wide range of stakeholders from various disciplines within and beyond company borders. Collaboration in such networks require the exchange of knowledge that is manifested in digital artefacts and consequently in data. As the utilization of that "hidden" knowledge has become increasingly important, the provision of relevant data in sufficient quality has also become crucial. This article proposes a reference model for knowledge driven data provision processes that is developed within a research project at the Virtual Vehicle Research Center GmbH for a future networked engineering environment. It describes a systematic process to drive operationalization of data provision from knowledge requirements to identify, extract and provide raw data until the application of such data sets. Still, the model in its current state is only applicable by descriptive means and needs further development and validation in practical use cases.

Keywords: data provision, reference model, knowledge discovery, networked engineering,

1 Introduction

Engineering is increasingly driven by data of various types and categories, especially in industry sectors with highly complex products such as the automotive and the railroad sector. Market demand and legal requirement are transformed into technical requirements. Product architectures are described in models and the maturity of the complex products is managed by KPIs. Numerical simulations that enable a deep insight in the behavior of a product long time before the first part is physically build requires various input data and produces a huge amount of result data. Moreover, an increasing number of development partners are involved, which have to be tracked adequately [1]. Above all Digital Twins of the product require an extensive dossier of numerous data and information about a specific real vehicle along the complete lifecycle [2],[3].

Communication and cross-disciplinary collaboration are crucial aspects of current vehicle development – either inside a company as well cross-enterprise. Looking at the

same base of relevant and consistent data from different perspectives is relevant for high effectivity and quality [4]. Generic data provision services, context enrichment, and continuous aggregation are key aspects for future effective development and lifecycle integration [5]. A high quality of consistent data is the necessary base for agile engineering methodology, for networking collaboration as well as for wide-spread implementation of supporting data analytics and artificial intelligence(AI)-based methods for knowledge discovery (KD) [2]. Cross disciplinary development teams have to have efficient access to related data, to analogies from other development activities, to knowledge by analyzing available information [6] and to effect models from specific topics, domains and disciplines.

Hence, the extraction and utilization of potential knowledge from such large data sets has become even more important to industrial practitioners as they expect considerable potentials to optimize current processes and to enable novel business models [7]. In order to extract potential knowledge from large data sets, these have to be evaluated, e.g. by statistical analysis during data mining (DM) endeavors [6]. Therefore, data have to be identified, prepared and delivered. However, most existing reference models for Knowledge Discovery and Data Mining (KDDM) do not include detailed descriptions of such a data provision process per se, but rather point out single aspects (e.g. data prospecting [8]) or summarize it in few steps (e.g. data understanding and data preparation [9]). Additionally, existing KDDM reference model mainly focus on single projects. Hence, they lack the ability to support the establishment of novel data-driven product concepts such as product-service systems (PSS) [10] or smart products [11], which require a constant supply of operational data to enable customer specific service adaptions.

The efficient application of KDDM in engineering is also a subject to the K2 research project V-Lab – Future Engineering Lab at the Virtual Vehicle Research Center GmbH (V2C2). Collaboratively with industrial companies from the automotive and the railroad sector, innovative processes and tools for networked engineering in future are investigated. A key aspect of data provision of this research is to integrate KDDM activities as continuous activities within the organizational processes to enable streamlined and automated (or respectively AI-aided) knowledge exchanges. Therefore, the authors developed a generic reference model for data provision processes that considers the data provision as a part of the process organization. It refers to elements of current frameworks for KDDM and data provision (see section 2) and suggests detailed process steps to for a systematic data provision process that drive operationalization of data provision from knowledge requirements to identify, extract and provide raw data until the application of such data sets (see section 3.2). Moreover, technical and organizational prerequisites for the proposed model are pointed out (see section 3.1).

2 State of the Art

In context of KDDM in product development, the role of data provision processes is to identify, collect and prepare relevant and reusable data in order to support knowledge (re)use in PDP [12]. For KDDM in industrial practice, the CRISP-DM (Cross Industry

Standard Process for Data Mining) reference model has been found to be the most widely used model compared to other models such as SEMMA (Sample, Explore, Modify, Model, and Assess) or KDD (Knowledge Discovery in Databases) [13],[14]. Therefore, the proposed reference model will only refer to CRISP-DM.

Within CRISP-DM, six phases are described: 1) Business Understanding, 2) Data Understanding, 3) Data Preparation, 4) Modelling, 5) Evaluation and 6) Deployment [9]. For the reference model, the phases 1-3 are most relevant, because those phases define the data need for the data mining project. The phases 4-6 are also considered, but as they highly focus on an actual data-mining model, there is a greater difference to the purpose of the reference model (data provision vs. data mining). Nevertheless, the phases 1-3 present a good overall framework as a starting point for creating a reference model for data provision. A clear understanding of the business objectives (phase 1) helps to focus on the right data in order to have a reasonable return on invest for real world applications. Data understanding (phase 2) is crucial for every data provision and data mining project: The available data must be understood from a domain knowledge perspective and needs to be linked to the identified business case from phase 1. Data preparation (phase 3) is all about making the data applicable for the business case. This is achieved by aggregating, cleaning and transforming the relevant data into a data structure, which is usable for subsequent phases. While CRISP-DM is a great starting point for the reference model for data provision processes, it also has notable limitations [15],[16]: Firstly, CRISP-DM is a framework for individual projects, whereas the reference model for data provision aims for continuous provision of relevant data in sufficient quality within business processes [6]. Secondly, CRISP-DM is a general framework suitable for many industries and use cases with a high abstraction level [17]. In contrast, the proposed reference model focuses on specific phases for data provision and the context of product development.

To evaluate further existing reference models for data provision processes, a literature review was conducted on Scopus (https://www.scopus.com) based on the keyword matrix presented in table 1. The initial query results were reduced according to the limiting topic keywords (see table 1) and to the period from 2009 to 2019.

Table 1 – Keyword matrix applied for literature review

Search Keywords (OR)	**Limiting Topic Keywords (OR)**
Data provision	Engineering
Data supply	Product development
Data retrieval	Data science
Data procurement	Data analytics
Data preprocessing	Produktentwicklung (German)
Information retrieval	Konstruktion (German)
Datenbereitstellung (German)	
Datenversorgung (German)	

The abstracts of the 57 remaining matches were then evaluated for direct (explicit mentions in abstract) or indirect (e.g. as paraphrase) references to the topic of data provision. After this step, 18 candidates remained of which 15 were available online.

As already noted in KDDM reference models, the evaluation of these sources also revealed that the data provision process per se is rarely considered specifically, but rather briefly mentioned as a component of application cases (e.g. data stream mining) or data management concepts [18]–[22]. Some authors use application cases to describe the potential benefits of analyzing data and its properties [23]–[25], but stay quite unspecific about how data is actually provided. Only one article by Thoben and Lewandowski provides a generic data provision framework for the utilization of operating data for product development [12]. Based on a closed-loop PLM approach [26], they describe four major requirements for a data provision process as well as a three-pillar concept comprising "Technical prerequisites", "Methodological concepts" and "Procedural implications" [12].

3 Reference Model for Data Provision Processes

As mentioned above, the proposed reference model draws on existing frameworks and integrates elements of those with typical stages of a generic data provision process. In contrast to CRISP-DM, the proposed model particularly addresses the data provision process and the causal relationships therein. Hence, it focuses is on the chronological and logical order of process steps as well as necessary input and output artefacts.

The proposed model is intended to support the establishment of a continuous data provision process within organizational process structures. To describe the elements of the reference model, the detailed steps are described in context of a hypothetical use case from the research project at V2C2. In that case, the simulation of a high-speed test (HST) has to be carried out during an engineering process in the automotive sector. In practice, HSTs are carried out to estimate the maximum speed of a vehicle. They can either be numerical simulations, physical simulations (road testing), or even a hybrid approach by combining numerical and physical simulation. The adequate method approach is selected according to the phase of development process and the availability of data or real parts. However, within the hypothetical scenario is assumed that there is a working simulation model for HSTs and that all relevant models, formulas and data are available. The implications for the proposed reference model for data provision process are described in the following section.

3.1 Technological and Organizational Prerequisites

To enable a streamlined and (partially) automated end-to-end data provision process, some essential technological and organizational requirements have to be met. For the reference model, the authors especially consider four clusters of prerequisites, which are described below. However, is has to be mentioned, that some of them are not realized yet or show a future scenario. In the context of the model development it is assumed that these prerequisites are already fulfilled.

Seamless integration of the processes: In order to enable an efficient and end-to-end data provision process, all organizational processes have to be modelled at a sufficient level of detail and implemented in operative business. Only then is it possible to create sufficient transparency regarding the need for knowledge, information and data [12]. Such a process modeling can be realized with an activity-based approach [27]. In this approach, all activities in the business context are regarded as a sequence of atomic activities (smallest, non-divisible activity), which are carried out by specific actors in order to achieve a specific goal in the process. These activities are described in the context of the environment in which they are embedded, namely the organization and its specific processes, the tools and IT systems used, and the physical and virtual artefacts used as input or produced as output. The activity-based approach enables a holistic description of the value-creating processes in the company and the identification of their reciprocity and dependencies. [27]

Consistency of data and models: In order for data to be identified and extracted, the data itself or the artifacts in which they are manifested have to be available in the respective value network. This requires both vertical (e.g. across company departments) and horizontal (e.g. over the product lifecycle) consistency and traceability of the data. According to the Model-based Systems Engineering (MBSE) approach, digital models can be used to link development data from different disciplines and development phases. The consistent deployment of digital models in all organizational processes and their company-specific orchestration can reduce the complexity of distributed development tasks and ensure the traceability of information and data. [28]

Integration of IT systems and their respective data sources: In order to realize a comprehensive data supply process, it is also necessary to integrate the IT systems used to manage the meta and user data [12]. Following the Product Lifecycle Management (PLM) approach, a PLM backbone can be used for that purpose, e.g. by means of a PDM/PLM system. This system would play the role of a central data management instance for all engineering processes and integrate other IT systems (e.g. ERP, PPS), tools (e.g. CAx) and specialized data warehouses from other business units (e.g. sales) via defined interfaces. For this purpose, it must also be ensured that the organizational processes are sufficiently digitalized and that data arising from any organizational activities are entered into the corresponding system [12],[28].

Company-specific knowledge model: The knowledge of a company constitutes the basic semantic framework on which all data provision and data application processes are based. This knowledge base must include knowledge about the company's own products, processes and tools as well as about the collaboration partners in the value creation network [12]. From a strategic point of view, it represents the background architecture of a company’s value creation activities and therefore should

1. be reflected in the processes, models and tools
2. be implemented in the IT infrastructure (e.g. PLM backbone) and
3. be accessible to all internal company IT systems - and to some degree also to external collaboration partners (e.g. via standardized interfaces).

The implementation of such a knowledge base for linking different knowledge areas can be achieved, among other things, by means of semantic technologies (e.g. ontologies) [29]–[31]. Also, methods and tools from research areas such as Business

Intelligence can provide suitable guidance to develop a basic semantic framework of company data (e.g. the Kimball bus matrix [32]).

3.2 Model Description

The proposed reference model for data provision processes can be divided into three generic phases: *Clarify need*, *Data acquisition* and *Data application* (Figure 1).

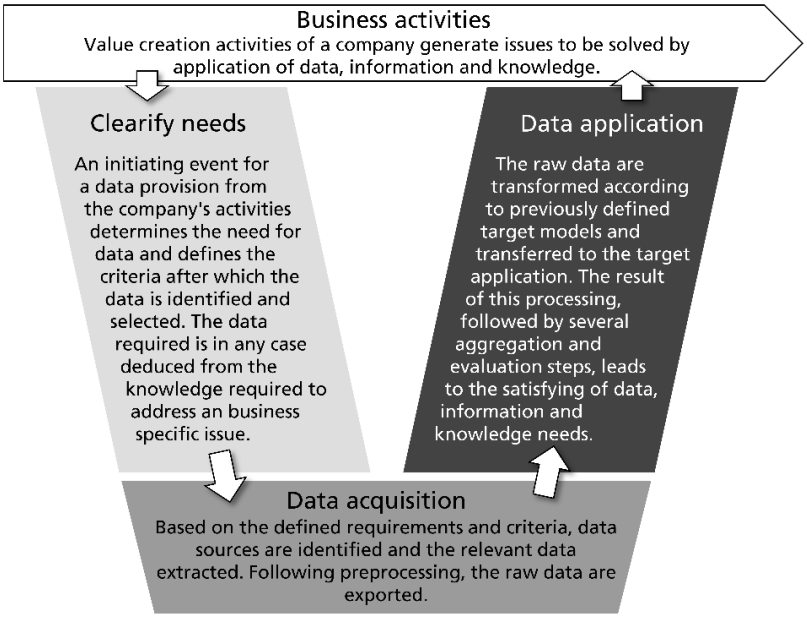

Figure 1: Generic phases of the proposed reference model for data provision processes

Each of these phases comprises specific stages as well as artifacts that serve as input or result as output. The detailed stages of the model are illustrated in Figure 2 and described in this section. In the phase of *Clarify needs*, the objectives for the data provision process are defined. As for any engineering activities, the initiator of a data provision process is always an issue that arises from a value creation activity of the respective organization [33]. By operationalizing this issue, the respective knowledge need can be identified (cf. business understanding in CRISP-DM [9]). In the present use case, the theoretical maximum speed of a vehicle at time X should be determined by a simulated HST in order to provide initial data for the variant definition of the engine and the transmission control. The knowledge need of a person who is hypothetically assigned with that task (e.g. a simulation engineer), can be described as: "What is the maximum speed that the vehicle can achieve at the given point in time X, according to the valid status of the relevant development data and simulation models, when linear acceleration is applied?" With the help of the company-specific knowledge base, this person can derive the relevant information from this question, e.g. which components and models need to be regarded for such a simulation. After identifying

the relevant information carriers (e.g. documents or CAD models), relevant attributes (e.g. max. torque) and quality criteria (e.g. completeness, correctness) have to be determined. Based on this defined data need, the *Data acquisition* activities can be initiated. In the first step, the corresponding data (e.g. valid CAD models in the PDM system) have to be identified and localized. The type of source system (e.g. company-owned vs. external) also specifies the data provenance, which has a great influence on the quality of the data (e.g. reliability, correctness). The properties of the source systems further determine the effort to utilize the data. The source data model as well as the source data formats and types determine how attributes are encoded in the source system and how they can be extracted or transformed (cf. data understanding in CRISP-DM [9]). The identified data has to be extracted from the source systems then and stored temporarily for data pre-processing. During preprocessing, transformations take place with the aim of preparing the extracted data specifically for later use, for example, by deleting invalid data sets or scaling data with different dimensions. The pre-processed collection of raw data can then be loaded to dedicated storage locations (e.g. a data warehouse) (cf. data preparation in CRISP-DM [9]).

This raw data can then be further processed in *Data application*, depending on the requirements of the exploration method. This results in data that meet the defined data requirements and can be fed into an exploration model. This can be either an already existing analysis model (i.e. already trained and in use) or a new model that is to be which experts can assess whether the business issues have been adequately answered and thus the knowledge need have been satisfied.

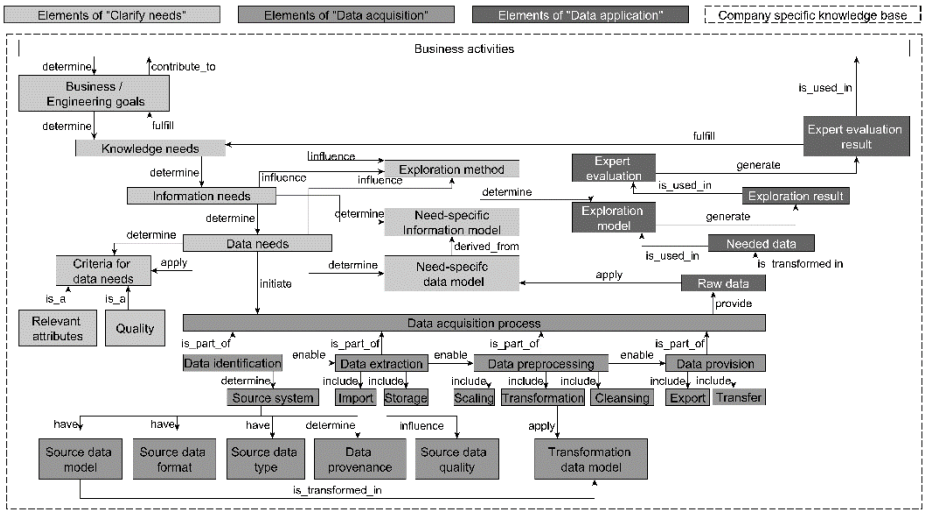

Figure 2. Detailed steps of the proposed reference model for data provision

4 Conclusion and Outlook

The proposed reference model for data provision processes suggests detailed generic stages to support and facilitate the operative implementation of such processes. Based on the HST example, the model elements were described and explained. In contrast to classical KDDM reference models such as CRISP-DM that are limited to support the implementation of individual DM projects, the detailed stages of the proposed model (see Figure 2) allow companies to understand how they can integrate the data provision process as a continuous process in their business organization. Especially, in the light of novel data-driven product concepts and business models as well as the growing importance of digital twins, it is necessary to understand data provision not only as isolated activities within DM projects, but also as a continuous element of value creation activities. By establishing routines for knowledge-driven data provision processes, companies can exploit the potentials of advanced analytics as well as AI-capabilities to improve their engineering processes (e.g. shorten development time) as well as their products and services (e.g. by continuous adaption to customer requirements) [34].

However, the reference model so far can only be used as a descriptive model or a canvas to plan or analyze data provision processes. The stages within the model need to be further complemented with technological solutions and methodological approaches. In particular, the transformations from knowledge needs to information and data needs as well as vice versa pose great challenges and will require further research [12],[34]. Furthermore, the reference model has to be further validated on practical use cases to identify potentials for improvement both on a conceptual and methodological level.

Acknowledgements
The publication was written during the research within K2 research project V-Lab – Future Engineering Lab (Virtual Vehicle Research GmbH). The authors would like to acknowledge the financial support within the COMET K2 Competence Centers for Excellent Technologies from the Austrian Federal Ministry for Climate Action (BMK), the Austrian Federal Ministry for Digital and Economic Affairs (BMDW), the Province of Styria (Dept. 12) and the Styrian Business Promotion Agency (SFG). The Austrian Research Promotion Agency (FFG) has been authorized for the program management.

References

1. Vajna S. (ed.): Integrated design engineering. Springer Berlin Heidelberg, Berlin, Heidelberg (2014).
2. Barricelli B. R., Casiraghi E., Fogli D.: A survey on digital twin: definitions, characteristics, applications, and design implications. IEEE Access, **7**, 167653–167671 (2019). DOI: 10.1109/ACCESS.2019.2953499
3. Stark R., Damerau T.: Digital twin. In 'cirp encyclopedia of production engineering' (eds.: Chatti S., Tolio T.) Springer Berlin Heidelberg, Berlin, Heidelberg, 1–8 (2019).
4. Neumann F. (ed.): Analyzing and modeling interdisciplinary product development. Springer Fachmedien Wiesbaden, Wiesbaden (2015).

5. Voet H., Altenhof M., Ellerich M., Schmitt R. H., Linke B.: A framework for the capture and analysis of product usage data for continuous product improvement. Journal of Manufacturing Science and Engineering, **141**, (2019).
6. Kurgan L. A., Musilek P.: A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review, **21**, 1–24 (2006). DOI: 10.1017/S0269888906000737
7. Markus M. L.: Toward a theory of knowledge reuse. Types of knowledge reuse situations and factors in reuse success. Journal of Management Information Systems, **18**, 57–93 (2015). DOI: 10.1080/07421222.2001.11045671
8. Anand S. S., Büchner A. G.: Decision support using data mining. Financial Times Management, London, (etc.) (1998).
9. IBM: Ibm spss modeler crisp-dm guide (2016).
10. Exner K., Stark R., Kim J. Y.: Data-driven business model a methodology to develop smart services. in 'Proceeding of the 2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC). pp.146–154 (2017).
11. Porter M. E., Heppelmann J. E.: How smart, connected products are transforming companies. Harvard business review, **93**, 96–114 (2015).
12. Thoben K.-D., Lewandowski M.: Information and data provision of operational data for the improvement of product development. In 'product lifecycle management in the era of internet of things' (eds.: Bouras A., Eynard B., Foufou S., Thoben K.-D.) Springer International Publishing, Cham, 3–12 (2016).
13. Ana Azevedo, Manuel Filipe dos Santos: Kdd, semma and crisp-dm: a parallel overview. (2008).
14. Piatetsky G.: Crisp-dm, still the top methodology for analytics, data mining, or data science projects - kdnuggets. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, accessed 17-04-2020 (2014).
15. Four problems in using crisp-dm and how to fix them - kdnuggets. https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html, accessed 17-04-2020 (2020).
16. Wiemer H., Drowatzky L., Ihlenfeldt S.: Data mining methodology for engineering applications (dmme)—a holistic extension to the crisp-dm model. Applied Sciences, **9**, 2407 (2019).DOI: 10.3390/app9122407
17. Huber S., Wiemer H., Schneider D., Ihlenfeldt S.: Dmme: data mining methodology for engineering applications – a holistic extension to the crisp-dm model. Procedia CIRP, **79**, 403–408 (2019).
18. Shentu J., Zheng M.: Mechanism design of data management system for nuclear power, 21–29.
19. Tan J. S. K., Ang A. K., Lu L., Gan S. W. Q., Corral M. G.: Quality analytics in a big data supply chain: commodity data analytics for quality engineering. in 'Proceeding of the TENCON 2016 - 2016 IEEE Region 10 Conference. Singapore', 3455–3463 (2016).
20. Ramírez-Gallego S., Krawczyk B., García S., Woźniak M., Herrera F.: A survey on data preprocessing for data stream mining: current status and future directions. Neurocomputing, **239**, 39–57 (2017).
21. Madenas N., Tiwari A., Turner C. J., Peachey S., Broome S.: Improving root cause analysis through the integration of plm systems with cross supply chain maintenance data. The International Journal of Advanced Manufacturing Technology, **44**, 2749 (2015).
22. Lin H.-T., Chi N.-W., Hsieh S.-H.: A concept-based information retrieval approach for engineering domain-specific technical documents, 349–360.

23. Al-Utaibi K. A., El-Alfy E.-S. M.: Intrusion detection taxonomy and data preprocessing mechanisms. Journal of Intelligent & Fuzzy Systems, **34**, 1369–1383 (2018). DOI: 10.3233/JIFS-169432
24. Alkhalil A., Ramadan R. A.: Iot data provenance implementation challenges. Procedia Computer Science, **109**, 1134–1139 (2017).
25. Hassler A. P., Menasalvas E., García-García F. J., Rodríguez-Mañas L., Holzinger A.: Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Medical Informatics and Decision Making, **19**, 33 (2019).
26. H.-B. Jun, D. Kiritsis, P. Xirouchakis: Closed-loop plm. In 'advanced manufacturing. an ict and systems perspective' (eds.: Taisch M., Thoben K.-D., Montorio M.) CRC Press, 79–87 (2007).
27. Lünnemann P., Stark R., Wang W. M., Manteca P. I.: Engineering activities — considering value creation from a holistic perspective. In '2017 international conference on engineering (ice/itmc)', 315–323 (2017).
28. Eigner M., Gilz T., Zafirov R.: Interdisziplinäre produktentwicklung - modellbasiertes systems engineering. PLMportal, München (2012).
29. Akmal S., Shih L.-H., Batres R.: Ontology-based similarity for product information retrieval. Computers in Industry, **65**, 91–107 (2014).
30. Borsato M.: Bridging the gap between product lifecycle management and sustainability in manufacturing through ontology building. Computers in Industry, **65**, 258–269 (2014). DOI: 10.1016/j.compind.2013.11.003
31. Stark R., Wang W. M., Pförtner A., Hayka H.: Einsatz von ontologien zur vernetzung von wissensdomänen in der nachhaltigen produktentstehung am beispiel des sonderforschungsbereiches 1026 – sustainable manufacturing (2014).
32. Kimball R., Ross M.: The data warehouse toolkit. The definitive guide to dimensional modeling. John Wiley & Sons, Inc, Indianapolis, IN (2013).
33. Wang W. M., Lünnemann P., Preidel M., Stark R.: Wissen in Produktentwicklungsprozessen – Ein Aktivitäten-basierter Analyseansatz. In '15. gemeinsames kolloquium konstruktionstechnik. interdisziplinäre produktentwicklung' (eds.: Brökel K., Grote K.-H., Stelzer R., Rieg F., Feldhusen J., Müller N., Köhler P.) Universität Duisburg-Essen, Essen, 183–192 (2017).
34. Klein P., van der Vegte W. F., Hribernik K., Klaus-Dieter T.: Towards an approach integrating various levels of data analytics to exploit product-usage information in product development. Proceedings of the Design Society: International Conference on Engineering Design, **1**, 2627–2636 (2019).