ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/380005365

A TAXONOMY OF DATA SCIENCE PROCESS MODELS: INSIGHTS FROM SCIENCE AND PRACTICE

Conference Paper · June 2024

CITATIONS
0

READS
111

2 authors:

Stefan Rösl
Ostbayerische Technische Hochschule Amberg-Weiden
8 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Christian Schieder
Ostbayerische Technische Hochschule Amberg-Weiden
75 PUBLICATIONS **223** CITATIONS
SEE PROFILE

All content following this page was uploaded by Stefan Rösl on 23 April 2024.

The user has requested enhancement of the downloaded file.

A TAXONOMY OF DATA SCIENCE PROCESS MODELS: INSIGHTS FROM SCIENCE AND PRACTICE

Completed Research Paper

Stefan Rösl, Technical University of Applied Sciences Amberg-Weiden, Germany, s.roesl@oth-aw.de

Christian Schieder, Technical University of Applied Sciences Amberg-Weiden, Germany, c.schieder@oth-aw.de

Abstract

Data science (DS) projects have become indispensable across a wide range of industries. Various data science process models (DSPM) have been proposed to provide structure and guidance for the project process. Despite recognizing the efficiency of DSPM, their practical adoption has been limited. However, the increasing frequency of data science projects and the associated repetitive process execution highlight the importance of this gap and justify further investigations. To support the application of DSPM by practitioners and guide further research, a comprehensive overview and systematic categorization of DSPM is necessary. Our research addresses this need by creating a taxonomy of DSPM. We utilized both inductive and deductive research methods to develop a taxonomy that includes three categories, 13 dimensions, and 67 characteristics. The taxonomy is demonstrated using a sample of 35 DSPM from the literature and evaluated through twelve expert interviews with DS experts and researchers.

Keywords: Process Model, Data Science, Taxonomy, Data Science Project Management.

1 Introduction

Data science (DS) has become a critical aspect of modern business strategies. It enables organizations to derive valuable insights from data, which can improve performance (Cao, 2017; Martinez et al., 2021). A growing number of DS projects show that organizations are eager to use insights from DS projects to make better decisions, reduce costs, and improve business processes (Saltz and Krasteva, 2022). Data science process models (DSPM) ensure a standardized and systematic approach to executing DS projects (Haertel et al., 2022a). Saltz et al. (2018) state that 85 % of DS professionals acknowledge the efficiency of DSPM, yet only 18 % follow a documented methodology. This observation indicates the limited applicability of the existing approaches (Das et al., 2015). Several studies suggest that the existing body of DSPM requires further investigation in the field of project management (Martinez et al., 2021; Saltz and Krasteva, 2022).

Prior research has mainly examined DSPM from a descriptive or theoretical perspective and has focused on analyzing only a limited number of common DSPM. This has led to partially redundant studies, with none offering a comprehensive classification framework. The research on different types of DSPM is insufficient, and the relevant dimensions and characteristics are still unclear. There is a lack of a structured and validated framework for effectively categorizing DSPM. This is a significant gap for two reasons: (1) the evolution from decision support systems to data mining, analytics, and DS, in conjunction with process models that have existed for decades, underscores the need to systematically bundle and organize DSPM knowledge, and (2) the increasing number of DS projects, coupled with their repetitive execution, emphasizes the importance of addressing the process perspective. A taxonomy is essential because it allows researchers to rely on a solid knowledge base by comprehensively categorizing DSPM. Practitioners can benefit from a better understanding of DSPM and their characteristics to select an appropriate model for their specific needs and contexts. Researchers can benefit from a structured analysis of DSPM to identify potential research opportunities and contribute to future advancements.

Therefore, our inquiry is guided by the following research question (RQ):

RQ: What are the key dimensions and characteristics that structure a Data Science Process Model (DSPM) as offered in current practice and academic literature?

We chose a taxonomy as our artifact because it is an important tool that provides both researchers and practitioners with valuable insights about complex domains through description, classification, or analysis (Nickerson et al., 2013). To answer the research question, we adopt the updated taxonomy development process by Kundisch et al. (2022). This approach has already been recognized in the information systems (IS) domain and builds upon the widely used method from Nickerson et al. (2013).

In the following section, we provide a brief overview of the theoretical foundations of DSPM, DS-related taxonomies, and related work in the field. Section 3 presents our methodological approach, followed by the data collection, development iterations, and evaluation of our taxonomy in Section 4. Section 5 presents the final taxonomy of DSPM and a detailed discussion of its dimensions. Finally, we discuss the impact of our research, limitations, and future work.

2 Research Background

This section introduces our understanding of DS based on scientific literature. We also examine DS-related taxonomies and differentiate our work from other studies on DSPM.

2.1 Data science and process models

In academic literature, there is no consistent distinction between the various terms used to describe activities related to data analytics (Bichler et al., 2017). In our research, we follow Bichler et al. (2017), who use a definition formulated by van der Aalst (2016) and summarize DS as an interdisciplinary field that focuses on creating value from different types of data and encompasses processes from extraction and transformation to analysis and presentation of insights, considering ethical, social, legal, and economic aspects. The ongoing digitalization has brought DS to the forefront of business strategies, resulting in a continuously increasing number of DS projects (Cao, 2017; Saltz and Krasteva, 2022). A DSPM provides systematic and structured support for efficiently executing DS projects (Saltz and Shamshurin, 2015). Adapted from Saltz and Hotz (2021), our working definition of DSPM is: a DSPM is a system of systematic and structured approaches, processes, or frameworks used to manage and execute DS projects while encompassing a range of stages, activities, and techniques to collect and analyze data by extracting valuable insights. Processes or organizational routines consist of two key elements: the formal definition or outline and the actual execution (Feldman and Pentland, 2003). Based on the provided definitions, we argue that a DSPM is the formal outline of a project, while the increasing frequency makes DS projects repetitive in actual process execution. Table 2 (c.f. Subsection 4.2) presents an overview of the most common DSPM, such as Knowledge Discovery in Databases (KDD), Sampling, Exploring, Modifying, Modelling, and Assessing (SEMMA), or Cross Industry Standard Process for Data Mining (CRISP-DM). We present the widely recognized CRISP-DM and the recently developed and promising Data Science Process Model (DASC-PM) as illustrative examples.

CRISP-DM offers a comprehensive and industry-independent structure for data mining projects. This reference model highlights the iterative nature of DS projects, covering the phases of *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, and *deployment* (Chapman et al., 2000; Shearer, 2000). The first four phases of CRISP-DM feature iterative interactions and are further detailed by generic tasks, specialized tasks, and outputs. The *business understanding* phase aims to clarify project objectives and business requirements, which are then translated into a data mining problem. The *data understanding* phase involves data collection and exploratory analysis to provide initial insights, and to identify data quality issues. *Data preparation* transforms raw data into a format suitable for modeling. This phase contains data selection, transformation, and cleaning activities and may require different preparations for each specific model. The *modeling* phase involves creating a model to describe a given problem and enclosing the model performance assessment. The evaluation phase focuses on the reviewing the achievement of the business objectives. *Deployment* can vary in scope, from generating a report of the findings to implementing a fully operational system. The balance between detailed guidance and a conceptual framework makes CRISP-DM versatile and generally applicable.

DASC-PM, introduced by Schulz et al. (2020), is an innovative framework developed collaboratively by over twenty experts to guide DS initiatives within the phases of *project order*, *data provision*, *analysis*, *deployment*, and *application*. The framework conceptualizes the process model under four overarching key areas: scientific procedures, application domains, IT infrastructures, and impact. Integrating the application domain into all project phases ensures alignment with scientific principles and project outlines. The analysis stage involves developing new methods or applying existing ones, resulting in the creation of valuable artifacts that are integrated into the application domain. Each phase can be further subdivided into specific activities that take into account relevant data, problem, and goal-oriented tasks. Schulz et al. (2020) highlight that DASC-PM successfully addresses critical DS challenges and outperforms existing models in terms of completeness, guidance, and practical impact.

2.2 Data science taxonomies and related literature

Taxonomies are a well-established approach in the IS domain for understanding, analyzing, and identifying similarities and differences within complex domains (Kundisch et al., 2022; Nickerson et al., 2013). We identified several specific taxonomies in the context of DS by searching the basket of eight IS journals and the proceedings of major IS conferences. These taxonomies cover technical, economic, or managerial perspectives. Table 1 provides a brief overview of the DS taxonomies.

Perspective	Purpose of the taxonomy	Reference	Source
Technical	Classify analytics use cases in auditing.	Krieger and Drews (2018)	ICIS
	Extend existing classification of machine learning-based fraud detection systems.	Matschak et al. (2022)	ECIS
Economic	Understand machine learning-driven business models in the business-to-business segment.	Vetter et al. (2022)	ECIS
	Describing, analyzing, and classifying approaches of data-based value creation.	Baecker et al. (2021)	ECIS
	Provide guidance through comprehensive structuring of data-driven business models.	Dehnert et al. (2021)	ECIS
Managerial	Understand, differentiate, and select AI-driven cybersecurity service based on the business model.	Gerlach et al. (2022)	ICIS
	Exploring and understanding different data cooperatives.	Zhu and Marjanovic (2022)	PACIS
	Understand and describe data ecosystems.	Gelhaar et al. (2021)	PACIS

Table 1. Data science taxonomies and their purpose.

Existing taxonomies typically categorize technical applications within specific industries, usage of distinct technologies, economic aspects such as value creation and business models, or data tools for strategic decisions. To the best of our knowledge, there is currently no taxonomy available that instantly addresses the structural or procedural elements of executing a DS project. Therefore, further investigation is necessary.

Some studies investigate DSPM directly. Haertel et al. (2022a) conducted a systematic literature review that identified 28 process models and focused on the aspects of project management by proposing a general Data Science Lifecycle (DSLC). The work of Saltz and Krasteva (2022) clustered 27 workflows into the types of standard, new, and adapted and investigated agile principles. Martinez et al. (2021) performed an in-depth analysis of 19 DS methodologies, examining their alignment with the proposed fundamental management pillars: project, team, and data and information management. Additional studies compared smaller sets consisting of common DSPM. Schulz et al. (2020) compared four well-known models based on to their horizontal completeness, vertical completeness, guidance, and reality and impact. As a result, weaknesses were identified and the introduction of DASC-PM was justified. Kutzias et al. (2023) compared seven existing models by analyzing literature and discussing them based on challenges identified through interviews. Another study examined eight DS methodologies by evaluating their strengths and limitations within the context of project execution (Oliveira and Brito, 2022). Existing research on DSPM has often focused on a limited number of established models, with each study concentrating on specific aspects based on their research focus. As a result, the knowledge base remains fragmented and incomplete, as individual studies fail to consider all available DSPM or comprehensively investigate all possible dimensions. The literature therefore lacks comprehensive and systematic understanding of DSPM.

Taxonomies are crucial in enabling researchers and practitioners to understand, analyze, and organize knowledge within a specific field (Nickerson et al., 2013; Kundisch et al., 2022). To address the gap in the missing taxonomic investigation of DS process execution, our objective is to enrich the current knowledge base by developing a scientifically grounded taxonomy of DSPM. By systematically determining their dimensions and characteristics, we contribute further to a more structured DSPM knowledge and improve the comprehensive understanding of DS process execution.

3 Research Method

We applied the extended taxonomy design process (ETDP) proposed by Kundisch et al. (2022) to design and build our taxonomy. This methodology builds upon the foundational taxonomy design approach in the IS research introduced by Nickerson et al. (2013), a method utilized in roughly two-thirds of all IS taxonomy developments post-2013 (Kundisch et al., 2022). While the guideline from Nickerson et al. (2013) is widely recognized as a standard in IS taxonomy creation, it has been criticized for lacking consistency and transparency and for not providing guidance on the evaluation (Kundisch et al., 2022). The ETDP addresses these shortcomings by incorporating additional design and evaluation steps, offering a more comprehensive and refined process. This recently refined taxonomy development approach has already been recognized and applied in the IS domain. The iterative ETDP aims to organize knowledge by categorizing and structuring it in a deductive and inductive way, as it is useful for scholars and practitioners (Kundisch et al., 2022). The process consists of 18 steps aligned with the six primary activities of the design science research (DSR) methodology proposed by Peffers et al. (2007). These stages serve as a roadmap for researchers to design a systematic and rigorous artifact in the form of a taxonomy. We describe the DSR activities below by mapping the ETDP steps.

In the first DSR activity, **identify the problem and motivate** (1-3), the ETDP specifies the observed phenomenon (1) that requires classification. This initial step outlines the problem and establishes the motivation for the taxonomy. The context of the taxonomy development will be further framed by specifying the target user groups (2) and the intended purpose(s) of the taxonomy (3). Notably, all steps in this DSR activity were added in the update by Kundisch et al. (2022).

The second stage, **defining the objectives of a solution** (4-5), involves determining the meta-characteristics (4) of the taxonomy. Additionally, researchers establish the subjective and objective ending conditions and the evaluation goals (5). These overarching attributes and criteria will guide the development of the taxonomy and evaluate its success. Kundisch et al. (2022) only included the additional evaluation goals in this stage.

The **design and development** stage (8-10) of the ETDP is characterized by the iterative development process, which starts by selecting a building approach (6). Researchers must decide between an inductive, empirical-to-conceptual (E2C) or a deductive, conceptual-to-empirical (C2E) approach based on the availability of objects or relevant insights in the knowledge base (Nickerson et al., 2013). In the E2C method, objects are first identified (Step 7e), followed by the identification of characteristics and grouping of objects (Step 8e) and grouping of characteristics (Step 9e) to dimensions. Conversely, in the C2E approach, dimensions and characteristics are conceptualized first (Step 7c), and then objects are aligned with these dimensions and characteristics (Step 8c). At the end of each approach, the actual taxonomy draft is created or revised (10). Design and development have been adopted from Nickerson et al. (2013); only (10) is a separate step in the ETDP, which was previously integrated into other steps.

In the **demonstration** stage (11-12), the drafted taxonomy is demonstrated by checking the objective ending conditions (11). The evaluation starts if the conditions are met (Step 12); otherwise, the taxonomy is revised and refined. Kundisch et al. (2022) consider the following four entry points for design iterations in the ETDP: (2) specify target user group(s), (4) determine meta-characteristic, (6) choose building approach, and (10) create/revise taxonomy. The guideline from Nickerson et al. (2013) only checks the ending conditions after the iterative building approaches. The procedure ends if they are met, but if not, it restarts with the selection of the building approach. Therefore, all subsequent activities are updated steps of the ETDP.

The next DSR activity is the **evaluation** stage (13-17). Researchers now need to assess the subjective ending conditions (13,14). If the criteria are not met, the entry points initiate a restart of another iteration. With an agreement reached, the ETDP process moves to its subsequent steps. First, configure the evaluation (15) considering the taxonomy´s evaluation goals (5), purpose (3), and target user group (2). Next, the evaluation is performed (16) to determine if the final taxonomy achieves the evaluation goals (17) and criteria to be a useful taxonomy. While failing the criteria requires a return to one of the entry points, satisfaction leads to the next DSR activity.

Finally, the **communication** stage (18) presents the result and outcome of the ETDP. Researchers report on the taxonomy (18) by specifying their dimensions and characteristics. This is crucial for the taxonomy to be adopted by the intended user groups.

By following the ETDP in line with the DSR methodology, researchers can ensure that a taxonomy is developed systematically and contributes to the knowledge base.

4 Taxonomy Development Process

Our research follows the approach proposed by Kundisch et al. (2022). Below we outline how we adapted the ETDP and structured the process based on the DSR activities into four elements: (1) problem and solution objectives, (2) design, development, and demonstration, (3) evaluation, and Section 5 will present (4) the communication.

4.1 Problem and solution objectives

DSPM are a complex and multifaceted phenomenon, making a taxonomy a suitable approach for systematic classification and more precise understanding. As discussed in Sections 1 and 2, there is currently no existing taxonomy of DSPM. Therefore, we propose a taxonomy to support researchers and practitioners, enabling a sufficient classification and aiding the selection and adaptation of methodologies in practical and academic settings. The meta-characteristics must be set concerning the overall purpose of the taxonomy and serve as a foundation for determining the subsequent characteristics (Nickerson et al., 2013). We defined the meta-characteristics as *specification*, *assistance*, and *requirements and goals* of DSPM to address the previously specified purpose. The ending conditions serve as checkpoints in the iterative development process. The procedure can be stopped if the taxonomy fulfills all the conditions and achieves stability. We adopted the objective ending conditions proposed by Nickerson et al. (2013) and formulated the following: i) all relevant objects have been examined, ii) no merge or split of objects has occurred, iii) each characteristic of each dimension has been selected by at least one object, iv) no new dimensions or characteristics have been merged, split, or added, v) every dimension is unique, and vi) each characteristic is unique within its dimension. Additionally, the authors must agree to the subjective ending conditions to evaluate whether the taxonomy is concise, robust, comprehensive, extendible, and explanatory (Nickerson et al., 2013). We formulated evaluation goals, as the taxonomy should support users by describing, classifying, and analyzing the phenomenon of DSPM to ensure a reliable taxonomy design with an ex-post evaluation (Kundisch et al., 2022).

4.2 Design, development, and demonstration

Our iterative research design is illustrated in Figure 1. Starting with a C2E approach, we iteratively refined our taxonomy through E2C iterations based on the organized sub-samples of objects. Five iterations, four inductive and one deductive, were needed until the taxonomy achieved stability with no further modifications required and all ending conditions being satisfied (Nickerson et al., 2013).

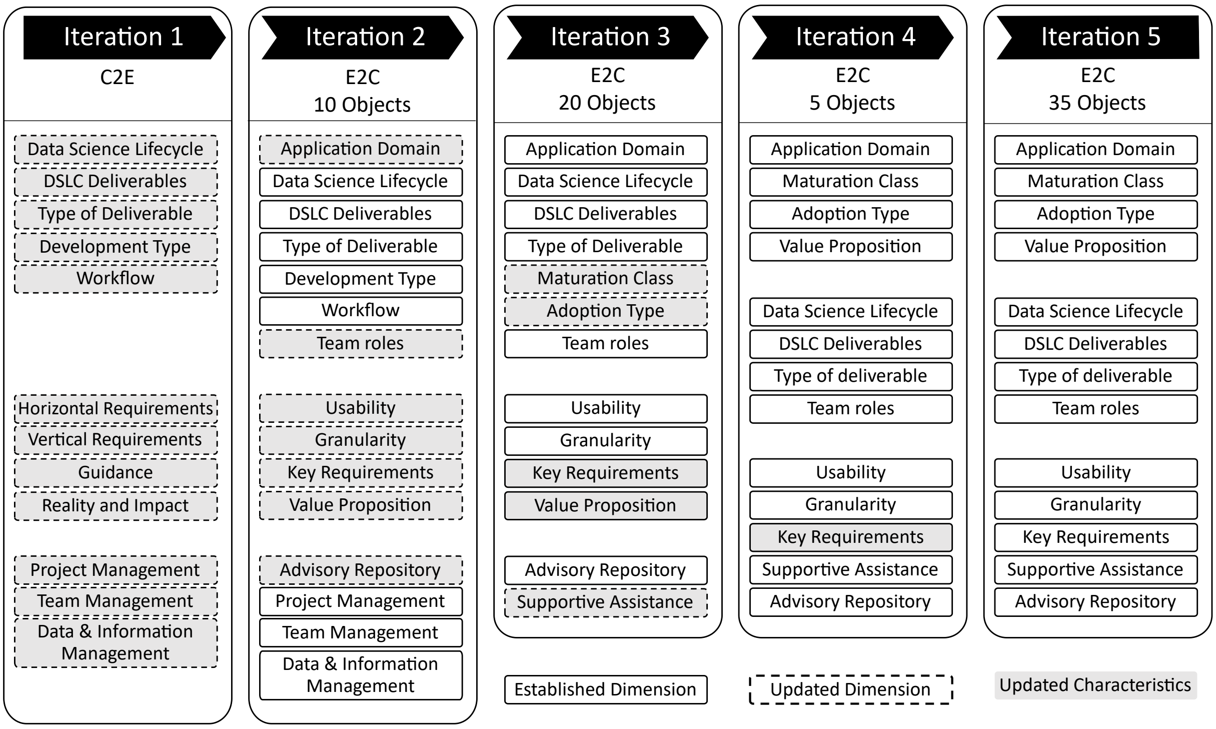

Figure 1. Iterative taxonomy development process.

According to the recommendations by Kundisch et al. (2022), we started with a C2E approach for the **first iteration**, as the existing knowledge base holds relevant insights about the phenomenon under consideration. In combination with the authors domain knowledge, we build the foundation for our taxonomy from extant research. Thus, we collected the phases of the general *Data Science Lifecycle* (Haertel et al., 2022a), their *deliverable type* (Haertel et al., 2022b), as well as the *development type* of DSPM adopted from Saltz and Krasteva (2022). The *workflow* between the phases and the overall support score in terms of *project management, team management*, and *data & information management* were derived by Martinez et al. (2021). The requirements and goals dimensions were adopted from Schulz et al. (2020) – *horizontal requirements, vertical requirements, guidance*, and *reality and impact.* The researcher finally conceptualizes the resulting Taxonomy T1, which consists of 12 dimensions and 50 characteristics and served as a starting point for the next iteration.

For the **second iteration**, we used the E2C approach as a number of objects were available (Kundisch et al., 2022). To identify relevant objects, we synchronized the three literature reviews: (1) Haertel et al. (2022a) with 27 DSPM, (2) Saltz and Krasteva (2022) with 28 DSPM, and (3) Martinez et al. (2021) with 19 DSPM. After synthesizing all identified methodologies, process models, or workflows and excluding duplicates, we identified 60 distinct articles. To enhance the quality of the taxonomy, we established three eligibility criteria: i) scientific rigor, ii) practical relevance, and iii) subject-matter relevance. We ensured adherence to these criteria by selecting only conference papers, journal articles, and professional material from leading enterprises. The subject-matter criteria were evaluated through abstract reading and full-text screening, focused on identifying DSPM by their process flow diagrams or descriptions. We excluded articles that did not provide valuable content. The examined final set of 35 eligible objects for the forthcoming E2C approaches is provided in the online appendix.

We extracted a top-ten sample from the identified objects to obtain a meaningful orientation in the investigation. In this subset, we incorporated the most relevant models highlighted in the literature (Saltz and Krasteva, 2022; Haertel et al., 2022a; Schulz et al., 2020) as well as the three most cited of the remaining objects articles (cite-search performed via Google Scholar on October 24, 2023). Table 2 lists the DSPM of the top-ten subset with reference and their inclusion criteria.

Year of publication	**Process model**	**Reference**	**Inclusion criteria**	**No. of citations**
1996	Knowledge Discovery in Databases (KDD)	(Feyyad, 1996)	(Saltz and Krasteva, 2022; Schulz et al., 2020)	729
1996	Sample, Explore, Modify, Model, Assess (SEMMA)	(SAS Institute Inc., 2017)	(Schulz et al., 2020)	N/A
2000	Cross-Industry Standard Process for Data Mining (CRISP-DM)	(Chapman et al., 2000; Shearer, 2000)	(Haertel et al., 2022a; Saltz and Krasteva, 2022; Schulz et al., 2020)	2492, 1920
2015	Foundational Methodology for Data Science (FMDS)	(Rollins and IBM Corporation, 2015)	(Saltz and Krasteva, 2022)	52
2015	Big Data Management Framework (BDMF)	(Dutta and Bose, 2015)	Top cites	626
2016	Analytics Solutions Unified Method-Data Mining (ASUM-DM)	(IBM Corporation, 2016)	(Haertel et al., 2022a)	N/A
2016	Agile Delivery Framework (ADF)	(Larson and Chang, 2016)	Top cites	728
2017	Domino Data Science Lifecycle (Domino DS)	(Domino Data Lab, 2017)	(Haertel et al., 2022a)	N/A
2019	The nine-stages of the machine learning workflow	(Amershi et al., 2019)	Top cites	761
2020	Team Data Science Process (TDSP)	(Microsoft, 2020)	(Saltz and Krasteva, 2022; Schulz et al., 2020)	N/A

Table 2. Top-ten subset of DSPM.

The remaining sample of 25 objects was randomly divided into two subsets of 20 and five objects for further iterations. We used the qualitative coding techniques outlined by Strauss and Corbin (1997) to revise the taxonomy and analyze the subsets. The coding techniques are well-established among IS taxonomy designers to facilitate the derivation of supplementary taxonomy elements. Leveraging this method enables us to further develop our taxonomy in the subsequent iteration.

We identified the dimension *usability* by performing the coding stages. This describes the accessibility and applicability in real-world scenarios and addresses the user. The previous dimensions of horizontal and vertical requirements get entirely refined for an even more detailed classification scheme. We promoted the characteristics *domain, different level of abstraction,* and *team roles* to dimensions, including renaming and adding elements. The remaining requirements were merged with new findings to the dimension *key requirements*. Furthermore, we refined the characteristics of the dimensions *guidance* and *reality and impact* and renamed them to *advisory repository* and *value proposition*. Lastly, some characteristics were added, split, or merged into new ones. The refined taxonomy T2 included 15 dimensions and 74 characteristics. The development procedure continues with that taxonomy T2, as we didn´t meet the defined ending conditions and have objects to classify left.

For the **next E2C iteration**, we relied on the randomly selected subsample of 20 objects and classified them into the existing taxonomy T2. This led to updates of characteristics. The dimension value proposition was newly configured with structural, organizational, economic, strategic, and operational characteristics. The dimension development type was split into *maturation class* and *adoption type*. The maturation defines whether an object is novel or based on existing DSPM, while the adoption categories refer to different ways of modifying the workflow. Further, three dimensions (*project management*, *team management*, and *data & information management*) were demoted to characteristics of the introduced *supportive assistance* dimension. The most significant change from T2 to T3 is the reorganization of the dimension *key requirements*. This was achieved by integrating the *workflow* dimension and with various other operations into the characteristics (adding, renaming, merging). As a result, we ended up with seven updated key requirements. As we have added further elements in this iteration, the procedure continues with the taxonomy T3, which consists of 13 dimensions and 68 characteristics.

In the **fourth iteration**, we started at step four of the ETDP as a possible entry point to refine the meta-characteristics, as listed in the design recommendations (Kundisch et al., 2022). We refined the previously defined meta-characteristics (specification, support, requirements, and goals) with our categories. The ***non-functional category*** typically concerns the broad characteristics that determine the process model´s operation but do not directly impact its specific functions. The ***process category*** relates to the activities and assets that underlie process execution and management within a project. The ***user category*** bundles the user's needs and focuses on usage options and support mechanisms to ensure smooth interaction. The ending conditions were not changed, and the development process was continued. In this E2C iteration, we categorized the final five objects using taxonomy T3. We executed only one modification within the key requirements dimension during this process. Specifically, we combined *collaboration and transfer* with *transparency* and renamed this merged characteristic to *collaboration and transparency*. With the help of swapping and assigning the dimensions to the new categories, we created taxonomy T4, consisting of 3 categories, 13 dimensions, and 67 characteristics.

In the **fifth iteration**, we recorded the complete sample of 35 DSPM and classified all relevant objects with the help of the taxonomy T4. As a result of this E2C cycle, our taxonomy met all defined objective ending conditions and reached stability. According to Kundisch et al. (2022), successfully assessing the objective ending conditions serves as a demonstration of the taxonomy in terms of the DSR principles. Thus, the development process ended with five iterations and proposed the final taxonomy T4.

4.3 Evaluation

The evaluation starts with an ex-ante verification of the subjective ending conditions, followed by an ex-post evaluation based on their configuration (Kundisch et al., 2022). Upon assessing the taxonomy independently, the authors concurred that it meets the subjective ending conditions of being concise, comprehensive, robust, extendible, and explanatory. Considering the intended purpose of the taxonomy, we configured the evaluation and specified their goals of describing, classifying, and analyzing DSPM. To achieve these objectives and determine the usefulness of our taxonomy, we establish two criteria: *completeness* and *robustness*. Completeness refers to the extent to which the taxonomy covers the essential dimensions and characteristics to classify all objects, thus ensuring their functionality (Prat et al., 2015). Robustness characterizes the ability of the taxonomy to process varying inputs and environment conditions (Prat et al., 2015).

As Kundisch et al. (2022) recommended, we perform the ex-post evaluation with new objects not used in the taxonomy-building process. The objects for the building process were identified by synchronizing three literature reviews, as described in iteration two. These reviews only cover the knowledge base until the beginning of 2022. To identify new objects, we conducted a systematic literature review based on the established procedure proposed by vom Brocke et al. (2009). The specific search string *("data science" OR "data mining" OR "data analytics") AND ("process model" OR "methodology" OR "reference model")* was formulated to this end, enhancing the accuracy of the outcomes. We conducted searches in relevant databases of the underlying research field, limiting our query to articles published from 2022 onwards. To refine our search results, we incorporated title-based filtering, which yielded the following results (initial search/title filtering): ACM Digital Library (3,535/1), AISeL (733/1), IEEE Xplore (907/5), ScienceDirect (26,847/13), and SpringerLink (17,000/12). The execution of the search strategy resulted in a pool of 32 relevant articles as of October 2023. The eligibility criteria from the second iteration were checked through subsequent manual abstract reading and full-paper screening. This resulted in four eligible evaluation objects that were not used in development and represent the phenomenon of DSPM. The evaluation objects are included in the online appendix.

For the evaluation, we conducted semi-structured interviews with experts. We contacted individuals from our professional, academic, and LinkedIn networks who possess experience in executing or consulting DS projects, are familiar with DSPM, and have at least three years of relevant experience. To ensure scientific rigor, the final panel included employees from eleven companies across diverse industries and positions, including start-ups, small, medium, and multinational corporations, as well as universities. Table 3 presents an overview of the 12 recruited experts.

ID	Industry	Company size	Job title	Experience
Exp1	Electronics	>2500	Team Lead Data Science	3-5 Years
Exp2	Consulting	1-50	Data Scientist	<10 Years
Exp3	Marketing	501-1000	Senior Data Analyst	<10 Years
Exp4	Consulting	101-500	Lead Data Scientist	<10 Years
Exp5	Consulting	101-500	Team Lead Data Analytics	>=10 Years
Exp6	University	101-500	Research Assistant	3-5 Years
Exp7	Software Development	1-50	Data Scientist	3-5 Years
Exp8	Electronics	501-1000	Data & Knowledge Manager	3-5 Years
Exp9	University	>2500	Research Assistant	3-5 Years
Exp10	University	501-1000	Research Assistant	3-5 Years
Exp11	University	501-1000	Professor Data Science	<10 Years
Exp12	Manufacturing	51-100	AI Developer	3-5 Years

Table 3. Expert profiles.

We explicitly selected only Research Assistants with recent publications on DSPM as interviewees, indicating a specialized knowledge base. All interviews, except Exp4 and Exp5, work for different companies. Their selection as experts, despite coming from the same company, adds two key factors to the assessment: (i) leadership of different teams within the same consulting company provides them with unique insights and experiences, and (ii) consulting serves as a hub of diversity, bringing valuable cross-sectoral perspectives and innovative solutions to the evaluation. Additionally, we ensure that the participants have not had any prior contact with the taxonomy. The evaluation assessments took place between October 2023 and March 2024 and were structured into three phases: i) preparation, ii) interviews, and iii) post-evaluation. Participants received detailed information on the taxonomy, including its dimensions and characteristics, and an example of a categorized DSPM, to prepare them for the evaluation phase. The experts were then responsible for categorizing the evaluation objects using the final taxonomy. The interview phase typically occurs one to three weeks later. Initially, the experts were introduced to the evaluation statements, followed by additional time to reflect on the taxonomy's usage. Finally, we discussed the completeness and robustness of the taxonomy using predefined statements. The statements used to assess the evaluation criteria are based on the subjective final conditions outlined by Nickerson et al. (2013). The criteria were ranked with the experts, and additional comments were noted for each statement. The insights from the interviews were then aggregated and presented in Table 4 alongside the evaluation statements. The values in Table 4 represent the percentage of experts who assigned a certain level of support to each statement.

Criteria | Statement		Full Support	Support	Neutral	Reject	Full Reject
Completeness	**A)** The clarity of the dimensions and characteristics ensures the ease of understanding without unnecessary complexity.	34 %	58 %	8 %	-	-
	B) The taxonomy encompasses all relevant dimensions and characteristics, ensuring no object is left unclassified.	42 %	50 %	8 %	-	-
Robustness	**C)** Each dimension and characteristic are defined with enough detail to ensure accurate categorization across diverse or complex objects.	50 %	42 %	8 %	-	-
	D) The taxonomy serves as a useful tool that simplifies the description and classification of objects within its scope.	92 %	8 %	-	-	-

Table 4. Evaluation results.

Over 90 % of the participants responded positively to the completeness statements (A and B). However, three experts (Exp3, Exp10, Exp11) expressed concerns that the dimensions of ***usability*** (D_9) and ***granularity*** (D_{10}) may be difficult to comprehend and could complicate the classification, reducing ease of understanding (A). The statement that the taxonomy includes all relevant dimensions and characteristics to classify the objects (B) received positive responses from all participants except one. Exp3 noted that the characteristic *impact measurement and valuation* ($C_{11,5}$) only superficially covers the practical need to justify the required resources against the value generated. In contrast, Exp9 stated that no necessary additions were seen ad-hoc. Furthermore, 92 % of respondents supported that the dimensions and characteristics were sufficiently detailed for accurate categorization (C). Exp9 requested more detailed definitions in certain dimensions, accompanied by illustrative examples, aiming to enhance the intuitive classification among non-experts. One researcher (Exp10) found the characteristic *iterative and agile character* ($C_{11.2}$) challenging and suggested providing more details or splitting it into separate elements. On the other hand, industry experts mentioned that the taxonomy is well-balanced (Exp1) or may have too many characteristics (Exp3). All experts agree that the taxonomy is useful for describing and classifying DSPM, even 92 % agreed with full support. The experts confirmed that the developed taxonomy serves as a helpful tool for comparing DSPM (Exp5, Exp8, Exp9), gaining an overview and better understanding (Exp4, Exp10, Exp11), and selecting a DSPM (Exp6, Exp12). In addition, Exp7 highlighted the effectiveness of the taxonomy in reflecting on one's own applied process model. Exp1 emphasized its adaptability, allowing it to focus on elements based on personal needs, while Exp2 argued for better visualization of the taxonomy instead of the table form. The high level of support (92-100 %) across these statements, paired with the experts´ profiles and the consequent multi-perspective, indicates that the actual taxonomy draft is complete. The balance between full support and support for statements A, B, and C can be attributed to the general complexity of the phenomenon under consideration. We conclude that the criteria were met and that taxonomy T4 reached the formulated evaluation goals. These results confirm that the taxonomy is a well-founded artifact that enables researchers and practitioners to systematically describe, classify, and analyze DSPM.

5 Taxonomy of Data Science Process Models

This section presents the taxonomy for DSPM using a tabular approach commonly used by taxonomy designers. Table 5 shows the final taxonomy and the classification results from the last iteration of the development process. The calculated ratio indicates the percentage of DSPM that possess a particular characteristic. The final taxonomy comprises three categories, 13 dimensions, and 67 characteristics.

Subsequently, dimensions and characteristics are defined to provide sufficient detail to the reader, ensuring accurate object categorization. The classification results are discussed in Subsection 5.4.

Category/Dimension D_i		**Characteristics** $D_{i,j}$			
Non-functional	D_1 **Application domain** (ME)	$C_{1,1}$ General (69%)	$C_{1,2}$ Software development (11%)	$C_{1,3}$ Domain specific (11%)	$C_{1,4}$ Application specific (9%)
	D_2 **Maturation class**	$C_{2,1}$ New (31%)	$C_{2,2}$ Based on CRISP-DM (57%)	$C_{2,3}$ Based on KDD (17%)	$C_{2,4}$ Based on other (29%)
	D_3 **Adoption type** (ME)	$C_{3,1}$ Novel (51%)	$C_{3,2}$ Extension (23%)	$C_{3,3}$ Specialization (14%)	$C_{3,4}$ Enrichment (11%)
	D_4 **Value proposition**	$C_{4,1}$ Structural (69%)	$C_{4,2}$ Organizational (43%)	$C_{4,3}$ Economic (29%)	$C_{4,4}$ Strategic (66%)
		$C_{4,5}$ Operational (54%)			
Process	D_5 **Data Science Lifecycle (DSLC)**	$C_{5,1}$ Business understanding (80%)	$C_{5,2}$ Data collection, exploration & preparation (100%)	$C_{5,3}$ Analysis (91%)	$C_{5,4}$ Evaluation (89%)
		$C_{5,5}$ Deployment (80%)	$C_{5,6}$ Utilization (51%)		
	D_6 **DSLC deliverables**	$C_{6,1}$ Business understanding (37%)	$C_{6,2}$ Data collection, exploration & preparation (23%)	$C_{6,3}$ Analysis (17%)	$C_{6,4}$ Evaluation (20%)
		$C_{6,5}$ Deployment (37%)	$C_{6,6}$ Utilization (23%)	$C_{6,7}$ Multiple (31%)	$C_{6,8}$ Not defined (54%)
	D_7 **Type of deliverable**	$C_{7,1}$ Specific documents (34%)	$C_{7,2}$ Final report (29%)	$C_{7,3}$ Strategic blueprints (31%)	$C_{7,4}$ other (26%)
		$C_{7,5}$ not defined (51%)			
	D_8 **Team roles**	$C_{8,1}$ Project Manager (31%)	$C_{8,2}$ Data Scientist (40%)	$C_{8,3}$ Data Engineer (14%)	$C_{8,4}$ Domain Expert (40%)
		$C_{8,5}$ Stakeholder (23%)	$C_{8,6}$ Technical support (23%)	$C_{8,7}$ other (37%)	$C_{8,8}$ Not defined (43%)
User	D_9 **Usability** (ME)	$C_{9,1}$ Limited (11%)	$C_{9,2}$ Basic (51%)	$C_{9,3}$ Intermediate (29%)	$C_{9,4}$ Advanced (9%)
	D_{10} **Granularity** (ME)	$C_{10,1}$ Makro (20%)	$C_{10,2}$ Meso (54%)	$C_{10,3}$ Micro (26%)	
	D_{11} **Key requirements**	$C_{11,1}$ Goal orientation (83%)	$C_{11,2}$ Iterative & agile character (71%)	$C_{11,3}$ Collaboration & transparency (40%)	$C_{11,4}$ Continuous transformation & modification (69%)
		$C_{11,5}$ Impact measurement & valuation (46%)	$C_{11,6}$ Scientific integrity (14%)		
	D_{12} **Supportive assistance**	$C_{12,1}$ Project management (89%)	$C_{12,2}$ Team management (40%)	$C_{12,3}$ Technology management (43%)	
	D_{13} **Advisory repository**	$C_{13,1}$ Process flow diagram (86%)	$C_{13,2}$ Documentation (91%)	$C_{13,3}$ Guideline (26%)	$C_{13,4}$ Template (11%)
		$C_{13,5}$ Tool (14%)	$C_{13,6}$ Technology recommendation (20%)	$C_{13,7}$ Demonstration (54%)	

Table 5. Final taxonomy for DSPM and classification results.

5.1 Non-functional category

The non-functional category serves to catalog general information on the process model. Within this category, we provide four dimensions and 17 characteristics. ***Application domain*** (D_1) determines whether the model fits *general*, *domain-specific*, or *application-specific* use. As some papers address this domain specifically, we added the option *software development*. ***Maturation class*** (D_2) indicates whether the model is *new* or derived from established DSPM such as *CRISP-DM, KDD*, or *others*. ***Adoption type*** (D_3) details the modifications made to the original model. This can be *novel* or the adjustment types defined by Saltz and Krasteva (2022): *specialization* (tailoring workflows to optimize for specific technologies or domains), *extension* (incorporating additional steps or tasks to broaden workflow phases), *enrichment* (expanding workflow scope for a more thorough project execution). The main driver for the practical use of DSPM is the potential ***value proposition*** (D_4), which encompasses six primary categories. *Structural* value is given by modifications to the workflow of the process, e.g., additional flexibility (Martinez-Plumed et al., 2021), improved guidance (Ahmed et al., 2018), and additional or more detailed steps (Vernickel et al., 2019). An *organizational* value proposition is delivered by defining roles and responsibilities in DS projects and promoting teamwork and communication (Schulz et al., 2020). Business impacts like time and cost efficiency (Martinez-Plumed et al., 2021), improved project success rates (Marbán et al., 2009), or enhanced analytics quality (Vernickel et al., 2019) are summarized in an *economic* value proposition. *Strategic* benefit is derived from facilitating the implementation of DS initiatives (Vanauer et al., 2015) or focusing on the business objectives and their alignments (Kaufmann, 2019). Improved execution of DS projects is presented as an *operational* value proposition. This can be achieved with improvements in project management (Marbán et al., 2009), emphasizing the technology layer (Grady, 2016), or incorporating insights from alternative approaches such as design thinking (Ahmed et al., 2018).

5.2 Process category

The nature of process models results from the underlying process elements. Therefore, the fifth dimension of the taxonomy presents the general ***Data Science Lifecycle*** (DSLC) adopted from Haertel et al. (2022a) with its six phases as characteristics: (1) *Business Understanding*, (2) *Data Collection, Exploration and Preparation*, (3) *Analysis*, (4) *Evaluation*, (5) *Deployment* and (6) *Utilization*. The ***DSLC deliverables*** (D_6) represent the next dimension: artifacts developed in the corresponding DSLC phases. The associated characteristics are the *six phases*. ***Types of deliverables*** (D_7) can be categorized as *not defined*, *specific document*, *final report*, *strategic blueprint*, or *others*. Strategic blueprints can occur through roadmaps, project plans, or similar. Others rely on diverse elements, such as images of the final architecture or other representations. We have not categorized deliverables such as data, models, or project insights since we define these as prerequisites for DS projects. Finally, ***team roles*** (D_8) round off the process category and are assigned to *project management, data scientist, data engineer, domain expert, stakeholder, technical support, others*, and *not defined*.

5.3 User category

The potential capabilities of the DSPM can be organized in three dimensions: *usability, granularity*, and *key requirements*. The ***usability*** (D_9) dimension is structured by a mutually exclusive four-tiered scale from *limited* to *basic*, *intermediate,* and *advanced*. The usability level is assessed based on the availability of information about the DSPM, its accessibility, the ease of understanding for the user, and whether the guidelines are practical and easy to apply in real-world scenarios. The ***granularity*** (D_{10}) dimension is divided into three levels: *macro, meso,* and *micro*. During the mapping process, an evaluation was conducted to determine the presence of different abstraction levels and the depth of the guideline's description. *Micro* level indicates that the activities are detailed down to individual tasks, as in CRISP-DM (Chapman et al., 2000). A *macro* granularity is presented in Rollins and IBM Corporation (2015) or Amershi et al. (2019). The remaining capabilities are allocated by adding the dimension ***key requirements*** (D_{11}). We state six characteristics that paraphrase the requirements and guide for an appropriate selection. *Goal orientation* refers to the identification and articulation of objectives, while the DSPM ensures that subsequent activities are aligned with these defined goals. The *iterative and agile character* of DSPM, with feedback loops or any kind of prototyping, presents the ability to react to changes in the environment based on preliminary insights and support different project sizes (do Nascimento and Oliveira, 2012). *Collaboration and transparency* involve concepts of exchanging information and knowledge within the team and with stakeholders. It stresses the importance of working together and promotes reproducibility and traceability, as Marbán et al. (2009) and Kaufmann (2019) highlighted. The characteristic *continuous transformation and modification* inherently includes the concept of cyclic workflows, emphasizing the ongoing improvement and evolution through the application and refinement of existing knowledge. Revisiting the goals aligned with the business problems and further usage of the gained insights to create value are summarized in the *impact measurement and valuation*. Finally, *scientific integrity* assesses whether scientific methods underpin the DSPM. For example, this could be case study evaluations (Kaufmann, 2019) or development following the DSR principles (Asamoah and Sharda, 2019). The dimension ***supportive assistance*** (D_{12}) specifically outlines the focused area where the DSPM provides support. In response to the primary challenges of DS projects identified by Martinez et al. (2021), we have articulated the characteristics of *project, team,* and *technology management*. How the DSPM provides guidance is systematically organized in the ***advisory repository*** (D_{13}). The guidance is typically manifested through various forms, such as *process flow diagrams, documentation*, or comprehensive *guidelines*. In addition, other models furnish *templates* (Vanauer et al., 2015), *tools* (Microsoft, 2020), or *technology recommendations* (Asamoah and Sharda, 2019). Some DSPM even present a *demonstration* through illustrative examples, case studies, or best practices (Amershi et al., 2019; Dutta and Bose, 2015).

5.4 Classification results

To provide a deeper understanding, we have calculated the percentages of each characteristic based on the 35 classified DSPM (refer to values in Table 5). In some dimensions, the total may exceed the maximum due to non-exclusivity. The dimensions *application domain* (D_1), *adoption type* (D_4), *usability* (D_9), and *granularity* (D_{10}) appear to be mutually exclusive, while all other dimensions are non-exclusive. The non-functional category indicates that over two-thirds of the DSPM are intended for general use, and more than half of the models are based on the CRISP-DM but still result in novel models rather than adoptions. Only 29 % comprise an economic value proposition, while over 66 % promote structural and strategic improvements. The CRISP-DM phases of the DSLC have a high ratio of over 80 %. About 50 % of the DSPM improve these phases and expand the workflow concerning the phase utilization. More than 50 % did not define their deliverables either according to a DSLC phase or according to their type. The project deliverable ratios are generally low, with a maximum of 37 % in the deployment phase. 43 % of the DSPM do not consider team roles, which aligns with organizational challenges identified by Martinez et al. (2021) or Saltz and Krasteva (2022). The roles data scientist and domain expert are required by 40 % of the models. Further, only a few DSPM provide good *usability* (D_9) and a high *granularity* (D_{10}). Although most models consider *key requirements* (D_{11}) such as goal orientation, iterative and agile character, and continuous transformation and modification, only 14 % demonstrate scientific integrity. While project management is the major *supportive assistance* (D_{12}), approximately 40 % focus on team or technology management. Over 86 % of the models provide process flow diagrams and documentation of the process. More than 50 % of DSPM are presented through a demonstration, but less than 15 % provide templates or tools.

The resulting statistics can be used to derive valuable insights from a research perspective. Our study highlights the absence of project deliverables (cf. D_6 and D_7) in existing DSPM, as Haertel et al. (2022a) mentioned. Only a minority of DSPM offer guidance for process execution through templates or tools (cf. D_{13}), resulting in lower usability with only 9 % achieving the advanced characteristic (cf. D_9). Table 5 shows that existing process models neglect the economic value proposition (cf. D_4, less than 30 %). Based on the further distribution and variance of the ratios, there is only a limited need for further research into new process models. In combination with the insights derived, there is a challenge in providing support systems to enhance the practical usability of DSPM, which warrants further investigation. Advisory or assistance systems, when paired with tools or templates, can address the gap in practical usability and economic value proposition of the original DSPM, even retrospectively. Practitioners can use the classification results to gain an overview of the current DSPM knowledge base and assess their workflow against it.

6 Conclusion

DSPM have been available for several decades, and new ones have been consistently introduced. While their general relevance is undisputed, they still need to gain wider adoption in practice. Currently, there is little research on the structures and systematics of DSPM, resulting in a limited understanding of their different dimensions and characteristics, and their potential fit for different application scenarios. As DS projects have become increasingly important for companies to gain competitive advantage, our research aims to close this gap by developing and evaluating an empirically grounded taxonomy of DSPM. We collected 35 unique DSPM from three literature reviews to develop our taxonomy. The development procedure followed the established approaches of Nickerson et al. (2013) and its extension by Kundisch et al. (2022). The final taxonomy comprises three categories, 13 dimensions, and 67 characteristics. Our study contributes a comprehensive overview of DSPM and their relevant elements, answering the research question of this paper.

Several conclusions can be drawn from our research. Our taxonomy proposes a structured framework to characterize DSPM and aid in the adaptation and selection of an appropriate model for specific needs in practical and academic settings. From a *managerial perspective,* the taxonomy primarily serves as a guidance tool by providing a holistic, simple, and precise overview of possible characteristics that DSPM can offer. Our taxonomy helps data scientists and project managers better understand DSPM. Additionally, practitioners can use the framework to evaluate, score, and benchmark their workflows against the current knowledge base. Decision-makers can use the resulting benchmarks or search for specific DSPM characteristics according to their organization's needs to facilitate strategic realignment in executing DS projects. Regarding the *scientific implications*, we have developed a rigorous and comprehensive taxonomy of DSPM through a combination of deductive and inductive research approaches. We contribute to the existing knowledge base by adding a coherent understanding of DSPM from an IS perspective. Researchers can benefit from the conceptual structure we have created for analyzing and classifying DSPM, which promotes general expertise. Despite the inherent diversity and complexity of DSPM, which pose challenges for future research on implementation or application success, we argue that our taxonomy makes this kind of research more accessible. Additionally, we suggest avenues for further investigation that will improve the practical usability of DSPM and initiate an economic value proposition retrospectively.

Despite a rigorous taxonomy design, it is important to acknowledge the limitations of our research. Unlike most taxonomy design projects, our development was limited to a small number of DSPM objects. The resulting limitation of our research lies in its reliance on existing literature reviews and collected DSPM rather than incorporating real-world DSPM instances. While this approach allows for a comprehensive and scientifically grounded taxonomy, the applicability of the taxonomic investigation of DSPM is debatable. Nevertheless, this approach represents a scientific innovation by providing insights into existing phenomena and expanding the scope of taxonomy design in IS research. In addition, selecting different development approaches or criteria can lead to alternative taxonomies. This variability is typical in DSR, allowing for different artifacts under different conditions (Hevner et al., 2004). The limited number of available objects also affects the scope of the evaluation. Only a few objects not used for the taxonomy-building process were available as evaluation samples. Therefore, they may not fully represent the diversity of DSPM. Future research could aim to identify potential DSPM archetypes and apply the taxonomy to real-world instances. This would offer a pathway for further validation and refinement of its applicability and effectiveness.

References

Ahmed, B., Dannhauser, T. and Philip, N. (2018). "A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects," in: *Proceedings of the 10th Computer Science and Electronic Engineering Conference,* Colchester, United Kingdom.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B. and Zimmermann, T. (2019). "Software Engineering for Machine Learning: A Case Study," in: *Proceedings of the IEEE/ACM International Conference on Software Engineering 2019,* Montreal, Canada.

Asamoah, D. A. and Sharda, R. (2019). "CRISP-eSNeP: Towards a data-driven knowledge discovery process for electronic social networks," *Journal of Decision Systems* 28 (4), 286–308.

Baecker, J., Böttcher, T. and Weking, J. (2021). "How Companies Create Value From Data – A Taxonomy on Data, Approaches, and Resulting Business Value," *ECIS 2021 Proceedings*.

Bichler, M., Heinzl, A. and van der Aalst, W. M. P. (2017). "Business Analytics and Data Science: Once Again?," *Business & Information Systems Engineering* 59 (2), 77–79.

Cao, L. (2017). "Data science: challenges and directions," *Communications of the ACM* 60 (8), 59–68.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. P., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.

Das, M., Cui, R., Campbell, D. R., Agrawal, G. and Ramnath, R. (2015). "Towards methods for systematic research on big data," in: *Proceedings of the IEEE International Conference on Big Data 2015,* Santa Clara, USA.

Dehnert, M., Gleiss, A. and Reiss, F. (2021). "What makes a data-driven business model? A consolidated taxonomy," *ECIS 2021 Proceedings*.

do Nascimento, G. S. and Oliveira, A. A. de (2012). "An Agile Knowledge Discovery in Databases Software Process," in: *Proceedings of the 3rd Data and Knowledge Engineering,* Wuyishan, China.

Domino Data Lab (2017). *Managing Data Science Projects*. URL: https://domino.ai/resources/field-guide/managing-data-science-projects (visited on March 29, 2024).

Dutta, D. and Bose, I. (2015). "Managing a Big Data project: The case of Ramco Cements Limited," *International Journal of Production Economics* 165, 293–306.

Feldman, M. S. and Pentland, B. T. (2003). "Reconceptualizing Organizational Routines as a Source of Flexibility and Change," *Administrative Science Quarterly* 48 (1), 94–118.

Feyyad, U. M. (1996). "Data mining and knowledge discovery: making sense out of data," *IEEE Expert* 11 (5), 20–25.

Gelhaar, J., Gürpinar, T., Henke, M. and Otto, B. (2021). "Towards a taxonomy of incentive mechanisms for data sharing in data ecosystems," *PACIS 2021 Proceedings*.

Gerlach, J., Werth, O. and Breitner, M. H. (2022). "Artificial Intelligence for Cybersecurity: Towards Taxonomy-based Archetypes and Decision Support," *ICIS 2022 Proceedings*.

Grady, N. W. (2016). "KDD meets Big Data," in: *Proceedings of the IEEE International Conference on Big Data 2016,* Washington DC, USA.

Haertel, C., Pohl, M., Nahhas, A., Staegemann, D. and Turowski, K. (2022a). "Toward A Lifecycle for Data Science: A Literature Review of Data Science Process Models," *PACIS 2022 Proceedings*.

Haertel, C., Pohl, M., Staegemann, D. and Turowski, K. (2022b). "Project Artifacts for the Data Science Lifecycle: A Comprehensive Overview," in: *Proceedings of the IEEE International Conference on Big Data 2022,* Osaka, Japan.

Hevner, March, Park and Ram (2004). "Design Science in Information Systems Research," *MIS Quarterly* 28 (1), 75.

IBM Corporation (2016). *Analytics Solutions Unified Method (ASUM). Datasheet*. URL: https://public.dhe.ibm.com/software/data/sw-library/services/ASUM.pdf (visited on March 29, 2024).

Kaufmann, M. (2019). "Big Data Management Canvas: A Reference Model for Value Creation from Data," *Big Data and Cognitive Computing* 3 (1), 19.

Krieger, F. and Drews, P. (2018). "Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy," *ICIS 2018 Proceedings*.

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T. and Szopinski, D. (2022). "An Update for Taxonomy Designers," *Business & Information Systems Engineering* 64 (4), 421–439.

Kutzias, D., Dukino, C., Kötter, F. and Kett, H. (2023). "Comparative Analysis of Process Models for Data Science Projects," in: *Proceedings of the 15th International Conference on Agents and Artificial Intelligence,* Lisbon, Portugal.

Larson, D. and Chang, V. (2016). "A review and future direction of agile, business intelligence, analytics and data science," *International Journal of Information Management* 36 (5), 700–710.

Marbán, O., Segovia, J., Menasalvas, E. and Fernández-Baizán, C. (2009). "Toward data mining engineering: A software engineering approach," *Information Systems* 34 (1), 87–107.

Martinez, I., Viles, E. and G. Olaizola, I. (2021). "Data Science Methodologies: Current Challenges and Future Approaches," *Big Data Research* 24.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J. and Flach, P. (2021). "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering* 33 (8), 3048–3061.

Matschak, T., Trang, S. and Prinz, C. (2022). "A Taxonomy of Machine Learning-Based Fraud Detection Systems," *ECIS 2022 Proceedings*.

Microsoft (2020). *What is the Team Data Science Process?* URL: https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview (visited on March 29, 2024).

Nickerson, R. C., Varshney, U. and Muntermann, J. (2013). "A method for taxonomy development and its application in information systems," *European Journal of Information Systems* 22 (3), 336–359.

Oliveira, D. F. and Brito, M. A. (2022). "Development of Deep Learning Systems: A Data Science Project Approach," in: *Proceedings of the 10th World Conference on Information Systems and Technologies*, Budva, Montenegro.

Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. (2007). "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* 24 (3), 45–77.

Prat, N., Comyn-Wattiau, I. and Akoka, J. (2015). "A Taxonomy of Evaluation Methods for Information Systems Artifacts," *Journal of Management Information Systems* 32 (3), 229–267.

Rollins, J. B. and IBM Corporation (2015). *Foundational Methodology for Data Science*. URL: https://tdwi.org/whitepapers/2017/12/adv-all-ibm-foundational-methodology-for-data-science.aspx&ntb=1 (visited on March 29, 2024).

Saltz, J. and Hotz, N. (2021). "Factors that Influence the Selection of a Data Science Process Management Methodology: An Exploratory Study," in: *Proceedings of the 54th Hawaii International Conference on System Sciences,* Honolulu, Hawaii.

Saltz, J., Hotz, N., Wild, D. and Stirling, K. (2018). "Exploring Project Management Methodologies Used Within Data Science Teams," *AMCIS 2018 Proceedings*.

Saltz, J. and Krasteva, I. (2022). "Current approaches for executing big data science projects-a systematic literature review," *PeerJ Computer Science* 8, e862.

Saltz, J. and Shamshurin, I. (2015). "Exploring the process of doing data science via an ethnographic study of a media advertising company," in: *Proceedings of the IEEE International Conference on Big Data 2015,* Santa Clara, USA.

SAS Institute Inc. (2017). *Introduction to SEMMA*. URL: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm (visited on March 29, 2024).

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kuehnel, S., Badwitz, W., Dann, D., Kloker, S., Alekozai, E. and Lanquillon, C. (2020). "Introducing DASC-PM: A Data Science Process Model," *ACIS 2020 Proceedings*.

Shearer, C. (2000). "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing* 5 (4), 13–22.

Strauss, A. L. and Corbin, J. M. (1997). *Grounded theory in practice.* 1st Edition. New York City: SAGE Publications.

van der Aalst, W. (2016). "Data Science in Action," in: van der Aalst, W. (ed.) *Process Mining: Data Science in Action,* 2nd Edition. Berlin, Heidelberg: Springer.

Vanauer, M., Bohle, C. and Hellingrath, B. (2015). "Guiding the Introduction of Big Data in Organizations: A Methodology with Business- and Data-Driven Ideation and Enterprise Architecture Management-Based Implementation," in: *Proceedings of the 48th Hawaii International Conference on System Sciences,* Honolulu, Hawaii.

Vernickel, K., Weber, J., Li, X., Berg, J. and Reinhart, G. (2019). "A Revised KDD Procedure for the Modeling of Continuous Production in Powder Processing," in: *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management 2019,* Macao, China.

Vetter, O., Hoffmann, F., Pumplun, L. and Buxmann, P. (2022). "What constitutes a machine-learning-driven business model? A taxonomy of B2B start-ups with machine learning at their core," *ECIS 2022 Proceedings*.

vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R. and Cleven, A. (2009). "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," *ECIS 2009 Proceedings*.

Zhu, J. and Marjanovic, O. (2022). "A Taxonomy of Data Cooperatives," *PACIS 2022 Proceedings*.

View publication stats