A Methodology for Constructing Patterns for the Management of Data Science Projects

Christian Haertel[a], Sarah Schramm, Matthias Pohl[b], Sascha Bosse[c], Daniel Staegemann[d], Christian Daase[e] and Klaus Turowski[f]

Institute of Technical and Business Information Systems, Otto-von-Guericke University, Magdeburg, Germany
{christian.haertel, sarah.schramm, matthias.pohl, sascha.bosse, daniel.staegemann, christian.daase,

Keywords: Data Science, Project Management, Pattern, Design Science Research.

Abstract: In the era of Big Data, the successful completion of Data Science (DS) projects is crucial. However, DS project management is quite challenging due to its interdisciplinary nature. Existing DS process models, such as CRISP-DM, have limitations, resulting in low success rates for these undertakings. To address this issue, a novel methodology for the construction of patterns in DS project management has been proposed, using the Design Science Research methodology. The design draws inspiration from existing pattern concepts to address common problems in DS project execution. The methodology is demonstrated through the creation of patterns for best practices in DS project management, synthesized from scientific literature. The goal of this approach is to provide a platform for exchanging and standardizing best practices in DS project management. While initial demonstrations show the general applicability of the methodology, further evaluations and case studies are necessary to assess its effectiveness and areas for improvement. The study identifies potential ambiguities in certain activities within the process, suggesting opportunities for refinement. Overall, this research contributes to the field of DS project management by offering a structured method to encapsulate and disseminate effective practices, supporting the successful execution of data projects in organizations.

1 INTRODUCTION

In a world where the amount of generated data is steadily increasing, businesses of various domains aim to derive potential advantages and enhance their competitive positioning (de Medeiros et al., 2020). Accordingly, Data Science (DS) as a discipline to extract knowledge and insights from data using various methods and techniques (Chang and Grady, 2019), has gained increasing significance (Cao, 2017). With its growing importance, the adequate management of DS projects becomes crucial. However, organizations often encounter challenges in implementing data-driven projects (Martinez et al., 2021a). DS is considered an interdisciplinary field that combines various areas such as statistics, computer science, machine learning, and domain-specific knowledge, which poses unique challenges to project management (Martinez et al., 2021a). Various DS process models have been developed to support the implementation and management of DS projects (e.g., CRISP-DM) (Saltz, 2015). However, research indicated several weaknesses in these methodologies (Martinez et al., 2021a). Hence, there is no widely accepted and applied approach, and custom methods are derived instead (Saltz, 2015; Saltz et al., 2018). These issues are reflected in the low DS project success rate (VentureBeat, 2019), demanding improvements for DS project management (Saltz and Krasteva, 2022). As the literature identifies common problems in the execution of DS projects (Martinez et al., 2021a), the adaptation of the *pattern* concept to DS appears promising. Patterns capture solutions to recurring problems in a domain in a simple and straightforward form (Fehling et al., 2014). An overview and structured presentation within patterns could ensure alleviated and methodology-independent access to common problems and solutions and, thus, contribute to the improvement of DS project management

[a] https://orcid.org/0009-0001-4904-5643
[b] https://orcid.org/0000-0002-6241-7675
[c] https://orcid.org/0000-0002-2490-363X
[d] https://orcid.org/0000-0001-9957-1003
[e] https://orcid.org/0000-0003-4662-7055
[f] https://orcid.org/0000-0002-4388-8914

Haertel, C., Schramm, S., Pohl, M., Bosse, S., Staegemann, D., Daase, C. and Turowski, K.
A Methodology for Constructing Patterns for the Management of Data Science Projects.
DOI: 10.5220/0012705300003690
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 1*, pages 354-365
ISBN: 978-989-758-692-7; ISSN: 2184-4992
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

activities. To the best of our knowledge, the pattern concept has not been applied before to DS (project management). Therefore, this research proposes a methodology for pattern creation for the field of DS project management, using the pattern identification, authoring, and application process of (Fehling et al., 2014) as a basis. Therefore, the following research question (RQ) will be examined:

RQ: How can a methodology for the construction of patterns for DS project management be designed and applied?

For both, researchers and practitioners, this artifact could form a platform for the exchange and standardization of best practices in the DS domain. Patterns can be created, expanded, modified, and linked to related areas. Overall, this study is intended to contribute to supporting and conducting data projects in organizations and, in turn, work toward improving the success rates of these projects. The Design Science Research (DSR) Methodology of (Peffers et al., 2007) will be leveraged to develop the pattern construction methodology. Therefore, in the next section, the application of the DSR paradigm in the context of this paper is described. Afterward, the relevant theoretical concepts for this work are discussed, including DS and related terms, project management, and fundamentals concerning the pattern concept. Consequently, the pattern development process for DS project management, an adaption of the process proposed by (Fehling et al., 2014), is outlined. After demonstrating the application of the artifact, this contribution is concluded with a summary and an outlook on future research endeavors.

2 METHODOLOGY

Scientific research seeks new insights and relationships in a specific field by applying scientific methods (Eisend and Kuß, 2023). In the context of information systems, DSR has emerged as a design-oriented approach to support the creation of artifacts to address practically relevant problems (Hevner et al., 2004). A pattern language construction methodology for DS can also be categorized as an artifact (method), and DSR can be considered suitable for its development. The Design Science Research Methodology (DSRM), according to (Peffers et al., 2007), is widely adopted in the DSR context. According to the DSRM, an artifact is designed, developed, and evaluated in six phases. The individual steps for the work at hand are explained in the following.

Problem Identification and Motivation: DS projects suffer from high failure rates (VentureBeat, 2019). Hence, the development of new or revised approaches for DS project management is necessary (Saltz, 2022). Pattern languages are utilized to document solutions to recurring problems in a given domain (Fehling et al., 2014) and have not yet been applied to DS. Accordingly, adapting this concept to this field by proposing a specific methodology for the development of patterns for DS can support addressing common issues in the execution of these undertakings.

Objectives of a Solution: The next step of the DSRM involves deriving goals for a solution. Conducting a DS project often involves encountering challenges related to team, project, and data and information management (Martinez et al., 2021a). The artifact of this research, a methodology for the construction of DS project management patterns, enables the capture of solutions for these often-faced problems in DS from the respective body of knowledge. Therefore, these patterns can be utilized to overcome these obstacles and, ultimately, lead to improved success rates in DS undertakings.

Design and Development: Based on the definitions by (Hevner et al., 2004), the developed artifact is characterized as a method in the DSR context, as the proposed pattern construction methodology provides a process on how to synthesize best practices in the shape of patterns in the context of DS. For this purpose, the general pattern identification, authoring, and application procedure of (Fehling et al., 2014) is adapted to DS project management. Therefore, foundations and methodologies from the DS knowledge base are used.

Demonstration: The application of the artifact to develop patterns for best practices in the management of DS projects is demonstrated. Because of page restrictions, only snippets from the created patterns can be shown in this paper.

Evaluation: This step aims to observe and measure the ability of the artifact with regards to providing a solution to the initially stated problem (Peffers et al., 2007). Thus, the Build-Evaluate pattern of (Sonnenberg and vom Brocke, 2012) is applied for this DSR project, consisting of the four steps Eval 1 (justified research gap), Eval 2 (validated design specification), Eval 3 (proof of applicability), and Eval 4 (proof of usefulness).

Communication: The final step within the DSRM is achieved through writing, submitting, and presenting this paper to the scientific community and interested practitioners of the field.

3 THEORETICAL BACKGROUND

The development of a methodology for the construction of patterns for DS project management initially requires establishing a sufficient knowledge base of the underlying concepts. First, this implies discussing the terminology around DS and its similarities as well as delimitations to related disciplines. Additionally, key aspects and existing research concerning (DS) project management need to be covered. Finally, the fundamentals of the pattern concept are introduced.

3.1 Data Science and Related Concepts

A widely accepted interpretation describes DS as the methodology for synthesizing useful knowledge from data through a process of discovery or formulation and testing of hypotheses (Chang and Grady, 2019). It is also characterized as an interdisciplinary field (Cao, 2018). Therefore, knowledge and methods from various disciplines are brought together to utilize data effectively (Schulz et al., 2020). DS has emerged as a unique discipline where a deep understanding of the application-specific domain, mathematical knowledge, and a solid technological background are essential (Schulz et al., 2020). DS has been used since the mid-2000s, focusing on gaining insights from data (Chang and Grady, 2019). However, similar goals were pursued earlier under different names. These related concepts are closely intertwined and cannot be easily separated from one another (Chang and Grady, 2019). Examples are Data Mining and Big Data.

Data Mining is the exploration of patterns and relationships in data using specialized algorithms (Chang and Grady, 2019). Skills in statistics, mathematics, machine learning, algorithms, and domain knowledge are applied in this process (Chang and Grady, 2019). Data Mining is known as a part of Knowledge Discovery in Databases (KDD), which constitutes a comprehensive process for extracting valuable knowledge from data (Fayyad et al., 1996; Chang and Grady, 2019). Data Mining is a step in this process (Fayyad et al., 1996). The proximity to DS is evident, explaining the sometimes synonymous use of the terms (Schulz et al., 2020). Generally, Data Mining can be understood as a subfield of DS (Chang and Grady, 2019).

Big Data can be defined by the four dimensions (Vs) Volume (enormous size of datasets), Velocity (speed of data generation and capturing), Variety (in data sources and formats), and Variability (changes in data flow, format, or volume) (Jeble et al., 2018; Chang and Grady, 2019). Additionally, while other characteristics such as Value (value of results), Veracity (data quality), or Complexity (data complexity) can be defined (Jeble et al., 2018), these dimensions can be considered drivers for new scalable architectures for data-intensive applications. Big Data posed a challenge for data processing as well as analysis and gave rise to the term DS to explore new techniques for this matter (Chang and Grady, 2019). Thus, DS encompasses Big Data and its analysis (Chang and Grady, 2019). Based on the outlined overlaps between these disciplines, approaches for these domains tend to be overarching.

3.2 Data Science Project Management

Project management can be defined as the ”application of knowledge, skills, tools, and techniques to project activities to meet the project requirements” (PMI, 2017). Within project management, various tasks such as project planning, risk management, and many more are fulfilled (Wack, 2007). During project planning, the scope, schedules, budget, and resource allocation to achieve the project goals are defined (Aichele and Schönberger, 2014). In IT projects, the principles, procedures, methods, techniques, and tools necessary for planning, controlling, and monitoring are summarized within IT project management (Heinrich, 1997; Wieczorrek and Mertens, 2007). In comparison, DS undertakings differ because of the data focus and the resulting explorative character (Das et al., 2015). Accordingly, both typical IT project management methods and more specific process models are used in DS (Saltz and Hotz, 2020). Several DS process models can be identified in the literature, including KDD, CRISP-DM, TDSP, and SEMMA. Typically, according to the DS lifecycle of (Haertel et al., 2022), a DS project can be structured in the broad phases *Business Understanding, Data Collection, Exploration and Preparation, Analysis, Evaluation, Deployment*, and *Utilization*. However, research suggests that the current DS methodologies are partially unsuitable for tackling common challenges in DS projects related to team, project, and data and information management (Martinez et al., 2021a). Hence, revised and new approaches are needed (Saltz and Krasteva, 2022).

3.3 Pattern

For experts working on a problem, it is unusual to develop a new solution that is entirely different from existing ones (Buschmann, 1996). Often, there is a tendency to rely on a similar and already solved problem, using the essential elements of the solution to address the new problem (Buschmann, 1996). An established concept for the structured documentation of proven solutions is *patterns* (Fehling et al., 2014). This principle originates from the architect Christopher Alexander, who laid its foundation back in 1977 (Alexander, 1979; Coplien and Harrison, 2005). Patterns describe a frequently occurring problem and the core of its solution so that it can be used repeatedly in different ways (Alexander et al., 1977). Patterns are recorded in text form following a specific structure and generally consist of a particular context, a problem, and a solution (Alexander, 1979). They are hierarchically organized, documented, and interconnected. Connections can be drawn within and to related subject areas, thus creating a complex and abstract solution system that can be individually accessed (Fehling et al., 2014; Coplien, 2000). The connections highlight relationships and capture hidden structures (Coplien, 2000). Accordingly, navigation is facilitated, and references to the application of patterns are provided, resulting in the creation of a pattern language (Fehling et al., 2014). In essence, patterns can be understood as a concept for the structured documentation of proven solutions to recurring problems in a specific domain (Fehling et al., 2014).

Although initially developed for architecture, the concept was successfully transferred to other domains such as the software field (Coplien and Harrison, 2005). Other authors have adopted, extended, modified, and influenced the initial context-problem-solution structure shaped by (Alexander et al., 1977). Ultimately, the compilation, content, and sequence of pattern sections are fundamentally left to the authors. There are no predefined rules for the identification and documentation of patterns, but there are guidelines for orientation like in (Wellhausen and Fiesser, 2011) and (Harrison, 2003). The creation of patterns is often performed through expertise and experience or in collaboration with experts (Iba and Isaku, 2012). Another possibility is the extraction from literature (Fehling et al., 2014; Günther and Knote, 2017). For instance, (Fehling et al., 2014) describe a pattern-creation process that is used in this work. Application of the pattern concept to (DS) project management also appears sensible since several common challenges need to be addressed in the course of a project. Accordingly, the synthesis of knowledge from experts and the literature for summarizing corresponding solutions in the form of patterns can provide added value.

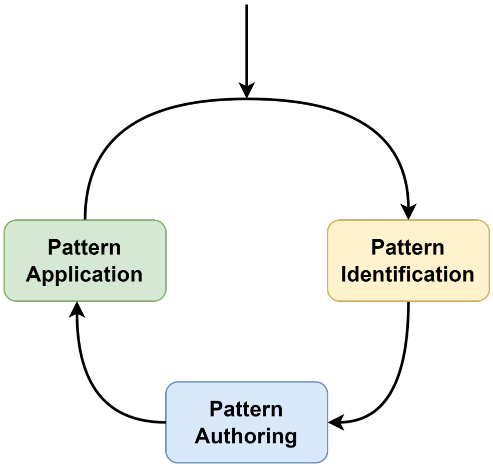

Figure 1: Pattern construction process, adopted from (Fehling et al., 2014).

4 PATTERN CONSTRUCTION PROCESS FOR DATA SCIENCE PROJECT MANAGEMENT

In the literature, there are different approaches to the development of patterns. For the application in this context, the process according to (Fehling et al., 2014) was selected and adapted to the field of DS project management since its applicability to various domains has already been demonstrated. Despite the weaknesses outlined in DS methodologies, several publications discuss success factors and best practices for DS project management. Therefore, the consolidation of this knowledge is beneficial. (Fehling et al., 2014) describe a detailed approach to the identification and authoring of patterns and, thus, allows the creation of patterns from the knowledge and expertise conveyed through the literature. The following subsections delve into the individual steps of this process, based on which DS patterns shall be developed. The method consists of three phases: pattern identification, pattern authoring, and pattern application, as illustrated in Figure 1. Each stage involves several iteratively traversed activities to continuously improve and adapt the developed results (Fehling et al., 2014). In particular, the first two phases are repeated multiple times to discover and form patterns. Finally, the third phase involves refining the patterns for specific use cases or application environments.

4.1 Pattern Identification

The first phase of the process, according to (Fehling et al., 2014), is carried out through five iteratively traversed activities as depicted in Figure 2. The exam-

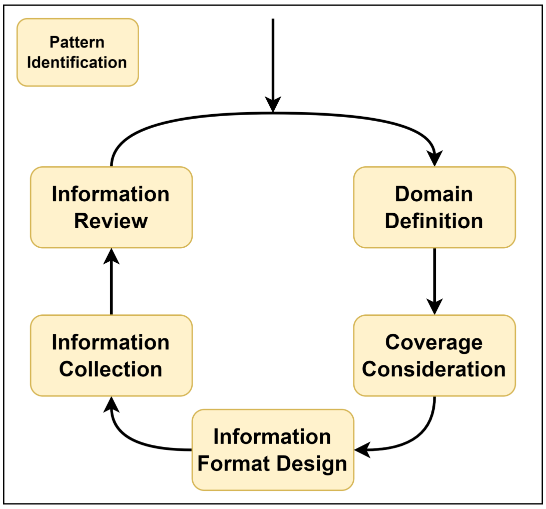

Figure 2: Pattern Identification phase, adopted from (Fehling et al., 2014).

ined domain is initially structured in this phase, and relevant information is gathered. In the first activity (*Domain Definition*), significant fundamentals of the investigated field are elaborated and shared with the group working on the patterns to build common knowledge and understanding of the domain as establishing a foundation for the field is considered important (Fehling et al., 2014). Initially, this requires creating a joint understanding of key terminology and concepts of DS (e.g., process models, ML algorithms) as seen in Section 3. Furthermore, since DS is a complex and interdisciplinary field, the focus is purely on DS project management.

The next activity, *Coverage Consideration*, focuses on assessing and narrowing down the scope of the chosen domain (Fehling et al., 2014). The domain can be extensive, making it impossible to consider the entire scope. Accordingly, the scope is adjusted to the size of the research group. Relevant topics are identified and aligned with characteristic problems of the domain. Since DS project management is extensive itself, it might be sensible to further limit the scope to the issues for the group (e.g., best practices, certain DS project stages) that mainly impact its DS project success. Therefore, considered information sources can be limited to a subset.

The subsequent task, *Information Format Design*, aims explicitly at team collaboration and establishing a unified structure for information capture and processing. As patterns will be identified based on information extracted from literature, uniform tools and templates for the collection, filtering, and analysis should be defined to achieve an efficient and transparent process.

The following activity, *Information Collection*, involves gathering information and coordinating its processing within the research group (Fehling et al., 2014). Since not all required information will be part of human memory, other information sources are required. Therefore, we propose the use of a literature review to acquire relevant material in a structured manner. The search process can involve both academic and non-academic databases. Depending on the chosen focus within the Coverage Consideration, the inclusion of grey/white literature might be helpful. It is recommended to use an established literature reviewing methodology (e.g., (vom Brocke et al., 2009)) to ensure rigor in this process. The search terms and inclusion/exclusion criteria should be determined within the group based on the previously selected DS scope. To identify patterns in the next phase, common themes and solutions from the literature have to be synthesized based on the defined structures from the Information Format Design.

In the last activity of this phase, *Information Review*, the information sources and the solutions to be considered therein are reviewed concerning their further processing by the research group (Fehling et al., 2014). The obtained volume of literature may be too large for the group to handle, refining the domain structure to create smaller and achievable sets of solutions. For example, limiting the scope to certain DS project activities might be useful. The criteria for literature review, both in terms of content and the possibilities for pattern development, can also be sharpened accordingly. The result set is further narrowed down in different filtering stages until a feasible set related to the specific problems is identified. If needed, forward and backward searches can be appended to further strengthen the pool of information sources.

4.2 Pattern Authoring

In the second phase, five activities are similarly undertaken (see Figure 3), and the patterns are formulated based on the similarities of existing solutions from the information basis (Fehling et al., 2014). Therefore, in the *Pattern Language Design*, the structure and sections of the patterns are defined (Fehling et al., 2014). As there is no universally valid format, these are individually determined by the pattern authors and adapted to their specific needs. Here, we recommend using a format using the sections *name, problem, context, challenges, solution, results, and links*, mainly built on the work of (Wellhausen and Fiesser, 2011). The *problem* is documented to support users in identifying and evaluating its suitability for their specific issues. *Context* defines the setting in which the pattern occurs. The difficulties encountered in addressing a

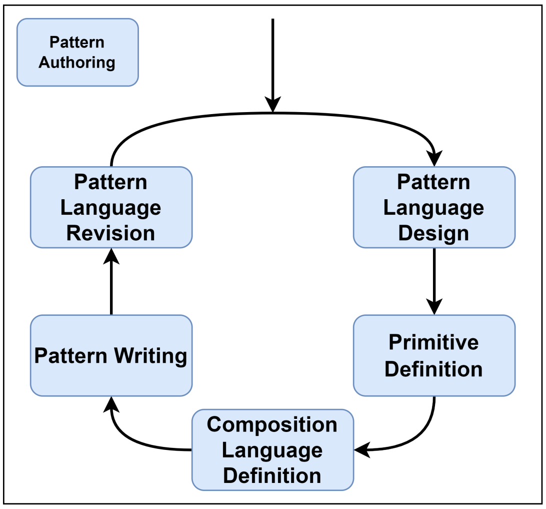

Figure 3: Pattern Authoring phase, adopted from (Fehling et al., 2014).

problem are also captured under *challenges* (Wellhausen and Fiesser, 2011). The application of the solution leads to *results*. To facilitate the application of patterns and illustrate their interconnections, relationships to other patterns are captured in the *links* (Wellhausen and Fiesser, 2011). Additionally, patterns are given a *name* by which they are identified. Each pattern in this context should be uniformly built upon the mentioned elements.

With the next activity, *Primitive Definition*, the definitions from Information Format Design can be further elaborated if needed (Fehling et al., 2014), depending on the insights gained from the literature. This aims to ensure consistent usage and homogenization of primitives within the research group.

In *Composition Language Design*, guidelines for sketches and formal specifications of the composition language are determined. For DS project management, this could involve the use of a DS process model (e.g., CRISP-DM) to clarify further the applicability of a pattern solution within a DS project.

The next activity, *Pattern Writing*, involves the actual documentation of the patterns based on the previously defined structure. The identified patterns from the literature are abstracted to a degree so that sufficient information is provided while remaining abstract enough to apply to various cases (Fehling et al., 2014). The process is iteratively repeated, with the patterns and individual sections revised and harmonized repeatedly. Discussions with other pattern authors or users are essential. Documentation is performed based on the approach presented in (Wellhausen and Fiesser, 2011) and begins with the solution of the pattern. Notes for the identified patterns and the information retrieved from the DS literature review are used for this purpose. Next, the problem is formulated concerning the described solution. The problem should not be trivial and is elucidated by questioning the relevance of the solution and the actual problem being addressed (Wellhausen and Fiesser, 2011). The result is then formulated to assess the impact of applying the solution. Subsequently, the challenges related to the results and the solution are identified. This involves examining why the described problem is more challenging to solve than it might initially appear (Wellhausen and Fiesser, 2011). Only afterward is the context described, determining the circumstances under which the problem arises. There is no optimal time to specify the name of the pattern, and it can emerge during the development process. In this case, the focus is on what supports the recall of the solution (Wellhausen and Fiesser, 2011). The links section is filled at the end once the interrelationships with other patterns have been clarified.

In the last activity of this phase, *Pattern Language Revision*, the created pattern language is evaluated and revised (Fehling et al., 2014). This can also involve the incorporation of DS practitioners to assess the usefulness of the individual patterns. Additionally, the structure and links between the patterns should be investigated. Patterns written at the beginning may have fewer connections and require revision.

4.3 Pattern Application

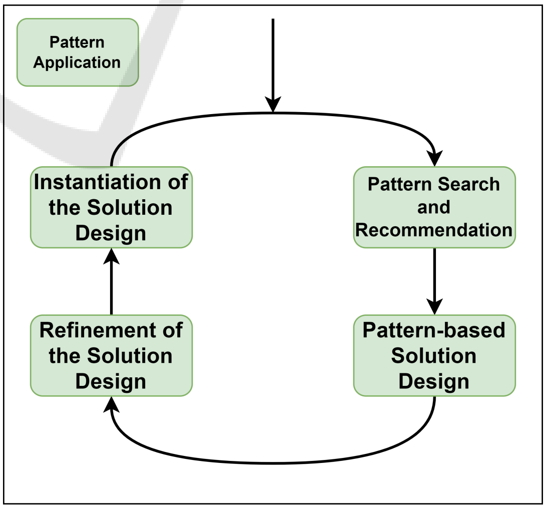
Pattern Application
Instantiation of the Solution Design
Pattern Search and Recommendation
Refinement of the Solution Design
Pattern-based Solution Design

Figure 4: Pattern Application phase, adopted from (Fehling et al., 2014).

The third phase, according to (Fehling et al., 2014), consisting of four subactivities (see Figure 4), pertains to the application of the developed patterns and can be considered and worked on independently of the other two phases. First, *Pattern Search and Recommendation* aims to facilitate users' navigation and identify suitable patterns for the specific use case at hand (Fehling et al., 2014). Accordingly, the derived DS patterns shall be accompanied by a summary section containing a brief description of the problem and solution of each of the patterns of the pattern language within a sentence (e.g., in tabular form) (Meszaros and Doble, 1997; Manns and Rising, 2012). Therefore, relevant patterns can be selected and further examined. For improved navigation, the patterns contain a *links* section, which indicates the relationship to other patterns. Moreover, it is recommended that this connection is graphically represented, too.

In *Pattern-based Solution Design*, support is provided for translating the abstract solution of a pattern into a specific use case (Fehling et al., 2014). The original literature, from which the solutions were abstracted, can be revisited to include concrete, existing solutions in the patterns to facilitate their usability (Fehling et al., 2014). Accordingly, the pattern notation is expanded by the section *"Examples"*, which provides a detailed reference solution of the DS problem described within the given pattern.

In the next activity, *Refinement of the Solution Design*, patterns are constrained and adapted to a specific environment where they should be applied (Fehling et al., 2014). In the DS context, this could involve a limitation to certain types of DS projects where the patterns are used to combat the possible differences in the manual pattern implementations.

Because of the determined pattern reference implementations and limitation to viable use case type(s), in the last activity, *Instantiation of the Solution Design*, the means to manage, configure, and deploy the patterns are determined (Fehling et al., 2014). Afterward, a specific refinement of the DS patterns based on the previously selected focus is enabled.

5 DEMONSTRATION

In alignment with the adopted DSR methodology, the application of the artifact to develop patterns for the management of DS projects is demonstrated in the following. This section describes how the authors applied the individual steps of the introduced method for identifying and notating patterns in the DS project management domain. Due to the page and time restrictions for this research, not all activities can be outlined to the fullest extent in this paper.

5.1 Pattern Identification

Domain Definition: As the goal of this research is centered around the development of patterns for DS project management, a joint understanding of theoretical fundamentals in this domain had to be established. Since some research group members already acquired some experience in the field, reference literature on conceptualization (e.g., NIST definitions) and common approaches (e.g., process models) were exchanged and discussed. The common understanding of the team was recorded. Section 3 constitutes the result of this activity.

Coverage Consideration: Due to the extensive nature of DS project management, it was necessary to limit the scope. Therefore, the focus was set on general successful approaches and the associated best practices regarding DS project management to address common problems and work toward mitigating the high failure rates encountered in the execution of data science projects (VentureBeat, 2019). The patterns should guide common challenges that arise during DS project phases and contribute to the development of effective data science project management.

Information Format Design: To adequately prepare for the Information Collection step, the research group jointly agreed to the joint use of certain tools and templates to facilitate collaboration for the following activities. For example, Citavi was used as a reference management tool for the literature. Here, the individual filtering stages were also depicted using the category feature. Accordingly, grouping related publications and mapping the inclusion/exclusion criteria was possible. Furthermore, to further organize the analysis of the obtained material, it was decided to use the concept matrix of (Webster and Watson, 2002). A corresponding template was created in Microsoft Excel, which the group jointly used.

Information Collection: The Information Collection for creating a DS project management pattern language was performed through a structured literature review according to the guidelines of (vom Brocke et al., 2009), consisting of the five phases definition of review scope, conceptualization of topic, literature search, literature analysis and synthesis, and research agenda. After completing the first two steps of this framework during Domain Definition and Coverage Consideration, Scopus and SpringerLink were queried with the search terms shown in Table 1, corresponding to the previously defined scope. The search yielded 286 results for *Scopus* and 205 publications for *SpringerLink*, where merely articles and conference papers were considered. In alignment with the outcome of Coverage Consideration, the re-

Table 1: Applied search terms for Information Collection.

Database	Search string
Scopus	TITLE((”Data Science” OR ”Big Data” OR ”Data Mining” OR analytics) AND (project OR management OR method* OR framework OR process OR model OR cycle)) AND TITLE-ABS-KEY ((”best practice” OR success OR pattern) AND project)
SpringerLink	1) *Title:* ”data science”, *All the words:* project, *Exact phrase:* best practice 2) *Title:* ”data science”, *All the words:* project, *One word:* success 3) *Title:* ”data science”, *Exact phrase:* project management

search group agreed to include articles that describe approaches or best practices for the execution and management of DS projects. Publications with an unfitting thematic focus, such as particular DS application scenarios and predominant coverage of technical questions (e.g., algorithms), were excluded. Each group member was assigned a literature subset for review.

Information Review: The obtained material base was filtered through multiple stages based on the inclusion and exclusion criteria. After the title assessment and removal of duplicates, 96 papers remained. Following the abstract evaluation, 38 articles were left in the pool. The full examination led to the removal of an additional 19 papers, resulting in an intermediate result set of 19 contributions directly relevant to successful approaches and best practices in DS project management. To expand the literature base for pattern writing, a backward and forward search (Webster and Watson, 2002) was also performed. This step added 13 more papers (32 in total). Next, a concept matrix was utilized to capture and analyze the common topics covered in the works to facilitate the extraction of the relevant information for the patterns.

5.2 Pattern Authoring

Pattern Language Design: Based on the introduced methodology for pattern construction, the research group employed the prescribed pattern structure, consisting of the sections name, problem, context, challenges, solution, results, and links (see Table 3). In the Pattern Writing step, these sections are completed based on the literature base.

Primitive Definition: The design decisions made during Information Format Design were reviewed and largely confirmed. Next to the concept matrix used to group the obtained information from the literature, the authors further agreed to use the joint Citavi repository for making more detailed annotations of important information within the included material since this can alleviate the pattern writing process.

Composition Language Definition: In this phase, it was determined to position patterns in the context of DS lifecycle stages proposed by (Haertel et al., 2022) to better clarify the applicability of the created patterns. Because of the focus on DS project management best practices, it was expected that most patterns would address the phase of *Business Understanding*.

Pattern Writing: Following the described approach, the sections of the patterns are completed in a specific order and as briefly as possible. Because of the page limitations, this step is described based on one example pattern (*”Alignment of Expectations”*) that was created during this process. The full pattern is depicted below in Table 3 and is written based on the inputs derived from the articles of (Cato et al., 2015; Gökay et al., 2023; Martinez et al., 2021b; Saltz and Shamshurin, 2016; Soukaina et al., 2019; Sun et al., 2018; Varela and Domingues, 2021; Yeoh and Koronios, 2010; Yeoh and Popovič, 2016; Schulz et al., 2020). Documentation begins with the solution of the pattern. An alignment regarding the potential of the to-be-developed DS application is required to raise awareness for realistic DS project expectations. This involves the project team, management, domain users, and other stakeholders. A situation assessment is needed to evaluate the feasibility of the set objectives. Based on similar problems, relevant resources (e.g., data, budget, skills) and their availability are discussed. Afterward, the problem section is elaborated to outline the relevance of the solution. Because of the data focus, DS projects have an explorative nature, which increases the difficulty of establishing goals and timelines. Additionally, management and users tend to have high expectations in DS applications. The results are written next to clarify the impact of the solution. This leads to a joint understanding of suitable expectations and the roughly required resources in the project. Based on the situation assessment, confidence is established regarding the feasibility of the project and its added value for the organization. Challenges in the context of this pattern mainly relate to the resources, especially data. A significant challenge in DS undertakings is data access. Additionally, the data exploration might reveal the unsuitability of the available data for achieving the initially set goals. Hence, a modification of the expectations could be necessary. This also applies to other resources like computing infrastructure or personnel.

Table 2: Abstracts for the created patterns.

Pattern name	Summary
Alignment of Expectations	Coordination of all stakeholders to sensitize regarding specific requirements, relevant resources, and challenges of the DS project.
Involvement of Senior Management	Upper management support, encouragement, and guidance are essential for the successful execution of data science projects, considering their distinctive demands and uncertainties.
Strategic Alignment of the Project	By aligning with the organizational strategy, the project enables the generation of valuable outcomes for the organization.
Scope	The project scope is carefully derived and maintained to facilitate the implementation of data science projects.
Process Organization	Processes are defined and established to facilitate controlled and targeted project execution, meeting project challenges.
Implementation of Change Management	DS projects requires a corresponding willingness and acceptance of changes that need to be incorporated development process.
Team Composition	Forming a team with members from different areas is crucial for managing DS projects, requiring comprehensive competencies and skills.
Project Team Competencies	Various technical and social competencies are required to manage aspects and requirements of DS projects.
Team Management	Promoting high productivity and cooperation among team members through effective leadership and coordination for successful project completion.
Ensuring Data Security	To ensure data security in data processing, security measures are integrated into project infrastructure and processes.
IT Infrastructure	To enhance productivity in the project, the IT infrastructure has to be set up under consideration of the unique requirements of the DS project.
Ensuring Data Quality	To be able to achieve the business objectives, suitable data must be integrated, prepared, and monitored to ensure high data quality.
Creation and Maintenance of Documentation	Documentation provides access to project procedures and (DS) results.
Project Performance Monitoring	Continuous monitoring of applications and processes enables efficient project implementation and infrastructure, facilitating project success.

Finally, the context of the pattern is described, which constitutes a specification of the problem to detail the circumstances under which it arises. Finally, the section on the links to related patterns is filled out and briefly elaborated (see Table 3). The Pattern Writing step is repeated multiple times to repeatedly revise and harmonize the sections and patterns. Using this procedure with the obtained material base, a total of 14 patterns for DS project management best practices were created. These are summarized in Table 2. The repeated occurrence and observation of identified solutions in the literature were significant for the formation of the patterns.

Pattern Language Revision: The components of the derived patterns were subject to multiple joint revisions within the research group. A further evaluation, including external experts, is planned to obtain further insights regarding the usefulness and completeness of the DS patterns for actual application.

5.3 Pattern Application

As the application of the developed patterns in practice in a DS project is still pending, this phase has only been partially completed so far, and thus, intermediate results are reported.

Pattern Search and Recommendation: This first activity focuses on improved search and navigation through the patterns. Therefore, each pattern was summarized in a sentence. The summary for the pattern ”Alignment of Expectations” is shown in Table 2. Furthermore, the links between the patterns were visually highlighted with the means of a cross matrix.

Pattern-based Solution Design: The pattern language notation was expanded with an ”Example” section to facilitate usability, using inputs from the material base. The result for the example pattern can be traced in the full notation in Table 3.

Refinement of the Solution Design: As the scope

Table 3: Pattern *Alignment of Expectations*.

Name	Alignment of Expectations
Problem	The explorative nature of DS projects increases the difficulty of establishing goals and timelines that confirm with expectations of management and domain users.
Context	DS project expectations are frequently not realized. Oftentimes, possibilities and results strongly depend on available resources, data access, and quality.
Challenges	The availability of resources impact the project outcome. A significant challenge in DS undertakings is data access. Additionally, the data exploration might reveal the unsuitability of the available data for achieving the business objectives. Hence, because of the inherent risks and uncertainties in DS projects, flexibility regarding the modification of the expectations might be necessary. This also applies to other resources like computing infrastructure or personnel.
Solution	The project team, management, domain users, and other stakeholders perform an alignment regarding the potential and limitations of the envisioned DS application. A situation assessment evaluates the feasibility of the set objectives and their added value for the organization. Based on detected similar problems and the corresponding solutions, the relevant resources (e.g., data, budget, competencies) and their availability are discussed.
Result	Development of a joint understanding of appropriate expectations and the approximately required resources and timelines. Based on the situation assessment, confidence is established regarding the feasibility of the DS project and its added value for the organization.
Links	Project expectations result from the *Strategic Alignment of the Project* and *Involvement of Senior Management*. Objectives have to be aligned with requirements to the project execution, including the *IT Infrastructure* and *Team Composition* to determine a realistic *Scope*. Moreover, expectations are defined regarding *Project Team Competencies* to complete the project tasks. During project execution, based on *Project Performance Monitoring* new or revised requirements and goals can arise.
Example	This pattern can be assigned to Business Understanding, which is a common phase in various DS process models. Here, the project circumstances are communicated with involved stakeholder groups to elaborate opportunities, requirements, and functionalities of the DS application (Schulz et al., 2020). A feasibility study can be used to evaluate the likelihood of fulfilling project requirements and objectives (Schulz et al., 2020).

of the Coverage Consideration and resulting Information Collection was set on general best practices and successful approaches for DS project management, no further limitations for application were defined in this step. However, based on the feedback of the revision with DS practitioners, this is subject to change.

Instantiation of the Solution Design: Due to the potential for various changes to the patterns in the future, the patterns are stored in a repository with shared access for each member of the research group. Version control is enabled to track changes that might be necessary based on the evaluation results and possible extensions to the material base.

6 CONCLUSION

DS projects suffer from high failure rates (VentureBeat, 2019), which indicates the need for new approaches for DS project management (Saltz and Krasteva, 2022). Therefore, in this work, we applied the pattern concept to DS since it allows for the structured summarization of solutions for common problems in a domain. Using the DSR methodology of (Peffers et al., 2007), the pattern creation process of (Fehling et al., 2014), consisting of Pattern Identification, Authoring, and Application, was adapted to DS project management. The functionality of the proposed method was demonstrated by the creation of patterns for DS project management best practices from the synthesis of scientific literature. Accordingly, researchers and practitioners can apply the introduced pattern construction method to synthesize solutions to frequently occurring issues in the execution and management of DS undertakings. Nevertheless, the study at hand is subject to certain limitations. Because of page restrictions, the evaluation within the DSR methodology following the guidelines of (Sonnenberg and vom Brocke, 2012) was not covered in this paper. While the ex-ante evaluations (Eval 1 and 2) were briefly touched upon in the background, further depth should be provided through literature reviews and expert interviews. The demonstration in Section 5 showed initial tendencies toward the artifact's general applicability, but a detailed assessment is still pending. Additionally, further case studies are needed to conclusively evaluate the usefulness of the proposed method and determine areas for improvement. For instance, the delimitation between certain activities (e.g., Primitive Definition and Composition Language Design) was not always clear, indicating the potential for consolidating some of the (sub)stages.

REFERENCES

Aichele, C. and Schönberger, M. (2014). *IT-Projektmanagement: Effiziente Einführung in das Management von Projekten*. SpringerLink Bücher. Springer Vieweg, Berlin.

Alexander, C. (1979). *The timeless way of building*. Oxford Univ. Pr, New York.

Alexander, C., Ishikawa, S., and Silverstein, M. (1977). *A Pattern Language: Towns, Buildings, Construction*.

Buschmann, F. (1996). *Pattern-oriented Software Architecture: A System of Patterns*. Wiley, Chichester.

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

Cao, L. (2018). Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3):1–42.

Cato, P., Golzer, P., and Demmelhuber, W. (2015). An investigation into the implementation factors affecting the success of big data systems. In *2015 11th International Conference on Innovations in Information Technology (IIT)*, pages 134–139. IEEE.

Chang, W. and Grady, N. (2019). Nist big data interoperability framework: Volume 1, definitions.

Coplien, J. O. (2000). *Software Patterns*. SIGS Books & multimedia, New York.

Coplien, J. O. and Harrison, N. B. (2005). *Organizational Patterns of Agile Software Development*. Pearson Prentice Hall, Upper Saddle River.

Das, M., Cui, R., Campbell, D. R., Agrawal, G., and Ramnath, R. (2015). Towards methods for systematic research on big data. *2015 IEEE International Conference on Big Data*, pages 2072–2081.

de Medeiros, M. M., Hoppen, N., and Maçada, A. C. G. (2020). Data science for business: benefits, challenges and opportunities. *The Bottom Line*.

Eisend, M. and Kuß, A. (2023). *Grundlagen empirischer Forschung: Zur Methodologie in der Betriebswirtschaftslehre*. Springer Fachmedien Wiesbaden and Imprint Springer Gabler, Wiesbaden, 3., überarbeitete aufage edition.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

Fehling, C., Barzen, J., Breitenbücher, U., and Leymann, F. (2014). A process for pattern identification, authoring, and application. In Eloranta, V.-P. and van Heesch, U., editors, *Proceedings of the 19th European Conference on Pattern Languages of Programs*, pages 1–9, New York, NY, USA. ACM.

Gökay, G. T., Nazlıel, K., Şener, U., Gökalp, E., Gökalp, M. O., Gençal, N., Dağdaş, G., and Eren, P. E. (2023). What drives success in data science projects: A taxonomy of antecedents. In *Computational Intelligence, Data Analytics and Applications*, pages 448–462, Cham. Springer International Publishing.

Günther, A. and Knote, R. (2017). How to design patterns in is research – a state-of-the-art analysis. *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*, pages 1393–1404.

Haertel, C., Pohl, M., Nahhas, A., Staegemann, D., and Turowski, K. (2022). Toward a lifecycle for data science: A literature review of data science process models. *PACIS 2022 Proceedings*.

Harrison, N. B. (2003). Advanced pattern writing: Patterns for experienced pattern authors. *EuroPLoP*, pages 809–828.

Heinrich, L. J. (1997). *Management von Informatik-Projekten*. R. Oldenbourg Verlag München Wien.

Hevner, A. R., March, S. T., and Park, J. (2004). Design science in information systems research. *MIS Quarterly*.

Iba, T. and Isaku, T. (2012). Holistic pattern-mining patterns: A pattern language for pattern mining on a holistic approach. *19th Pattern Languages of Programs conference*.

Jeble, S., Kumari, S., and Patil, Y. (2018). Role of big data in decision making. *Operations and Supply Chain Management*, Vol. 11(No. 1):36–44.

Manns, M. L. and Rising, L. (2012). *Fearless Change: Patterns for Introducing New Ideas*. Addison-Wesley.

Martinez, I., Viles, E., and Olaizola, I. G. (2021a). Data science methodologies: Current challenges and future approaches. *Big Data Research 24*.

Martinez, I., Viles, E., and Olaizola, I. G. (2021b). A survey study of success factors in data science projects. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2313–2318.

Meszaros, G. and Doble, J. (1997). A pattern language for pattern writing. *Pattern languages of program design*, pages 529–574.

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.

PMI (2017). *A guide to the project management body of knowledge (PMBOK guide)*. Sixth edition edition.

Saltz, J. (2022). Nine questions to evaluate a data science team's process: Exploring a big data science team process evaluation framework via a delphi study. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2667–2672. IEEE.

Saltz, J., Hotz, N., Wild, D., and Stirling, K. (2018). Exploring project management methodologies used within data science teams. *AMCIS 2018*.

Saltz, J. S. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. *IEEE International Conference on Big Data 2015*.

Saltz, J. S. and Hotz, N. (2020). Identifying the most common frameworks data science teams use to structure and coordinate their projects. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2038–2042. IEEE.

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

Saltz, J. S. and Shamshurin, I. (2016). Big data team process methodologies: A literature review and the identification of key factors for a project's success. In

2016 IEEE International Conference on Big Data (Big Data), pages 2872–2879.

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kuehnel, S., Badewitz, W., Dann, D., Kloker, S., Alekozai, E. M., and Lanquillon, C. (2020). Introducing dasc-pm: A data science process model. *ACIS 2020*.

Sonnenberg, C. and vom Brocke, J. (2012). Evaluations in the science of the artificial – reconsidering the build-evaluate pattern in design science research. In Peffers, K., Rothenberger, M., and Kuechler, B., editors, *Design science research in information systems*, SpringerLink Bücher, pages 381–397, Berlin. Springer.

Soukaina, M., Anoun, H., Ridouani, M., and Hassouni, L., editors (2019). *A study of the factors and methodologies to drive successfully a big data project*.

Sun, S., Cegielski, C. G., Jia, L., and Hall, D. J. (2018). Understanding the factors affecting the organizational adoption of big data. *Journal of Computer Information Systems*, 58(3):193–203.

Varela, C. and Domingues, L. (2021). Risks of data science projects - a delphi study. pages 982–989.

VentureBeat (2019). Why do 87% of data science projects never make it into production?

vom Brocke, J., Simons, A., 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*.

Wack, J. (2007). *Risikomanagement für IT-Projekte: Zugl.: Hamburg, Univ., Diss., 2006*, volume 54 of *Betriebswirtschaftliche Forschung zur Unternehmensführung*. Dt. Univ.-Verl., Wiesbaden, 1. aufl. edition.

Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, Vol. 26, No. 2 (Jun. 2002)*, pages 13–23.

Wellhausen, T. and Fiesser, A. (2011). How to write a pattern? a rough guide for first-time pattern authors. *Proceedings of the 16th European Conference on Pattern Languages of Programs*.

Wieczorrek, H. W. and Mertens, P. (2007). *Management von IT-Projekten: Von der Planung zur Realisierung*. Xpert.press. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2., überarbeitete und erweiterte auflage edition.

Yeoh, W. and Koronios, A. (2010). Critical success factors for business intelligence systems. *Journal of Computer Information Systems*, Vol. 50(No. 3):23–32.

Yeoh, W. and Popovič, A. (2016). Extending the understanding of critical success factors for implementing business intelligence systems. *Journal of the Association for Information Science and Technology*, 67(1):134–147.