Original article

INCREASING CONSISTENCY, TRACEABILITY AND TRANSPARENCY IN DATA SCIENCE PROJECTS: ANALYSIS AND FRAMEWORK

David J. Wolf[1], Adrian Specker[2]

[1,2]*University of Applied Sciences and Arts Northwestern Switzerland, School of Engineering, Aargau, Switzerland*

Received: 27 July 2023
Revised: 21 August 2023
Accepted: 24 January 2024

Abstract: Based on experiences in data-based projects, it was hypothesized that traditional project approaches often fail to ensure consistency, traceability, and transparency, contributing to a low success rate of such projects. This hypothesis was tested by compiling documented challenges from various data-based projects and comparing methods from literature and practice. The comparison enabled the formulation of objectives and led to the development of a novel method, focusing on standardization, regular exchange, and accountability to enhance consistency, traceability, and transparency of project-relevant objects. It also accommodates existing procedures for handling data-based projects. Application of this method allows for meticulous planning on multiple levels and iterative progress. Findings support the initial hypothesis, suggesting the method's potential to improve success rates in data-based projects.

Keywords: Consistency; Databased Projects; Data Science Method; Project Management; Traceability; Transparency.

1. INTRODUCTION

1.1 Motivation

This study addresses consistent issues in data science projects across various sectors, which often hinder their success and full realization despite the transformative potential of data analysis. Investigating the root causes of common problems like inconsistent data handling and lack of process transparency, our research aims to develop methodologies to enhance project efficiency and success rates. By identifying and strategically counteracting these challenges, we aspire to optimize resource utilization and amplify the benefits of data-driven decision-making in diverse applications.

1.2 Primary goal of the study

This study seeks to identify and address critical issues affecting data science project success through detailed secondary research involving academic articles, industry reports, and diverse case studies. Besides identifying consistent problems like inconsistency and lack of transparency, the research aims to propose effective strategies and methodologies as solutions, enhancing the likelihood of project success and furthering the potential of data-driven research and practice.

1.3 Hypothesis

Based on previous engagements in data science projects, we hypothesize that the absence of three elements: Consistency, Transparency, and Traceability, substantially contributes to project challenges.

- Consistency: Consistency refers to maintaining uniformity in processes, procedures, and information across the project's lifespan. A lack of consistency could lead to incongruities and errors, potentially derailing the project.

Corresponding author. Email: david.jan.wolf@gmail.com
ISSN 2560-4961 (online)

https://doi.org/10.56889/twsg4973

- Transparency: Transparency, on the other hand, involves clarity and openness in all stages of the project, allowing stakeholders to understand the project's progress and any emerging issues. Without transparency, problems could remain unidentified or unaddressed, jeopardizing the project's success.
- Traceability: Traceability refers to the ability to track the project's progress and changes, providing crucial context for decision-making. A lack of traceability could obscure the causes of issues, making them harder to resolve.

This hypothesis will guide our research and will be validated against empirical evidence, potentially identifying these elements as critical issues impacting data science project success or leading us to explore alternative contributing factors.

2. LITERATURE REVIEW

2.1. Challenges in data science projects

Numerous studies illustrate various challenges in data-based projects, predominantly focusing on issues with consistency, traceability, and transparency. Studies by Martinez et al. (2021), Larson and Chang (2016), and Kühn et al. (2018), pinpoint common issues such as poor stakeholder management, ineffective communication, unclear goal definitions, and inadequate focus on coordination and collaboration, all adversely affecting project traceability and data quality. Research by Weber et al. (2019), Stoudt et al. (2021), and Schock (2018), highlight additional problems like the lack of traceability, poor data quality, and the absence of high-quality data, impacting basic project requirements and outcomes. Various other sources, including Byrne (2017), FDFMS (2015), and Fayyad et al. (1996), underscore the significant impact of poor data quality, problematic definition of success criteria, and inadequate collaboration on the sustainability and traceability of project results. Furthermore, Merkelbach et al. (2022), Dietrich et al. (2015), and Schulz et al. (2020) identified challenges related to aligning business and IT goals, lack of a standardized analysis environment, and inadequacies in documentation and method use, respectively. Additional insights from DSPA (2019), Domino (2018), PMI (2020), and Microsoft (2016) also discuss challenges regarding goal definition, stakeholder management, and collaboration. Table 1 shows the summary of the problem collection from the research.

Table 1: Collection and categorization of problems in data science projects

Problems	Consistency	Traceability	Transparency	Others
Unclear role distribution	X			
Lack of standards, rules and concrete tools	X			
Unclear definition of goals and metrics	X	X	X	
Insufficient coordination		X		
Inadequate collaboration			X	
Poor communication			X	
Insufficient knowledge maintenance		X		
Poor sustainability of project results		X		
Lack of project documentation		X		
Inadequate stakeholder management		X	X	
Poor data quality				X
Inconsistent implementation of project procedure	X	X		

Although the comparison does not allow a clear assignment of the problems to one of the 3 properties, practically all of the problems mentioned concern at least one of them. This shows the close connection between consistency, traceability, and transparency, and it becomes clear that the three properties are decisive for the success of a data-based project. In the following part of this paper, we will

mainly focus on the solution of the mentioned problems in these three areas.

Thus, the thesis stated at the beginning can be confirmed. The proposed countermeasures from the literature are to increase the consistency of project-relevant objects within a data-based project by defining standards and rules. These also seem to have a positive effect on traceability. Increased collaboration and communication should also increase transparency and traceability. In this context, a definition of project-relevant objects leads to greater consistency. The monitoring and assurance serve the traceability. The communication of the current status can be related to transparency.

2.2. Countermeasures

In analyzing solutions for data-based project management, we categorized strategies into four primary improvement areas: consistency, traceability, transparency, and additional miscellaneous factors, all detailed in.

Table 2: Collection and categorization of solutions to problems in data science projects

Proposed solutions	Consistency	Traceability	Transparency	Others
Development of a standardized workflow	X	X		
Standardization of folder structure	X	X		
Standardization of processes	X	X		
Standardization of the documentation procedure	X	X		
Definition & monitoring of deliverables	X	X		
Continuous documentation		X		
Visualization of results		X	X	
Reconciliation of goals		X		
Consolidation of success metrics	X	X		
Communication of results			X	
Role definition	X	X		
Increasing collaboration between project participants		X	X	
Development of a unified knowledge base	X	X		
Monitoring of data sets	X	X		
Monitoring of models	X	X		
Knowledge management	X	X		
Ensuring the required data quality				X
Monitoring and maintenance of goals		X	X	
Standards for presentation of results	X	X	X	
Tool extensions to methods	X			
Extension of phases by taxonomies	X			
Definition of an analysis environment	X	X		
Definition of interfaces within a data-based project	X	X		
Combination of different models for execution of data-based projects				X

Solutions to augment consistency incorporate standardized workflows, structured folders and processes, consolidated metrics, continuous documentation, and defined roles. Enhanced traceability involves standardized workflows and documentation, continuous documentation, defined & monitored deliverables, result visualization, and ensured data quality, among

others. Improving transparency is achievable through standardized workflows, structured documentation, continuous documentation, result visualization, goal reconciliation, and increased collaboration. The miscellaneous category highlights increased project participant collaboration and defined analysis environments as crucial solutions. While maturity models for methods can enhance all three areas, the actual impact may depend on project-specific implementations. Thus, while providing a foundational framework, application of these solutions should be adapted per the distinct project requirements (Table 2).

2.3. Research gaps

Our literature review and practice examination in data-based projects highlight several crucial research gaps affecting project efficiency, effectiveness, and transparency, which could guide future research and methodology enhancement. A notable gap is the lack of a comprehensive framework that encapsulates all components from goal definition to final evaluation in implementing data-based projects, hindering project consistency and traceability. Additionally, there is an absence of a synthesized overview that demonstrates the interconnectedness of various project activities and their necessary conditions, obstructing teams from effectively navigating project dynamics. Also identified was a gap in defining specific tool utilization guidelines for activity execution within projects, affecting practical implementation effectiveness and efficiency. Lastly, the unavailability of technology- and tool-independent methods, which restricts method applicability across various project contexts, constitutes a significant research void. Addressing these gaps could substantially advance both theory and practice in data-based projects, enhancing project implementation in effectiveness, efficiency, and transparency.

2.4. Data science methodologies

In the rapidly evolving landscape of data science, a multitude of methods exist for conducting data-based projects. The efficacy and applicability of these methods, however, can be as diverse as the techniques themselves. To systematically comprehend and dissect the dynamics of these varying methods, we undertook an extensive review, seeking to highlight the relative strengths and weaknesses of each approach. A rigorous qualitative analysis was performed on these methods, focusing primarily on the processes adopted and the utilization of specific tools and measures. The objective of this analysis was to elucidate the degree of support these methods provide towards three fundamental aspects of any data-based project - consistency, traceability, and transparency. Interestingly, the review revealed a significant dichotomy. Certain methods overtly address the triad of consistency, traceability, and transparency, outlining explicit strategies to combat the lack thereof. In contrast, other methods do not overtly tackle these issues, and in-stead, their relevance and effectiveness in dealing with these challenges seem to be implicitly woven into the methodological fabric. This paper aims to elucidate these differences and provide a comparative perspective to better inform the selection and application of methods for future data-driven projects. The following is a list of the methods investigated:

- DMEPM - Data Mining Engineering Process Model (Marbán et al., 2009)
- Scr: Scrum (Schwaber & Sutherland, 2017)
- DDScr: Data-driven Scrum (DSPA, 2019)
- FMDS - Foundational Methodology for Data Science (FDFMS, 2015)
- GMLDSW - GitHub-ML Data Science Workflow (Byrne, 2017)
- KDD – Knowledge Discovery in Databases (Fayyad et al., 2014)
- BSPF - Business Science Problem Framework (Dametreus, 2019)
- DASC-PM – Data Science Process Model (Schulz et al., 2020)
- CRISP - Cross Industry Standard Process for Data Mining (IBM, 2011)
- eCRISP - Extended Cross Industry Standard Process for Data Mining (Schock, 2018)
- ASUM - Agile and Unified Method for Data Mining (Haertel, 2021)
- RAMSYS – Rapid Collaborative Data Mining System (Moyle & Jorge, 2001)
- DDSL - Domino Data Science Lifecycle (Domino, 2018)

- CMMLM - Comprehensive Management of Machine Learning Models (Weber et al., 2019)
- DAL - Data Analytics Lifecycle (Dietrich et al., 2015)
- TDSP – Team Data Science Process (Microsoft, 2019)
- ABIDF - Agile BI Delivery Framework (Larson & Chang, 2016)
- BDMC - Big Data Management Canvas (Kaufmann, 2019)
- SEMMA – Sampling, Exploring, Modifying, Modeling & Assessing (SAS Institute, 1999)
- HDSP – Harvard Data Science Process (Byrne, 2017)
- MLOps - Machine Learning Operations (PMI, 2020)
- DSPyW - Data Science with Python Workflow (Hathaway, 2021)
- DSRW - Data Science with R Workflow (Çetinkaya-Rundel et al., 2022)
- BEDSW - BinaryEdge Data Science Workflow (Byrne, 2017)
- BDMF - Big Data Managing Framework (Haertel, 2021)
- CDAW - Conceptual Data Analysis Workflow (Stoudt et al., 2021)
- BDIAI - Big Data Ideation, Assessment and Implementation (Haertel, 2021)

For the referencing of the models the abbreviations mentioned here are used from now on.

3. METHODS

3.1. Literature review process

In order to substantiate our hypothesis, we conducted an exhaustive literature review. This procedure entailed an organized exploration of multiple databases comprising academic journals, conference articles, and germane books, with a concentrated emphasis on literature addressing the difficulties and triumphs linked to data science projects. We established a targeted search strategy utilizing keywords linked to our hypothesis such as “failure in data science projects”, “consistency”, “transparency”, “traceability” and “project management.” Our aim was to explore a comprehensive spectrum of perspectives, both aligning and contrasting with our proposed hypothesis. Upon extraction, the gathered literature was meticulously analyzed to pinpoint recurring themes and valuable insights that correlate with our hypothesis. We adopted a critical stance, particularly towards contradictory evidence or discrepancies that could potentially challenge our proposed understanding of the problem. This diligent investigation of literature served as a crucial preliminary step towards validating or refuting our initial hypothesis and setting the stage for subsequent research steps.

3.2. Data collection and comparison of approaches

Post-literature review, a wide-ranging data collection on methodologies used in data science projects, from both academic and industrial sources, was initiated. The data collection process adhered to the following criteria for a rigorous content analysis:

- Source Selection: Data was gathered from a diverse range of academic and industrial sources, including peer-reviewed journals, conference proceedings, industry reports, and online repositories. This selection ensured a broad representation of methodologies.
- Inclusion and Exclusion Criteria: Inclusion criteria were defined to encompass relevant methodologies used in data science projects, while exclusion criteria were applied to exclude unrelated or outdated approaches.
- Data Retrieval: Systematic keyword searches and database queries were employed to retrieve relevant publications and materials. These searches were performed in multiple databases and sources to minimize bias.
- Data Coding: Each methodology identified during the data collection process was subject to qualitative evaluation. This evaluation focused on several key aspects:
    - Structure: Analyzing the overall framework and organization of the methodology.
    - Specific Methods: Examining the techniques and tools used within each methodology.
    - Consistency, Transparency, and Traceability: Assessing how each methodology addressed these critical factors.

- Detailed Documentation: Detailed records were maintained for each methodology, including publication details, authors, publication date, and a summary of its key features.
- Thematic Analysis: A thematic analysis approach was employed to identify recurring themes and novel ideas within the collected data. This allowed for the identification of common patterns and emerging trends.

The rigorous content analysis enabled us to identify recurring themes and novel ideas, providing a comprehensive perspective on varied strategies and how effectively they manage key issues identified in our hypothesis. This systematic approach ensured that the data collection process was robust, unbiased, and capable of generating valuable insights for our research.

3.3. Comparison methodology

An in-depth qualitative analysis was executed to compare collected approaches, guided by our hypothesis and focusing on how each method handles consistency, transparency, and traceability. A comparative framework, built on scholarly literature and our research objectives, was systematically applied, enabling a critical assessment of each approach's merits and drawbacks. The comparative analysis provided insights into the strengths and weaknesses of each methodology and played a vital role in validating or refuting our hypothesis, highlighting the problematic aspects affecting data science project success.

4. RESULTS

4.1. Comparison

4.1.1. Phases, processes and activities

Data-driven projects commonly utilize an iterative approach, organized under overarching thematic milestones. Through analysis, activities and phases from various methods have been categorized into broad segments: Preceding, Domain, Data, Development, Deployment, Subsequent, and Parallel Processes (refer to Table 3). Preceding Processes focus on foundational and definitional aspects, ensuring project readiness. Domain Processes establish clear project goals and align subsequent activities. Data Processes curate and clean the dataset, vital for ensuing transformations and output accuracy. Development Processes detail the blueprint for implementing the chosen solution, while Deployment Processes actualize the devised solution, ensuring it fulfills established objectives. Subsequent Processes pertain to solution maintenance, ensuring its continued effectiveness, and Parallel Processes involve concurrent, project-spanning tasks such as monitoring and communication. These categorizations enable structured conceptualization of diverse and complex activities in data-driven project execution.

Table 3: Categorization of the processes of different data science methods

	Preceding	Domain	Data	Development	Deployment	Subsequent	Parallel
DMEPM	X	X	X	X	X	X	X
ASUM		X	X	X	X	X	X
D	X	X	X	X	X		
DASC-PM	X	X	X	X	X		
DDSL		X	X	X	X	X	
FMDS		X	X	X	X	X	
ABIDF		X	X	X	X	X	X
BDMC		X	X	X	X	X	
CMMLM		X	X	X	X	X	X
MLOps		X	X	X	X	X	X
GMLDSW		X	X	X	X		
KDD		X	X	X	X		
BSPF		X	X	X	X		
eCRISP		X	X	X	X		
BDIAI		X	X	X	X		
CRISP		X	X	X	X		
RAMSYS		X	X	X	X		
DAL		X	X	X	X		
TDSP		X	X	X	X		
SEMMA		X	X	X			
HDSP		X	X	X			
CDAW			X	X	X		
BDMF		X	X				
DSPyW			X	X			
DSRW			X	X			
BEDSW			X	X			

4.1.2. Project-relevant objects

The relevance of objects in data-driven projects is pivotal, and for the purposes of method comparison, objects are classified into: Project-relevant, Phase-relevant, and Time-relevant categories. Project-relevant objects are crucial throughout all phases, Phase-relevant objects are significant in specific phases, and Time-relevant objects have importance at particular milestones. For example, in Preceding Processes, key objects include rules and roles, while in Parallel Processes, objects like project goals and risks persistently support the project. Domain Processes chiefly concern project goals, Data Processes prioritize datasets, Development Processes hinge on models and evaluations, Deployment Processes concentrate on operationalizing the model, and Subsequent Processes attend to the ongoing support of the deployed solution. This categorization and mapping underscore the interconnected nature of elements in data projects, offering a nuanced framework for method evaluation.

4.1.3. Tools

Data-driven projects leverage several methods aiming for enhanced consistency, traceability, and transparency throughout their execution. Consistency involves standardized processes and roles, utilizing tools and practices from various methods like ABIDF, DMEPM, and DAL, to ensure structured and accountable project flow. Traceability, encompassing the progression tracking of project tasks, employs tools and strategies from methods such as KDD, TDSP, and CMMLM to maintain continuous documentation and monitoring. Transparency seeks clarity in processes and objectives by using various tools and methods, including those from TDSP, KDD, and CMMLM, to facilitate clear communication and documentation throughout the project. Collectively, tools from these methods answer pivotal W-questions, reinforcing project methodology and facilitating its efficient and effective implementation.

4.2. Conclusion of analysis

Navigating through the challenges of data-based projects, particularly regarding inconsistent definitions, traceability, and transparency of relevant objects, demands the development of adept methods and strategies. Analyzing various approaches reveals methods that address these challenges by modifying existing strategies or devising new ones, all sharing a commonality of process flow aimed at defining relevant objects. Our findings advocate a generic procedure for managing project-relevant objects, involving their determination, consistent definition, clear documentation, and transparent communication, with standards, rules, and roles positively impacting project-relevant objects' properties. Identifying four generic project phases - Domain, Data, Development, and Deployment - each with specific activities and overarching goals, facilitates understanding project progression. Consequently, our insights into challenges and solution strategies in data-based projects lay a foundation for evolving more efficient methods in the field.

4.3. Matching problems and countermeasures

The research emphasizes the imperative of visual representations and detail variances in data-driven project methods, contributing to user accessibility and method acceptance. Integrating this, a generic extension needs to accommodate method variations and offer a standard procedure, ensuring universal project approach applicability. Addressing the complexity in defining relevant objects in data-driven projects and acknowledging the need for context-specific extensions is vital. Significance is found in meticulous planning and implementing phases and defining roles to ensure object consistency, traceability, and transparency, contributing to overall project success. Employing templates and iterations optimize transparency, traceability, and project team flexibility, while independent milestones and checklists enhance object quality, transparency, and traceability. Domain knowledge acquisition is vital, with a backlog facilitating ongoing activity and objective maintenance.

Table 4: Matching of problems and proposed solutions in data science projects

	Unclear role distribution	Lack of standards, rules and concrete tools	Unclear definition of goals and metrics	Insufficient coordination	Inadequate collaboration	Poor communication	Insufficient knowledge maintenance	Poor sustainability of project results	Lack of project documentation	Inadequate stakeholder management	Poor data quality	Inconsistent implementation of project procedure
Visual and detailed presentation of the method		X	X	X	X	X						X
Standardized approach for aligning the method with the project process		X		X				X				X
Planning and implementation phase for objects		X	X	X	X			X	X		X	X
Definition of roles	X	X								X		
Usage of templates		X			X				X			X
Iterative procedure			X	X	X	X	X	X			X	X
Definition of project-independent milestones		X	X	X	X	X		X				X
Definition of communication rules and fixed exchanged dates		X		X	X	X				X		X
Use of customer appointments for Feedback			X	X	X	X	X			X		X
Maintaining tasks and goals throughout the project			X	X	X		X	X	X			X
Definition of tasks in the backlog		X	X	X	X	X	X					X
Combining the backlog with generic activity areas and a folder structure		X	X	X	X	X	X	X	X		X	X
Phased data processing and model development		X		X	X		X	X	X		X	X
Documenting and communicating processing steps according to a defined standard		X		X	X	X	X		X		X	

The study further identifies persistent issues in data-driven project management, including unclear role distribution, lack of standards, and inefficient collaboration. Proposed countermeasures, derived from various method studies and compiled into a comprehensive approach, aim to address these issues robustly and strategically (Table 4). The ensuing chapter will elaborate on a conceptual methodology that integrates best practices from varied data-driven project methods, emphasizing clear role distribution, iterative procedures, template incorporation, phased data processing, and the innovative use of a backlog. Ultimately, this aims to significantly elevate the consistency, traceability, and transparency throughout data-based project implementations, providing a structured and comprehensive solution to prevailing data-management challenges.

4.4. DW-Model

4.4.1. Roles

In implementing data-based projects via the proposed method, defining roles is paramount for maintaining project object relevance and integrity, with each role possessing specific responsibilities. The method suggests four roles:

- Method Master, ensuring method implementation and managing project structure and deadlines;
- Product Owner, overseeing project backlog and ensuring alignment with customer needs through all project stages;
- Development Team, executing backlog activities and documenting results; and
- Customer, providing crucial feedback for aligning outcomes with expectations, without being involved in execution.

In essence, each role plays a vital part in facilitating consistent, traceable, and transparent project execution, highlighting the necessity of distinct role definitions in data-based project implementation.

4.4.2. Artifacts

In the proposed method for data-based project implementation, the use of artifacts plays a significant role in ensuring consistency and clarity of project-relevant objects. Artifacts, in this context, refer to the tangible outcomes or components that contribute to the project's organization, coherence, and alignment with its objectives. Key artifacts within this method include the folder structure and the project backlog.

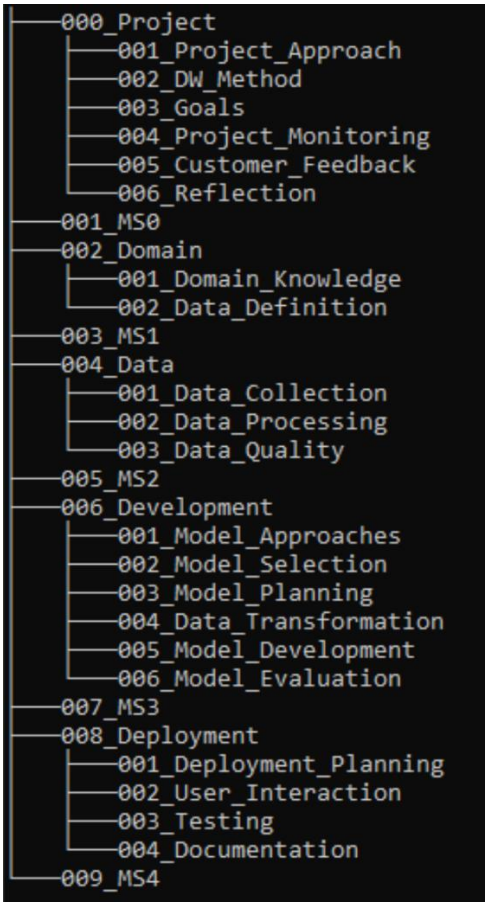

Figure 1: Generic folder structure for data science projects

The folder structure serves as a centralized repository for storing all information, rules, and knowledge produced throughout the project and its various phases. (Figure 1) The primary aim of centralizing the folder structure is to prevent the creation of information silos in private computer systems of project participants, fostering a shared knowledge base. Standardizing the folder structure not only facilitates easier access to information but also enhances

traceability for the project participants. The Method Master takes responsibility for maintaining the repository and ensuring adherence to standards.

Goals	Goal 1						
	Goal 2						
	Goal 3						
	....						
	Milestone	**Phase**	**What for?**	**What?**	**When?**	**Who?**	**Status**
MS0	**Basic requirements**						
		Project	Definition	Project Approach			
		Project	Definition	Project Roles			
		Project	Definition	Extension			
		Project	Definition	Adjustment			
		Project	Project Monitoring	...			
		...	...	...			
MS1	**Data definition**						
		Domain	Domain Knowledge				
		Domain	Data Definition				
		...	...				
MS2	**Complete and cleansed dataset**						
		Data	Data Collection				
		Data	Data Processing				
		Data	Data Quality				
		...	...				
MS3	**Developed preferred model**						
		Development	Model Approaches				
		Development	Model Selection				
		Development	Model Planning				
		Development	Data Transformation				
		Development	Model Development				
		Development	Model Evaluation				
		...	...				
MS4	**Operationalized solution**						
		Deployment	Deployment Planning				
		Deployment	User Interaction				
		Deployment	Testing				
		Deployment	Documentation				
		Deployment	Knowledge Transfer				

Figure 2: Components of the generic backlog for data science projects

The project backlog, maintained centrally by the Product Owner, encapsulates the project's goals and activities (Figure 2). The backlog includes the overarching project goal, intermediate milestone goals, activity areas, and the specific tasks necessary to achieve these milestones. Milestone backlogs form when activities related to a particular milestone goal are collated. Once all activities of a milestone backlog are completed, the corresponding milestone is deemed achieved. The iteration backlog, on the other hand, comprises activities selected from the milestone backlog at the onset of an iteration. The three-tiered backlog structure allows for a step-by-step refinement of activities. Importantly, the project backlog is directly linked to the folder structure, reinforcing the consistency of project-relevant objects. The artifacts serve as a medium to answer essential project questions like 'what for,' 'what,' 'who,' 'when,' and 'where' initially, followed by 'why,' 'how,' and 'with what' during the progression of the project. Thus, the proposed method's artifacts ensure consistent object definition and provide a robust structure for project management.

4.4.3. General procedure

Data-based project implementation is structured across Project, Phase, and Iteration levels, each with specific objectives, procedures, and deliverables.

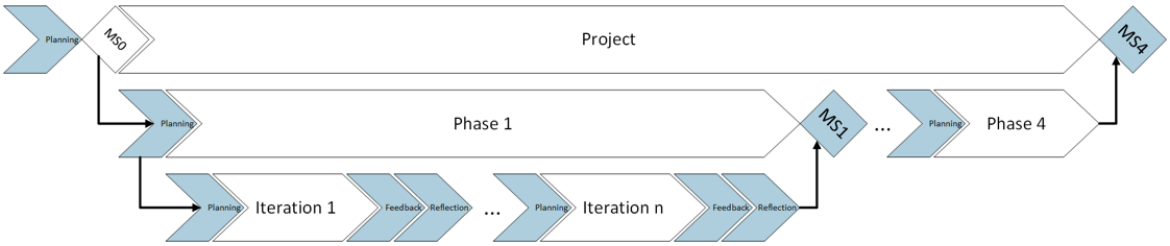

Figure 3: Generic three level process for a data science project

The Project level establishes object definitions; the Phase level strategizes activities for milestones; and the Iteration level focuses on detailed planning, execution, and continuous improvement through feedback and adjustments. Methodology uses a project backlog and centralized folder structure across all levels, ensuring planning and documentation are traceable (Figure 3).

4.4.4. Milestones

For systematic project execution, key milestones are established:

- MS0: Ensures basic requirements and foundational preparations.
- MS1: Involves comprehensive data definition, facilitating targeted data processing.
- MS2: Necessitates a cleansed data set for model development.
- MS3: Requires identifying a preferred model aligned with objectives.
- MS4: Achieved when the preferred model is operationalized in the real production environment.

Accomplishing these milestones ensures systematic, traceable, and transparent project outcomes, adhering to a structured approach for consistent implementation and monitoring.

4.4.5. Rules

Team members adhere to specified rules and utilize templates from a central repository for backlog activity execution and documentation. Upon task assignment, a uniquely named folder (date and activity name) is created for saving all relevant documents and standard-compliant documentation. Only after the method master's approval of the folder and documentation, tasks can be marked as "Done".

Throughout milestones 1 and 2, teams systematically organize specific objects (data sets, models, evaluations) in designated folder structures to ensure clarity and avoid redundancy. Raw and processed data are stored in suitably named subfolders to facilitate chronological sorting and easy identification of the most recent data set. Strict adherence to naming conventions and storing protocols ensures clear management and traceability of data and models without overwriting or deleting inputs.

Meetings, organized and led by the Method Master, ensure alignment with DW method rules and achievement of meeting objectives. The Product Owner inputs and modifies backlog entries during meetings with team consensus. The Method Master, while able to delegate documentation responsibilities, ensures execution, filing, and post-meeting central folder structure alignment adhere to specified rules. Additionally, adherence to rules concerning folder and documentation creation is verified upon task completion.

4.4.6. Events

The method enforces structured events or deadlines with specific timings and participants, ensuring communication and project monitoring.

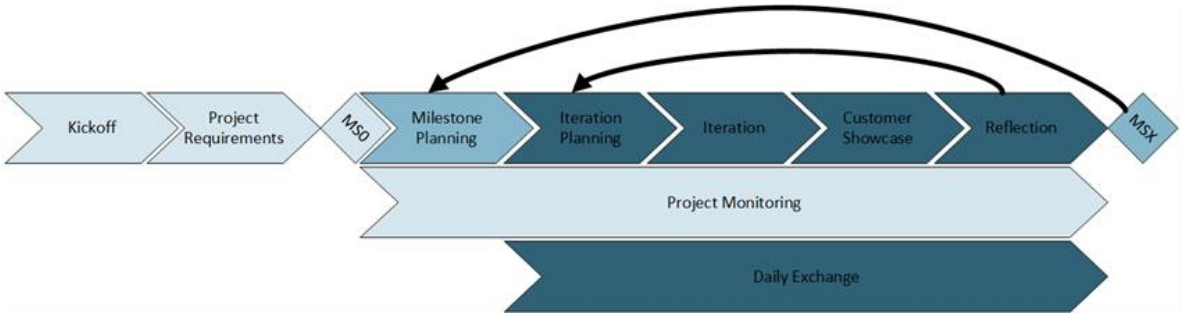

Figure 4: Generic events & activities for a data science project

Figure 4 illustrates how these structured checkpoints, such as "Kickoff" and "Project Foundations", streamline communication and transparency through defined steps like project approach selection, role clarification, and creation of a centralized data repository.
The "Kickoff" event initiates the project, aligning stakeholders such as the Method Master,

Product Owner, and Development Team. Essential steps undertaken include:

- Project Approach Selection: The team, collectively, chooses a project approach, which also informs the creation of the project's folder structure.
- Project Goal Definition: Stakeholders collaboratively define a clear project goal to ensure uniform understanding.
- Role Clarification: Team members are assigned distinct roles and responsibilities using a template.
- Data Repository and Backlog: The team selects a centralized data storage location and a backlog maintenance method, with tool choice remaining flexible.

Post project approach selection, team members acquaint themselves with the guide, and the Method Master initiates specific tasks originating from the kickoff, including creating a centralized folder structure to manage project information. Key tasks before the subsequent team meeting involve:

- Method Familiarization: The Method Master thoroughly studies the extension guide, understanding components and processes.
- Centralized Repository Creation & Access: A central repository is developed, and access is provided to all project members, using a predetermined tool.
- Aligning Milestones with Procedure: Integration of a generic method extension with a specific project structure occurs, focusing on designated hierarchy levels. Specific phases and milestones (e.g., MS1 to MS4) are identified and mapped with generic method's phases and activity areas.
- Folder Structure Implementation: Folders for each phase and milestone, aligning with the generic structure, are created by the Method Master.
- Activity Area Alignment: Activity areas are mapped to specific project procedure phases.
- Populating Folder Structure: The Method Master fills the structure with pertinent information, like kickoff documentation and milestone checklists.
- Scheduling Subsequent Steps: Following repository completion, the Method Master organizes the next meeting.

In the "Project Requirements" event, key stakeholders, including the Method Master and Product Owner, collaboratively establish and agree on the project's requirements. Essential steps include:

- Centralized Repository and Backlog Introduction: The Method Master introduces and explains the centralized data repository and project backlog, handing over backlog maintenance to the Product Owner.
- Objective Transfer to Backlog: The Product Owner moves goals set during the kickoff into the backlog, witnessed by all project participants.
- Project-Relevant Object Identification: Collectively, the team identifies and agrees upon critical project objects from a list, ensuring continual monitoring.
- Defining Objects in the Backlog: Project objects are uniformly defined and entered into the backlog by the Product Owner, ensuring clarity and designated responsibilities among team members.
- Metrics and Template Creation: Documentation templates are completed for each object, answering all pertinent W-questions.
- MS0 Achievement: Upon all stakeholders approving the MS0 checklist items, planning for MS1 commences.

Subsequently, the Method Master organizes project-relevant object subfolders and templates, ensures MS0 requirements are met, and coordinates the first phase-level deadline while maintaining oversight of project object monitoring deadlines.

During periodically scheduled "Project Monitoring" events, key stakeholders track and adjust project progress. Essential steps include:

- Project Objects Overview: Initiating with a synopsis of defined project-relevant objects.
- Status Updates: Designated individuals present each object's status and history using monitoring templates.

- Risk Mitigation: The meeting addresses identified project risks and discusses potential corrective measures.
- Adding Objects: Additional objects can be proposed and, if accepted, are added to the Backlog and assigned monitoring metrics and templates by the Product Owner and Method Master, respectively.

The “Milestone Planning” event at each phase's start aligns the Method Master, Product Owner, and Development Team towards phase goals through several key steps:

- Milestone Backlog Creation: Activities and tasks necessary for the next milestone are outlined in the backlog. Additional activity areas identified are added by the Product Owner and might adjust next milestone's checklist requirements.
- Task Execution & Documentation Rules: Method Master establishes and ensures understanding of activity execution and documentation standards, with the team agreeing on a uniform task process representation.
- Phase-Specific Standards: Particularly in Data and Development phases, the Method Master emphasizes data and model maintenance principles, underlining centralized rule storage.

Post-meeting, the Method Master adds any new activity areas to the central folder structure and oversees subsequent milestone planning adherence throughout the project.

The “Iteration Planning” event identifies and agrees on tasks for the upcoming iteration, consisting of several steps:

- Iteration Backlog Creation: Tasks from the Milestone Backlog move to the Iteration Backlog, marked with the iteration number in the 'When?' column by the Product Owner, considering feedback and reflections from the last iteration. Tasks are clarified or subdivided as needed.
- Daily Exchange Setup: The team sets a time and place for daily meetings to discuss and resolve issues throughout the iteration.
- Responsibility Assignment: Initial tasks and respective project members are assigned and noted in the backlog's 'Who?' and 'Status' columns.

No additional documentation outside the backlog is needed. The Method Master ensures adherence to Iteration Planning rules in subsequent iterations throughout the project.

The Daily Exchange is a brief meeting involving the Method Master, Product Owner, and Development Team to synchronize on the iteration's progress, tackle potential obstacles, and agree on the day’s plan. Key steps include:

- Status Update: Team members update on task progress. Completed tasks, complying with all rules, are marked “Done” in the backlog.
- Task Assignment: New tasks are assigned and updated to “In Progress” in the backlog.
- Problem Solving: Obstacles are identified and solutions are initiated. The Method Master oversees adherence to the rules throughout the iteration.

At the iteration's end, the Customer Showcase, led by the Method Master, displays the project's progress to the customer, aiming to gather feedback. Vital steps involve:

- Status Update: The Product Owner details the iteration's progress.
- Customer Feedback: Feedback, influencing the backlog or project goal, is gathered and documented.
- Examination of Milestone Requirements: Milestone achievement is evaluated and, if fulfilled, planning for the next is initiated.

Following the meeting, the Method Master archives the documentation and customer feedback in specific folders.

The Reflection meeting, post-Customer Showcase, involves reflecting on the past iteration's efficiency and communication. Key elements:

- Internal Reflection: The team evaluates the past iteration, discussing and documenting improvement areas.
- External Reflection: Customer communication is assessed and strategies for enhancement are considered.

Subsequent to the meeting, the Method Master archives the Reflection documentation with a date-named file path in the central repository.

5. DISCUSSION

The development of a generic guideline aimed to enhance consistency, traceability, and transparency in data-centric projects, leveraging insights from both practical and theoretical findings from analyzed project-relevant objects, phases, and activities. The constructed model blends elements of the Scrum method and original analysis, incorporating a Backlog and an iterative approach for phase execution, linked with a standardized folder structure, which facilitates a comprehensive and consistent definition of project activities across all dimensions. While its application has showcased the method's potential in preventing several issues by centralizing goals and information in the Backlog and maintaining a uniform understanding of tasks and milestones among team members, the method's success hinges on strict adherence to rules and role responsibilities. It remains adaptable in terms of tool usage, though this flexibility might introduce inconsistency and obscurity, warranting future studies to explore potential tool standardization within the method. Overall, despite requiring additional testing and optimization, the method emerges as a viable framework for augmenting consistency, traceability, and transparency in data-based projects.

6. CONCLUSION

The research substantiated that consistency, traceability, and transparency of pertinent objects are crucial for the triumph of data-driven projects, with their complexity best navigated through standardization and communication. An amalgamation of various methods into a generic approach, derived from analyzing different data-project strategies, augments project success and streamlines management of project-related objects.

While the current methodological form permits the use of specific tools and ensures detailed tracking and documentation through the backlog and central storage, there are several avenues for further research:

- Real-Condition Development: Exploring real-world conditions and their impact on the standardization process could lead to improved methodologies.
- Dedicated Tool Development: Investigating the creation of a dedicated software tool to enhance standardization and automate tasks such as folder generation and task duration calculation.
- Advanced Documentation Techniques: Researching advanced techniques for standardized definition and chronological documentation of activities, potentially enabling automated, user-controlled generation of extensive project documentation sorted by parameters like iteration number and start date.
- Extending Applicability: Studying the potential for introducing overarching milestones and related requirements to extend the method's applicability beyond data science
  projects.

This study, while offering a framework for data-project implementation and revealing areas for further refinement and expansion, underscores the vital role of continual development and adaptation in optimizing the method's efficiency and utility in varied project contexts. These research directions represent opportunities to enhance the methodology further and address specific challenges in data-driven projects.

REFERENCES

Byrne, C. (2017). Development Workflows for Data Scientists. O'Reilly Media, Inc.

Cetinkaya-Rundel, M., Hardin, J., Baumer, B., McNamara, A., Horton, N., & Rundel, C. (2022). An educator's perspective of the tidyverse. Technology Innovations in Statistics Education, 14(1). https://doi.org/10.5070/t514154352

Dametreus, V. (2019). Business Science Problem Framework. State College, PA.

Dietrich, D., Heller, B., Yang, B., & EMC Education Services. (2015). Data science & big data analytics: Discovering, analyzing, visualizing and presenting data. John Wiley & Sons.

Domino. (2018). The Practical Guide to Managing Data Science at Scale-Lessons from the field on managing data science projects

and portfolios. Data Lab, Inc.

DSPA. (2019). DATA DRIVEN SCRUM GUIDE: AN AGILE FRAMEWORK DESIGNED FOR DATA SCIENCE. Data Science Process Alliance.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Commun. ACM, 39(11), 27–34. https://doi.org/10.1145/240455.240464

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (2014). From Data Mining to Knowledge Discovery in Databases. Association for the Advancement of Artificial Intelligence (AAAI).

FDFMS. (2015). Foundational Methodology for Data Science. TDWI (The Data Warehousing Institute).

Haertel, C. (2021). Grundlagen Data Science und IT- Projektmanagement/-controlling (p. 65). Otto-von-Guericke-Universität Magdeburg.

Hathaway, J. (2021). 1 Introduction | Python for Data Science. University of Michigan. https://byuidatascience.github.io/python4ds/introduction.html

IBM. (2011). IBM SPSS Modeler CRISP-DM Guide (p. 53). IBM.

Kaufmann, M. (2019). Big Data Management Canvas: A Reference Model for Value Creation from Data. Big Data and Cognitive Computing, 3(1). https://doi.org/10.3390/bdcc3010019

Kühn, A., Joppen, R., Reinhart, F., Röltgen, D., von Enzberg, S., & Dumitrescu, R. (2018). Analytics Canvas – A Framework for the Design and Specification of Data Analytics Projects. Procedia CIRP, 70, 162–167. https://doi.org/10.1016/j.procir.2018.02.031

Larson, D., & 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. https://doi.org/10.1016/j.ijinfomgt.2016.04.013

Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. Information Systems, 34(1), 87–107. https://doi.org/10.1016/j.is.2008.04.003

Martinez, I., Viles, E., & G. Olaizola, I. (2021). Data Science Methodologies: Current Challenges and Future Approaches. Big Data Research, 24, 100183. https://doi.org/10.1016/j.bdr.2020.100183

Merkelbach, S., Von Enzberg, S., Kuhn, A., & Dumitrescu, R. (2022). Towards a Process Model to Enable Domain Experts to Become Citizen Data Scientists for Industrial Applications. 2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems, 1–6. https://doi.org/10.1109/ICPS51978.2022.9816871

Microsoft. (2016). Team Data Science Process: Productivity practices for collaborative data science. Microsoft.

Microsoft. (2019). What is the Team Data Science Process? (p. 403). Microsoft.

Moyle, S., & Jorge, A. (2001). RAMSYS-A methodology for supporting rapid remote collaborative data mining projects.

PMI. (2020). Playbook for Project Management in Data Science and Artificial Intelligence Projects. Project Management Institute.

SAS Institute. (1999). Data Mining and the Case for Sampling: Solving Business Problems Using SAS® Enterprise MinerTM Software. SAS Institute Inc.

Schock, C. (2019). CRISP-DM - Ein Ansatz zur Systematisierung von Digitalisierungsprojekten in Produktionsumfeldern. https://doi.org/10.24406/publica-fhg-299535

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kuehnel, S., Badewitz, W., Dann, D., Kloker, S., Alekozai, E. M., & Lanquillon, C. (2020). Introducing DASC-PM: A Data Science Process Model. AIS Electronic Library (AISeL). https://doi.org/10.25673/92266

Schwaber, K., & Sutherland, J. (2017). Der Scrum Guide: Der gültige Leitfaden für Scrum: Die Spielregeln. Creative Commons.

Stoudt, S., Vásquez, V. N., & Martinez, C. C. (2021). Principles for data analysis workflows. PLOS Computational Biology, 17(3), e1008770. https://doi.org/10.1371/journal.pcbi.1008770

Weber, C., Hirmer, P., Reimann, P., & Schwarz, H. (2019). A New Process Model for the Comprehensive Management of Machine Learning Models. Proceedings of the 21st International Conference on Enterprise Information Systems, 415–422. https://doi.org/10.5220/0007725304150422