Scaling Data Governance Through Automation and Dynamic Data Spaces

by
Bjarki Freyr Sveinbjarnarson

Dissertation submitted to the School of Computer Science
at Reykjavik University in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

2nd of May, 2025

Supervisors:	Anna Sigríður Islind, Professor, Reykjavik University
	Erna Sif Arnardóttir, Associate Professor, Reykjavik University
Examining Board:	Stefán Ólafsson, Assistant Professor, Reykjavik University
	Tomas Lindroth, Assistant Professor, University of Gothenburg
Examiner:	Olivia Benfeldt, Assistant Professor, Copenhagen Business School

Scaling Data Governance Through Automation and Dynamic Data Spaces

Bjarki Freyr Sveinbjarnarson

May 2nd, 2025

Abstract

Managing vast amounts of heterogeneous data is a growing challenge in healthcare research, where data must be structured, stored, and validated for diverse research purposes while minimizing the manual resources required for maintenance. This thesis explores scalable data management solutions, focusing on database design, data quality enforcement, and automated data governance to ensure usability across research disciplines.

We designed and developed a homogeneous database structure capable of accommodating a wide range of structured health data while minimizing complexity and ensuring flexibility for future research needs. This database architecture enables the seamless integration of new data sources without expanding storage requirements, making it a scalable solution for large-scale research projects. Additionally, we implemented a validation system embedded within a digital infrastructure to enforce predefined data standards, reducing errors at the point of entry and improving data quality.

The empirical work underlying this thesis is based on four studies conducted within a large sleep research project encompassing diverse datasets. First, we evaluated the feasibility of displaying key dataset insights through a digital infrastructure, providing researchers with an overview of stored data. Second, we developed a homogeneous database framework that eliminates the need for additional tables as new data sources are added. Third, we conducted a study to identify the root causes of poor data quality at the input stage, analyzing how researchers interact with data validation mechanisms and where errors emerge. Lastly, we examined data governance models that minimize human intervention while maintaining research integrity, ensuring sustainability in data management.

Findings from these studies reveal common causes of poor data

quality, including a lack of metadata prioritization, reliance on assumptions instead of documentation, inconsistent formatting, and diverse interpretations of what constitutes as good data. Researchers also expressed conflicting priorities regarding data governance, further complicating standardization efforts. Despite these challenges, our homogeneous database design and automated data governance framework provide a workable solution for multidisciplinary research projects by reducing manual oversight and improving long-term data usability.

This thesis contributes to the fields of data governance, database architecture, dynamic data spaces, information systems, and digital infrastructure by demonstrating how scalable, automated solutions can streamline data processes while ensuring high-quality, standardized data for diverse purposes.
Keywords: Data, Data Governance, Automation, Digital Infrastructure, Data Management, dynamic data spaces, Information Systems

Acknowledgements

I owe my deepest gratitude to my supervisors, Anna Sigríður Islind and Erna Sif Arnardóttir, whose expertise, encouragement, and unwavering guidance have been invaluable throughout this journey. Their insightful feedback and belief in my work have helped shape both this thesis and my academic growth.

Getting to know my sweetheart, Elena Richert, was the highlight of my PhD. Her love, patience, and unwavering support have been my anchor throughout this journey, and she has been my pillar to lean on through every high and low.

Special thanks to Daníel Emil Sigurðsson, a dear friend and invaluable colleague, whose dedication and hard work significantly lightened my PhD journey. His willingness to take on numerous projects and streamline critical processes in the Sleep Revolution not only improved the efficiency of our work but also allowed me to focus on completing this thesis. His support came at a crucial time, making an immense difference in ensuring that I could see this journey through. I am deeply grateful for his contributions and his unwavering commitment.

Lastly, my heartfelt thanks go to my family—my parents, Sveinbjörn Höskuldsson and Kolbrún Eydís Ottósdóttir, and my brother, Atli Þór Sveinbjarnarson—whose unwavering encouragement and belief in my abilities have been the foundation of my success. Your love and support have made this journey possible, and I am endlessly grateful to have you by my side.

This thesis is part of the Sleep Revolution project, funded by the European Union's Horizon 2020 research and innovation program under grant agreement no. 965417. I am deeply grateful for the opportunity to contribute to this important initiative and for the support that made this work possible.

Publications List

Paper 1. Layer Upon Layer: Developing Layered Modular Architectures for Data-Driven Health Platform

Full reference: B. F., Arnardottir, E. S., & Islind, A. S. (2024). Layer Upon Layer: Developing Layered Modular Architectures for Data-Driven Health Platform. In Digital (Eco) Systems and Societal Challenges: New Scenarios for Organizing (pp. 91-108). Cham: Springer Nature Switzerland.

Paper 2. The Sleep Revolution Platform: A Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data.

Full reference: Sveinbjarnarson, B. F., Schmitz, L., Arnardottir, E. S., & Islind, A. S. (2023). The sleep revolution platform: A dynamic data source pipeline and digital platform architecture for complex sleep data. Current Sleep Medicine Reports, 9(2), 91-100.

Paper 3. Data Work in Healthcare: Mediating Data Quality and Data Governance in a Data-Intensive World

Full reference: Sveinbjarnarson, B. F., Schmitz, L., Arnardottir, E. S., & Islind, A. S. (2025). Data Work in Healthcare: Mediating Data Quality and Data Governance in a Data-Intensive World. Submitted to *Scandinavian Journal of Information Systems*. (Accepted).

Paper 4. Let the System Handle It: Simplifying Data Governance Using Automation

Full reference: Sveinbjarnarson, B. F., Arnardottir, E. S., & Islind, A. S. (Forthcoming). Let the System Handle It: Simplifying Data Governance Using Automation. (Under construction).

Other Selected Publications by the Author

Schmitz, L., Sveinbjarnarson, B. F., Gunnarsson, G. N., Davíðsson, Ó. A., Davíðsson, Þ. B., Arnardóttir, E. S., ... & Islind, A. S. (2022, August). Towards a Digital Sleep Diary Standard. In Americas Conference on Information Systems (AMCIS), Minneapolis, USA, August 2022.

Sveinbjarnarson, B. F., Arnardóttir, E. S., Islind, A. S., (2023, October). Designing and Developing Layered-Modular Architecture for Data-driven Health Platforms. In Italian Conference of Information Systems (ITAIS), Turin, Italy, October 2023.

Contents

Chapter 1

Introduction

Modern research and organizations increasingly rely on digital data to function. But with data volumes growing nearly a hundredfold over the past 15 years [1], the situation has become overwhelming. In practice, this often means researchers working with hundreds of disconnected spreadsheets, files, and folders—some containing thousands or even millions of entries. Without proper structure, managing this data becomes slow, error-prone, and frustrating. This growing burden has been referred to as technical debt or data debt—terms that describe how incomplete, unprocessed, or poorly managed data accumulates over time, creating a backlog of unfinished data work across organizations [2].

Good data quality—meaning that data is complete, consistent, and usable—is critical for making research reliable and repeatable. But as data flows in greater amounts without the resources or methods to keep up with it, quality eventually suffers [3]. Mistakes multiply, documentation is forgotten, and valuable data becomes difficult or even impossible to reuse [3]. To help address these challenges, many organizations have started to turn to data governance [4]–[7].

However, data is often managed without formal data governance structures. Many organizations begin by creating informal data principles or standards—broad guidelines about how data should be collected, stored, and used. These are often shaped by current needs, available tools, or individual preferences. For example, teams may agree on which file formats to use, where to store files, or how to name variables. While this works in smaller

settings, such informal approaches become increasingly difficult to maintain as datasets grow. It is not uncommon for projects to manage data across hundreds of spreadsheets or files, often with millions of entries, spread across disconnected folders and cloud services.

Although such data standards are valuable, they are not always consistently followed—especially when short-term project needs or time pressures take priority. Even well-intentioned teams may fall back on inconsistent or makeshift data practices under short-term pressure, as noted by Chu and Dexter [8]. Many technological giants similarly notes that long-term success in managing data requires more than short-term wins; it depends on building structures that promote trust, accountability, and repeatability [9].

Data governance is a structured, team-based way of managing data to ensure it remains usable and trustworthy. It involves clearly defining roles and responsibilities, setting rules for how data should be collected, structured, and shared, and regularly checking whether those rules are followed [10], [11]. But traditional approaches to data governance often rely on manual oversight and require significant resources to maintain—resources that are often scarce in academic and clinical research settings [12], [13].

This is where data governance becomes essential. Rather than relying on individual judgment, data governance introduces structured roles, responsibilities, and procedures that clarify how data should be handled, validated, and shared. It shifts the focus from isolated tasks to collaborative efforts aimed at ensuring data quality and long-term sustainability [14]. Data governance typically also involves mechanisms for monitoring and feedback, helping teams stay aligned with agreed-upon standards even as data and staff change over time [15].

Effective data governance depends on several key factors. One of the most important is clarity—not only about the data itself, but about how it is meant to be used, by whom, and under what conditions [16]. As organizations manage increasingly large and complex datasets, the underlying processes often become harder to coordinate and sustain [10]. This complexity is not just technical [17]. It also arises when individuals struggle to understand their roles, responsibilities, and the rules guiding data use [18]–[20]. As a result, even well-intentioned efforts at managing data

may fail to last over time.

Another key challenge lies in monitoring—specifically, whether data-related procedures are actually followed. It is not enough to define standards—organizations also need systems to verify that data is entered correctly, that required metadata is present, and that users interact with the data as expected. This ongoing oversight is often cited as one of the most demanding aspects of data governance [21]–[23]. While many data governance strategies rely on documentation and review cycles, these can quickly become unmanageable as the number of data sources, systems, and users grows [24], [25].

This is why effective data governance requires more than comprehensive documentation or rule-setting. It must be socio-technical—that is, it must address both the technological systems that store and manage the data, and the people, practices, and routines that shape how data is created, shared, and interpreted [26]. A growing body of literature calls for simpler, more usable data governance frameworks that reduce barriers to participation and support sustainable, everyday use in real-world settings [27]–[29].

Despite increasing interest in data governance, much of the existing literature continues to focus on business-oriented cases, especially where data is standardized and structured around commercial goals. Less attention has been paid to contexts involving sensitive personal data—such as health, behavior, or biology—which often require stricter protections, greater flexibility, and higher trust among users [21], [30]. These are precisely the kinds of environments where more adaptive, automated, and people-centered approaches to data governance are most needed.

This thesis explores how this largely manual workload in data governance can be simplified and partially automated to reduce effort while still supporting reliable, long-term use of research data.

Today's digital systems are rarely built from scratch. Instead, they are assembled from many moving parts—different platforms, services, and components that exchange data behind the scenes. As data flows through these systems, it is stored, processed, and transformed in ways that are difficult to track or manage. This creates what is often described as a data space: a shared environment where data moves across organizational and technical boundaries [31], [32]. When this data flow is continuous and constantly evolving, it is referred to as a dynamic data space [33].

The term dynamic data space has emerged in response to growing challenges in managing data that crosses systems with different rules, formats, and update cycles [34]. In earlier systems, data was often kept in isolated, local repositories that could be controlled within a single organization. Today, however, services are commonly distributed. Key functions such as recruitment, consent management, and analytics are handled by different external platforms. While each of these may apply strong internal data governance, coordination between them is often lacking [32], [35], [36]. We lay out the main differences between dynamic data spaces and static systems in table 1.1.

This lack of coordination becomes especially problematic when sensitive personal data is involved, such as health or behavioral information [32], [33]. Regulations like the GDPR (General Data Protection Regulation) require that individuals must be able to request the deletion of their personal data—not just in one system, but everywhere it might be stored or processed [37]. In a fragmented environment, where copies of data exist across multiple services, ensuring full compliance with such requests becomes highly complex [33].

Dynamic data spaces aim to address these challenges by enabling more structured, secure, and accountable ways for data to move between systems [38]. Their goals include improving transparency, coordinating the lifecycle of data (from creation to deletion), and maintaining legal and ethical compliance, even when data is shared among many actors [39]. However, these environments bring new challenges of their own. Systems must be able to communicate despite technical differences, organizations must coordinate their decisions, and stakeholders must retain oversight even when relying on complex or opaque systems—often described as black boxes [33], [40].

A core reason for this difficulty is that dynamic data spaces are decentralized by nature [32]. No single actor controls the entire system. Instead, responsibility is distributed: each stakeholder contributes components, follows its own internal logic, and makes independent decisions. While this decentralization supports flexibility and scalability, it also creates friction for tasks that require system-wide coordination—such as ensuring the correct version of a dataset, validating metadata, or deleting data on request.

This thesis does not attempt to address every aspect of dynamic data spaces. Instead, it contributes targeted solutions re-

lated to data governance and automation that support the broader goals of dynamic data spaces. The proposed model introduces shared standards and practices that help streamline operations across distributed systems. By simplifying tasks like data deletion, metadata tracking, and validation, the solution helps improve trust, traceability, and coordination in settings where services are evolving and responsibilities are shared.

Key Differences		
Characteristic	Static Systems	Dynamic Data Spaces
Data movement	Centralized or limited	Decentralized and continuous
Update frequency	Infrequent batch updates	Real-time or frequent syncing
System architecture	Single platform or siloed	Distributed and interconnected
User interaction	Manual and isolated	Collaborative and adaptive
Compliance management	One-time validation	Ongoing monitoring and control

Table 1.1: Comparison of static systems and dynamic data spaces

Dynamic data spaces do not exist on their own—they rely on a strong digital foundation to function. This foundation is known as digital infrastructure, the underlying technological framework that supports how data is stored, moved, and processed across systems. Without reliable infrastructure, it becomes nearly impossible to maintain the fast, complex, and large-scale data flows that dynamic data spaces require [41].

The meaning of digital infrastructure has evolved over time. Today, it refers not just to technical tools, but to broader sociotechnical arrangements—the combination of digital technologies, users, data practices, and physical components that together enable digital operations [42], [43]. These infrastructures often span many organizations and systems, making them complex and interdependent.

Digital infrastructure is especially relevant in the context of dynamic data spaces because both depend on the ability to coordinate across diverse systems. Many infrastructures today include multiple platforms, each with different formats and standards, yet they are expected to exchange data reliably [44]. Supporting this kind of interoperability—the ability of different systems and technologies to exchange and use data effectively—is one of the central roles of digital infrastructure.

However, building and maintaining such infrastructure is challenging. It requires significant investment in time, money, and expertise. It can be difficult to adapt existing systems to new requirements, and the more complex the infrastructure becomes, the harder it is to oversee and manage it effectively [41], [45], [46]. These challenges are often amplified in resource-limited environments, where organizations must achieve high levels of coordination and security with limited technical or financial capacity.

This thesis was conducted within the large-scale research initiative Sleep Revolution, which received approximately 15 million euros in funding from the European Union's Horizon 2020 program [47]. The project seeks to advance the science of sleep by combining two main types of data collection: retrospective clinical data—existing health data gathered from hospitals across Europe—and prospective data, which is newly collected during the course of the project. These datasets include personal information from individuals across Europe and are drawn from a wide range of sources.

To enable large-scale analysis, all data is brought together in a centralized computing environment known as a high-performance computing cluster. A cluster is a powerful system composed of multiple connected computers, designed to process and store large amounts of data efficiently. Because the data involved is sensitive and pseudonymized, the data is not permitted to leave this secure environment, ruling out the use of public cloud services.

This setup required the development of a dynamic data space—a secure system for sharing data within the project without violating privacy or regulatory requirements. To support this, careful data governance was needed, along with a digital infrastructure capable of handling growing demands over time. The overall goal of Sleep Revolution is to improve the diagnosis and treatment of sleep disorders by standardizing scoring methods, digitalizing how people report their sleep experiences, improving

how sleep is measured, and developing new methods for interpreting this information [47], [48].

Given sleep's vital role in overall health—despite being historically under-researched—the project brings together experts from many fields, including neuropsychology, machine learning, medicine, sports science, and information systems [48], [49]. This disciplinary diversity results in a wide range of research approaches, tools, and data types, shaped by the needs and practices of different fields. A shared ambition across the project is to move sleep measurements beyond traditional clinical settings and into participants' everyday lives. This shift is supported through the use of wearables—devices like smartwatches and fitness trackers that record physiological data—and mobile apps that gather subjective feedback, such as how rested a person feels. These are combined with objective measurements like movement, heart rate, and breathing patterns to create a more holistic and continuous view of sleep.

This real-world, high-volume data poses serious challenges for infrastructure and long-term use. To be valuable, data must not only be collected—it must also be documented, findable, and reusable. That means building a digital infrastructure that can support a wide variety of data sources and keep everything organized and traceable over time. However, as is typical in many research environments, the resources allocated to data management are limited. The Sleep Revolution project is no exception. Budget constraints required practical solutions for data handling that remained compliant, efficient, and scalable. These included strategies to improve metadata quality, support future reuse, and reduce technical debt over time [50].

Like many dynamic data spaces, the project also had to respect individual rights under regulations like the GDPR [37]. Participants must be able to withdraw their data at any time, which introduces complex requirements for data deletion, auditing, and communication across distributed systems. These requirements called for a combination of strong data governance, well-coordinated infrastructure, and automation to reduce manual effort and risk.

Importantly, the project aims to leave behind more than just results—it seeks to build a sustainable infrastructure for future sleep research. Yet this goal can be at odds with short-term research needs. Each participating discipline has its own timeline

and priorities, which may not align with the long-term objective of creating reusable datasets. This highlights the importance of a socio-technical approach: one that considers not only the technical systems in place, but also the people, practices, and institutional context surrounding them. In such environments, data governance, digital infrastructure, and dynamic data spaces must work together—balancing flexibility for individual researchers with the shared responsibility of managing data in a secure, transparent, and durable way.

My work has focused on designing and developing the digital infrastructure of the Sleep Revolution project—conceptualized in this thesis as a dynamic data space—and studying its effects through use. This effort spanned four years of development work, culminating in the infrastructure that now supports data processing within the project. Its design and impact are analyzed across the four papers appended to this thesis.

The PhD journey has required adaptation, resilience, and continuous learning. It began with a focus on developing a digital platform to connect researchers and participants with their data. Early into the process, however, I encountered the realities of working with fragmented, inconsistent, and often incomplete real-world data. This practical exposure prompted a shift toward understanding the underlying structures that make—or break—data reusability. It led to research on layered modular architectures in digital platforms (papers 1 and 2), and later on data governance and the nature of everyday data work (papers 3 and 4). These shifts were not theoretical in nature; they were driven by the demands of managing diverse data inputs—from multilingual questionnaires to smartwatch sensor streams and digital sleep diaries captured through mobile apps.

The task of integrating, validating, and preparing these datasets for research use revealed the necessity of simplifying and, where possible, automating core data processes. Over time, I developed a more comprehensive digital infrastructure that allowed for more consistent data storage, automated validation, and easier access to high-quality data for researchers. In doing so, the work progressively took shape as a dynamic data space.

This process presented a number of challenges. Supporting other researchers, coordinating data across sources, and dealing with prototype tools such as early-stage APIs (Application Programming Interface) required extensive manual effort. These ex-

periences highlighted the often invisible—but crucial—data work that underlies research infrastructure. This aligns with what Barley called the "backrooms of science" [51], and more recent work on the invisibility of technical labor in digital research systems [52]. The amount of manual, behind-the-scenes work reinforced the need for a socio-technical approach—one that acknowledges not just the technical infrastructure, but also the people and practices needed to sustain it over time.

Throughout the project, the focus has remained on building digital solutions that directly address the complexity of handling large, heterogeneous datasets in an era of growing personal datafication. By automating core elements of data governance and developing scalable infrastructure, the work seeks to reduce dependence on manual processes and ensure that data is both trustworthy and reusable. These solutions support ongoing research and establish a foundation for future projects by offering design guidelines for others working on similar infrastructure challenges.

This PhD contributes to the field by proposing practical strategies for reducing complexity in dynamic data space without undermining their flexibility or compliance. It is guided by the following research question:

- **Research Question:** What socio-technical strategies can reduce complexity in data governance while maintaining functionality, scalability, and compliance in dynamic data spaces?

This thesis is based on four interrelated papers that together address the challenge of designing, developing, and governing large-scale data systems under constrained resources. Across these studies, the thesis presents a simplified dynamic data space paired with a streamlined data governance framework. The overall contribution demonstrates how complexity in data governance can be reduced without compromising functionality, scalability, or compliance. In doing so, the thesis contributes to the information systems field broadly, and to the literature on data governance, dynamic data space, and socio-technical aspects in digital environments more specifically.

Paper 1, *Layer Upon Layer: Developing Layered Modular Architecture for Data-Driven Health Platforms*, explores the architectural foundation for handling complex data systems. Initially presented at the Italian Conference on Information Systems and

later extended into a book chapter in Digital (Eco)systems and Challenges, this paper introduces a layered-modular architecture for data extraction and representation. The goal was to create a reusable, adaptable platform architecture capable of supporting multidisciplinary research and forming the basis for a unified data space. Evaluation involved expert usability assessments, semi-structured interviews with ten participants, and scalability testing with growing data volumes.

Paper 2, *The Sleep Revolution Platform: A Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data*, published in Current Sleep Medicine Reports, builds directly on the architectural foundation of Paper 1. It addresses the challenge of integrating heterogeneous sleep data—spanning subjective and objective measurements collected over time—into a standardized structure. The main contributions are the conceptualization of a homogeneous data space and the development of a dynamic data pipeline for transforming diverse inputs into a coherent, reusable format. This paper moves the technical foundation closer to a functioning dynamic data space by supporting consistent, long-term data use across research domains.

Paper 3, *Data Work in Healthcare: Mediating Data Quality and Data Governance in a Data-Intensive World*, currently in third-round review at the Scandinavian Journal of Information Systems, shifts the focus from architecture to everyday data practices. It investigates the root causes of data quality issues in healthcare research by examining the data work required to make data usable within the infrastructure developed in Papers 1 and 2. Using a combination of feasibility testing, surveys, and semi-structured interviews with 13 healthcare researchers, the study identifies recurring errors and underlying behaviors. Key findings include the neglect of metadata, overreliance on assumptions, inconsistent formatting, and variable interpretations of data quality. These insights point to the need for better validation, documentation, and training practices.

Paper 4 *Let the System Handle It: Simplifying Data Governance Using Automation*, builds on the same empirical foundation as Paper 3 but focuses on solutions rather than challenges. It introduces an automated data governance framework integrated into the dynamic data space developed throughout the thesis. This framework supports real-time validation, enforces metadata standards, and generates structured outputs with minimal manual

oversight. Drawing on the same study, the paper outlines how automation and embedded feedback mechanisms can ease the data governance burden in resource-limited environments. The proposed model promotes simplicity, improves data reliability, and fosters sustainable data governance practices across multidisciplinary projects.

Together, the four papers form a coherent progression. Paper 1 proposed a flexible architecture but revealed challenges in practical usability, leading to the development of the more refined infrastructure and data pipelines in Paper 2. Papers 3 and 4 then empirically evaluated this infrastructure through direct interaction with researchers using the system. Paper 3 analyzed persistent data quality issues and their socio-technical roots, while Paper 4 shifted toward an actionable data governance model grounded in automation and user-centered design.

This thesis advances existing literature by combining theory-informed analysis with practical design work. While much has been written about the challenges of data governance, dynamic data spaces, and digital infrastructure, this research contributes concrete methods for addressing those challenges in practice. By embedding data governance into infrastructure, automating routine tasks, and supporting user learning through feedback, the proposed model enhances interoperability, ensures data accuracy, and supports the creation of sustainable, evolving data ecosystems.

Chapter 2

Theoretical Background and Related Work

This chapter introduces the core concepts that form the foundation of this thesis: (i) data governance, (ii) dynamic data spaces, and (iii) digital infrastructure. These three areas are central to both the challenges addressed and the contributions made throughout the work. Each section begins by explaining the concept in simple terms, followed by a discussion of why it is relevant today. The chapter then outlines key findings from the literature, highlights ongoing challenges, and ends by describing how this thesis builds upon or extends existing knowledge. A fourth section on interoperability is also included, as it plays a critical role in connecting the other three concepts and was discussed throughout the introduction.

2.1 Data Governance

Data governance has been part of organizational discourse for several decades [53], [54], but its meaning remains somewhat ambiguous. As data-driven practices have become central to modern operations, the term has evolved from a theoretical framework into a practical necessity [15]. At its core, data governance provides a structure for how individuals and systems interact with data, with the goal of ensuring its accuracy, security, and usability. In this thesis, data governance is defined as "the framework of poli-

cies, standards, roles, and processes that ensure data is accurate, secure, and usable throughout its lifecycle" [17].

The concept gained increased attention after 2010 [54]–[56], as organizations began to understand the operational and economic risks of unmanaged data. Poor data governance can result in not just technical problems but real-world costs. In 2014, the economic impact of insufficient data management in the U.S. was estimated at over 600 billion dollars annually [57]. By 2023, this had risen to 3 trillion dollars [58], highlighting the compounding consequences of ungoverned or mismanaged data assets. As data has become the world's most valuable resource—surpassing even oil in economic significance—its proper data governance is more critical than ever [59].

As the volume, variety, and velocity of data increase, so too does the complexity of governing it [60]. Data governance is a socio-technical challenge—it involves not only systems and technologies but also people, roles, and organizational structures [17], [26], [54], [61]. Human factors such as unclear ownership, lack of accountability, and resistance to change are often as difficult to manage as the technical aspects, which include scalability, automation, and interoperability [16], [20], [62].

Recent studies have identified several success factors and barriers. A consistent theme in the literature is the lack of clarity around data governance roles and objectives [17], [20], [26]. Without defined responsibilities, organizations often experience fragmented efforts and inconsistent enforcement of policies. Relatedly, organizations struggle to adapt their data governance frameworks to change—whether due to shifting regulations or evolving technologies—resulting in resistance and inefficiencies [61], [63].

Monitoring and enforcement present another major hurdle. While setting up a data governance model is difficult, ensuring it is followed over time is even harder. Manual oversight quickly becomes unsustainable at scale, making automation an essential component of any modern data governance strategy [21], [23], [64]. The challenge is to strike a balance: overly rigid models can inhibit innovation, while too much flexibility can lead to data governance failure [65].

These issues are further complicated by the heterogeneity of data and users. Data may appear in multiple formats across systems, and individuals working with data bring diverse skill levels, priorities, and practices. In complex environments like research

institutions or healthcare systems, data governance often breaks down at the point of entry, where the need for standardization and validation is highest. For example, Brous et al. [28] found that success depends on monitoring data at intake, maintaining oversight of standards, and clearly defining ownership structures to avoid confusion and conflict. These findings reflect a broader consensus: effective data governance must be both technically robust and socially embedded [15].

Despite growing awareness of its importance, data governance is still highly manual in practice. Organizations often rely on human labor for validation, metadata management, and access control. This creates inefficiencies and bottlenecks that are difficult to scale [66]. Surveys show that many professionals spend a substantial part of their work week on data-related tasks, often at the expense of more strategic activities [67]. These inefficiencies are especially problematic in settings where resources are limited.

In response, data governance models are beginning to evolve. Organizations are moving toward more automated, integrated frameworks that pair policy with technical solutions [67]. Modern platforms can automate data lineage tracking, metadata standardization, workflow management, and compliance monitoring [14]. This shift is reflected in movements like data-centric AI, which seek to embed data governance principles directly into data workflows, improving trust and reducing the need for reactive fixes [66].

Even low-resource settings are beginning to benefit from automation. A recent case study in Zanzibar's health system showed how an automated data governance framework enabled the adoption of AI (Artificial Intelligence) technologies, despite limited infrastructure and technical capacity [68]. This underscores the broader relevance of structured, scalable data governance models—not only in corporate environments, but also in healthcare, academia, and public service sectors.

Still, major challenges remain. Interoperability issues, lack of standardization, and limited technical expertise continue to slow progress [26], [32], [64]. In particular, the ability to monitor data governance across distributed systems and to enforce policies consistently in dynamic, evolving environments remains an unresolved issue [21], [23].

This thesis contributes to this evolving field by designing and implementing a data governance model embedded directly into

the infrastructure of a large-scale research project. Drawing from real-world development, the work shows how automation, clear role definitions, and continuous validation can reduce complexity without compromising compliance or usability. By integrating data governance into the architecture of a dynamic data space, this thesis offers a practical and adaptable solution for managing data in resource-constrained, multidisciplinary environments.

2.2 Dynamic Data Spaces

Dynamic data spaces are secure, collaborative digital environments that enable organizations to share, integrate, and manage data while retaining full control over their data assets [31], [69]. Unlike traditional data warehouses, which rely on centralized storage and fixed schema designs, dynamic data space are decentralized and flexible. They are designed to support continuous data flows and seamless data exchange across different systems, organizations, and even entire industries. dynamic data spaces aim to support interoperability, safeguard data sovereignty, and maintain compliance with legal and ethical regulations. However, their decentralized and evolving nature often introduces new complexities in achieving these very goals [33], [70]. The relevance of dynamic data spaces has grown significantly with the increasing need for real-time data access, cross-border collaboration, and privacy-aware data use. One of their defining characteristics is the ability to enable real-time data exchange while preserving data privacy and security [71]–[74]. This makes them especially useful in sectors such as healthcare, government, and scientific research, where access to sensitive data must be balanced with strict regulatory requirements [75], [76]. In Europe, dynamic data spaces have gained policy traction due to growing regulatory scrutiny, particularly under initiatives like the European Data Governance Act and the broader European Data Strategy [77], [78]. Several major efforts are now underway to build dynamic data space at scale. The European Health Data Space (EHDS), for example, aims to create a unified network for secure medical data sharing among healthcare professionals, researchers, and patients [79]. Similar initiatives are emerging in fields like energy, mobility, and finance, each reflecting a broader push toward breaking down data silos and enabling shared infrastructures [77]. These initiatives promise

greater data reuse, interoperability, and innovation across sectors. However, as dynamic data spaces expand, so do the challenges. A central issue is interoperability—ensuring that data from diverse sources, domains, and formats can be seamlessly integrated and used [31], [32], [80]. Although the European Union's regulatory frameworks are attempting to address this through common standards [78], consistent implementation across regions remains difficult. Fragmentation caused by differences in national data laws and sector-specific policies slows progress toward unified data governance [81]. In addition to regulatory hurdles, there are deeper socio-technical and economic tensions. Interoperability, while critical for accessibility and collaboration, often increases complexity [82], [83]. Greater technical integration can require more sophisticated data governance structures, which in turn create barriers to adoption [31]. Scholars have also raised concerns over the risk of "winner-takes-all" scenarios, where dominant organizations define the terms of access and participation, limiting inclusivity and fairness [69], [84]. These issues highlight the importance of data governance models that are not only technically sound, but also transparent, equitable, and inclusive. The ethical dimension of dynamic data spaces is equally important. In domains like healthcare and government, ensuring participant control and data retractability is critical [37], [85]. Regulations such as the GDPR require that individuals can delete their personal data across all systems where it is used [37]. This becomes particularly complex in data environments where information is constantly flowing, duplicated, or transformed. Health-related research adds another layer of difficulty, as it involves heterogeneous medical data, evolving consent requirements, and strict standards for protecting patient and participant rights [49], [86]. Recent literature has called attention to these socio-technical challenges. Bacco and Kocian (2024) emphasize that as data spaces grow in scope, their complexity increases, slowing adoption [31]. They propose minimum interoperability mechanisms and shared components as ways to manage this complexity. Other scholars warn that without such frameworks, efforts to build interoperable data spaces may result in fragmented or incompatible ecosystems [87]. Deshmukh et al. further argue that most existing data space models are still limited to context-specific applications and lack flexibility for broader use [77]. Some have proposed query-based interoperability techniques to allow dynamic interactions without requiring

full standardization [88], while others stress the need for foundational advances in data standardization and data governance [89]. Another emerging issue is the question of research integrity within shared data environments. Recent increases in retracted scientific publications have raised concerns about data quality, traceability, and accountability in collaborative infrastructures [90]–[92]. Without robust provenance tracking and validation mechanisms, dynamic data spaces risk spreading errors or misinformation at scale [70]. Despite these emerging proposals, many existing frameworks focus primarily on system-level architecture and technical standards, often overlooking the practical realities of data governance in constrained environments. While recent work by Gieß et al. has begun mapping the design options for data space architectures [39], [40], these frameworks emphasize high-level components and offer limited guidance on how to implement data governance effectively in resource-limited or continuously evolving settings. This thesis contributes to these discussions by presenting a dynamic data space that has been developed, deployed, and tested within a large-scale, interdisciplinary research project. The model developed here simplifies many of the core tasks associated with dynamic data space—such as integration, validation, metadata tracking, and participant control—while ensuring compliance and scalability. Unlike much of the existing literature, this work focuses on the practical realities of building dynamic data space under constraints, offering a socio-technical approach that balances data governance, infrastructure, and usability.

2.3 Digital Infrastructure

Digital infrastructure forms the backbone of modern data governance and dynamic data spaces. It includes all the systems and services needed to store, process, and move digital data within and across organizations. This infrastructure is composed of both physical elements—such as servers, data centers, and network equipment—and software-based tools like cloud platforms, APIs, and cybersecurity frameworks [93]. As organizations face increasing volumes and complexity of digital data, modernizing digital infrastructure has become essential for maintaining operational efficiency, data security, and compliance [94], [95].

The importance of robust infrastructure is especially clear in

dynamic environments, where data flows constantly between platforms and stakeholders. However, many organizations still struggle to upgrade or maintain their systems [96], [97]. Manny et al. identified key barriers to infrastructure modernization, including unclear strategic direction, limited funding, and poor planning [98]. These socio-organizational issues often lead to fragmented systems that lack scalability, integration, and proper security. On the other hand, studies like Li et al.'s work in Tibet show the upside of infrastructure investment: improvements in digital infrastructure were strongly linked to higher productivity and economic performance [99]. These findings underline the role infrastructure plays not just in digital operations, but in broader social and economic development.

One of the most transformative shifts in recent years has been the widespread adoption of cloud computing. Cloud platforms provide scalable storage, computing power, and security tools on demand, allowing organizations to manage large datasets without building costly on-site infrastructure [100], [101]. This is particularly valuable in resource-constrained environments, where cloud services can lower barriers to participation in data-intensive projects. For instance, a study published in SoftwareX discusses the development of Crane Cloud, a multi-cloud service abstraction layer designed to address challenges in resource-limited settings. It highlights how cloud services can mitigate issues like frequent internet outages and limited computing resources [102]. Additionally, research in Geo-spatial Information Science explores the synergy between Big Data and cloud computing, noting that cloud platforms provide scalable resources that enable organizations with limited infrastructure to engage in complex data analysis tasks [103]. Further, many cloud platforms now offer automation features powered by AI and machine learning, such as automated data classification, metadata tagging, and access control enforcement [100]. These studies underscore the role of cloud services in democratizing access to data-intensive projects across various organizational contexts.

However, cloud-based infrastructure also raises legal and regulatory concerns. For example, storing data in cloud systems operated by U.S.-based providers may conflict with the European Union's GDPR [37]. The U.S. CLOUD Act allows American authorities to access data stored by U.S. companies, even when that data is physically located abroad [104], [105]. This creates a ten-

sion between technological convenience and data sovereignty, particularly when handling sensitive health information [106].

To manage growing complexity in digital systems, many organizations are exploring new architectural approaches to access and govern data more efficiently. One such approach is called data fabric. This refers to a design where a unified data layer connects multiple systems, making it possible to access and manage data in real time—even when that data is spread across different platforms. By using shared standards and automated processes, data fabric helps organizations improve interoperability (the ability of systems to work together) and maintain consistent data quality and structure across systems [100].

Another emerging model is the data lakehouse, which combines elements of data lakes (which store raw, unstructured data) with data warehouses (which organize data in structured formats for analysis). The lakehouse model aims to offer the flexibility of a data lake with the data governance and performance benefits of a warehouse. This helps organizations scale up their data operations while still enforcing important rules around data usage, security, and retention.

Together, these new infrastructure models support more automated and consistent data governance. They allow organizations to apply policies for how data should be handled—such as who can access it, how long it should be stored, or how it should be validated—without needing to manage each system separately. Despite these advancements, many organizations still rely on legacy infrastructure—outdated systems that are difficult to scale or integrate. Often, organizations operate a mix of older databases, newer cloud applications, and disconnected tools, each with different security models and access rules [107], [108]. This patchwork setup leads to interoperability issues and requires custom automation pipelines just to maintain consistent data flow.

Regulatory compliance further complicates the design and deployment of digital infrastructure. As laws around data privacy and sovereignty become stricter, organizations must ensure that their systems comply with regional and international requirements [109]. Designing infrastructure that can legally store, process, and move data across jurisdictions is a growing concern [104]–[106]. Failing to address these risks may result in fines, operational limits, or reputational damage [110], [111].

In response, many organizations are turning to automation as

a core strategy. By automating data governance tasks—such as validation, access control, and metadata management—they can improve data quality and reduce human error [107]. However, not all organizations have the expertise or resources to develop these tools in-house. As a result, there is increasing reliance on third-party, standardized data governance solutions that can be more easily integrated into existing infrastructure [112].

This thesis contributes to the conversation on digital infrastructure by showing how a scalable and compliant system can be built within a large-scale research project, despite resource constraints. The work demonstrates that infrastructure and data governance are not separate concerns but must be designed together. By integrating data governance mechanisms directly into the infrastructure—through validation systems, metadata enforcement, and data deletion features—the developed system supports compliance, usability, and long-term sustainability. This practical approach provides an alternative to more centralized or resource-heavy solutions, offering insights for organizations working in similarly constrained or distributed environments.

2.4 Interoperability

Interoperability refers to the ability of diverse systems, datasets, and technologies to communicate, integrate, and function cohesively [113]. While often treated as a technical problem, interoperability is increasingly understood as a socio-technical challenge. Achieving it requires not only aligned technical standards and protocols, but also coordinated data governance, shared vocabularies, and mutual understanding across organizational and disciplinary boundaries [83], [114]–[116].

The importance of interoperability spans across many domains, including data governance, dynamic data spaces [80], and digital infrastructure [41]. It is also a recurring concern in applied sectors such as healthcare [117]–[120], research data management [121], big data environments [120], and industrial systems [98], [122]. In each of these areas, organizations face similar problems: integrating diverse data sources, maintaining consistent data quality, and ensuring that information can be reused and understood across systems. Yet despite this shared challenge, the literature on interoperability remains highly fragmented—distributed

across domains that approach the topic using different assumptions, methods, and priorities.

This fragmentation is compounded by inconsistencies in terminology. Concepts such as data harmonization, data standardization, data portability, data compatibility, and data unification are frequently used alongside or in place of interoperability, often without clear distinctions between them [115]. As a result, similar challenges are framed differently across disciplines, making it difficult to compare approaches or develop generalizable solutions. Ironically, this lack of cohesion means that the literature on interoperability often lacks interoperability itself—a meta-level challenge that continues to hinder progress in this important area.

A major technical barrier to interoperability is the absence of shared data standards. As Deshmukh et al. note, many initiatives succeed in structuring metadata but fall short when it comes to aligning the data itself [77]. Without common vocabularies and interpretive frameworks, systems remain isolated, and organizations must rely on custom-built integrations that are costly and difficult to maintain [108].

Interoperability is also closely tied to data quality. In large-scale and fast-moving environments—such as streaming systems or automated data pipelines—poor-quality data can undermine the entire effort. Taleb et al. highlight how the volume, variety, and velocity of modern datasets introduce inconsistencies that are difficult to manage without robust validation and cleaning mechanisms [60]. When data quality fails, interoperability becomes ineffective, as errors and mismatches propagate across systems [28]. This highlights the interdependence between the two: without reliable data, even the best-designed interoperability frameworks fall short.

Data governance and organizational structures add another layer of complexity. Spagnoletti et al. emphasize that interoperability is not just about aligning technologies—it also requires coordination among people, departments, and legal responsibilities [123]. Regulatory concerns further shape what is possible. As Fast et al. show, data sharing is influenced heavily by legal frameworks, including cross-border restrictions and consent management, all of which must be addressed in any practical interoperability solution [124].

This thesis responds to this fragmentation by treating interoperability not as a standalone technical goal, but as a shared

underlying concern across the three key areas explored in this work: data governance, dynamic data spaces, and digital infrastructure. The developed solution automates several critical processes that support interoperability: enforcing data standards, validating structured inputs, and improving data quality through real-time feedback and metadata enforcement. While the solution is rooted in the context of large-scale research systems, its design is broadly relevant. Many fields—ranging from healthcare to public sector data management and AI systems—could benefit from simplified, data governance-aligned integration mechanisms.

However, given the wide variation in how interoperability is conceptualized across disciplines, this thesis limits its scope to the three domains most central to its empirical and design work. These are the areas in which the contributions are most clearly situated. At the same time, the approach developed here may serve as a basis for future work that further connects and integrates interoperability research across other literatures.

Improving interoperability remains essential for building flexible, resilient digital ecosystems. Without coordinated efforts—both technical and organizational—systems will continue to face siloed architectures, duplicated efforts, and rising integration costs [108]. By approaching interoperability through the lens of automation, data governance, and practical simplification, this thesis contributes both conceptual clarity and real-world mechanisms for easing integration across complex systems.

The thesis builds on the recognition that interoperability challenges are not only technical but also deeply embedded in social, organizational, and infrastructural contexts. While the focus is on developing practical, low-friction solutions, the work is grounded in principles from socio-technical systems thinking, which acknowledges the intertwined nature of people, technology, and processes in shaping data quality and system usability. In addition, the research follows the structure of Action Design Research, a methodology that explicitly supports the iterative development of solutions within real-world environments. Action Design Research emphasizes reflection, continuous refinement, and the formalization of learning—making it especially well suited for addressing persistent data challenges in dynamic, multi-stakeholder environments. Rather than starting from an abstract theoretical model, the thesis responds directly to a failing system, seeking to improve interoperability through focused design interventions that

are both minimal and broadly applicable. This approach ensures that the work remains grounded, adaptive, and relevant to both academic inquiry and practical implementation.

Chapter 3

Methods

This chapter outlines the methodological approach used to guide the design, development, and evaluation of the digital infrastructure and data governance model presented in this thesis. The research follows the Action Design Research methodology, which integrates principles from both action research and design science to address complex, real-world problems through iterative artifact development and evaluation. Across four interrelated studies, the research engaged in multiple cycles of building, intervention, and reflection. These efforts led to the creation of a simplified, automation-supported data infrastructure aimed at improving data quality and interoperability. The chapter details how different methods—ranging from scalability testing, usability evaluations, automated pipeline development, and mixed-method studies involving feasibility tasks, surveys, and interviews—were applied across the four papers. Together, these methods provided a robust foundation for refining both the technical system and the data governance model in close collaboration with stakeholders and real-world data challenges. In table 3.1 we outline the focus and methods in each paper.

Summary of Methods Across Papers		
Paper	Focus	Methods Used
Paper 1	Platform scalability and usability	Performance testing, usability testing, System Usability Scale and AttrakDiff questionnaires, semi-structured interviews
Paper 2	Infrastructure and pipeline design	Iterative database and pipeline development, user-driven refinement, metadata standardization
Paper 3	Data quality and researcher behavior	Feasibility testing, surveys, system logging, semi-structured interviews, thematic analysis
Paper 4	Automated data governance and system integration	Iterative design of data governance model, integration into infrastructure, feedback-driven adaptation

Table 3.1: Summary of methods used in each paper

3.1 Action Design Research Methodology

In this thesis, I adopted the Action Design Research methodology, originally proposed by Sein et al. [125] and further elaborated by Mullarkey et al. [126]. Action Design Research is a research approach that integrates principles from both Action Research (AR), as outlined by Susman and Evered [127], and Design Science Research (DSR), as conceptualized by Hevner et al. [128]. This methodological framework is particularly suited for addressing complex socio-technical challenges by iteratively designing, developing, and evaluating artifacts within their real-world contexts.

Action Design Research follows a cyclical process that con-

sists of four key stages: **i) Problem Formulation**: In this initial phase, researchers define the research problem and formulate key research questions. The problem is typically rooted in real-world challenges, ensuring the study remains practically relevant. **ii) Building, Intervention, and Evaluation**: This stage involves the iterative creation of an artifact, such as a system, framework, or model, followed by its implementation in a real-world setting. The intervention allows researchers to assess the artifact's effectiveness and refine its design based on empirical observations. **iii) Reflection and Learning**: The third phase focuses on analyzing the outcomes of the intervention, identifying key insights, and refining theoretical understanding. Researchers reflect on how the developed artifact influences the problem space and how contextual factors shape its use. **iv) Formalization of Learning**: In the final stage, the knowledge gained from the research is generalized into broader theoretical contributions, design principles, or best practices that can be applied beyond the specific study context.

The Action Design Research methodology is particularly effective for tackling complex, evolving problems where technological solutions must be both innovative and deeply embedded in their organizational and social environments. It enables researchers to develop practically useful artifacts while simultaneously contributing to theoretical advancements in information systems, dynamic data spaces, data governance, and digital infrastructure research [126]. By engaging in iterative cycles of problem formulation, artifact development, evaluation, and reflection, Action Design Research ensures that the solutions produced are not only functional but also continuously refined based on real-world application and stakeholder input [129].

3.2 Paper 1: Representing and Filtering Data Using Layered-Modular Architecture

This part of the research investigates the question: *How can layered-modular architecture be used to present and filter research data in a digital platform?* The work responds to one of the core infrastructure challenges identified in the early phases of the

Sleep Revolution project—namely, how to make large, heterogeneous datasets accessible to researchers through a user-friendly, scalable platform.

The goal was to evaluate both the technical performance of the system and the usability of its interface, with the underlying idea that good infrastructure must balance backend efficiency with frontend accessibility. This aligns with the broader aims of digital infrastructure and data governance discussed in the background: namely, that systems should make data easier to retrieve, filter, and interpret—especially in resource-limited environments.

3.2.1 Scalability Testing

To evaluate the platform's performance, we conducted a scalability test measuring its efficiency in handling large datasets. Since data filtering was identified as the most time-intensive operation, we focused on how the system responded to increasing data volumes within a MySQL database [130], [131]. The test began with a dataset of 100 rows and 50 columns and incrementally increased by 100 rows per iteration, reaching up to 100,000 rows—far beyond the expected operational volume. Both text-based and numerical data were included to assess the system's ability to handle different data types. Filtering operations applied string-matching and numerical range selection using an "OR" logic across all columns to simulate realistic data querying conditions.

3.2.2 Usability and User Experience Testing

To complement the technical evaluation, we conducted a usability test in sleep research with ten potential end-users. The goal was to determine how well the platform's interface supported data retrieval tasks in practice. Using a mock dataset (100 rows $\times$ 10 columns), participants were asked to filter simulated sleep-related and body measurement data and answer four structured queries—such as determining gender distributions, analyzing apnea-hypopnea index (AHI) trends, and calculating average weights for specific subgroups. Accuracy and task completion time were recorded as quantitative measures of system usability. We show in figure 3.1 the filters the participants engaged with and we show the overview in figure 3.2.

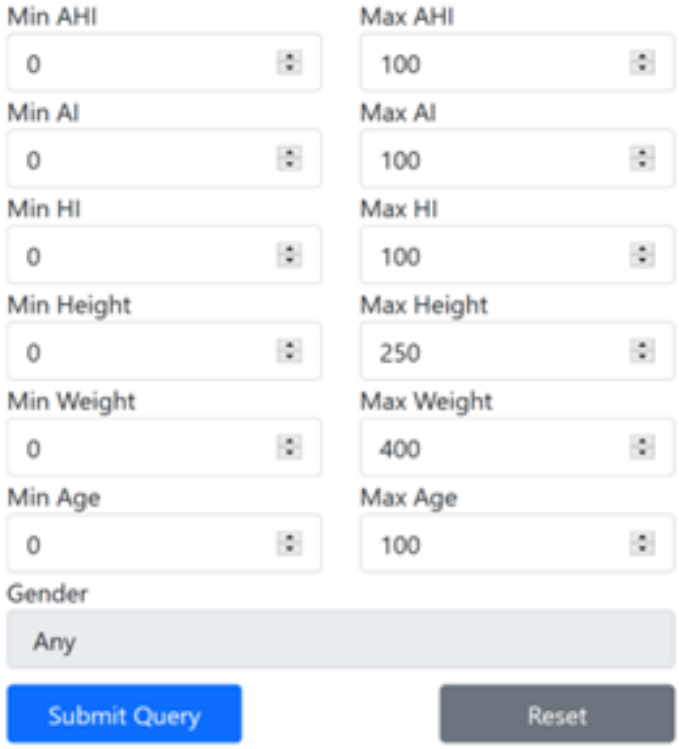
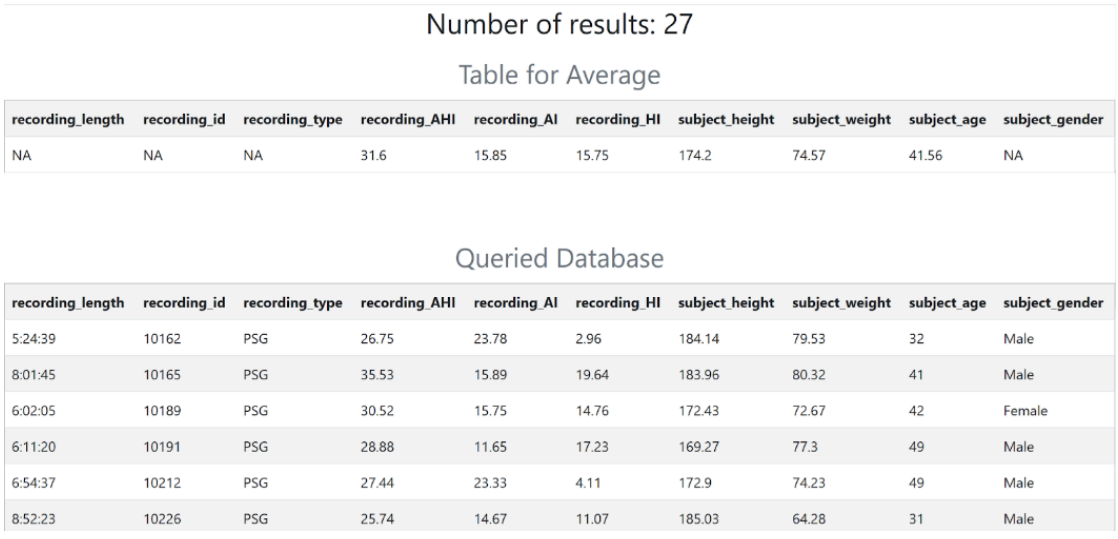


Figure 3.1: The filters the participants used in their tasks.

After completing the tasks, participants filled out two standardized usability questionnaires: the System Usability Scale (SUS) [132] and AttrakDiff [133], both administered via Google Forms. These were followed by semi-structured interviews to gather qualitative feedback on usability challenges, feature clarity, and user expectations.

Following the questionnaires, a semi-structured interview was conducted with each participant, lasting approximately 20 minutes. These interviews explored user perceptions of the system's strengths and weaknesses, with particular focus on features that were perceived as unhelpful or missing, as well as what improvements would be necessary for the platform to become valuable in their daily work. Participants were encouraged to elaborate on their expectations, frustrations, and suggestions for improvement.



Number of results: 27

Table for Average

recording_length	recording_id	recording_type	recording_AHI	recording_AI	recording_HI	subject_height	subject_weight	subject_age	subject_gender
NA	NA	NA	31.6	15.85	15.75	174.2	74.57	41.56	NA

Queried Database

recording_length	recording_id	recording_type	recording_AHI	recording_AI	recording_HI	subject_height	subject_weight	subject_age	subject_gender
5:24:39	10162	PSG	26.75	23.78	2.96	184.14	79.53	32	Male
8:01:45	10165	PSG	35.53	15.89	19.64	183.96	80.32	41	Male
6:02:05	10189	PSG	30.52	15.75	14.76	172.43	72.67	42	Female
6:11:20	10191	PSG	28.88	11.65	17.23	169.27	77.3	49	Male
6:54:37	10212	PSG	27.44	23.33	4.11	172.9	74.23	49	Male
8:52:23	10226	PSG	25.74	14.67	11.07	185.03	64.28	31	Male

Figure 3.2: The overview the participants viewed getting updated using the filters.

This combination of performance testing, structured usability scoring, and open-ended qualitative interviews provided a well-rounded understanding of both the system's scalability and its real-world usability. The insights from these evaluations directly informed the platform's refinement, setting the stage for the more extensive infrastructure and data governance developments described in subsequent papers.

3.3 Paper 2: Transforming Heterogeneous Sleep Data into a Homogeneous Platform Database

This part of the research explores two related questions that arose in response to the growing data complexity in the Sleep Revolution project: *(i) How can we represent sleep data from heterogeneous sources in a homogeneous digital platform database?* and *(ii) How can a data source pipeline transform various data sources into a homogeneous data format?* These questions align with the broader challenges of data governance and digital infrastructure in dynamic data spaces, where standardization, integration, and reuse are essential yet difficult to achieve.

This work was conducted in response to the growing complexity of managing personal, behavioral, and clinical data from multiple systems within the Sleep Revolution project. In line with the challenges of dynamic data spaces, this study focused on simplifying the underlying digital infrastructure to improve interoperability and data quality, while minimizing manual labor for researchers. The aim was to develop a database and pipeline system that could consistently transform fragmented datasets into a unified, structured format.

3.3.1 Evolution of the Database and Pipelines

The development of the database and pipeline system in this PhD research emerged out of necessity. As data collection expanded across multiple sources, we iteratively sought to simplify the database structure by identifying commonalities between the data. Through experimentation with various relational database designs, our primary goal was to reduce the workload and enhance

usability. Over time, we managed to simplify the structure significantly—ultimately representing all data using just a few tables and parameters.

At its core, the database was reduced to four essential tables, designed to function similarly to an Excel sheet with subsheets, headers, rows, and cells. To avoid unnecessary data duplication, study participation was maintained as a separate table. The structure was iteratively refined by adding metadata parameters such as study date, study description, form description, and entry descriptions. These metadata fields emerged naturally—whenever researchers identified missing contextual information, new metadata fields were introduced to ensure data completeness and usability. While some metadata fields, such as "description," remained free-format due to their ambiguous nature, the overall approach ensured that all relevant contextual details were available for future research. We show how a spreadsheet would be split into the database design in figure 3.3.

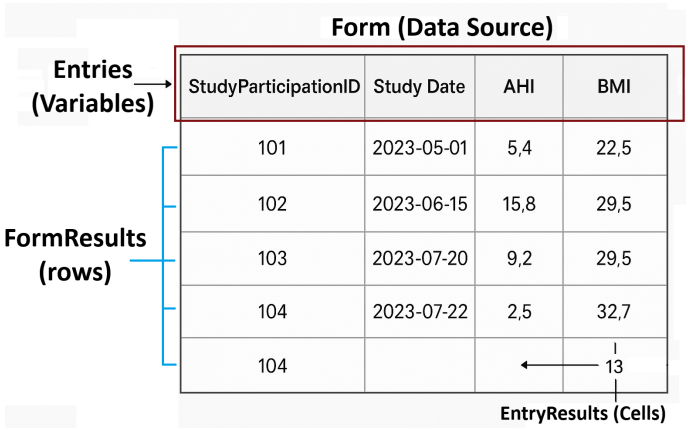


Figure 3.3: The database design explained through the lens of a spreadsheet.

3.3.2 Automated Data Pipelines

To maintain a homogeneous data structure and streamline data handling, we developed a set of automated pipelines: **i) Insertion Pipeline**: Responsible for inserting comma-separated values (CSV) data or API-derived data into the database. **ii) Data**

Conversion Pipelines: Custom-built for each data source, these pipelines transformed diverse datasets into a standardized format compatible with the insertion pipeline. **iii) Export Pipeline**: Designed to extract and organize structured data into a folder-based system, ensuring researchers could access the processed data efficiently.

The insertion pipeline evolved dynamically based on real-world errors that bypassed validation. It was continuously updated to strengthen error detection, enforce data integrity, and ensure completeness. Any validation failures within this pipeline influenced the conversion pipelines, as they were only designed to reformat data into an acceptable structure for insertion. Consequently, this process established a feedback loop—where errors that surfaced in one data source triggered improvements across all pipelines, reinforcing the system's robustness.

3.3.3 Iterative Refinement and User-Driven Development

The iterative nature of this system meant that researchers working with exported data in the folder structure played a key role in shaping improvements. Feedback from data users provided valuable insights, allowing us to fine-tune the structure over time. By continuously incorporating user-driven refinements and enhancing error-handling mechanisms, the system achieved an efficient and adaptable architecture. This ensured that large-scale, heterogeneous healthcare data could be made usable, structured, and accessible for multidisciplinary research purposes—supporting the overarching goals of dynamic data spaces and sustainable data governance.

3.4 Paper 3: Understanding Data Work in Curating Quality Healthcare Data

This part of the research addresses the question: *What kind of data work is involved in curating quality healthcare data?* The study was motivated by the observation that much of the data collected within the Sleep Revolution project required substantial

manual effort to clean, validate, and understand. These challenges are deeply connected to the broader issues of data governance and dynamic data spaces, where data often flows across organizational boundaries and formats without standardized oversight. This paper focuses on identifying the behaviors, assumptions, and constraints that shape how data users—who, in this context, are the researchers themselves—engage with data and contribute to its quality.

3.4.1 Database and Digital Infrastructure Development

Before conducting the study, a relational database framework was developed to support the project's complex data management needs across disciplines. This infrastructure served as the foundation for the work presented in this paper. It consolidated over 2,600 variables into seven core tables, each structured with 2 to 9 parameters. All data entries were linked to individual participants and their respective studies through a table named *StudyParticipation*, which served as the universal primary key.

The database formed the foundation of a digital infrastructure that was developed iteratively over a three-year period in collaboration with 39 partner institutions across Europe and Australia. It supports heterogeneous data sources, such as objective sleep measurements, subjective survey responses, and data collected through a mobile application.

To ensure scalability and interoperability, the database followed a modular architecture, structured as follows: **i) Form**: defines the foundational table structure. **ii) Entry**: represents individual parameters. **iii) FormResult**: corresponds to rows. **iv) EntryResult**: maps to individual data points (cells).

This structure, allows new data sources to be added without increasing the number of tables. Metadata fields were iteratively added in response to user needs, improving data completeness and usability.

A validation mechanism was embedded into the digital infrastructure to enforce data quality at the point of entry. This mechanism evolved into the Data Integrity Assurance System (DIAS), which was refined through co-design with researchers. DIAS enforced standardized formats, prompted users with feedback, and embedded early-stage data governance into data workflows. We

show the upload functionality component in figure 3.4 along with the feedback the users got after uploading.

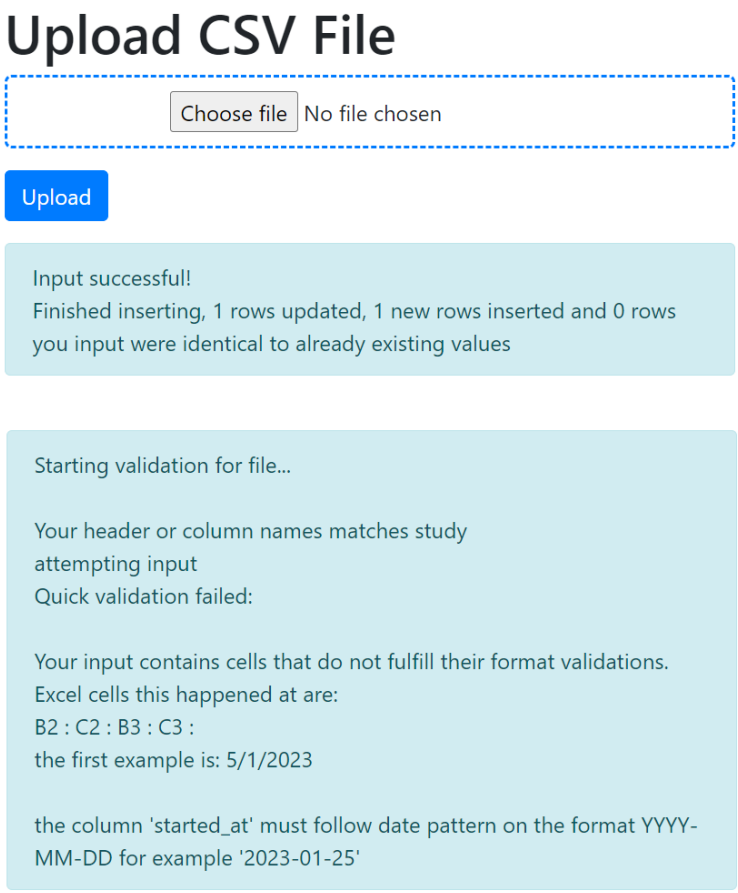


Figure 3.4: The upload functionality the participants used to complete their tasks along with the two possible feedback below. After upload finishes, one of the feedback text gets shown, either successful upload (upper), or unsuccessful upload (lower) with actionable errors.

A mixed-methods study was conducted to evaluate how researchers interact with the validation system and to understand common sources of data quality issues. The study combined: **i)**

Feasibility Testing: Evaluating the usability and effectiveness of the validation system. **ii) Surveys**: Capturing user behaviors, data work experiences, and time spent on data handling tasks. **iii) Semi-structured Interviews**: Exploring validation practices, challenges in data interpretation, and researcher engagement with data governance mechanisms.

The study covered both prospective and retrospective data work, examining how participants responded to system feedback and how they prioritized different aspects of data quality with minimal training.

3.4.2 Participant Engagement and Data Collection

Thirteen researchers participated in the study, drawn from sleep research, sports science, neuroscience, information systems, and machine learning—all working actively with healthcare-related data. Participants were recruited through professional networks and project partners.

Each session lasted between 65 and 120 minutes and followed a consistent structure: **i) Consent and Introduction**: Participants were informed of the study goals and data handling procedures. **ii) Feasibility Task**: Participants processed example data (12 rows $\times$ 8 columns) into structured CSV files and uploaded them to the validation system. **iii) System Interaction and Observations**: Participants received real-time feedback from the system; in remote sessions, screen-sharing allowed the researcher to observe user behavior and use of documentation. **iv) Survey and Interview**: Participants completed a background and behavior questionnaire, followed by a 17-question semi-structured interview exploring their experience with the system and general data work practices.

The collected data included system logs (completion rates, task times, and error types), survey responses, and interview transcripts. Interviews were recorded, transcribed, and translated when necessary. We show the participant engagement in figure 3.5.

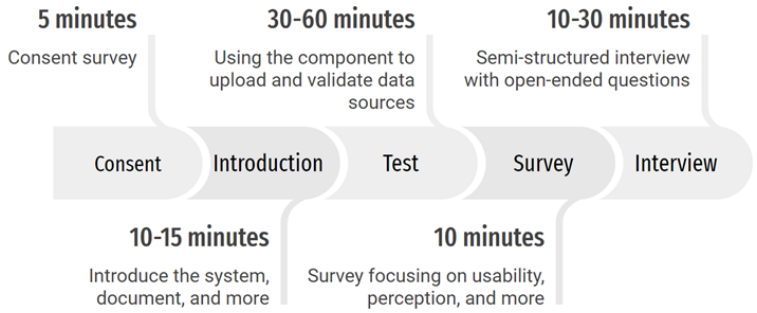


Figure 3.5: The setup of the participant engagement along with time taken and short description.

3.4.3 Evaluation and Analysis

All transcripts were thematically coded using an iterative coding process, aiming to identify patterns in data work practices, common obstacles to data quality, and how users adapted to the validation system. Key areas of analysis included: **i)** Usability and effectiveness of the DIAS validation system. **ii)** Patterns in researcher behavior and assumptions during data entry. **iii)** Socio-technical barriers to consistent data governance and quality control.

Interview responses were lightly edited for clarity while preserving intent. The findings contributed directly to refining both technical validation rules and broader data governance practices, and they highlight the human factors involved in creating sustainable, high-quality data infrastructures within dynamic data spaces.

3.5 Paper 4: Automating Validation for Simplified and Sustainable Data Governance

This part of the research addresses the question: *How can automated validation systems support simplified and sustainable data governance?* Building on the challenges identified in Paper 3, this paper shifts focus from diagnosing data quality issues to designing and evaluating a practical solution. The goal was to develop a

data governance model that reduces manual workload, improves data reliability, and aligns with real-world constraints—especially in resource-limited environments.

The empirical foundation and study design for this work are the same as those described in Paper 3. Participant recruitment, feasibility testing, survey distribution, and semi-structured interviews followed the same structure and timeline. However, whereas Paper 3 focused on understanding how data users interact with infrastructure and where quality issues emerge, Paper 4 centers on how data governance mechanisms can be embedded directly into that infrastructure.

3.5.1 Governance Model Design and Integration

The design of the data governance model followed an iterative, socio-technical approach grounded in the Action Design Research methodology. Insights from user behavior, validation errors, and observed data work practices were continuously incorporated into evolving data governance policies and technical rules. This led to the creation of a lightweight, automation-driven data governance framework designed to support user compliance through minimal intervention and real-time feedback.

Key components of the data governance model included: **i) Validation at the point of data entry**: Ensuring early error detection and compliance with format and completeness rules. **ii) Metadata enforcement**: Preserving structure, traceability, and context for future reuse. **iii) A structured folder-based output system**: Making validated data easily accessible and promoting standardized use. **iv) Feedback-as-governance mechanisms**: System responses act as teaching tools, helping users learn correct formatting and documentation practices over time. **v) Low user burden**: Achieved by embedding data governance rules naturally into workflows without requiring extensive training or manual oversight.

These components were refined over multiple development cycles and co-design sessions, drawing on input from real users and reflecting the iterative logic of Action Design Research. Rather than separating data governance from infrastructure, this model demonstrates how data validation, quality control, and policy enforcement can be encoded directly into the system itself. In do-

ing so, Paper 4 contributes a practical example of infrastructure-embedded data governance that supports sustainability, compliance, and user-aligned simplicity within dynamic data spaces.

3.6 Researcher's role

As per Reykjavik University guidelines, a declaration of authorship is provided alongside the thesis, outlining my contributions to each stage of the research and publication process for the papers in this thesis. Specific contributions are outlined in Table 3.2.

Paper name	**Idea**	**Related work & literature**	**Data gathering**	**Research design**	**Artifact design**	**Analysis & synthesis**	**Draft**	**Administration**
Layer Upon Layer: Developing Layered Modular Architectures for Data-Driven Health Platform	EE	EE	ME	EE	ME	ME	ME	EE
The Sleep Revolution Platform: A Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data	ME	EE	ME	ME	ME	ME	ME	ME
Data Work in Healthcare: Mediating Data Quality and Data Governance in a Data-Intensive World	ME	EE	ME	ME	ME	ME	ME	ME
Let the System Handle It: Simplifying Data Governance Using Automation	ME	EE	ME	ME	ME	ME	ME	ME

Table 3.2: Declaration of authorship contribution.
ME = Main effort, includes the main effort in the indicated column. EE = Equal efforts, includes that there was a shared equal effort between at least one other author of the paper. CE = Contributing effort, entails important effort, but there is someone else in the author list that delivered the main effort. LE = Learning effort, includes an effort of a learning character, for instance, by assisting with the data collection or the analysis.

Chapter 4

Results

The following chapter will briefly introduce the four appended papers through a summary of the main findings and contributions.

4.1 Paper 1: Representing and Filtering Data Using Layered-Modular Architecture

Title: Layer Upon Layer: Developing Layered Modular Architecture for Data-Driven Health Platforms

Status: Published.

Outlet: In Digital (Eco)Systems and Societal Challenges: New Scenarios for Organizing.

Type of outlet: Book chapter.

Full reference: Sveinbjarnarson B. F., Arnardóttir, E. S. and Islind, A. S. "Layer Upon Layer: Developing Layered Modular Architecture for Data-Driven Health Platforms" In Digital (Eco)Systems and Societal Challenges: New Scenarios for Organizing (pp. 91-108). Cham: Springer Nature Switzerland.

This paper addressed the question: *How can layered-modular architecture be used to present and filter research data in a digital*

platform? The work responded to early-stage infrastructure challenges in the Sleep Revolution project, where researchers needed a practical interface to access and filter data efficiently. These results contribute to the broader conversation on digital infrastructure within dynamic data spaces, where flexible, user-friendly systems are necessary for enabling data access across disciplines without relying on rigid data pipelines or central control.

The system was designed using a relational database framework and a layered-modular architecture, facilitating efficient data extraction, validation, and filtering for over 10,000 sleep studies. The architecture provided a scalable approach to managing heterogeneous data sources, ensuring that extracted data maintained consistency and adaptability across multiple research applications. A key feature was the validation module, which enforced data quality standards to prevent formatting inconsistencies and errors.

4.1.1 Scalability and Performance

To evaluate the system's scalability, filtering performance was tested on progressively larger datasets, ranging from 100 rows to 100,000 rows across 50 columns. The system demonstrated linear performance scaling, with filtering times remaining below 1.5 seconds for numerical data and 0.75 seconds for text-based queries at the largest dataset size. This confirmed that the architecture could efficiently handle large-scale data retrieval without significant performance degradation. Figures 4.1 and 4.2 show the resulting scalability for databases consisting of integers and strings, respectively.

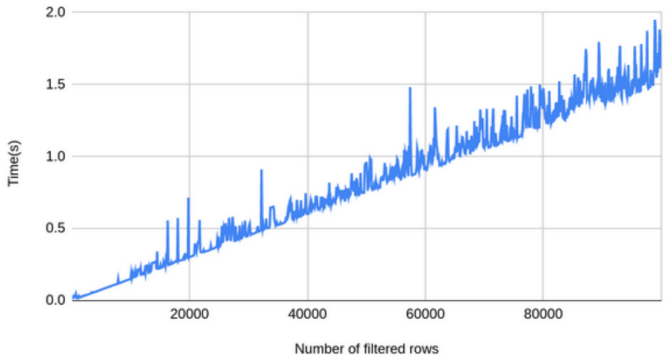


Figure 4.1: Scalability of the system when processing integer-based datasets.

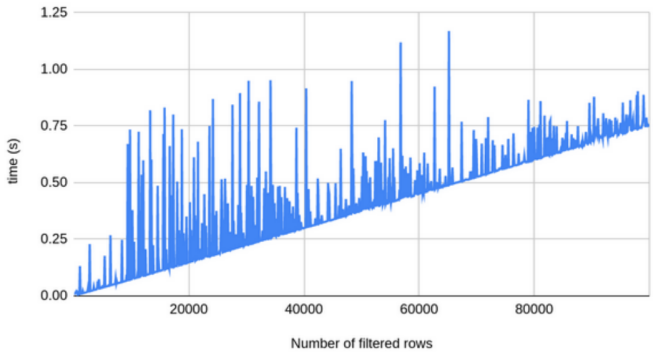


Figure 4.2: Scalability of the system when processing string-based datasets.

4.1.2 Usability and User Experience

A usability and user experience study was conducted with 10 domain experts, including researchers and data specialists, to assess the effectiveness of the user interface for data filtering and retrieval. Participants completed four structured tasks, including navigating the system, applying filters, and interpreting results. Task completion times averaged under 30 seconds, except for one more complex query, which took approximately 45 seconds on average.

Following task completion, participants completed the SUS and AttrakDiff questionnaires. The SUS score of 90/100 indicated that the system was perceived as highly usable, well above the usability benchmark of 68/100. The AttrakDiff results suggested that participants found the system intuitive and visually appealing, though some noted difficulties interpreting certain questionnaire items.

4.1.3 User Feedback and System Limitations

Semi-structured interviews provided deeper insights into user experiences and system limitations. While participants appreciated the interactive data filtering and structured database design, they emphasized the need for improved navigation, additional data integration, and a broader range of filtering parameters. Some users found the current implementation too limited for their specific research needs, advocating for enhancements in system overview and data exploration capabilities.

Overall, the study confirmed that the system is scalable, efficient, and user-friendly, with strong usability scores and positive feedback from domain experts. However, the results also highlighted key areas for further refinement, particularly in user interface enhancements and expanded data integration to accommodate more complex research demands. These findings underscore the value of layered-modular architecture as a flexible foundation for future digital infrastructure within dynamic and multidisciplinary research settings.

4.2 Paper 2: Transforming Heterogeneous Sleep Data into a Homogeneous Platform Database

Title: The Sleep Revolution Platform: A Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data

Status: Published.

Outlet: Current Sleep Medicine Reports.

Type of outlet: Journal article.

Full reference: Sveinbjarnarson, B. F., Schmitz, L., Arnardottir, E. S., & Islind, A. S. (2023). The Sleep Revolution Platform: A Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data. *Current Sleep Medicine Reports, 9*(2), 91-100.

This paper addressed two related research questions: *(i) How can we represent sleep data from heterogeneous sources in a homogeneous digital platform database?* and *(ii) How can a data source pipeline transform various data sources into a homogeneous data format?* The work was motivated by early-stage integration problems in the project's digital infrastructure. Researchers were working with inconsistent spreadsheets, API outputs, and database dumps from multiple partners—each using different formats, variable names, and levels of documentation. These results contribute to the broader effort of simplifying infrastructure design and achieving interoperability in dynamic data spaces, where datasets must remain reusable and meaningful across disciplinary boundaries.

4.2.1 Homogeneous Database and Digital Platform Design

The development of a homogeneous database and digital platform architecture aimed to create a flexible, scalable, and standardized approach to handling diverse research data sources. The database was structured around a five-fold core model, consisting of entries, forms, entry-results, form-results, and ownership metadata. This design eliminated the need to create additional tables and columns for each new data source, significantly reducing complexity and ensuring consistency across all stored data. We show the resulting database design in figure 4.3.

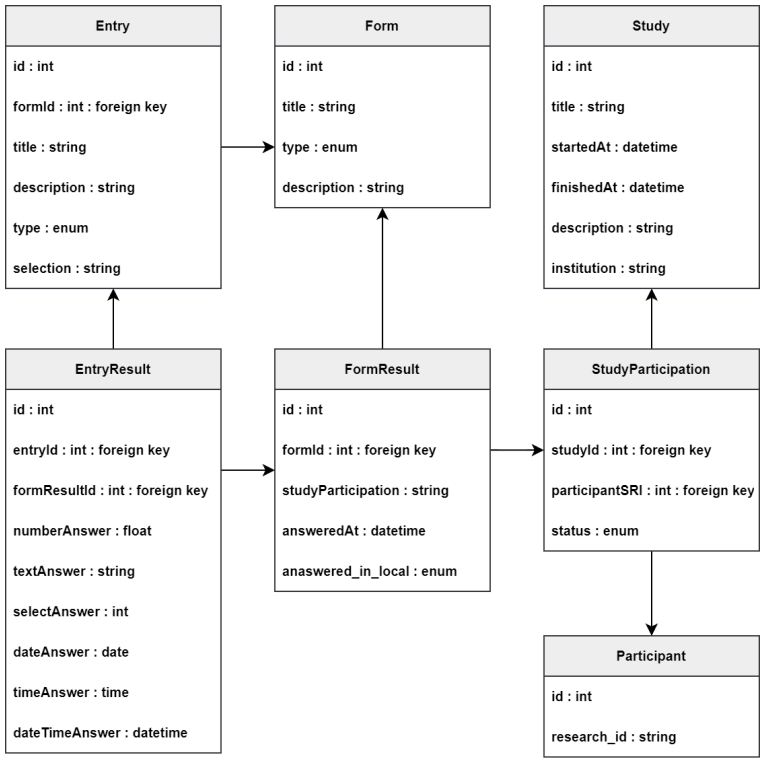


Figure 4.3: The resulting database, capable of representing all the structured data without requiring modifications.

A key strength of this model is its simplicity—new data sources can be integrated seamlessly without altering the database schema. The architecture also supports holistic digital platform development, allowing for flexible front-end applications and streamlined data querying. Furthermore, the design facilitates retrospective data entry, a crucial aspect for research environments where additional data points may be added after initial collection. Another novel aspect is the system's ability to export data into researcher-preferred formats (e.g., Python, R, or SPSS), ensuring interoperability between the platform and common data analysis tools.

The ownership metadata layer was implemented to maintain participant-level data integrity, ensuring researchers can analyze

both within-subject and between-subject data while preserving traceability across different data sources.

4.2.2 Data Source Pipeline

To integrate heterogeneous data into the homogeneous database, we developed a data source pipeline that automatically processes and standardizes incoming datasets. The pipeline follows a structured four-step process: **i) Parameter Formulation**: Data sources are examined to determine relevant parameters for end-users, often requiring collaboration between developers, data collectors, and researchers. **ii) Forms and Entries Creation**: Identified parameters are structured into standardized CSV files containing variable names, data types, and descriptions. This process ensures uniformity across data sources and minimizes manual errors. **iii) Process-Pipeline Transformation**: Unprocessed data is cleaned, validated, and formatted into a standardized structure through a series of six crucial steps. This includes: a) Combining fragmented datasets, b) Assigning ownership metadata, c) Handling duplicate and incomplete data, and d) Ensuring compatibility with downstream processing. **iv) Data Insertion**: Processed data is inserted into the homogeneous database using the insert-pipeline, allowing for seamless integration and immediate usability for research applications. Some data sources require API-based insertions, while others use batch-file processing methods.

The iterative refinement of the pipeline ensured that each new data source followed a shared, standardized workflow, making it possible to process and store heterogeneous datasets in a unified format. This not only improved data quality and consistency but also streamlined data retrieval and analysis for researchers.

4.2.3 Impact and Adaptability

By implementing this homogeneous database and data source pipeline, the project successfully: **i)** minimized complexity in handling multiple research datasets; **ii)** increased adaptability by ensuring any new data source could fit within the standardized framework; **iii)** improved efficiency in data processing, allowing for real-time validation and integration; and **iv)** enabled cross-disciplinary research, making data easily accessible and analyz-

able across different fields.

The framework provides a foundation for scalable data governance, ensuring that research data remains structured, interoperable, and reusable while reducing the manual workload associated with traditional database management. We show in figure 4.4 the overall pipeline for transforming data sources into the homogeneous database design.

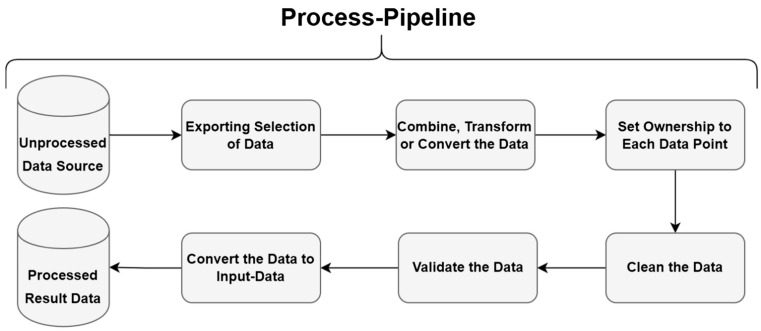


Figure 4.4: The different steps in the pipeline required to process a data source into the database.

4.3 Paper 3: Understanding Data Work in Curating Quality Healthcare Data

Title: Data Work in Healthcare: Mediating Data Quality and Data Governance in a Data-Intensive World

Status: Accepted

Outlet: Scandinavian Journal of Information Systems.

Type of outlet: Journal article.

Full reference: Sveinbjarnarson, B. F., Schmitz, L., Arnardottir, E. S., & Islind, A. S. (2025). Data Work in Healthcare: Mediating Data Quality and Data Governance in a Data-Intensive World. Submitted to *Scandinavian Journal of Information Sys-*

tems.
This paper addressed the question: *What kind of data work is involved in curating quality healthcare data?* The study was motivated by persistent issues in data consistency and quality during earlier infrastructure development. While technical systems were designed to standardize structure and format, it became clear that human behavior and engagement with data posed equally significant challenges. These findings contribute to the broader discourse on data governance in dynamic data spaces, where effective coordination depends not only on infrastructure but also on the people who enter, clean, and interpret data.

The findings from this study center around three key themes: (i) data feasibility as an integral part of data work, (ii) trust issues and their impact on data work, and (iii) cultivating excellence in data work. These insights stem from feasibility testing, survey responses, and semi-structured interviews with healthcare researchers interacting with the digital infrastructure.

4.3.1 Data Feasibility as Part of Data Work

Through feasibility testing, participants were asked to split a provided dataset into seven structured files, add metadata columns, and upload them into the digital infrastructure. Despite being given example input sheets and detailed documentation, participants introduced errors that the system subsequently flagged, including formatting issues, missing or additional columns, duplicate entries, and incorrect file types. The data was originally error-free, highlighting that issues arose during participant interactions with the system.

A key finding was that all participants required multiple attempts before successfully uploading their files. While mistakes such as spelling errors and incorrect date formatting were common, participants also misinterpreted metadata and mislabeled columns—errors that were not always caught by validation checks but became evident through observations. These findings emphasize the necessity of early intervention and validation mechanisms to prevent faulty data from entering research databases.

Participant behavior revealed a reactive rather than proactive approach to data work. Initially, many displayed low confidence, frequently seeking clarification rather than engaging with documentation or system feedback. Over time, they became faster

and more independent, but frustration was common when their uploads were rejected. Notably, many participants treated the process as a challenge to "get the system to accept their input" rather than an effort to ensure data was properly structured for future research. This highlights the importance of not just enforcing validation rules but also educating users on the significance of meaningful data preparation.

The results underscore that technical data governance mechanisms alone are insufficient—researchers require better training, improved guidance, and strategies to support meaningful user engagement if long-term data reusability is to be achieved.

4.3.2 Trust Issues and Their Impact on Data Work

Despite encountering validation errors, most participants reported in surveys that they found the system intuitive and valuable for their research tasks. The majority rated all of the tasks as 'very easy' or 'somewhat easy,' with only a few exceptions. There was strong agreement on the value of the digital infrastructure and embedded data governance mechanisms for improving research data quality.

However, a critical challenge emerged: significant disagreement on what constitutes "good" data quality. Some participants prioritized consistent formats, unique identifiers, and duplicate removal, while others deemed these factors less important. This reflects diverging priorities in data governance and the challenge of designing one-size-fits-all solutions that accommodate different research needs.

Participants also expressed skepticism about data they had not collected themselves. Many struggled to interpret ambiguous variable names, missing metadata, and inconsistent data structures, leading to distrust. The absence of clear metadata documentation exacerbated these issues, making data reuse challenging and reinforcing concerns about data reliability.

A recurring theme was that merging datasets was difficult due to missing identifiers and formatting inconsistencies, illustrating the complexities of ensuring interoperability in research data. Participants spent substantial time cleaning, merging, and converting data—tasks that could be streamlined with stronger data governance frameworks and standardized formats. However, distrust

in externally sourced data remained a persistent challenge, highlighting the need for transparency and clarity in data work.

4.3.3 Cultivating Excellence in Data Work

Interviews revealed that most participants had no formal training in data management or data quality. Instead, they relied on self-taught methods or informal guidance from colleagues. Even those with formal training typically focused on technical aspects like database design rather than broader data governance principles.

Participants widely acknowledged that data work remains invisible and underappreciated, with no standardized education or guidelines in place. This lack of structured training contributes to inconsistent approaches to data governance, as researchers develop their own fragmented methods. While some attempted to implement personal data standards, these were often incomplete or project-specific, reinforcing long-term inefficiencies in data management.

When asked about adopting the digital infrastructure for validation, most participants supported its use, particularly for ensuring data consistency and improving data governance. However, some raised concerns about rigid validation requirements hindering research flexibility. One participant noted that strict formatting rules could become an unnecessary hurdle depending on the research context, indicating a need for balancing enforcement with adaptability.

Another challenge was documentation engagement. Despite the availability of detailed guidelines, most participants rarely consulted them, preferring to rely on trial-and-error or direct questions. While some acknowledged the importance of documentation, they admitted only using it when absolutely necessary. This suggests that passive documentation alone is insufficient, and more interactive or embedded guidance may be needed to foster best practices in data work.

Participants also had varying preferences on data storage and organization. Some preferred databases with query functionalities, while others favored folder structures with browsing tools. This diversity highlights the need for adaptable digital infrastructures that accommodate different researcher workflows.

In the interviews, many participants reflected on broader challenges they had encountered in previous projects—challenges that

align closely with data governance concerns. They pointed to issues such as human error, limited understanding, poor documentation, and a reliance on short-term fixes rather than sustainable practices. Several emphasized that gaps in data management education exacerbated these problems, with one participant suggesting that data literacy should be considered a core skill. Others noted that the pressure to meet short-term research goals often leads to the neglect of long-term data structure and usability.

A key takeaway from these findings is that data governance models must go beyond technical validation mechanisms—they must also support training, encourage best practices, and provide adaptable frameworks that align with researchers' diverse needs.

4.3.4 Summary of Key Findings

The study revealed several key insights: **i)** *Data feasibility testing* revealed persistent errors in data preparation, emphasizing the need for robust validation mechanisms and improved user guidance; **ii)** *Trust issues* in data work emerged as a major barrier, particularly when researchers handled data they did not collect themselves. Ambiguous metadata and formatting inconsistencies reinforced skepticism and reduced reusability; **iii)** *Lack of formal training* in data governance led to inconsistent and fragmented approaches to handling research data; **iv)** *Validation-driven digital infrastructures* were supported by participants for improving data quality, though they emphasized that enforcement should be balanced with flexibility; **v)** *Low engagement with documentation* suggested a need for more integrated and interactive guidance in data governance models; **vi)** *Reactive data work* was common, with researchers focused more on satisfying validation requirements than ensuring meaningful data structuring for future use; and **vii)** *Educational gaps, human error, and inconsistent practices* were found to hinder data governance, highlighting the necessity of structured yet adaptable solutions.

These findings contribute to a broader understanding of the human aspects of data work and data governance. Within dynamic data spaces and other cross-disciplinary digital infrastructures, addressing these challenges is essential for ensuring data reliability, long-term usability, and the sustainability of shared data systems.

4.4 Paper 4: Automating Validation for Simplified and Sustainable Data Governance

Title: Let the System Handle It: Simplifying Data Governance Using Automation

Status: Under construction.

Outlet: TBD.

Type of outlet: Journal article (planned).

Full reference: Sveinbjarnarson, B. F., Arnardottir, E. S., & Islind, A. S. (Forthcoming). Let the System Handle It: Simplifying Data Governance Using Automation.

Building on the same empirical foundation as Paper 3, Paper 4 shifts the focus from identifying data quality challenges to addressing them through an automated data governance model. While the study design, including participant recruitment, feasibility testing, surveys, and interviews, remains consistent with the previous paper, this phase concentrated on the development and evaluation of a data governance solution embedded directly into the digital infrastructure.

The data governance model was developed iteratively using a socio-technical approach informed by Action Design Research. User feedback, observed behavior, and recurring validation errors guided continuous refinement of the model's design. The result was a lightweight, automation-driven framework that aimed to simplify data governance without reducing its effectiveness. https://www.overleaf.com/project/67e62b755bec02425a4a3f74
Key components of the data governance model included:

The key features of our automated governance approach included: i) **Validation at the point of data entry**, catching errors early and ensuring compliance with formatting and structure requirements; ii) **Metadata enforcement**, improving context, traceability, and interpretability of datasets; iii) **A structured folder-based output system**, which standardized how

data was delivered to researchers, improving reusability and consistency; iv) **Feedback-as-governance mechanisms**, where the system offered real-time responses to user input, helping users understand and adopt best practices over time; and v) **Minimized user burden**, by embedding data governance tasks naturally into the workflows rather than requiring separate processes or training.

Key Results

The implementation of this automated data governance model demonstrated that data validation and policy enforcement can be simplified without sacrificing control or quality. Embedding these mechanisms directly into data workflows improved compliance while reducing manual oversight. Participants reported that the system's real-time feedback helped them gradually adopt better data practices and increased their understanding of data governance requirements.

Survey and interview responses highlighted positive reactions to the structured output and validation features, especially the reduced workload and greater clarity around expectations. The standardized format also improved trust in data quality and reusability across different domains. These findings support the argument that sustainable data governance can be achieved through automation, provided the system is designed with users' needs and workflows in mind.

A visual summary of the final data governance model is shown in figure 4.5.

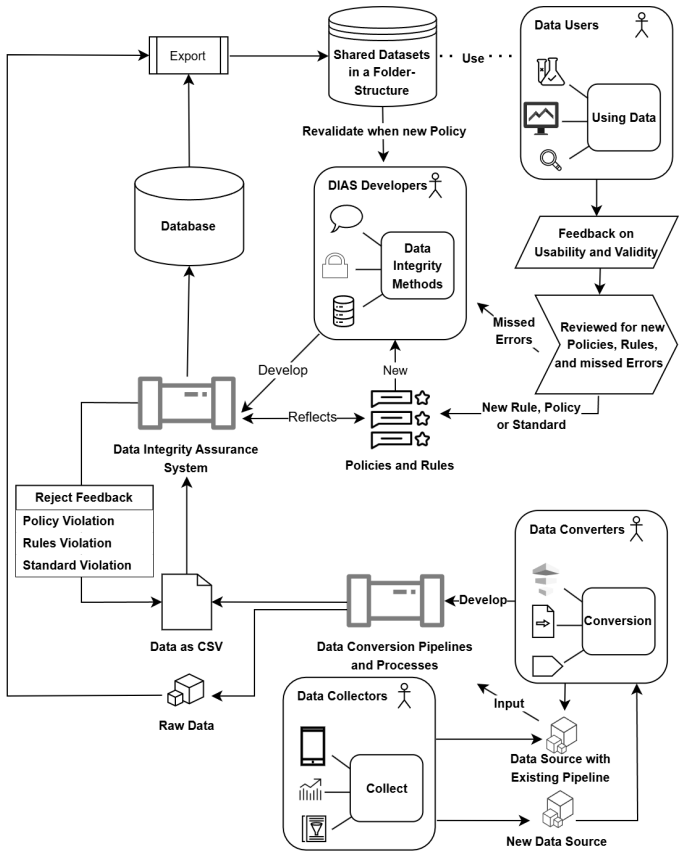


Figure 4.5: The automated data governance model.

Chapter 5

Discussion

This research set out to investigate a central challenge in information systems:

> **How can complexity in data governance be reduced without sacrificing functionality, scalability, or compliance?**

To address this, the thesis focused on three interlinked domains: data governance, dynamic data spaces, and digital infrastructure. Through four studies situated within the Sleep Revolution project, the thesis developed and evaluated a simplified, automation-driven data governance model embedded into a digital infrastructure. The aim was not only to reduce the burden of manual data governance, but also to support sustainable, real-world use in resource-constrained, multidisciplinary environments.

The research questions were addressed as follows: i) **Paper 1** asked: *How can layered-modular architecture be used to present and filter research data in a digital platform?* This study laid the foundation by evaluating platform scalability and usability. It highlighted the importance of a unified, flexible architecture to support data interaction at scale. ii) **Paper 2** asked: *How can we represent sleep data from heterogeneous sources in a homogeneous digital platform database?* and *How can a data source pipeline transform various data sources into a homogeneous data format?* This study contributed a simplified infrastructure and standardized pipeline for transforming diverse sources into a uniform, interoperable structure—an essential component for dynamic data

space design. iii) **Paper 3** asked: *What kind of data work is involved in curating quality healthcare data?* It revealed the socio-technical origins of poor data quality, surfacing key behavioral and structural factors that undermine data governance. iv) **Paper 4** asked: *How can automated validation systems support simplified and sustainable data governance?* Building on the previous study, it designed and tested a lightweight, automation-driven data governance model that embeds policy enforcement into system workflows.

Together, these studies answer the thesis' overarching research question by demonstrating that complexity in data governance can be reduced through a combination of simplification, automation, and carefully designed digital infrastructure. The work offers concrete contributions to the design of dynamic data spaces, particularly in resource-limited and multidisciplinary environments.

The following sections discuss these findings in relation to the theoretical concepts introduced in the earlier chapters. Each subsection connects the empirical results to ongoing debates in the literature on data governance, dynamic data spaces, and digital infrastructure. Throughout, special attention is given to the role of interoperability and socio-technical design, which emerged as cross-cutting concerns across all four papers.

Our approach showed that simplicity and automation are key factors in minimizing data governance burden without compromising functionality. By enforcing a single file format (CSV) at input, we significantly reduced the variability in data structures, thereby streamlining validation, ensuring interoperability, and minimizing human error. Additionally, rather than requiring custom database configurations for every dataset, our homogeneous database model accommodated 70 different data sources under a unified schema, preventing fragmentation and enhancing scalability.

Further, we embedded data governance policies at the earliest stages of data processing, ensuring that validation and metadata enforcement occurred at input rather than being an afterthought. This automated enforcement of standards, rather than relying on user diligence, directly addressed one of the primary causes of data quality issues—human error, neglect of documentation, and inconsistent adherence to data governance principles. Through automated metadata management, access control, and validation mechanisms, we demonstrated that data governance could im-

prove over time without requiring additional effort from users.

Our findings also highlighted the role of user behavior in data governance adoption. By integrating data governance mechanisms naturally into workflows—rather than introducing additional responsibilities—we reduced friction in compliance. Users gradually learned data governance policies through interactive system feedback, leading to a self-reinforcing data governance model that improves data quality and trustworthiness over time.

Ultimately, our study shows that automation, interoperability, and structured digital infrastructure are not independent elements but interconnected pillars of effective data governance. By balancing these principles, we successfully reduced complexity while maintaining compliance, ensuring scalability, and improving trust in data management systems. These results contribute to ongoing discussions in data governance, dynamic data space, and digital infrastructure by offering a scalable, automated data governance model applicable across various domains.

Modern research and organizational activity is increasingly shaped by digital data. As this thesis has shown, handling that data effectively requires more than just storage capacity or flexible systems—it requires a data governance approach that is both scalable and sustainable. Yet, the scale of today's data production presents a foundational challenge. With digital data volumes growing nearly a hundredfold over the past 15 years [1], the pace of generation has far outstripped traditional management practices. Personal data collection, in particular, has surged in both volume and sensitivity, raising the stakes for secure, traceable, and reusable data infrastructures [134]. This growing influx of data adds pressure to data governance systems, especially in settings like research and healthcare, where quality and compliance must be balanced with resource constraints [60], [67], [135]. It is within this context that dynamic data spaces have emerged as a promising model—providing mechanisms to govern data as it flows across disciplines, organizations, and systems.

As the scale and speed of data generation increase, many organizations have struggled to keep up—resulting in what is often described as data debt: the backlog of unfinished, unvalidated, or poorly documented data work that accumulates over time [67], [136]. This debt is not just technical; it has practical consequences for data quality, trust, and reuse. Problems with accuracy, completeness, and consistency persist across sectors, particularly in

complex or collaborative settings such as healthcare and academic research [137]. In response, many institutions rely on off-the-shelf software platforms to manage their data, hoping to reduce costs and avoid building custom infrastructure from scratch [138]–[141]. Yet these generic tools often introduce further complications. As platforms expand their feature sets to accommodate broad use cases, they become increasingly difficult to govern and integrate into existing workflows—especially when data must be reused across domains or adhere to evolving standards [142], [143]. This growing complexity contributes to inefficiencies, oversight gaps, and operational bottlenecks [67], [144].

The solution proposed in this thesis responds directly to these challenges. Rather than layering on more features or requiring manual oversight, the approach simplifies data governance through a streamlined, automation-driven model embedded within the system architecture. By integrating data validation, metadata enforcement, and standardized output directly into the infrastructure, the system reduces the manual workload typically associated with data governance while ensuring compliance and data quality. As demonstrated in Papers 2, 3, and 4, this model supports consistent data handling even across diverse sources and user practices—showing how automation and simplification can effectively counteract data debt and fragmentation in real-world environments.

One promising response to the growing complexity of data governance is the emergence of dynamic data spaces. Dynamic data spaces are designed to enable secure, scalable, and interoperable data sharing while preserving data sovereignty and regulatory compliance [14], [31], [145]. Unlike traditional static systems, dynamic data spaces support real-time data movement across organizational boundaries—making them especially suitable for collaborative, cross-disciplinary, or privacy-sensitive environments. However, while dynamic data spaces hold considerable promise, implementing them remains technically demanding, particularly when integrating diverse data formats and ensuring consistent data governance across systems.

This thesis contributes a pragmatic solution to these challenges by demonstrating how simplified infrastructure and standardized data input mechanisms can support the goals of dynamic data spaces without introducing unnecessary complexity. Rather than accommodating a wide array of formats such as XML or JSON,

which would require either intricate validation logic or redundant ingestion tools, we enforced a single input format (CSV) throughout the infrastructure. This design choice drastically reduced input variability and made the system easier to scale and maintain. By consolidating all structured data into one standardized pathway, we minimized potential data governance failures while improving interoperability. As shown in Paper 2, this standardization not only simplified data ingestion but also enabled seamless metadata validation and downstream reuse—illustrating how dynamic data space principles can be effectively supported even in resource-limited contexts.

Building on this, Paper 2 [86] demonstrated how consolidating structured data within a unified schema can significantly improve both scalability and automation in dynamic data spaces. Traditional databases typically create a new table for each dataset, a practice that increases complexity, introduces maintenance burdens, and heightens the risk of data fragmentation. In contrast, our approach structured nearly 2,000 parameters from 70 distinct data sources into a homogeneous database model, minimizing redundancy and manual configuration while supporting large-scale integration across domains.

This simplified schema forms the backbone of our digital infrastructure and directly supports key dynamic data space principles. By unifying data under a common structure and enforcing a single input format (CSV), we enabled automated validation, consistent metadata enforcement, and real-time data ingestion—all of which are critical for building trustworthy and interoperable data ecosystems [41], [80], [122], [146], [147]. Standardization at both the data and infrastructure level reduces variability and error propagation, a core challenge in big data governance efforts [60], [148].

To address the diverse needs of our data users working across disciplines, the system was designed to produce a consistent, navigable, and standards-compliant data output. All validated data was exported in CSV format and organized into an automated hierarchical folder structure, grouping data by study, source type, and content area. Each output folder was accompanied by auto-generated metadata files that provided critical context—such as study descriptions, parameter names, and participant identifiers—ensuring the data was interpretable and reusable even outside its original setting.

This structure not only improved usability but also supported one of the central goals of dynamic data spaces: to create a reusable and interoperable environment that supports long-term data use. In contrast to traditional approaches, where structured exports often require custom formatting or manual oversight, our folder system standardized output across all sources. This consistency allowed data users to quickly understand and navigate complex datasets without needing to familiarize themselves with internal system logic or data transformations.

The output model also aligned with widely recognized data governance frameworks that emphasize metadata completeness, traceability, and reusability as core principles for sustainable digital infrastructure [10], [23], [56], [149]. By automating these features directly into the architecture, the system minimized reliance on user discipline and instead relied on machine-enforced consistency. As observed in Papers 3 and 4, participants expressed greater trust in these curated outputs than in the original data files—describing the folder structure as not only clearer and easier to work with, but also more reliable. This perception reflects a broader shift: automation is not just a tool for reducing data governance burden—it is also a mechanism for building user trust.

The system further supported privacy-aware data governance by enabling structured, automated control over data access and deletion. When a participant withdrew from the study, their associated records could be efficiently excluded from future exports by flagging a single metadata field (e.g., Status = Quit). This ensured compliance with GDPR's right-to-be-forgotten requirements without requiring complex deletion workflows or system-wide data searches. In fragmented data environments, such functionality is often difficult or impossible to implement consistently—highlighting the value of structured outputs in supporting participant agency and compliance [37], [76], [85].

This metadata-rich structure also had implications for data integrity and quality assurance. By automating folder generation, metadata tagging, and file validation, the system reduced opportunities for human error while improving transparency around how data was organized and processed. These features—often highlighted in data governance literature as ideal but difficult to achieve—were made feasible through embedded design, rather than additional processes or documentation. This model reinforces the idea that good data governance can be infrastructural,

woven into the systems that shape everyday data work [16], [29], [41], [109].

By automating these data governance tasks and embedding them directly into the data infrastructure, the solution offers a practical implementation of dynamic data space principles—one that remains sustainable even in resource-limited contexts. As global data sharing increases in scope and regulatory complexity, systems like this one can serve as template models for scalable, low-friction data governance, supporting cross-border collaboration, legal compliance, and trust at scale [15], [58], [150].

While the structured output model demonstrated how automation can enhance consistency and trust in data reuse, the process of data input revealed a different set of challenges. In Paper 3, we examined user interactions during data entry and validation, and found that socio-technical frictions remained a major obstacle. Specifically, user behavior often diverged from the intended standards of the system, reflecting a disconnect between expectations and actual practice.

Participants frequently relied on assumptions rather than engaging with available documentation or validation feedback. This was particularly evident in metadata compliance—only two out of thirteen participants correctly completed a required "Description" field, despite having access to clear instructions. These behaviors highlight a core difficulty in data governance: when standards are not automatically enforced, they are often ignored, regardless of how well they are documented [26], [54].

Such findings reinforce the argument that manual oversight is not a sustainable path to data quality. Automated validation must be built into the system to detect and correct errors in real time [10], [23]. Yet, automation alone is not a silver bullet. As participants became more familiar with the system, they began to rely on it, but this also introduced a new risk—they sometimes overlooked contextual or semantic errors that the system could not detect. For example, entering "height = 191" instead of "181" passed structural validation, but introduced incorrect information. This demonstrates the need for hybrid data governance strategies that combine automation with human awareness and oversight [151].

Importantly, every participant made at least one system-detected mistake. These ranged from formatting issues to missing columns or invalid file types, underscoring that even experienced

users are prone to error without immediate feedback. At the same time, this feedback became a learning mechanism: over repeated interactions, users internalized the data governance policies embedded in the system. Rather than simply correcting their inputs, they began adjusting their expectations and behaviors. This interactive learning loop proved to be a powerful method for improving adherence to standards, especially in resource-constrained environments where formal training is limited [108].

These results align closely with literature emphasizing the inadequacy of relying solely on human diligence in data governance, particularly in domains like healthcare, finance, and research. Effective data governance requires not just policies, but embedded mechanisms for real-time validation, structured input, and contextual learning. Our findings suggest that automation can act not only as a compliance tool, but also as a pedagogical one—helping users build better data practices over time [15], [26], [136].

Our study revealed significant variation in how participants interpreted the concept of "good" data quality. While some emphasized metadata completeness, others focused more on format consistency or structural integrity. Crucially, only those who prioritized metadata actually completed the metadata fields correctly, suggesting that data governance errors often stem not from a lack of capability, but from differing values and priorities.

Although previous literature has commonly linked data governance challenges to gaps in skill or data literacy [17], [26], our findings point to an additional challenge: data governance issues can arise even when users understand the rules, simply because they do not prioritize certain aspects of data quality. Participants frequently understood the purpose of data governance requirements but still overlooked metadata or sought to “just get through” validation. This highlights the need for data governance models that account not only for users’ capabilities but also for their motivations, incentives, and behavioral tendencies.

The input system played a central role in operationalizing data governance within our infrastructure. Rather than shifting responsibility away from users, it embedded validation directly into the submission process, providing immediate feedback and enforcing standards without requiring additional oversight. This approach reflects broader trends in automation-driven data governance, where compliance mechanisms are integrated into routine workflows to minimize manual intervention and reduce data

governance fatigue [11], [16], [56], [66].

By ensuring that users interacted only with validated and structured data, the system helped prevent common role conflicts and inconsistencies. Data governance rules evolved alongside actual use, as the system surfaced errors, flagged inconsistencies, and nudged users toward better practices—all without increasing their workload [22], [108]. Participants engaged meaningfully with the system, and the structured design of the input tasks promoted a smoother experience without compromising data integrity.

A persistent difficulty in data governance implementation is securing user participation [26], [108], [152]. Our design addressed this by aligning with immediate user goals—particularly in scenarios involving multiple heterogeneous data sources, such as those in the Sleep Revolution project. The input system simplified interoperability and reduced ambiguity, making it easier for users to comply without extensive training or technical knowledge. In contrast to data governance models that impose additional demands, our approach embedded data governance within existing routines, supporting adoption through minimal friction and clear value [108], [153].

In relation to the broader concepts introduced earlier, this work offers a practical answer to long-standing challenges in data governance. While prior literature often emphasizes complexity, high resource needs, or centralization [56], [58], our approach demonstrates that effective data governance can emerge from minimal, well-placed structure. This simplicity can be especially valuable in resource-limited or cross-sectoral environments, where traditional data governance frameworks often fail to adapt [64], [154].

To summarize the core contributions of this research, Table 5.1 outlines the main empirical insights derived from the studies, while Table 5.2 presents the design guidelines that emerged through the iterative development and evaluation of the data governance framework. Together, these tables distill the findings into practical takeaways for future design and implementation efforts.

Key Findings from the Research	
Theme	Description
Assumptions	Data work often relies on assumptions rather than verification.
Documentation Limitations	Documentation is frequently ignored and unreliable for ensuring compliance.
Unavoidable Errors	Errors that cannot be technically prevented will eventually occur.
Subjective Data Quality	The definition of "data quality" varies between users, affecting how data is handled and validated.

Table 5.1: Key findings that emerged from the research

Design Guidelines	
Guideline	Description
Simplicity First	Simplicity should always be a primary design goal and should only be sacrificed when absolutely necessary.
Focused Functionality	Limiting features to a core purpose allows for deeper refinement and reduces complexity.
Automated Standards	Standards should be enforced through automated validation rather than relying on documentation.
Pre-Use Validation	Validation and metadata enforcement should occur before data is used.
Reducing Workload	Data governance should lighten, not add to, user responsibilities.
Common Structure	Structured data should follow a shared schema to enable dynamic data spaces.
Consistent Foundation	A consistent data structure streamlines future development and integration.
Automated Curation	Data validation and structuring should be largely automated.
Interactive Learning	Feedback mechanisms should help users learn correct data governance practices over time.

Table 5.2: Design guidelines developed from the study

Limitations

While this research demonstrates promising results in developing and applying a simplified, automation-driven data governance model, several limitations must be acknowledged.

First, although the framework is theoretically generalizable and designed to accommodate diverse data sources, it has only been tested within a single research context. Its applicability to other domains—such as public sector institutions or industry settings—remains to be evaluated.

Second, participant numbers in the empirical studies were limited: ten participants for the usability testing in Paper 1, and thirteen for the feasibility testing, surveys, and interviews in Papers

3 and 4. These participants were drawn from a relatively small pool of experts working in similar domains and cultural settings, which may introduce bias in the qualitative findings.

Third, although the data governance system was developed iteratively over four years, the results presented in Papers 3 and 4 reflect evaluations based on a mature version of the system. Participants interacted with example files rather than contributing to long-term, production-level data entry. As of the time of writing, responsibility for populating the system still primarily lies with the development team, limiting our ability to assess long-term adoption or organizational change.

Finally, the data governance model is designed for large-scale, structured data integration. Its value proposition may be less compelling in environments dealing with relatively small or infrequent datasets, where the overhead of implementation may outweigh the benefits of automation.

These limitations suggest that while the results are encouraging, additional testing across diverse organizational contexts and over longer timeframes is needed to fully validate the model's generalizability and long-term impact.

Conclusion

This thesis demonstrates that data governance can be simplified and automated without compromising interoperability, scalability, or compliance. By combining a dynamic data space with a robust digital infrastructure, we developed a system that reduces complexity while supporting consistent, high-quality data use across diverse domains. The work responds to multiple key challenges: how to reduce human error and workload in data governance, how to build interoperable systems from the ground up, and how to ensure that data governance frameworks remain adaptable and scalable in real-world use. Together, these contributions provide a foundation for rethinking data governance in both research and operational contexts.

Our central research question—"How can complexity in data governance be reduced without sacrificing functionality, scalability, or compliance?"—guided an exploration into the origins of data errors and how they might be prevented. By identifying core socio-technical challenges, we showed how automation, structured

interoperability, and digital infrastructure can simplify data management while still supporting high compliance and data integrity.

Our findings reveal that data work is often based on assumptions rather than verification, making human-driven compliance unreliable. Users frequently ignore documentation, suggesting that data governance frameworks should not rely solely on written policies. Instead, they should enforce standards automatically. Our approach applied validation and metadata enforcement before any data use, ensuring that errors were caught early—at the point of ingestion—rather than downstream. We also observed that data quality is a highly subjective concept. These differences in perception directly affect how users engage with data governance standards.

By embedding simplicity as a design principle, we showed that reducing complexity lowers both the likelihood of errors and user resistance. A consistent database structure across all data sources meant interoperability was built-in from the start—not added later. Rather than supporting many redundant features, the system emphasized a narrow focus with a shared structure, which enabled scalability and cross-domain integration.

Our automated data governance model minimized manual work rather than introducing new responsibilities. Importantly, we did not assign new roles or tasks to users interacting with data. Instead, data governance practices were learned gradually through interactive system feedback. This approach proved far more practical than relying on static documentation or formal training. These findings align with broader trends advocating for automated validation, real-time standard enforcement, and adaptive data governance structures.

By balancing interoperability, simplicity, and automation, this research makes contributions to key areas of information systems research—especially data governance, dynamic data spaces, and digital infrastructure. Looking ahead, future research should explore how dynamic data space models can be expanded beyond structured data. This includes adapting automation and data governance for raw, semi-structured, or unstructured data. Additionally, continued work on simplification strategies for cross-domain interoperability is needed, to ensure that data governance frameworks remain scalable, adaptable, and aligned with evolving digital ecosystems.

Future Work

Building on our findings, future work should explore the integration of raw and unstructured data into our data governance framework. While the current design centers on structured sources, extending it to handle raw data would significantly broaden its applicability across research and organizational contexts. However, this shift raises new challenges around processing complexity, validation mechanisms, and ensuring interoperability—prompting the question of whether automation strategies that proved effective for structured data can scale to more ambiguous formats.

A second promising direction is the development of a dynamically expanding database architecture. Whereas our present model enforces a consistent structure, future iterations could enable the system to scale and adapt automatically as new data sources emerge. This would involve designing a backend that manages evolving schema complexity while preserving a simple and intuitive user experience. Such adaptability would align with broader efforts in digital infrastructure to reconcile simplicity with scalability, allowing data governance models to grow organically with organizational needs.

Finally, future research should continue to investigate the boundaries and potential of automation in data governance. Our results confirm that automation is essential for reducing manual workload, increasing reliability, and enabling compliance at scale. However, further work is needed to explore how automated data governance can accommodate increasingly diverse data types, dynamic organizational contexts, and evolving regulatory requirements. Advancing these areas will be crucial for developing data governance frameworks that are not only robust but also sustainable and domain-agnostic.

Bibliography

[1] Statista, *Volume of data created, captured, copied, and consumed worldwide from 2010 to 2024*, Accessed: 2024-02-19, 2024. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/.

[2] S. Ambler, *Data debt: Addressing enterprise data quality problems*. [Online]. Available: https://agiledata.org/essays/datatechnicaldebt.html.

[3] C. Seaman and Y. Guo, "Measuring and monitoring technical debt", in *Advances in Computers*, vol. 82, Elsevier, 2011, pp. 25–46.

[4] Precisely, *Data governance adoption has risen dramatically - here's how*, Accessed: 2025-02-24, 2024. [Online]. Available: https://www.precisely.com/blog/data-integrity/2025-planning-insights-data-governance-adoption-has-risen-dramatically.

[5] J. J. Al Wahshi, J. Foster, and P. Abbott, "An investigation into the role of data governance in improving data quality: A case study of the omani banking sector", in *ECIS 2022 Research Papers*, AIS Electronic Library (AISeL), 2022.

[6] B. M. V. Bernardo, H. São Mamede, J. M. P. Barroso, and V. M. P. D. dos Santos, "Data governance & quality management—innovation and breakthroughs across different fields", *Journal of Innovation & Knowledge*, vol. 9, no. 4, p. 100 598, 2024.

[7] L. Elfman and Data.World, *10 modern data management best practices: An essential guide*, 2025.

[8] D. Chu and Secoda, *How deezer balances short and long-term data management goals*, 2024.

[9] P. Kutty, P. Christensen, and IBM, *A step-by-step guide to setting up a data governance program*, 2023.

[10] R. Stevens, E. Verbakel, D. van der Linden, and F. S. de Boer, "Operationalizing and automating data governance", *Journal of Big Data*, vol. 9, no. 1, pp. 1–25, 2022. DOI: 10.1186/s40537-022-00673-5. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00673-5.

[11] B. Rowan, "Data governance: What it is and how it enhances data management", *StateTech Magazine*, 2023. [Online]. Available: https://statetechmagazine.com/article/2023/11/data-governance-what-it-and-how-it-enhances-data-management-perfcon.

[12] D. Patel, "Research data management: A conceptual framework", *Library review*, vol. 65, no. 4/5, pp. 226–241, 2016.

[13] A. Surkis and K. Read, "Research data management", *Journal of the Medical Library Association: JMLA*, vol. 103, no. 3, p. 154, 2015.

[14] C. Stedman, *What is data governance and why does it matter?*

[15] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, "A systematic literature review of data governance and cloud data governance", *Personal and ubiquitous computing*, vol. 23, pp. 839–859, 2019.

[16] K. Ngcobo, S. Bhengu, A. Mudau, B. Thango, and M. Lerato, "Enterprise data management: Types, sources, and real-time applications to enhance business performance-a systematic review", *Systematic Review| September*, 2024.

[17] I. Alhassan, D. Sammon, and M. Daly, "Critical success factors for data governance: A theory building approach", *Information Systems Management*, vol. 36, no. 2, pp. 98–110, 2019. DOI: 10.1080/10580530.2019.1589670. [Online]. Available: https://doi.org/10.1080/10580530.2019.1589670.

[18] M. Sargo, *Data governance lessons learned for best practices*, Accessed: 2025-02-24, 2024. [Online]. Available: https://www.dataideology.com/data-governance-lessons-learned-for-best-practices.

[19] B. Otto, "Data governance: A conceptual framework, structured review, and research agenda", *International Journal of Information Management*, vol. 49, pp. 424–438, 2019. DOI: 10.1016/j.ijinfomgt.2019.07.008. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2019.07.008.

[20] O. B. Nielsen, J. S. Persson, and S. Madsen, "Why governing data is difficult: Findings from danish local government", in *Smart Working, Living and Organising: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2018, Portsmouth, UK, June 25, 2018, Proceedings*, Springer, 2019, pp. 15–29.

[21] I. Alhassan, D. Sammon, and M. Daly, "Data governance activities: A comparison between scientific and practice-oriented literature", *Journal of Enterprise Information Management*, vol. 31, no. 2, pp. 300–316, 2018. DOI: 10.1108/JEIM-01-2017-0014.

[22] O. B. Nielsen, "A comprehensive review of data governance literature", in *Selected Papers of the IRIS, Issue Nr 8 (2017)*, Association for Information Systems, 2017, pp. 120–133. [Online]. Available: https://aisel.aisnet.org/iris2017/3/.

[23] S. U. Lee, L. Zhu, and R. Jeffery, "Data governance for platform ecosystems: Critical factors and the state of practice", in *Pacific Asia Conference on Information Systems (PACIS)*, Accessed: 2025-02-24, 2017. [Online]. Available: https://arxiv.org/abs/1705.03509.

[24] H. M. Wonga and S. Norbainib, *The data governance: A comprehensive literature review from professional viewpoints*.

[25] S. O. Aderemi, "Exploring the impact of big data on data governance", Ph.D. dissertation, Walden University, 2024.

[26] O. Benfeldt, J. S. Persson, and S. Madsen, "Data governance as a collective action problem", *Information Systems Frontiers*, vol. 22, no. 2, pp. 299–313, 2020. DOI: 10.1007/s10796-019-09923-z.

[27] M. Galvin, M. Heverin, É. Mac Domhnaill, *et al.*, "Challenges and solutions to complex data governance issues in cross-national, cross-sectoral, multidisciplinary real world health research: A descriptive overview", *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, pp. 1–7, 2025.

[28] P. Brous, M. Janssen, and R. Krans, "Data governance as success factor for data science", in *Conference on e-Business, e-Services and e-Society*, Springer, 2020, pp. 431–442.

[29] G. Birkbeck, T. Nagle, and D. Sammon, "Challenges in research data management practices: A literature analysis", *Journal of Decision Systems*, vol. 31, no. sup1, pp. 153–167, 2022.

[30] Precisely, *Data governance 101: Moving past challenges to operationalization*, Accessed: 2025-02-24, 2024. [Online]. Available: `https://www.precisely.com/resource-center/ebooks/data-governance-101`.

[31] M. Bacco, A. Kocian, S. Chessa, A. Crivello, and P. Barsocchi, "What are data spaces? systematic survey and future outlook", *Data in Brief*, vol. 57, p. 110 969, 2024.

[32] E. Curry, S. Scerri, and T. Tuikka, *Data spaces: design, deployment and future directions*. Springer Nature, 2022.

[33] B. Otto, "The evolution of data spaces", in *Designing data spaces: The ecosystem approach to competitive advantage*, Springer International Publishing Cham, 2022, pp. 3–15.

[34] L. Nagel, J. J. Hierro, E. Perea, *et al.*, "Design principles for data spaces: Position paper", E. ON Energy Research Center, Tech. Rep., 2021.

[35] R. M. Baygi, L. D. Introna, and L. Hultin, "Everything flows: Studying continuous sociotechnological transformation in a fluid and dynamic digital world", *MIS quarterly*, vol. 45, no. 1, pp. 423–452, 2021.

[36] H. Benbya, N. Nan, H. Tanriverdi, and Y. Yoo, "Complexity and information systems research in the emerging digital world", *MIS quarterly*, vol. 44, no. 1, pp. 1–17, 2020.

[37] E. Union, *Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation)*, Accessed: 2025-02-24, 2016. [Online]. Available: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

[38] S. Franzoi, T. Grisold, and J. vom Brocke, "Studying dynamics and change with digital trace data: A systematic literature review.", in *ECIS*, 2023.

[39] A. Gieß, F. Möller, T. Schoormann, and B. Otto, "Design options for data spaces.", in *ECIS*, 2023.

[40] A. Gieß, T. Schoormann, F. Möller, and I. Gür, "Discovering data spaces: A classification of design options", *Computers in Industry*, vol. 164, p. 104 212, 2025.

[41] D. G. Broo and J. Schooling, "Digital twins in infrastructure: Definitions, current practices, challenges and strategies", *International Journal of Construction Management*, vol. 23, no. 7, pp. 1254–1263, 2023.

[42] D. Tilson, K. Lyytinen, and C. Sørensen, "Digital infrastructures: The missing is research agenda", *Information Systems Research*, vol. 21, no. 4, pp. 748–759, 2010. DOI: `10.1287/isre.1100.0318`.

[43] P. Constantinides, O. Henfridsson, and G. G. Parker, "Digital infrastructure evolution: A digital trace data study", in *Proceedings of the 39th International Conference on Information Systems (ICIS)*, Accessed: 2025-02-24, 2018. [Online]. Available: `https://www.researchgate.net/publication/364199024_Digital_Infrastructure_Evolution_A_Digital_Trace_Data_Study`.

[44] J. Bodenhausen, C. Sorgatz, T. Vogt, *et al.*, "Securing wireless communication in critical infrastructure: Challenges and opportunities", *arXiv preprint arXiv:2311.01338*, 2023, Accessed: 2025-02-24. [Online]. Available: `https://arxiv.org/abs/2311.01338`.

[45] M. Schulte-Althoff, K. Schewina, D. Fürstenau, and G. M. Lee, "On the heterogeneity of digital infrastructure in entrepreneurial ecosystems", in *53rd Annual Hawaii International Conference on System Sciences, HICSS 2020*, Hawaii International Conference on System Sciences (HICSS), 2020, pp. 5728–5737.

[46] K. Wang, S. Cao, and P. Morita, "Data governance challenges on smart home involving multiple research sites", *European Journal of Public Health*, vol. 34, no. Supplement_3, ckae144–1025, 2024.

[47] E. Commission, *Revolution of sleep diagnostics and personalized health care based on digital diagnostics and therapeutics with health data integration*, Funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 965417, Accessed: 2025-02-24, 2020. [Online]. Available: `https://cordis.europa.eu/project/id/965417`.

[48] E. S. Arnardottir, A. S. Islind, M. Óskarsdóttir, *et al.*, "The sleep revolution project: The concept and objectives", *Journal of Sleep Research*, vol. 31, no. 4, e13630, 2022.

[49] E. S. Arnardottir, A. S. Islind, and M. Óskarsdóttir, "The future of sleep measurements: A review and perspective", *Sleep medicine clinics*, vol. 16, no. 3, pp. 447–464, 2021.

[50] M. T. Tulinius and A. Márton, "Flowing with digital currents: Exploring the flow-dynamics of digital innovation ecosystems", in *ICIS 2024 Proceedings*, Accessed: 2025-02-24, 2024. [Online]. Available: `https://aisel.aisnet.org/icis2024/diginnoventren/diginnoventren/32`.

[51] S. R. Barley and B. A. Bechky, "In the backrooms of science: The work of technicians in science labs", *Work and occupations*, vol. 21, no. 1, pp. 85–126, 1994.

[52] A. S. Islind, H. Vallo Hult, V. Johansson, E. Angenete, and M. Gellerstedt, "Invisible work meets visible work: Infrastructuring from the perspective of patients and healthcare professionals", in *54th Hawaii International Conference on System Sciences (HICSS9, Tuesday, January 5, 2021 to Friday, January 8, 2021*, Hawaii International Conference on System Sciences, 2021, pp. 3556–3565.

[53] R. A. W. Rhodes, "The new governance: Governing without government", *Political studies*, vol. 44, no. 4, pp. 652–667, 1996.

[54] M. Micheli, M. Ponti, M. Craglia, and A. Berti Suman, "Emerging models of data governance in the age of datafication", *Big Data & Society*, vol. 7, no. 2, p. 2 053 951 720 948 087, 2020.

[55] V. Khatri and C. V. Brown, "Designing data governance", *Communications of the ACM*, vol. 53, no. 1, pp. 148–152, 2010.

[56] K. Bližnák, M. Munk, and A. Pilková, "A systematic review of recent literature on data governance (2017-2023)", *IEEE Access*, 2024.

[57] S. Alofaysan, B. Alhaqbani, R. Alseghayyir, and M. Omar, "The significance of data governance in healthcare", in *BIOSTEC 2014: Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, vol. 5, 2014, pp. 178–187.

[58] S. Marcucci, N. G. Alarcon, S. G. Verhulst, and E. Wullhorst, "Mapping and comparing data governance frameworks: A benchmarking exercise to inform global data governance deliberations", *arXiv preprint arXiv:2302.13731*, 2023.

[59] S. S. Give, "The world's most valuable resource is no longer oil, but data", *The Economist*, 2017.

[60] I. Taleb, M. A. Serhani, and R. Dssouli, "Big data quality: A survey", in *2018 IEEE International Congress on Big Data (BigData Congress)*, 2018, pp. 166–173. DOI: 10.1109/BigDataCongress.2018.00029.

[61] J. Ladley, *Data governance: How to design, deploy, and sustain an effective data governance program.* Academic Press, 2019.

[62] E. Parmiggiani and M. Grisot, "Data curation as governance practice", *Scandinavian Journal of Information Systems*, 2020.

[63] B. Petzold, M. Roggendorf, K. Rowshankish, and C. Sporleder, "Designing data governance that delivers value", *McKinsey Digital*, vol. 26, 2020.

[64] R. Mahanti, *Data governance and data management*. Springer, 2021.

[65] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy artificial intelligence", *Government Information Quarterly*, vol. 37, no. 3, p. 101 493, 2020.

[66] J. R. Talburt, L. Ehrlinger, and J. Magruder, *Automated data curation and data governance automation*, 2023.

[67] Dataversity and Erwin, *2020 state of data governance and automation*, 2020.

[68] T. Li, A. Wandella, R. Gomer, and M. H. Al-Mafazy, "Operationalizing health data governance for ai innovation in low-resource government health systems: A practical implementation perspective from zanzibar", *Data & Policy*, vol. 6, e63, 2024.

[69] T. Schurig, A. Kari, and D. Fürstenau, "The symphony of orchestrated participatory data space governance: A systematic review", in *Proceedings of the Thirty-Second European Conference on Information Systems (ECIS 2024)*, Accessed: 2025-02-24, Paphos, Cyprus, 2024.

[70] M. Hupperz and A. Gieß, *The interplay of data-driven organizations and data spaces: Unlocking capabilities for transforming organizations in the era of data spaces*, 2024.

[71] T. Wang, X. Zhang, J. Feng, and X. Yang, "A comprehensive survey on local differential privacy toward data statistics and analysis", *Sensors*, vol. 20, no. 24, p. 7030, 2020.

[72] S. Paiho, P. Tuominen, J. Rökman, M. Ylikerälä, J. Pajula, and H. Siikavirta, "Opportunities of collected city data for smart cities", *IET Smart Cities*, vol. 4, no. 4, pp. 275–291, 2022.

[73] H. Ren, H. Li, X. Liang, S. He, Y. Dai, and L. Zhao, "Privacy-enhanced and multifunctional health data aggregation under differential privacy guarantees", *Sensors*, vol. 16, no. 9, p. 1463, 2016.

[74] E. Illman, *Avoiding growing pains in the development and use of digital twins*, 2024.

[75] E. Curry, T. Tuikka, A. Metzger, *et al.*, "Data sharing spaces: The bdva perspective", in *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*, Springer International Publishing Cham, 2022, pp. 365–382.

[76] M. Ryan, P. Gürtler, and A. Bogucki, "Will the real data sovereign please stand up? an eu policy response to sovereignty in data spaces", *International Journal of Law and Information Technology*, vol. 32, no. 1, eaae006, 2024.

[77] R. A. Deshmukh, D. Collarana, J. Gelhaar, *et al.*, "Challenges and opportunities for enabling the next generation of cross-domain dataspaces", in *The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference*, 2024.

[78] E. Commission, *European data governance act*, 2020.

[79] E. Union, *Pioneering the eu's sector-specific data spaces: The european health data space*, 2024.

[80] G. Solmaz, F. Cirillo, J. Fürst, *et al.*, "Enabling data spaces: Existing developments and challenges", in *Proceedings of the 1st International Workshop on Data Economy*, 2022, pp. 42–48.

[81] I. Borchert, "Interoperability of data governance regimes: Challenges for digital trade policy", *CITP*, 2024.

[82] K. McBride, S. Kamalanathan, S.-M. Valdma, T. Toomere, and M. Freudenthal, "Digital government interoperability and data exchange platforms: Insights from a twenty country comparative study", in *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, 2022, pp. 90–97.

[83] D. Hodapp and A. Hanelt, "Interoperability in the era of digital innovation: An information systems research agenda", *Journal of Information Technology*, vol. 37, no. 4, pp. 407–427, 2022.

[84] A. P. Rodriguez Müller, J. Martin Bosch, and L. Tangi, "Interoperability for emerging technologies: A systematic scoping review", in *International Conference on Electronic Government*, Springer, 2024, pp. 387–401.

[85] C. Staunton, M. Shabani, D. Mascalzoni, S. Mežinska, and S. Slokenberga, "Ethical and social reflections on the proposed european health data space", *European Journal of Human Genetics*, vol. 32, no. 5, pp. 498–505, 2024.

[86] B. F. Sveinbjarnarson, L. Schmitz, E. S. Arnardottir, and A. S. Islind, "The sleep revolution platform: A dynamic data source pipeline and digital platform architecture for complex sleep data", *Current Sleep Medicine Reports*, vol. 9, no. 2, pp. 91–100, 2023.

[87] L. Stegemann, R. Gubser, M. Gersch, *et al.*, "Future-oriented and patient-centric? a qualitative analysis of digital therapeutics and their interoperability", in *ECIS 2023*, 2023.

[88] D. Le Phuoc, S. Schimmler, A. Le-Tuan, U. A. Kuehn, and M. Hauswirth, "Towards a decentralized data hub and query system for federated dynamic data spaces", in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 1452–1457.

[89] F. Möller, I. Jussen, V. Springer, *et al.*, "Industrial data ecosystems and data spaces", *Electronic Markets*, vol. 34, no. 1, p. 41, 2024.

[90] S. Ishihara and T. Matsutsuka, "Towards interoperable data spaces: Comparative analysis of data space implementations between japan and europe", *arXiv preprint arXiv:2501.15738*, 2025.

[91] R. Van Noorden, "More than 10,000 research papers were retracted in 2023 — a new record", *Nature*, vol. 624, no. 7992, pp. 479–481, Dec. 12, 2023, Bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Scientific community, Publishing. DOI: 10.1038/d41586-023-03974-8. [Online]. Available: https://www.nature.com/articles/d41586-023-03974-8 (visited on 02/26/2025).

[92] R. Van Noorden, "Exclusive: These universities have the most retracted scientific articles", *Nature*, vol. 638, no. 8051, pp. 596–599, Feb. 19, 2025, Bandiera_abtest: a Cg_type: News Feature Publisher: Nature Publishing Group Subject_term: Ethics, Scientific community, Publishing, ISSN: 1476-4687. DOI: 10.1038/d41586-025-

00455-y. [Online]. Available: https://www.nature.com/articles/d41586-025-00455-y (visited on 02/26/2025).

[93] S. Coutts and S. Gagnon-Turcotte, "Data governance and digital infrastructure: Analysis and key considerations for the city of toronto", *Open North*, 2020.

[94] B. Metin, F. G. Özhan, and M. Wynn, "Digitalisation and cybersecurity: Towards an operational framework", *Electronics*, vol. 13, no. 21, p. 4226, 2024.

[95] V. Duvvur, "Modernizing government it systems: A case study on enhancing operational efficiency and data integrity", *International Journal of Computational and Experimental Science and Engineering*, 2025.

[96] M. Brommeyer, M. Whittaker, and Z. Liang, "Organizational factors driving the realization of digital health transformation benefits from health service managers: A qualitative study", *Journal of Healthcare Leadership*, pp. 455–472, 2024.

[97] D. Jackson and C. Allen, "Enablers, barriers and strategies for adopting new technology in accounting", *International Journal of Accounting Information Systems*, vol. 52, p. 100 666, 2024.

[98] L. Manny, M. Duygan, M. Fischer, and J. Rieckermann, "Barriers to the digital transformation of infrastructure sectors", *Policy Sciences*, vol. 54, pp. 943–983, 2021.

[99] Y. Li, Q. Du, G. Ma, and C. Gazang, "Can strengthening digital infrastructure enhance productivity in the cultural industry? evidence from tibet", *PloS one*, vol. 20, no. 1, e0317366, 2025.

[100] P. Ghosh, *Data architecture trends in 2022*, 2022.

[101] S. Sachdeva, S. Bhatia, A. Al Harrasi, *et al.*, "Unraveling the role of cloud computing in health care system and biomedical sciences", *Heliyon*, 2024.

[102] E. Bainomugisha and A. Mwotil, "Crane cloud: A resilient multi-cloud service abstraction layer for resource-constrained settings", *Development Engineering*, vol. 7, p. 100 102, 2022.

[103] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big data and cloud computing: Innovation opportunities and challenges", *International Journal of Digital Earth*, vol. 10, no. 1, pp. 13–53, 2017.

[104] L. Dorn and Conceptboard, *The us cloud act: Threatening european data protection*, 2024.

[105] A. Mathew, *Cloud data sovereignty governance and risk implications of cross-border cloud storage*, 2024.

[106] K. Foitzick and activeMind.legal, *U.s. cloud act vs. gdpr*, 2020.

[107] IESBA, *Technology landscape - focus on data governance*, 2023.

[108] C. A. Bassi and S. N. A. d. Souza, "Challenges to implementing effective data governance: A literature review", *IC3K 2023: Proceedings*, 2023.

[109] M. Savona, "Data governance: Main challenges", *CESifo*, 2024.

[110] Scalecomputing, *Data sovereignty, data residency, and data localization: An introduction*, 2023.

[111] K. Linask-Goode, *Data sovereignty and privacy in financial services: Adapting faster with data-centric architecture*, 2024.

[112] R. Vinogradov, *Enterprise data integration: Building a unified data ecosystem*, 2025.

[113] P. Pagano, L. Candela, and D. Castelli, "Data interoperability", *Data Science Journal*, vol. 12, GRDI19–GRDI25, 2013.

[114] J. A. Balch, M. M. Ruppert, T. J. Loftus, *et al.*, "Machine learning–enabled clinical information systems using fast healthcare interoperability resources data standards: Scoping review", *JMIR Medical Informatics*, vol. 11, e48297, 2023.

[115] A. Torab-Miandoab, T. Samad-Soltani, A. Jodati, and P. Rezaei-Hachesu, "Interoperability of heterogeneous health information systems: A systematic literature review", *BMC medical informatics and decision making*, vol. 23, no. 1, p. 18, 2023.

[116] B. Otjacques, P. Hitzelberger, and F. Feltz, "Interoperability of e-government information systems: Issues of identification and data sharing", *Journal of management information systems*, vol. 23, no. 4, pp. 29–51, 2007.

[117] L. McVey, N. Alvarado, J. Greenhalgh, *et al.*, "Hidden labour: The skilful work of clinical audit data collection and its implications for secondary use of data via integrated health it", *BMC Health Services Research*, vol. 21, pp. 1–11, 2021.

[118] O. of the National Coordinator for Health Information Technology, *Connecting health and care for the nation: A 10-year vision to achieve an interoperable health it infrastructure*, 2014.

[119] B. Schmidt, A. Hohlfeld, and N. Leon, *Defining and conceptualising data harmonisation: A scoping review protocol', systematic reviews 7 (1), 1-6*, 2018.

[120] E. P. Adeghe, C. A. Okolo, and O. T. Ojeyinka, "The role of big data in healthcare: A review of implications for patient outcomes and treatment personalization", *World Journal of Biology Pharmacy and Health Sciences*, vol. 17, no. 3, pp. 198–204, 2024.

[121] M. Ashiq, M. H. Usmani, and M. Naeem, "A systematic literature review on research data management practices and services", *Global Knowledge, Memory and Communication*, vol. 71, no. 8/9, pp. 649–671, 2022.

[122] A. Amjad, F. Azam, M. W. Anwar, and W. H. Butt, "A systematic review on the data interoperability of application layer protocols in industrial iot", *Ieee Access*, vol. 9, pp. 96 528–96 545, 2021.

[123] P. Spagnoletti, N. Kazemargi, P. Constantinides, A. Prencipe, *et al.*, "Data control coordination in the formation of ecosystems in highly regulated sectors", *Journal of the Association for Information Systems*, 2024.

[124] V. Fast, D. Schnurr, and M. Wohlfarth, "Regulation of data-driven market power in the digital economy: Business value creation and competitive advantages from big data", *Journal of Information Technology*, vol. 38, no. 2, pp. 202–229, 2023.

[125] M. K. Sein, O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren, "Action design research", *MIS quarterly*, pp. 37–56, 2011.

[126] M. T. Mullarkey and A. R. Hevner, "An elaborated action design research process model", *European journal of information systems*, vol. 28, no. 1, pp. 6–20, 2019.

[127] G. I. Susman and R. D. Evered, "An assessment of the scientific merits of action research", *Studi organizzativi*, no. 2022/2, 2023.

[128] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research", *MIS quarterly*, pp. 75–105, 2004.

[129] A. Q. Gill and E. Chew, "Configuration information system architecture: Insights from applied action design research", *Information & Management*, vol. 56, no. 4, pp. 507–525, 2019.

[130] B. F. Sveinbjarnarson, E. S. Arnardottir, and A. S. Islind, "Layer upon layer: Developing layered modular architectures for data-driven health platform", in *Digital (Eco) Systems and Societal Challenges: New Scenarios for Organizing*, Springer, 2024, pp. 91–108.

[131] B. F. Sveinbjarnarson, E. S. Arnardottir, and A. S. Islind, "Designing and developing layered-modular architecture for data-driven health platforms", *Italian Conference of Information Systems*, 2023.

[132] J. Brooke *et al.*, "Sus-a quick and dirty usability scale", *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[133] M. Hassenzahl, M. Burmester, and F. Koller, "Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität", *Mensch & Computer 2003: Interaktion in Bewegung*, pp. 187–196, 2003.

[134] L. Biedebach, M. Óskarsdóttir, E. S. Arnardottir, and A. S. Islind, "Two sides of the same pillow: Unfolding the relationship between objective and subjective sleep quality with unsupervised learning", *international conference of information systems (ICIS)*, 2023.

[135] F. L. Susanto, A. Maulani, and S. N. Nuri, "Implementation of big data analytics and its challenges in digital transformation era: A literature review", *International Journal of Informatics and Information Systems*, vol. 7, no. 2, pp. 90–99, 2024.

[136] T. C. Redman, A. M. Smith, J. Ladley, M. Vercauteren, M. Hawker, and A. Wilkerson, *Data governance is failing: Here's why*, Accessed: 2025-02-24, 2024. [Online]. Available: `https://www.cdomagazine.tech/opinion-analysis/data-governance-is-failing-heres-why`.

[137] A. ALSHAMMARI, M. NASSER, and M. N. YUSUF, "Big data governance challenges arising from data generated by intelligent systems technologies: A systematic", *IEEE Access*, 2025.

[138] E. Christiawan, *Comparative assessment of custom developed and packaged software in accomplishing business's objective*, 2016.

[139] D. T. Mirzoev and R. Alvarez, "Leveraging vmware vcloud director virtual applications (vapps) for operational expense (opex) efficiency", *arXiv preprint arXiv:1404.2157*, 2014.

[140] S. Deochake, "Cloud cost optimization: A comprehensive review of strategies and case studies", *arXiv preprint arXiv:2307.12479*, 2023.

[141] Ourwitly, *Off-the-shelf software and custom software: Advantages and disadvantages*, 2024.

[142] A. Fawzy, A. Tahir, M. Galster, and P. Liang, "Data management challenges in agile software projects: A systematic literature review", *arXiv e-prints*, arXiv–2402, 2024.

[143] A. Marie Smith and Dataversity, *Data governance trends in 2025*, 2024.

[144] N. Shadbolt, *Can better data save the nhs?*, 2024.

[145] A. Hutterer and B. Krumay, *The adoption of data spaces: Drivers toward federated data sharing*, 2024.

[146] ITSA, *Digital infrastructure strategy report*, 2023.

[147] R. Hussain Shaikh, *Data interoperability: Key principles, challenges, and best practices*, 2024.

[148] A. Abbasi, S. Sarker, and R. H. Chiang, "Big data research in information systems: Toward an inclusive research agenda", *Journal of the Association for Information Systems*, vol. 17, no. 2, p. 3, 2016.

[149] J. MERKUS, R. W. HELMS, and R. Kusters, "Data governance capabilities; empirical validation in case studies of large organisations", *36th Bled eConference: Digital Economy and Society: The Balancing Act for Digital Innovation in Times of Instability, BLED*, pp. 35–48, 2023.

[150] W. B. Group, *Data for better lives*, 2022. [Online]. Available: `https://openknowledge.worldbank.org/bitstream/handle/10986/35218/211600ov.pdf`.

[151] A. Alsaad, "Governmental data governance frameworks: A systematic literature review", in *2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, IEEE, 2023, pp. 150–156.

[152] G. Maren and Dataversity, *Understanding the potential failures of a data governance program*, 2024.

[153] A. Procter and Spark, *Why data governance keeps holding enterprises back*, 2025.

[154] P. Vassilakopoulou, E. Skorve, and M. Aanestad, "Enabling openness of valuable information resources: Curbing data subtractability and exclusion", *Information Systems Journal*, vol. 29, no. 4, pp. 768–786, 2019.

Glossary

Action Design Research A research methodology combining action research and design science, emphasizing iterative cycles of building, intervention, evaluation, and learning within real-world contexts.

Automation The use of computational systems to execute tasks without continuous human intervention, often employed to reduce manual labor and increase consistency in data handling.

Data Fabric An architectural approach that enables seamless access and sharing of data across a distributed data environment, using metadata, automation, and integration technologies to unify data management, governance, and integration.

Data Curation The active management of data throughout its lifecycle to ensure it is discoverable, accessible, high-quality, and reusable.

Data Debt Accumulated inefficiencies or issues in data quality, documentation, or structure that create increasing effort and risk over time.

Data Governance A framework of roles, responsibilities, standards, and processes that ensure

	the accuracy, integrity, and usability of data across its lifecycle.
Data Infrastructure	The underlying technical and organizational systems that enable the storage, movement, validation, and use of data.
Data Lake	A centralized repository that stores large volumes of raw data in its native format, including structured, semi-structured, and unstructured data, enabling flexible access and analysis.
Data Pipeline	A sequence of automated steps that move and transform data from its raw state to a structured, usable format.
Data Quality	The condition of a dataset as measured by factors such as accuracy, completeness, consistency, and relevance.
Data Validation	The process of ensuring that data meets predefined formats, rules, and logical consistency before being accepted into a system.
Data Warehouse	A structured and centralized data storage system designed for reporting and analysis, where data is cleaned, transformed, and organized into predefined schemas to support decision-making.
Digital Infrastructure	A socio-technical system consisting of hardware, software, standards, and human processes that collectively enable digital services and data flows.
Dynamic Data Space	A flexible, scalable, and secure digital environment that supports the integration, management, and data governance of heterogeneous data across disciplines, organizations, or systems.

Feasibility Testing A method for assessing the practicality and usability of a system or concept through hands-on interaction and feedback from users.

FormResult FormResult and EntryResult are elements within the homogeneous database structure representing, respectively, a completed form (row) and a specific value within that form (cell).

Framework A structured approach or model that outlines key elements, relationships, and principles guiding a process or system.

Data Governance Model A conceptual structure outlining how rules, responsibilities, and processes are applied to manage data effectively.

Homogeneous Database A database model where different data sources follow the same structural format, allowing for unified data storage, access, and validation.

Interoperability The ability of different systems, organizations, or software to exchange, interpret, and use data effectively, enabling seamless integration and communication across platforms or domains.

Metadata Data that provides context or descriptive information about other data, such as source, structure, or meaning.

Platform A software system or environment that provides services and capabilities for managing and accessing data or applications.

Research Data Data collected, observed, or created in the course of conducting research.

Validation System The validation system, called Data Integrity Assurance System (DIAS), embedded into the digital infrastructure, designed to validate data input and enforce standard compliance.

Appendix A

Publication I

Designing and Developing Layered-Modular Architecture for Data-driven Health Platforms

Conference Paper · September 2023

CITATIONS
0

READS
96

3 authors:

Bjarki Freyr Sveinbjarnarson
Reykjavík University
6 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Erna S Arnardottir
Reykjavík University
190 PUBLICATIONS **5,091** CITATIONS
SEE PROFILE

Anna Sigridur Islind
Reykjavík University
143 PUBLICATIONS **904** CITATIONS
SEE PROFILE

Designing and Developing Layered-Modular Architecture for Data-driven Health Platforms

Bjarki Freyr Sveinbjarnarson[1, 2 [0009-0001-7063-5744]], Erna Sif Arnardottir[1, 2 [0000-0003-0877-3529]] and Anna Sigridur Islind[1, 2 [0000-0002-4563-0001]]

[1] Reykjavik University, Department of Computer Science, Reykjavik, Iceland
[2] Reykjavik University, Sleep Institute, Reykjavik, Iceland
bjarkis@ru.is

Abstract. Data collection is a foundational aspect of research, as most research projects ultimately rely on their data. Quick and effective data overview is crucial, especially when working with extensive datasets. This paper seeks to add to the ongoing discourse about designing and developing layered-modular architectures for data-driven health platforms. We particularly emphasize data representation and address the research question: *How can layered-modular architecture be used to present and filter research data in a digital platform?* We designed and developed a layered-modular architecture for a digital platform aimed at data extraction and representation. We argue that this approach renders digital platforms more reusable, resilient to changes, and easily extendable. To evaluate its effectiveness, we evaluated the layered-modular architecture of the platform with 10 experts and collected data on user experience and usability, along with semi-structured interviews. Furthermore, we examined the system's scalability with increasing data loads. The paper's key contribution revolves around the layered-modular architecture and design principles tailored for digital platforms, aiming to function effectively in socio-technical environments, particularly within intricate multidisciplinary research projects.

Keywords: Digital Platform, Design Principles, Layered-Modular Architecture, Data-Driven Health Platforms, Sleep Revolution.

1 Introduction

The modern world greatly depends on information, as we gather various types of data in larger quantities each year across society. This is evident in how swiftly we are collecting data. According to Statista [1], from 2010 to 2020, the total data created, captured, copied, or used grew from about 2 to 59 zettabytes (59 trillion gigabytes). Since then, global data has continued to grow, and it is expected to keep growing in the coming years [1]. At the same time, research is relying more on data, with some fields needing diverse data types due to the increased variety of data available [2–4]. Sleep studies show this trend, often using multiple sensors simultaneously, resulting in a wide variety of data [4, 5]. Additionally, sleep has a large impact on multiple factors, which makes data collection complex and a wide variety of data on sleep, from different

sensors, outlines the essence of sleep research. This means that data for sleep research can be both of objective and subjective nature and come from different sensor types, collected at different times, including physiological signals from wearables, cognitive tests, digital self-assessments, questionnaires, etc. [5]. When this data is collected at first, it is usually not in a usable format [6]. To make it usable, it is vital to first organize the collected data [6]. That means that the data is not usable until it has been processed into information [7].

The Sleep Revolution project, a sizable initiative funded by the European Union and conducted at Reykjavik University in partnership with 38 collaborators across Europe, collects a wide variety of sleep data [8]. The collected data is consolidated within a single digital platform. Here, the diverse data sources are transformed into a consistent structure and displayed for participants. In the digital platform's backend, filters are crucial to researchers to navigate and comprehend the gathered data. Moreover, as the Sleep Revolution project progresses, it is likely that the types of data, extraction methods, and the requirements of the platform will evolve. This underscores the need for adaptability in the platform's extraction and filtering features.

The main goal of this paper is to illuminate the socio-technical perspective during the design and development phase of the digital platform. This focus is particularly on extracting metadata from multiple sleep measurements, aiming to make it easier for end-users within the digital platform context to understand the data's content. To achieve this effectively, considering the different origins and characteristics of available data, end-users should be able to examine the data in a manner that lets them recognize the information it holds, using variables like statistical power, correlation, and information coverage. Consequently, the paper outlines the design and development of the digital platform, which facilitates the extraction of various data into a standardized structure. This structure showcases the data's information and enables the isolation and effective visualization of data with specific traits. Given the future research purposes that remain undiscovered, a highly adaptable architecture is necessary for the digital platform's data exploration. Furthermore, an adaptive architecture simplifies future designs for expanding and integrating the digital platform. In this paper, we leverage a layered-modular architecture to achieve adaptability in design and development. This raises the following question: *How can layered-modular architecture be used to present and filter research data in a digital platform*? The paper's key contribution revolves around the layered-modular architecture and design principles tailored for digital platforms, aiming to function effectively in socio-technical environments, particularly within intricate multidisciplinary research projects.

2 Related Research

2.1 Data and Data-Driven Healthcare

Collecting data forms the foundation of every research project. However, researchers often employ multiple methods for data collection, leading to the accumulation of diverse data types [9]. Common systematic data collection approaches encompass questionnaires, interviews, focus groups, observations, extraction of data from sensors, and

utilization of secondary data sources such as electronic health records [9]. Moreover, when research necessitates substantial volumes of both subjective and objective data, as seen in studies involving sensor data (objective data) and sleep diaries (subjective data), comprehending the data's underlying information can pose challenges [4]. Similarly, individuals might encounter difficulties in understanding information when dealing with extensive data entries or complex data formats. Additionally, if data originates from various sources, it can adopt multiple formats. In such instances, the organization and comprehension of information within the data can be enhanced using modular metadata structures [10]. Consequently, in digital platforms that accommodate diverse or new data types, adaptability, extensibility, and scalability become imperative [10]. During metadata creation, two valuable refinements contribute to improving metadata quality [10]. Firstly, consolidating various elements that signify the same entity can enhance clarity. For instance, different data sources might use terms like "participant," "patient," and "ID," all referring to the same subject's name. Such elements can be unified, using "ID" universally. Secondly, defining the range or format of a data type can be beneficial. For example, time encoding could include formats such as "yyyy-mm-dd," "dd/mm/yy," or "July 5th, 2002, 15:30." Additionally, certain types of metadata, like name order or dates, might be interpreted differently across cultures. For instance, 01/02/2020 signifies February 1st in some countries and January 2nd in others, highlighting the importance of addressing multicultural considerations [9]. Combining diverse formats into a unified element enhances the usability of the digital platform [10]. Consequently, the method of associating metadata with collected data is pivotal, as its effectiveness may vary based on the association model [10].

Data structures categorize data collections into three types: i) unstructured, ii) semi-structured, and iii) (fully) structured. Unstructured data, such as texts, images, and videos, lacks relational database descriptions, leading to challenging navigational control [11]. Semi-structured data incorporates tags as markers, straying from structured data models while retaining flexibility and the potential for transformation into structured data through schemas [12]. Fully structured data is organized beforehand, as commonly found in relational databases. This setup ensures better performance and navigation but comes with less flexibility and scalability [11].

Effective digital platform design necessitates careful consideration of core components like the database and backend. We appraised the six most prevalent database management systems (DBMS) [13]. DB-Engines, a knowledge hub for relational and NoSQL DBMS, showcases the most popular DBMS, focusing on relational types for this review [14]. Among them, Oracle ranks highest, lauded for its system mentions, general interest, and professional networks [14]. Oracle excels in managing vast data with concurrent user access via distributed processing in a client/server architecture [15]. MySQL, the second most popular, is an open-source, high-performance, multi-user DBMS available under GNU General Public License or a commercial license [16]. Microsoft SQL, the third most used, offers editions catering to performance and runtime requirements, with good scalability, availability, and security [17]. PostgreSQL, the fourth most popular, is an open-source object-relational DBMS, adhering to the SQL standard and boasting a liberal license [18]. IBM Db2, fifth in popularity, is an open-source relational DBMS with AI-driven capabilities, data management empowerment,

and essential tools [19]. Finally, SQLite, the sixth most common, serves a wide range of applications, operating systems, and browsers as a free and open-source database engine [20].

2.2 Data-driven Health Platforms for Sleep Data

Sleep constitutes a fundamental activity, accounting for roughly a third of human lives [3]. Remarkably, the endurance limit for sleep deprivation is shorter than for food, with the record being about 11 days [21]. Our evolving understanding of sleep has revealed an array of sleep disorders, characterized by frequent interruptions, diminished duration, and compromised quality, is often without the individual's awareness [4, 22]. Among these, obstructive sleep apnea, a prevalent disorder in which breathing ceases periodically throughout the night, stands out. Although approximately 20% of the general population is affected while a mere 15% of those are aware of their condition [22, 23]. Insufficient and poor-quality sleep is linked to impaired cognitive function, reduced quality of life, heightened disease susceptibility, including cancer, hindered physical recovery and growth, weakened immune response, and numerous other repercussions. Sleep studies encompass the collection and analysis of multiple parameters, including oxygen saturation, sleep fragmentation, sleeping position, duration across different sleep stages, heart rate, and more [3, 4, 22, 24]. Over time, researchers have designed questionnaires, sleep diaries, and cognitive tests to both identify individuals at risk of sleep disorders and comprehend the extent it affects quality of life [4].

Organizing components within a digital platform is crucial for various reasons. These reasons include: i) justifying implementations and decisions, ii) providing a clearer overview and purpose, iii) reducing platform complexity, iv) enhancing code reusability and adaptability to changes, and v) potentially achieving scalability, flexibility, and better performance depending on the chosen structure [13, 25]. Hence, selecting an architecture that aligns with the platform's design is essential to meet its requirements. Modular architecture is a widely used approach in software design and development that involves creating independent modules as part of a larger system [26]. This approach aims to decrease system complexity while increasing flexibility [26]. Additionally, due to the ease of replacing, updating, or removing components in modular architecture, it is considered adaptable [13, 26].

Likewise, layered-modular architecture divides code modules into distinct layers or packages [25]. Each layer serves a specific purpose within the system, forming a collection of components [25]. The three common layers in layered-modular architecture are the presentation/contents layer, business/service layer, and database/device layer [13, 25]. The presentation layer handles user interaction and request communication [25]. The business layer processes these requests into logical actions, often involving communication with the database layer [25]. The database layer manages data queries and access [25]. Layers are considered ‘open’ if the layer above can communicate with layers below them; conversely, if layers above cannot directly communicate with those below, the layer is ‘closed’ [25]. Generally, designers and developers aim to maintain closed layers to minimize dependencies within the system [25]. Another notable architecture is the Microkernel Architecture, featuring a core application [27]. Plug-in

components extend the core system's capabilities [27]. The design focus for the core digital platform is to make minimal adjustments necessary to seamlessly integrate added plugins [27].

3 Research Approach

Action Design Research (ADR) constitutes a research approach aimed at assessing digital products [28]. The ADR process involves four distinct phases: i) Problem Formulation, ii) Building, Intervention, and Evaluation, iii) Reflection and Learning, and iv) Formalization of Learning [28]. During the Problem Formulation phase, aspects like design creation, research question formulation, and role definition are addressed for subsequent evaluation [28]. The second phase encompasses design implementation, intervention, and evaluation. The third phase, executed concurrently with the first two, involves ongoing reflection and observation [28]. Ultimately, the last phase formalizes the acquired insights from the design process [28]. In this study, the ADR framework is applied as continuous learning informed and refined the design, akin to the third ADR stage [28]. Moreover, the design was developed with evaluation as a core consideration, aligning with the pivotal aspects of the first ADR stage [28]. Additionally, usability tests and evaluations were crafted to contribute to future system enhancements, which resonates with the essence of the second ADR stage [28]. Lastly, design principles were formulated to guide analogous problem-solving endeavors, mirroring the purpose of the fourth ADR stage [28]. However, it is worth noting that this research method does not encompass all the tasks and principles outlined in ADR [28]. For instance, tasks such as identifying the initial knowledge-creation target and creating mutually influential roles are not explicitly incorporated in this study [28].

3.1 Research Context

The primary objective of the Sleep Revolution project is to amass a comprehensive collection of sleep data, enabling researchers to extract insights into fundamental aspects of sleep, sleep disorders, and alternative methods of measurement. The core focus of the project lies in the collection and processing of sleep data, necessitating certain key considerations for delivering dependable results. These critical factors include: i) the ability to select relevant parameters to characterize data, ii) comprehending data coverage across various subjects, and iii) recognizing the significance of coverage for subject-specific characteristics. Typical parameters used to describe sleep data encompass factors like age, body mass index, gender, weight, height, and sleep study result parameters like sleep apnea events per hour (apnea-hypopnea, AHI), sleep profile across different sleep stages and total sleep time [3].

The underlying concept of the project is to utilize these parameters to extract research, training, and testing datasets from the amassed data to draw exploratory conclusions. To ensure efficiency and prevent unfruitful outcomes, a thorough examination of the data is imperative. Furthermore, understanding the information content of the data is crucial as it may lack statistical significance. This process also facilitates the

identification of dependent and independent parameters. For instance, if we collect a lot of data across gender, age, and people suffering from obstructive sleep apnea, the participant with apnea might mostly fall onto a certain gender or age, creating a possible bias in studies. Recognizing the significance of parameter descriptions could potentially optimize the creation of informative training and testing datasets. Moreover, the establishment of a coverage measure enables the identification of data not included in the training and test datasets, directing such data toward human-in-the-loop processing in real-world applications. As the raw data is reprocessed to mine additional parameters, new data types are integrated into the database. However, it's important to note that the database's purpose is not to directly address research questions; rather, it aims to locate relevant sleep studies for reprocessing in order to address research inquiries. Consequently, criteria for additional database columns should prioritize searchability within the database, rather than attempting to answer all conceivable future inquiries.

The Sleep Revolution project's main users encompass researchers from diverse fields like psychology, sports science, and machine learning [8]. Providing these researchers with a clear understanding of the collected data holds merit, aiding their future project decisions and understanding of data limitations. In clinical research, the use of inclusion and exclusion criteria is standard practice to align data samples with research goals, and this can be facilitated through data overviews and filtering. Besides assisting researchers in removing irrelevant data, filtering can also prove valuable in acquiring pertinent insights for both ongoing and future projects. In clinical research, diversity is pivotal for result credibility, necessitating broad age and gender representation. With the project spanning multiple disciplines, various relevant data types are present. Thus, study coordinators can employ data filtering and overviews to ensure participants represent a diverse array of data types, thereby enhancing overall data distribution.

3.2 Scalability Test to Technically Evaluate the Architecture

Every digital platform needs a performance evaluation to understand its limitations. Since the most time-consuming functionality of the architecture is filtering the data, the objective of the tests is to see how long it takes to filter an increasing amount of data. It is expected that the MySQL database can handle the maximum number of filters of the maximum amount of expected data within a few seconds [29]. We decided to start at 100 rows with 50 columns of simulated data and measure the speed. In each iteration we added 100 rows of data and measured the time. We decided to stop the test at 1,000 iterations where the final dataset contains 100,000 rows of data which is well beyond the expected number of measurements. The scalability test was conducted twice: once with a database storing text exclusively, and once with the database containing only numbers. For text inputs, the filter utilized was "%z%z%" (featuring two "z" characters). The simulated string input data followed the format "column_x_row_y," where "x" and "y" were replaced by the respective row and column numbers. Conversely, the number filter focused on identifying values between 49.0 and 50.0. The simulated numerical data encompassed random numbers ranging from 0.0 to 99.9, with no values falling between 49 and 50. Both the string and number filters employed the "OR"

operator among column filters to ensure that all 50 columns underwent the filtering process for each row.

3.3 User Experience and Usability Test to Evaluate the Architecture

We conducted an usability and user experience test to assess the digital platform's functionality, with the participation of ten potential end-users. The primary aim of this evaluation is to determine the effectiveness of a user interface featuring filterable results, specifically in supporting end-users' tasks. To ensure a controlled testing environment, we opted for simulated data as opposed to actual data, as certain participants had limited access to the full set of extracted results. The test comprised a filterable user interface containing 100 simulated rows and 10 columns, replicating real-world data parameters. Among these parameters, we selected seven sleep-related result parameters and three body measurement parameters that were handpicked for their relevance to the project. Our design featured a user-friendly filter interface where participants could input values, as illustrated in figure 2. We delivered a brief introduction to the system's features, explaining its functionalities to the participants. Their task encompassed responding to four specific questions that required them to use the filters and the displayed table. These questions were as follows: i) How many subjects in the database are male, and how many are female? ii) Is the severity of Apnea-Hypopnea Index (AHI) higher for individuals aged 40 and above, or for those under 40? iii) How many male participants in the database exhibit a severe AHI score of 30 or higher? iv) What are the average weights for males and females with a moderate apnea (AHI 15 to 30)? Participants recorded their answers for each question using Google Forms. Additionally, we measured the time taken for each task to gauge its level of complexity. Figures 2 and 3 provide visual representations of the table and the filter interface.

Number of results: 27

Table for Average

recording_length	recording_id	recording_type	recording_AHI	recording_AI	recording_HI	subject_height	subject_weight	subject_age	subject_gender
NA	NA	NA	31.6	15.85	15.75	174.2	74.57	41.56	NA

Queried Database

recording_length	recording_id	recording_type	recording_AHI	recording_AI	recording_HI	subject_height	subject_weight	subject_age	subject_gender
5:24:39	10162	PSG	26.75	23.78	2.96	184.14	79.53	32	Male
8:01:45	10165	PSG	35.53	15.89	19.64	183.96	80.32	41	Male
6:02:05	10189	PSG	30.52	15.75	14.76	172.43	72.67	42	Female
6:11:20	10191	PSG	28.88	11.65	17.23	169.27	77.3	49	Male
6:54:37	10212	PSG	27.44	23.33	4.11	172.9	74.23	49	Male
8:52:23	10226	PSG	25.74	14.67	11.07	185.03	64.28	31	Male

Fig. 2. Tables within the user interface employed by participants during the usability test. The "Queried Database" table showcases rows that meet specified filtering criteria, while the "Average Table" presents the calculated averages of the queried database beneath. The tables' upper sections display the total count of database entries that satisfy the applied filters.

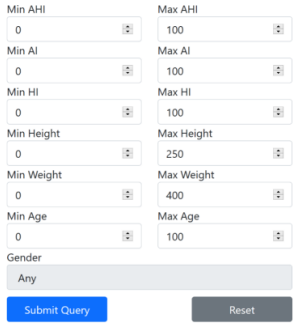


Fig. 3. The accessible filters within the user interface employed for the usability test. The showcased example filters utilized in the usability test encompass six numerical filters and one selection filter.

Upon completing their tasks, participants engaged in a reflective process concerning their system experience. This involved responding to two user experience questionnaires, namely the System Usability Scale [30] and AttrakDiff [31], both administered through Google Forms. Following the completion of the questionnaires, a semi-structured interview lasting approximately 20 minutes was conducted. The interview primarily focused on aspects that participants found unfavorable in the system.

4 Results

4.1 Architectural Constraints

In the database management system designed for extracted data, a relational database was chosen to accommodate information from over 10,000 sleep studies and provide search and filtering capabilities. Out of the discussed options, MySQL was preferred due to its compatibility and documentation support. For the development of the database-driven information extraction system, software architecture was pivotal. Layered-modular architecture was selected due to its adaptability across applications and systems [25]. It efficiently presents, reads, and filters data, making it suitable for various platforms, such as websites, desktop, or mobile applications. Other architectures like microkernel and modular, although excelling in certain aspects, lacked the agility required for this project's design.

The architecture facilitates information extraction from diverse data sources, transforming them into relevant parameters. Extractors process and standardize data formats, generating result sets that provide comprehensive insights. A critical aspect involves generating descriptive results addressing key questions. An example is the custom extractor designed for Noxturnal (sleep scoring software, Nox Medical, Reykjavík, Iceland) data within the Sleep Revolution project. However, accommodating various data sources like smartwatches and sleep diaries requires independent components within

the architecture. Format validation ensures extracted data aligns with the architecture's database format. The relationship among data sources, extractors, and the format validator is depicted in Figure 1.

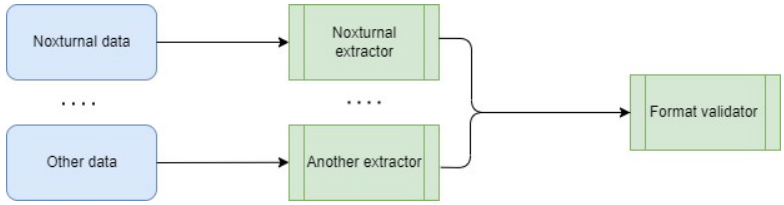


Fig. 1. The sequential stages of data extraction followed by validation. For every distinct data source, a dedicated extraction component is created. However, these components collectively utilize a common format validation module to ensure the data's adaptability.

4.2 Layered-modular Architecture

The digital platform encompasses the extraction of information from various sources, storage of this information in a database, and the presentation of the resultant database in an interactive manner. Additionally, a layered-modular architecture was initially developed for the digital platform to ensure its flexibility. However, due to the platform's division into two separate subsystems – data extraction and filtering/display – two distinct layered-modular architectures were required. These layered-modular architectures are visualized in Figure 4.

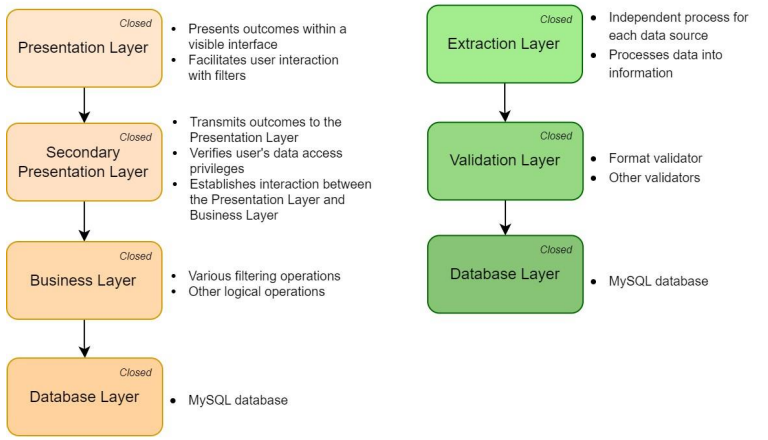


Fig. 4. The left side depicts the layered-modular architecture of the data filtering and display system. On the right, we demonstrate the layered-modular architecture for extracting information from a data source.

In the architecture, the presentation layer is responsible for creating the visual representation of data. While different applications can possess distinct presentation layers, they all access a shared secondary presentation layer. This secondary presentation layer comprises functions employed by various components within the presentation layer. For instance, it houses a function to convert filtered results into a JSON (JavaScript Object Notation) format, which may be preferred by web or mobile applications. Consequently, the decision to place a new function within the presentation layer or the secondary presentation layer depends on its potential for reuse. Concurrently, the business layer encompasses a collection of logical functions. The business layer's reusability is crucial, as new software or external data might interface directly with it rather than through the secondary presentation layer. The decision of whether a functionality belongs in the business layer or the secondary presentation layer is entirely contingent on the specific role that the function plays within the system.

4.3 Scalability

To evaluate the scalability of the digital platform, we found it essential to assess its performance. Given that the backend of the platform revolves around filters, a deliberate choice was made to test its capabilities with both floating-point and string entries. The tests were conducted according to the following specifications; Programming language: Python, Development environment: PyCharm, Operating system: Linux, Motherboard: Z270X-Gaming K5, Processor: Intel(R) Core(™) i7-7700K CPU @ 4.20GHz, 4201 Mhz, 4 Cores, 8 Logical Processors, Graphics card: NVIDIA GeForce RTX 2080 TI. Figure 5 shows how the scalability assessment of using filters for strings and floats.

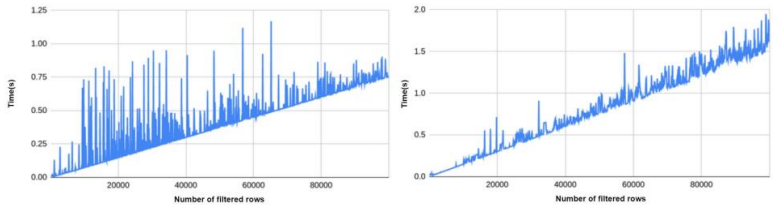


Fig. 5. The scalability assessment of filtering for strings (left) and numbers (right). In both evaluations, there were 50 columns, and the testing involved measuring the filtering time for each input iteratively. The iteration range extended from 100 to 100,000, increasing in increments of 100.

The statistical analysis demonstrates a linear rise in the processing time for both types of data. Each scalability test concluded at 100,000 rows with 50 columns, resulting in a total of five million inputs tested in the final iteration. The processing time was slightly above 1.5 seconds for numbers and around 0.75 seconds for strings.

4.4 Usability and User Experience

Ten participants tested the usability and user experience where they: i) performed four timed tasks, ii) answered the System Usability Scale questionnaire, iii) answered the AttrakDiff questionnaire, and iv) participated in a semi-structured interview. The participants included six members of the Sleep Revolution project, two sleep experts, and two novice users familiar with research and databases. Eight participants completed the test in person, while two participants took the test online using screen-sharing features on Discord and Messenger. All participants accurately completed the four tasks, except for one instance where the participant confused the AHI and AI parameters. The average completion time for each task was less than 30 seconds, apart from the final task taking approximately 45 seconds.

The System Usability Scale comprises 10 questions with five response options, ranging from 1 (strongly disagree) to 5 (strongly agree) [30]. Unfortunately, due to an error in Google Forms, one of the 10 questions was omitted. This missing question pertained to whether the user identified any irregularities in the system. The results are presented in Table 1.

Table 1. The outcomes of the System Usability Scale questionnaire, presenting each question along with the average rating and standard deviation assigned by the participants. The ratings reflect the extent to which participants agreed with the statements, ranging from one (strongly disagree) to five (strongly agree).

Question	Average Rating	Standard Deviation
I think I would like to use the system frequently	4.1	0.8
I found the system to be simple	4.7	0.5
I thought the system was easy to use	4.8	0.4
I think that I can use the system without the support of a technical person	4.9	0.3
I found the various function in the website were well integrated	4.2	0.6
I would imagine that most people would learn to use the system very quickly	4.5	0.5
I found the system very intuitive	4.1	0.7
I felt very confident using the system	4.3	0.6
I could use the system without having to learn anything new	4.6	0.5

The results suggest that participants generally viewed the system as easy to use and showed a willingness to use it in the future. Overall, the usability score was 90 on the System Usability Scale, surpassing the threshold of 68 that signifies high usability.

The AttrakDiff questionnaire comprises 25 questions, each featuring a pair of opposite words. Participants are presented with 7 radio buttons for each question they use like a slider to which word is the closer to their experience.

Table 2. The results of all 25 AttrakDiff questions in the Usability Test, their average score, and their standard deviation. The answers had seven radio button options where 1 represents the first word and 7 represents the latter word.

Question	Average Rating	Standard Deviation
Human (1) or technical (7)	4.0	1.2
Isolating (1) or connective (7)	4.4	1.0
Pleasant (1) or unpleasant (7)	2,1	0.8
Inventive (1) or conventional (7)	4.3	1.8
Simple (1) or complicated (7)	1.6	0.7
Professional (1) or unprofessional (7)	1.7	0.9
Ugly (1) or attractive (7)	4.9	1.4
Practical (1) or impractical (7)	1.7	0.5
Likeable (1) or disagreeable (7)	1.6	0.7
Cumbersome (1) or straightforward (7)	6.3	0.6
Stylish (1) or tacky (7)	3.7	1.3
Predictable (1) or unpredictable (7)	1.9	1.2
Cheap (1) or premium (7)	4.5	1.0
Alienating (1) or integrating (7)	6.0	0.9
Brings me closer to people (1) or separates me from people (7)	3.5	1.0
Unpresentable (1) or presentable (7)	5.6	0.9
Rejecting (1) or inviting (7)	5.5	1.1
Unimaginative (1) or creative (7)	4.5	1.1
Good (1) or bad (7)	2.1	1.1
Innovative (1) or conservative (7)	2.1	1.6
Dull (1) or captivating (7)	4.5	1.3
Undemanding (1) or challenging (7)	2.4	1.0
Motivating (1) or discouraging (7)	3.2	1.2
Novel (1) or ordinary (7)	4.0	1.3
Unruly (1) or manageable (7)	6.4	0.7

Some participants found it challenging to answer certain questions due to multiple possible interpretations or questions that did not directly align with their experiences. Despite this, most answers leaned towards the more positive option out of the two available for each question.

In the semi-structured interviews, we focused on asking the participants what was missing and what would need to be added for them to be able to use it for their own tasks. Most participants expressed the need for a system with an overview and navigation, which was already a cumbersome and time-consuming task for them. However, it

was very unison for the participants, stating they would need both more data sources, parameters, and functionalities, before they would start using it.

5 Discussion

Our world is becoming increasingly information-centric, collecting a myriad of data types and expanding volumes across society. The rise of data-driven research, especially in healthcare settings, necessitates the development of systems capable of extracting meaningful insights from abundant and diverse data [2–4].

Our evaluation primarily concentrated on the backend of the digital platform, utilizing a filterable database and an extraction mechanism. Our approach employs fully structured data due to its performance advantages, albeit at the expense of some flexibility [11]. This choice aligns with the importance of performance and usability in our context. Furthermore, unstructured data, like raw sleep measurements, poses storage and navigation challenges [11].

Our findings illustrate the applicability of layered-modular architecture in designing complex digital platforms suitable for socio-technical and multidisciplinary scenarios. The layered structure enhances adaptability and integrates validation layers, facilitating the incorporation of new features while maintaining system credibility. Usability tests and performance evaluations confirm the expandability, utility, and scalability of the architecture, highlighting its strong usability and user experience aspects. Insights from the semi-structured interview prompted various enhancements, including additional data types, more engaging visualization options, and improved interactions.

Prior research on platforms has explored various facets of platform dynamics [32]. Some studies have delved into platform architecture [33, 34], platform economics [35], leadership within platforms [36], and the generative nature of digital platforms [37]. Additionally, digital platforms have found applications across diverse industries, including finance [38], healthcare [39, 40], transportation [41], and electricity distribution [42]. A recent wave of research has emerged concerning data-driven digital health platforms [43–45], resulting in an expanding literature encompassing different application domains and scopes [46]. Nonetheless, scant attention has been directed toward identifying suitable architectures for intricate research projects that are both multidisciplinary and socio-technical in nature.

Furthermore, we acknowledge the limitations of our evaluation. The architecture's testing was confined to a subset of data sources and volumes, and its performance with larger datasets requires further exploration. Scalability tests demonstrated efficient filtering of entries, yet participant feedback indicated potential resource-intensive functionality gaps. A more comprehensive user-centric assessment is necessary to thoroughly gauge scalability and potential bottlenecks. While the AttrakDiff and System Usability Scale surveys affirmed the platform's positive aspects, feedback from semi-structured interviews introduced concerns about missing functionalities and dataset limitations, influencing usability perceptions. Moreover, the predefined tasks might not accurately reflect real-world scenarios, potentially leading to differing usability insights. In summary, the design, development, and utilization of data-driven

digital health platforms significantly rely on architectural choices. Our exploration of layered-modular architecture underscores its value in managing intricate data requirements. However, further research is essential to fully unravel its real-world usability implications.

6 Conclusion

In conclusion, the designed layered-modular architecture holds substantial promise. This architecture, defined by distinct code layers or packages, has been successfully applied in our study. The results confirm that our layered-modular approach establishes a resilient framework, crucial in complex multidisciplinary environments. Creating data-driven digital health platforms hinges on factors like robust data availability, diverse filtration capabilities and responsive hosting. Our architecture's validation via usability tests, surveys, and interviews demonstrates its competence in effectively handling data. Furthermore, its linear scalability makes it adaptable for larger datasets. However, the architecture's application to more extensive datasets remains unexplored, and pre-determined tasks in usability tests could introduce bias. While promising, further research is needed for real-world usability insights. From these findings, three design principles emerge: First, the design and development of digital platforms should center around components with singular purposes, fostering easier replacement, updates, or removal of elements in the future. Second, the integration of validation layers simplifies additions to the platform, enhances result credibility, scalability, and reliability. Lastly, the utilization of usability tests in tandem with semi-structured interviews emerges as a valuable methodology for evaluating products, identifying user challenges, and gaining insights into prospective enhancements and expansions. Addressing our research question, we successfully created an architecture for extracting sleep data, optimizing modularity and enhancing usability. This layered-modular approach significantly improves data presentation and filtering, ensuring usability, scalability, and performance.

References

1. Holst, A.: Amount of information globally 2010-2024. Statista. Erişim Adresi: https://www. statista. com/statistics/871513/worldwide-datacreated/ (2020).
2. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health information science and systems. 2, 1–10 (2014).
3. Pevernagie, D.A., Gnidovec-Strazisar, B., Grote, L., Heinzer, R., McNicholas, W.T., Penzel, T., Randerath, W., Schiza, S., Verbraecken, J., Arnardottir, E.S.: On the rise and fall of the apnea−hypopnea index: A historical review and critical appraisal. Journal of Sleep Research. 29, e13066 (2020). https://doi.org/10.1111/jsr.13066.
4. Arnardottir, E.S., Islind, A.S., Óskarsdóttir, M.: The Future of Sleep Measurements: A Review and Perspective. Sleep Medicine Clinics. 16, 447–464 (2021). https://doi.org/10.1016/j.jsmc.2021.05.004.

5. Mehra, R., Stone, K.L., Varosy, P.D., Hoffman, A.R., Marcus, G.M., Blackwell, T., Ibrahim, O.A., Salem, R., Redline, S.: Nocturnal Arrhythmias Across a Spectrum of Obstructive and Central Sleep-Disordered Breathing in Older Men: Outcomes of Sleep Disorders in Older Men (MrOS Sleep) Study. Archives of Internal Medicine. 169, 1147–1155 (2009). https://doi.org/10.1001/archinternmed.2009.138.
6. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., Jagadish, H.V., Labrinidis, A., Madden, S., Papakonstantinou, Y., Patel, J., Ramakrishnan, R., Ross, K., Shahabi, C., Suciu, D., Vaithyanathan, S., Widom, J.: Challenges and Opportunities with Big Data 2011-1. Cyber Center Technical Reports. (2011).
7. Bellinger, G., Castro, D., Mills, A.: Data, information, knowledge, and wisdom, (2004).
8. Arnardottir, E.S., Islind, A.S., Óskarsdóttir, M., Ólafsdóttir, K.A., August, E., Jónasdóttir, L., Hrubos-Strøm, H., Saavedra, J.M., Grote, L., Hedner, J., Höskuldsson, S., Ágústsson, J.S., Jóhannsdóttir, K.R., McNicholas, W.T., Pevernagie, D., Sund, R., Töyräs, J., Leppänen, T., Revolution, S.: The Sleep Revolution project: the concept and objectives. Journal of Sleep Research. 31, e13630 (2022). https://doi.org/10.1111/jsr.13630.
9. Harrell, M.C., Bradley, M.: Data collection methods: Semi-structured interviews and focus groups. (2009).
10. Duval, E., Hodgins, W., Sutton, S., Weibel, S.L.: Metadata principles and practicalities. D-lib Magazine. 8, 1–10 (2002).
11. Sint, R., Schaffert, S., Stroka, S., Ferstl, R.: Combining unstructured, fully structured and semi-structured information in semantic wikis. In: CEUR Workshop Proceedings. pp. 73–87. Citeseer (2009).
12. Buneman, P.: Semistructured data. In: Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. pp. 117–121 (1997).
13. Yoo, Y., Henfridsson, O., Lyytinen, K.: Research Commentary—The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research. Information Systems Research. 21, 724–735 (2010). https://doi.org/10.1287/isre.1100.0322.
14. DB-ENGINES: DB-Engines Ranking of Relational DBMS, https://db-engines.com/en/ranking/relational+dbms, last accessed 2021/03/25.
15. Cyran, M., Lane, P., Polk, J., Cheevers, S., Colrain, C., Goorah, V., Hartstein, M., Haydu, J., Hu, W., Krishnan, R., others: Oracle Database Concepts, 10g Release 2 (10.2) B14220-02. (2005).
16. Widenius, M., Axmark, D.: MySQL reference manual: documentation from the source. O'Reilly Media, Inc. (2002).
17. Gorman, K., Hirt, A., Noderer, D., Pearson, M., Rowland-Jones, J., Ryan, D., Sirpal, A., Woody, B.: Introducing Microsoft SQL Server 2019: Reliability, scalability, and security both on premises and in the cloud. Packt Publishing Ltd (2020).
18. PostgreSQL: PostgreSQL 13.2 Documentation, https://www.postgresql.org/docs/current/, last accessed 2021/04/25.
19. IBM: IBM Db2 – Data Management Software, https://www.ibm.com/analytics/db2, last accessed 2021/04/21.
20. SQLite: SQLite, https://www.sqlite.org/about.html, last accessed 2021/04/25.
21. Coren, S.: Sleep deprivation, psychosis and mental efficiency. Psychiatric Times. 15, 1–3 (1998).

22. Chokroverty, S.: Overview of sleep & sleep disorders. Indian Journal of Medical Research. 131, 126 (2010).
23. Epstein, J.B., Rea, G., Wong, F.L., Spinelli, J., Stevenson-Moore, P.: Osteonecrosis: study of the relationship of dental extractions in patients receiving radiotherapy. Head Neck Surg. 10, 48–54 (1987). https://doi.org/10.1002/hed.2890100108.
24. Rundo, J.V., Downey, R.: Chapter 25 - Polysomnography. In: Levin, K.H. and Chauvel, P. (eds.) Handbook of Clinical Neurology. pp. 381–392. Elsevier (2019). https://doi.org/10.1016/B978-0-444-64032-1.00025-4.
25. Richards, M.: Event-driven architecture. Software Architecture Patterns; O'REILLY: Newton, MA, USA. 18–19 (2015).
26. Simon, H.A.: The Sciences of the Artificial, reissue of the third edition with a new introduction by John Laird. MIT press (2019).
27. Gubser, R., Schulte-Althoff, M., Heinemann, N., Pohle, J., Fürstenau, D.: DATA GOVERNANCE STRATEGIES FOR DATA PLATFORMS – A MULTIPLE CASE STUDY IN NURSING CARE. ECIS 2023 Research-in-Progress Papers. (2023).
28. Sein, M.K., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R.: Action Design Research. MIS Quarterly. 35, 37–56 (2011). https://doi.org/10.2307/23043488.
29. Schwartz, B., Zaitsev, P., Tkachenko, V.: High Performance MySQL: Optimization, Backups, and Replication. O'Reilly Media, Inc. (2012).
30. Bangor, A., Kortum, P.T., Miller, J.T.: An Empirical Evaluation of the System Usability Scale. International Journal of Human–Computer Interaction. 24, 574–594 (2008). https://doi.org/10.1080/10447310802205776.
31. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Szwillus, G. and Ziegler, J. (eds.) Mensch & Computer 2003: Interaktion in Bewegung. pp. 187–196. Vieweg+Teubner Verlag, Wiesbaden (2003). https://doi.org/10.1007/978-3-322-80058-9_19.
32. de Reuver, M., Sørensen, C., Basole, R.C.: The digital platform: a research agenda. J Inf Technol. 33, 124–135 (2018). https://doi.org/10.1057/s41265-016-0033-3.
33. Gawer, A.: Platforms, Markets and Innovation. Edward Elgar Publishing (2011).
34. Tiwana, A., Konsynski, B., Bush, A.A.: Research Commentary—Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics. Information Systems Research. 21, 675–687 (2010). https://doi.org/10.1287/isre.1100.0323.
35. GAWER, A.R., Srnicek, N.: Online platforms: Economic and societal effects. Panel for the Future of Science and Technology (STOA) European Parliament (2021).
36. Leong, C., Pan, S., Leidner, D., Huang, J.-S.: Platform Leadership: Managing Boundaries for the Network Growth of Digital Platforms. Journal of the Association for Information Systems. 20, (2019). https://doi.org/10.17705/1jais.00577.
37. Fürstenau, D., Baiyere, A., Schewina, K., Schulte-Althoff, M., Rothe, H.: Extended Generativity Theory on Digital Platforms. Information Systems Research. (2023). https://doi.org/10.1287/isre.2023.1209.
38. de Reuver, M., Verschuur, E., Nikayin, F., Cerpa, N., Bouwman, H.: Collective action for mobile payment platforms: A case study on collaboration issues between banks and telecom operators. Electronic Commerce Research and Applications. 14, 331–344 (2015). https://doi.org/10.1016/j.elerap.2014.08.004.

39. Bahmani, A., Alavi, A., Buergel, T., Upadhyayula, S., Wang, Q., Ananthakrishnan, S.K., Alavi, A., Celis, D., Gillespie, D., Young, G., Xing, Z., Nguyen, M.H.H., Haque, A., Mathur, A., Payne, J., Mazaheri, G., Li, J.K., Kotipalli, P., Liao, L., Bhasin, R., Cha, K., Rolnik, B., Celli, A., Dagan-Rosenfeld, O., Higgs, E., Zhou, W., Berry, C.L., Van Winkle, K.G., Contrepois, K., Ray, U., Bettinger, K., Datta, S., Li, X., Snyder, M.P.: A scalable, secure, and interoperable platform for deep data-driven health management. Nat Commun. 12, 5757 (2021). https://doi.org/10.1038/s41467-021-26040-1.
40. Islind, A.S.: Platformization : Co-Designing Digital Platforms in Practice. (2018).
41. Svahn, F., Lindgren, R., Mathiassen, L.: Applying Options Thinking to Shape Generativity in Digital Innovation: An Action Research into Connected Cars. In: 2015 48th Hawaii International Conference on System Sciences. pp. 4141–4150 (2015). https://doi.org/10.1109/HICSS.2015.497.
42. Kiesling, L.L.: Implications of Smart Grid Innovation for Organizational Models in Electricity Distribution, https://papers.ssrn.com/abstract=2571251, (2015).
43. Sigurðardóttir, S.G., Islind, A.S., Óskarsdóttir, M.: Collecting data from a mobile app and a smartwatch supports treatment of Schizophrenia and bipolar disorder. In: Challenges of Trustable AI and Added-Value on Health. pp. 239–243. IOS Press (2022).
44. Sveinbjarnarson, B.F., Schmitz, L., Arnardottir, E.S., Islind, A.S.: The Sleep Revolution Platform: a Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data. Current Sleep Medicine Reports. (2023). https://doi.org/10.1007/s40675-023-00252-x.
45. Schmitz, L., Sveinbjarnarson, B., Gunnarsson, G., Davíðsson, Ó., Davíðsson, Þ., Arnardottir, E., Óskarsdóttir, M., Islind, A.: Towards a Digital Sleep Diary Standard. (2022).
46. Islind, A.S., Lindroth, T., Snis, U.L., Sørensen, C.: Co-creation and Fine-Tuning of Boundary Resources in Small-Scale Platformization. In: Lundh Snis, U. (ed.) Nordic Contributions in IS Research. pp. 149–162. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-43597-8_11.

Appendix B

Publication II

Current Sleep Medicine Reports (2023) 9:91–100
https://doi.org/10.1007/s40675-023-00252-x

The Sleep Revolution Platform: a Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data

Bjarki Freyr Sveinbjarnarson[1,2] · Lisa Schmitz[1,2] · Erna Sif Arnardottir[2] · Anna Sigridur Islind[1,2]

Accepted: 9 March 2023 / Published online: 10 April 2023

Abstract

Purpose of Review The complexity of the data collected for sleep research is increasing, and the focal point of sleep research is dependent on a higher number of data sources. Data collected for sleep studies often includes both subjective and objective measurements of sleep quality and is gathered over a more extended period, e.g., for weeks, months, or even years. However, this variety and volume of data make it challenging and time-consuming for researchers to utilize. Therefore, sophisticated data structures are necessary to utilize data in sleep research.

Recent Findings This paper explores how heterogeneous data sources can be represented in a homogeneous database design. The following research questions drove our work: (i) How can we represent sleep data from heterogeneous sources in a homogenous digital platform database? and (ii) How can a data source pipeline transform various data sources into a homogeneous data format?

Summary This paper's main contributions are conceptualizing the design and development of a homogeneous database and digital platform architecture and a data source pipeline that fits well for sleep research in particular and healthcare research in general.

Keywords Sleep Revolution · Information Systems · Digital Platform Architecture · Database Design · Data Source Pipeline

Introduction

Sleep research and medicine involve data collection from a variety of sources [1••]. The gold standard for measuring and assessing more complex sleep disorders such as sleep-disordered breathing is polysomnography (PSG), a method that traditionally involves collecting processed signals from connected sensors [2, 3]. Driven by a growing need for longitudinal sleep assessment, both for research, clinical assessment, and treatment response, additional methods to collect subject and objective sleep data have been developed such as digital symptom trackers and wearables as well as treatment compliance assessments [4]. Moreover, signal formats generally differ between devices and manufacturers, contributing further to the data's variety.

Researching and improving longitudinal sleep assessments is one of the goals of the Sleep Revolution [1••]. The project has multiple cornerstones in different research fields, with the aim of collecting data in multiple prospective studies and processing retrospective data from thousands of completed sleep studies from different sources. Due to the quantity and complexity of the data sources, the collected data must be translated into a homogeneous data format to ensure that researchers can work efficiently with the data. Therefore, it is vital to design and develop a digital platform with a novel database architecture to store, process, and combine the various data sources for the Sleep Revolution [4].

The growing complexity of modern digital platforms due to an increase in data variety and the resulting challenges are not unique to sleep research and medicine [5•, 6]. To address these challenges, researchers have focused on advancing digital platform design by providing architecture guidelines on an abstract level, such as the layered modular architecture [7•]. While such high-level concepts

✉ Bjarki Freyr Sveinbjarnarson
bjarkis@ru.is

[1] Department of Computer Science, Reykjavik University, Reykjavik, Iceland

[2] Reykjavik University, Sleep Institute, Reykjavik, Iceland

enable digital platform designers to organize the complete digital platform ecosystem, De Reuver et al. emphasize the need to expand the research on digital platform architecture on different levels [5•]. In this paper, we contribute to this call by focusing on the processing and organization of data found at the data level of digital platform architectures. We present a homogeneous digital platform and database design to represent the heterogeneous data sources and a data source pipeline design for processing the data sources into a homogeneous data format. Our contributions result from a detailed analysis and categorization of the data. The following research questions drove our work: (i) How can we represent sleep data from heterogeneous sources in a homogenous digital platform database? and (ii) How can a data source pipeline transform various data sources into a homogeneous data format? The main contributions of this paper are through the conceptualization of the design and development of a homogeneous database and digital platform architecture and a data source pipeline to convert and process heterogeneous data sources into a validated homogenous format that fits well for sleep research in particular and healthcare research in general. In this paper, the Sleep Revolution provides an illustrative example to show the flexibility and dynamic capacity of the digital platform design. Based on our illustrative example, we argue that this particular digital platform design can be generalized and utilized in other contexts as it provides a dynamic approach for modern data.

Related Work

Designing and developing an information system that combines and adapts various data sources for multiple end-user groups is challenging. During the design and development processes, it is vital to preserve digital platform characteristics that have been identified as essential features for a successful information system design.

In their editorial paper, Constantinides et al. highlight key insights on the architecture of digital platforms [8]. They identify abundant data collection due to the rise of machine learning methodology and the increased digitization of diverse processes, such as those in healthcare. Therefore, the way digital platforms are designed to facilitate abundant data collection is a key element going forward. This phenomenon leads to new challenges for platformization and creates a need for research about digital platform architecture and data organization. Both should fulfill the criteria of being "stable and evolving" [9] in order to design a robust and long-lived digital platform ecosystem.

Yoo et al. [7•] illustrate how product innovation through digital capabilities of traditionally analog products influences the requirements of platform architecture. One of the digital platform characteristics emphasized by this development is data homogenization. They argue that "unlike analog data, digital data originate from heterogenous sources and can be combined easily with other digital data to deliver diverse services" [7•]. Our paper follows this philosophy by applying it to the data structure and architecture of the SleepWell, the Sleep Revolution platform. Though collected through a variety of digital products and physical devices (mobile applications and sensors), we aim to bring the data together in a homogenous format. In a layered modular architecture, as introduced by Yoo et al., this feature contributes to the flexibility of a digital platform, keeping it open to the option of adding new digital elements to it, such as interfaces [7•].

The recombinability of digital elements is highlighted by the research conducted by De Reuver et al. [5•]. To ensure that multiple applications can connect to the interface of a digital platform, data presented through that interface must not be subject to constant changes to its format. Therefore, it is advantageous if data changes and new data points can be represented within the existing data structure of the digital platform.

In recent years, digital platforms have been emerging in the healthcare industry [10]. Driven by the increasing availability of sensors and wearables for the consumer market, the wide variety of data sources drives the development of healthcare applications. However, the issue of dealing with heterogeneous data from various sources within a unified platform ecosystem arises from this situation, and research is scarce on that topic. In this paper, we contribute to that gap in the literature. Bache et al. encountered this problem when defining an architecture to combine multiple heterogeneous data sources and query them efficiently [11]. Their solution focused on the development of an abstract, reusable query model. This query model hides the underlying structure of data and enables interfaces to connect to it efficiently [11]. A new data source can be added to their architecture with a lightweight adapter. Nevertheless, this adapter must be independently developed for each new data source. The notion of lightweight adapters is similar to the seminal paper by Bygstad, which argues for the architectural vision of lightweight versus heavyweight modules [12]. However, their paper is of conceptual nature and outside of healthcare, whereas our paper contributes with digital platform architecture within healthcare specifically and with a special user case.

Methods

Action Design Research

We aimed to design and develop a digital platform that could function as a bridge between healthcare professionals, researchers, and participants and include heterogeneous data in a homogenous architecture. As we see it, it is important

to have multiple feedback loops to ensure that the digital platform design and development are aligned with the needs of the different end-user groups. Because the Sleep Revolution project size calls for a large-scale SleepWell platform, in which the abovementioned researchers are from multiple disciplines alongside the participants (that have a wide variety of needs) and the healthcare professionals, from multiple sectors, the formulation of the digital platform requirements has been a complex, iterative process. Action design research (ADR) is a method that fits well for complex and iterative research projects. ADR has four scopes: (i) problem formulation, (ii) designing solutions, (iii) reflecting upon the solutions, and (iv) learning outcomes [13]. We created an ADR workflow that builds on the four scopes for our research project (Fig. 1). Through the design and development phase, interviews were conducted continuously with different end-users.

Sleep Revolution

The Sleep Revolution is a European Union Horizon 2020 project across multiple countries and different beneficiaries. Moreover, it is a multi-disciplinary consortium with a cornerstone in multiple fields, such as sleep medicine and research, computer science, biomedical science, psychology, engineering, and sports science. One of the major objectives of the Sleep Revolution is to transform the current diagnostic methods and treatment follow-up for sleep-disordered breathing [1••]. This objective utilizes the retrospective sleep study data pool of tens of thousands of sleep recordings and health information [1••]. Another major objective is to promote participatory healthcare with technological solutions, where it aims to design a digital platform to promote participatory healthcare used in numerous prospective studies [1••]. Coupled with that, an important step is to centralize a wide variety of sleep data from thousands of patients and research participants into unified data sets that can be accessed digitally through a novel digital platform [1••].

Data Sources

Due to the project's aforementioned magnitude, the data comes in a variety of different formats and has been collected with different devices; ergo, the data is heterogeneous. Therefore, unifying the heterogeneous data sources into a homogenous format while also representing the data feasibly in our digital platform is a significant challenge. Before designing and developing a digital platform architecture that combines multiple sleep data sources, we mapped out key data sources collected throughout our different Sleep Revolution projects.

Sleep Studies

During an overnight sleep study, a variety of signals are collected using sensors to measure changes in physiological states that occur while a person sleeps. The study includes channels to measure electroencephalography for brain wave activity, electrooculography to measure eye movements, and chin electromyography for muscle tone. Together, these channels allow for the assessment of different sleep stages and wake periods measured in 30 s epochs throughout the night and arousals from sleep. The sleep studies also include respiratory flow assessment to assess breathing, as well as respiratory movements via thorax and abdomen belts and blood oxygen levels and pulse via a pulse oximeter. Together, these measurements allow for

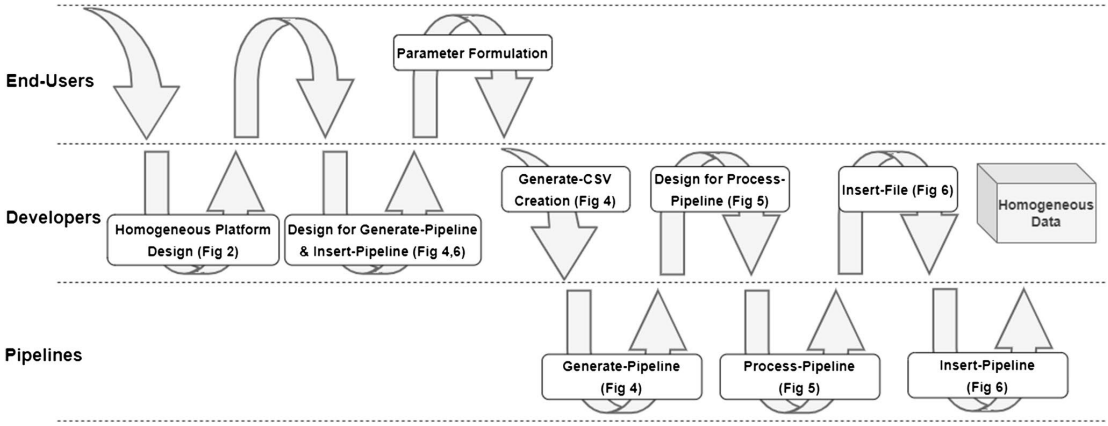


Fig. 1 An illustration of our action design research workflow for the designing of the digital platform design and the pipelines

the assessment of sleep apnea severity. Additionally, an electrocardiography, leg electromyography (for periodic leg movement assessment), body position and activity, and possibly synchronized video and audio (e.g., for snore measurement) are included. Therefore, during a sleep study, multiple sensors and devices are used to capture sleep objectively. The prospective data collection, processing, and device formatting are done in Sleep Revolution through Noxturnal (Nox Medical, Reykjavik, Iceland). Furthermore, the sleep studies are manually scored using Noxturnal as well. The results from the scored PSG are split into 16 different categories, e.g., "position activity" or "respiratory." The 16 categories represent over 1700 unique classifications of data. Examples of classification are "SleepTotalN3Duration," the sleep duration in the N3 sleep stage, and "SnoringTrainsPerHourSupine," the number of "snoring trains" per hour in a supine sleeping position. The high number of unique classifications is due to the high number of scored events, like apnea, hypopnea, snoring, and arousal, which can be further distinguished by, e.g., the sleep stage and the sleeping position. Therefore, it adds up to a multitude of data. The sleep studies are exported and include (i) raw signal files using a European Data Format (EDF) and (ii) parameter files in a semi-structured Extensible Markup Language (XML) format.

Wearables

The digital platform allows wearable solutions to be connected to it. One of the most used wearables in the Sleep Revolution research is the smartwatch Scanwatch (Withings, Paris, France). The Scanwatch has a photoplethysmography sensor and a 3-axis accelerometer, enabling it to track exercise, sleep, heart rate, and more. The smartwatch is connected to the user's phone via BlueTooth and gathers data from all connected Withings devices into a database that is accessible via an application programming interface (API). The SleepWell platform connects to the API to retrieve data, including exercise sessions, step counts, elevation, and sleep together with raw sensor data values. While other wearables can be integrated, the Withings watch is used in all prospective studies.

Questionnaires

Research Electronic Data Capture (REDCap) [14] is a secure web application used to manage the majority of questionnaires in the Sleep Revolution, forms for staff entry as well as informed consent by research participants. This web application allows for the creation of multiple types of questions, with a wide array of answering options, e.g., checkboxes, radio buttons, time fields, numerical input, and open-ended text answers. In the Sleep Revolution, over twenty different questionnaires have been created, sent out, and collected, with over 3000 responses from participants.

The questionnaires have two types of data sources: (i) the setup of the questionnaire, i.e., the choices of questions and types of answers, and (ii) the results of each question for each collected answer, for each participant. The exported output files from questionnaires are semi-structured comma-separated values (CSV) files. Moreover, the European Sleep Questionnaire (ESQ) is currently being digitally designed into the SleepWell platform, with access in 15 different languages [1••].

Digital Sleep Diaries

Sleep diaries are a valuable tool for gathering subjective data for providing an overview of people's sleep quality and habits over an extended period of time [15]. The Sleep Revolution designed and developed a mobile application (an app) with an adapted version of the Consensus Sleep Diary [15] with both a morning and evening sleep diary, also in 15 different languages. The Sleep Revolution app feeds the data directly to our digital platform architecture. The sleep diary within the app is used to collect longitudinal data on subjective sleep quality and habits over a period of 3 months [16].

Cognitive Tests

The Sleep Revolution app feeds data directly into SleepWell. The app also includes a cognitive battery to measure cognitive function over an extended period of time. A cognitive battery is a collection of different cognitive tasks done in a row where each cognitive task targets specific cognitive processes or domains, i.e., perceptual skills, processing speed, episodic memory, and reasoning. The cognitive tasks and batteries are used to document current cognitive ability as well as changes in cognitive ability over time in the Sleep Revolution project.

The Sleep Revolution currently uses three sources for cognitive tests: (i) an in-lab cognitive battery, (ii) an at-home cognitive battery completed directly within the SleepWell platform by the participant, and (iii) four cognitive tests that are included in the Sleep Diary app. The data format varies for the in-lab cognitive battery since some tasks make use of photos and other types of complex input. The in-lab cognitive battery was designed in the software Inquisit (Millisecond, Seattle, USA) and outputs a semi-structured CSV file. The latter two sources send their processed results using our API directly through our digital platform, to the database.

Other Data Sources

In addition to the data sources mentioned above, there is a wide variety of other data sources collected in Sleep Revolution. We further mapped these out in Table 1.

Each data source requires a unique preprocessing approach in order to be functional, depending on how it was collected and the original format. The difficulty of preprocessing depends on its given format, if it includes manual inputs, is reliant on other data sources, and other similar factors. To provide an overview of the challenges, we have mapped them out in Table 2.

Results

First, we present the homogenous database and digital platform design and thereafter present the data source pipeline. The conceptualization of the two parts outlines the main contribution of this paper.

Homogeneous Database and Digital Platform Design

We arrived at a digital platform design that is simple to use and understand, yet flexible and dynamic enough to incorporate the various data sources. The generalized design of the database, on which the digital platform rests, is represented in Fig. 2. A five-fold design is used as the core model to represent the data: (i) entry, (ii) form, (iii) entry-result, (iv) form-result, and (v) owner. An entry represents a data point, and a form is a collection of the entries, e.g., one entry can be one question in a form which is one questionnaire. Therefore, a data source can have multiple forms, like multiple questionnaires. The entry-result and form-result tables, on the other hand, store the individual answers and total answers to those questions and questionnaires, respectively. Therefore, every entry and form are only created once, but each entry-result and form-result can have none or multiple inputs.

One of the design strengths of this digital platform is its simplicity, since no additional tables are required to extend or add new data sources, independent of the data source. The design's simplicity makes it easy to apply, for example, a wide variety of data analysis types to the data in the project's next phase. Moreover, the simplicity allows for flexible and dynamic front-end development due to the consistency of the querying and the limited number of tables. This way, we can fit new data sources of heterogeneous quality into this dynamic design. Furthermore, the entry and entry-result tables make adding new data points to already existing forms effortless. That makes the design revolutionary and novel for research data collection since, for many data sources, it is common to add a new additional data point to a form retrospectively. Moreover, the flexible database design allows for holistic digital platform architecture. An additional novel feature is derived from the fact that the design allows exporting the data into a format or file that fits the different researcher end-users work environment or data analysis software, e.g., Python, R, or SPSS (Statistical Package

Table 1 Examples of the data sources and their format as collected in the Sleep Revolution

Source	Examples	Data format
Spreadsheets	Body measurements, complementary data, participant schedules, data ownerships, etc	XLSX, CSV, ODS
Audio recordings	Interviews, speech recordings, etc	Audio file formats
Video recordings	Sleep laboratory recordings	Video file formats
Mobile applications	Subjective data for chronic pain, menstrual cycle etc	JSON
Other wearable devices	Empatica smartwatch, Nukute collar, Dreem headband, etc	Multiple formats for each wearable

XLSX, Excel Microsoft Office Open XML Format Spreadsheet file; *CSV*, comma-separated values file; *ODS*, OpenDocument Spreadsheet; *JSON*, JavaScript Object Notation

Table 2 Examples of the data source categories collected in the Sleep Revolution as well as their environment's dependencies and constrictions

Category	Uses web-app, SDK, or software	Complementary resources	API	Continuous
Questionnaires	Some	Yes	No	No
Wearables	Yes	Yes	Some	Yes
Sleep studies	Yes	Yes	No	Yes
Cognitive tests	Yes	Some	Some	No
Physical measurements	No	Some	No	No

API, application program interface; *SDK*, software development kit

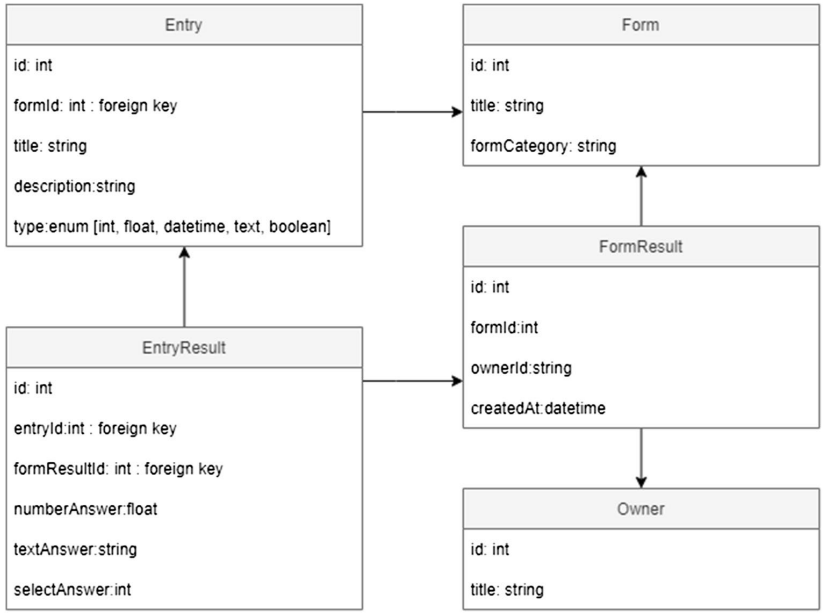


Fig. 2 The homogeneous digital platform design fits all the different data sources. Instead of adding new tables and columns for every data source, the five-fold core model adds them as inputs, therefore reducing the size and complexity of the database considerably

for the Social Sciences). The design uses the ownership table to connect the data sources, a crucial element for research, i.e., the ability to connect all data points within a research participant vs. between research participant for analysis.

Data Source Pipeline

The diverse data sources need to fit into the homogeneous digital platform design, which requires processing of every data source into a homogeneous format. Since the project has a great number of data sources, we designed a generic high-level diagram for the data source pipeline that fits the data sources (Fig. 3).

The *parameter formulation creation* step is about understanding the data source and choosing the relevant data points for end-users. The process commonly requires semi-structured interviews or collaboration between the developers, data collection team, and end-users. The selected data points are then split into one or more forms. Data sources can have multiple forms, for example, a questionnaire web application can have multiple questionnaires.

The *forms and entries creation* step consists of taking the decided parameters in the previous step and creating a comma-separated values (CSV) file for each form containing: (i) parameters, (ii) data type (integers [int], Boolean [true/false], date, etc.), and (iii) description of parameter. The step uses its own pipeline to create the form and entries (Fig. 4). The pipeline simplifies the creation of new forms and entries, only requiring a CSV containing entry names, data types, and optional descriptions. To ensure homogeneity between forms and entries, all data sources use the same

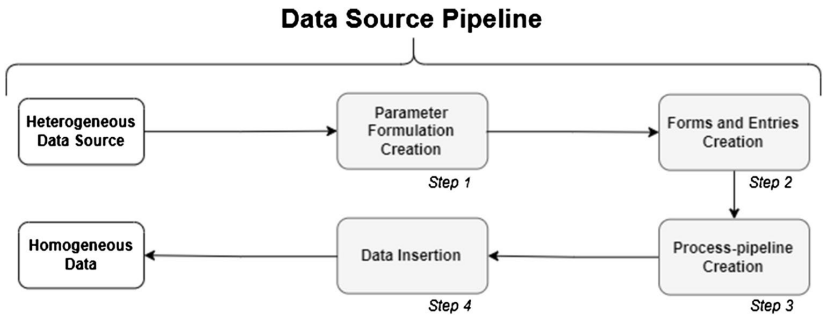


Fig. 3 The data source pipeline converts a heterogeneous data source into homogeneous data in four steps

script with a strict evaluation. The script uses the form's title, the form's category, and CSV as input to create the form and entries.

The *process-pipeline creation* step revolves around creating a pipeline to transform the unprocessed data source into a processed result data (Fig. 5) so that it can be automatically input (Fig. 6). The complexity of each sub-step in the process-pipeline depends on the data source environment, as stressed in the examples given in the "Data Sources" chapter. Therefore, each data source requires different approaches. In the process-pipeline, we illustrate six crucial steps to translate the data sources into a homogeneous format (Fig. 5). As shown in Table 2, some data sources rely on other software, web-apps, APIs, or SDKs for collecting the data. That makes it necessary to *export a selection of data* in the data source's given data format. The exported data can be in multiple files or in difficult-to-use formats, requiring scripts or manual work to *combine, transform, or convert the data* into the developer's preferred data formats. In this research context, the data sources require ownership, that is, which participant the data result belongs to. However, the software, digital platforms, apps, and other data sources often do not contain a feature to add ownership. Therefore, it is essential to use additional resources, such as spreadsheets, to *set ownership to each data point* and to ensure data integrity. The collected research data may contain duplicates and incomplete or unusable entries, making it necessary to *clean the data*. Moreover, data sources often rely on additional resources, manual work, or manual inputs, making it essential to review and *validate the data*. After the data points have been preprocessed, cleaned, validated, and added ownership onto, it is necessary to *convert the data to input-data* that is accepted by the insert-pipeline (see Fig. 6).

The *data insertion* step is about using the insert-pipeline (Fig. 6) to transform the processed result data into form results and entry results. The insert-pipeline outlines a collection of pipelines to insert processed data results into the

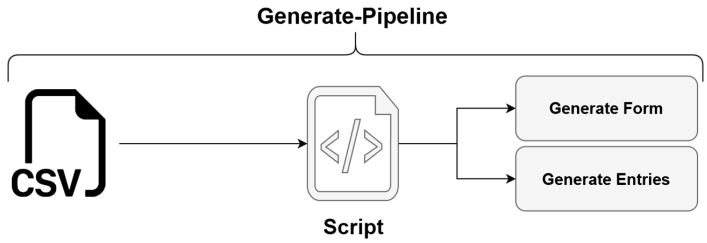


Fig. 4 The generate-pipeline takes comma-separated values (CSV) as input to create forms and individual entries for data sources

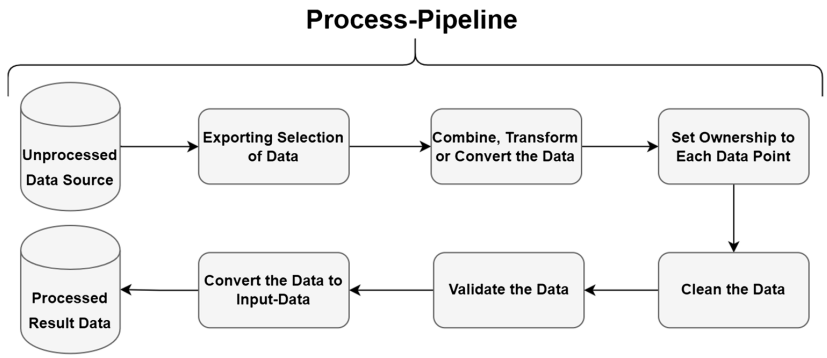


Fig. 5 The process-pipeline consists of six crucial steps to convert unprocessed data source into processed result data

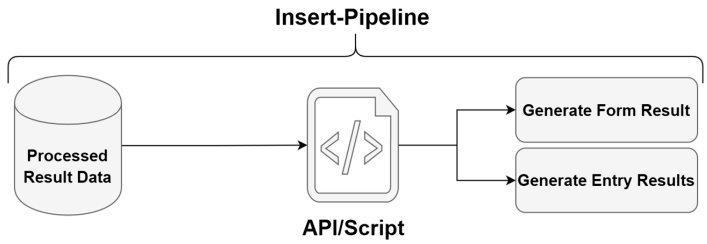


Fig. 6 The insert-pipeline converts processed result data into form results and entry results using a script or an application programming interface (API)

database, therefore connecting the data source pipeline to the homogeneous database and digital platform architecture. Some data sources consist of local files, whereas others are only accessible using an API (Table 2). In those cases, it is necessary to create its own insert-pipeline.

Discussion

In this paper, we present the design and development of a database and digital platform architecture alongside a data source pipeline to streamline data for sleep research [4]. The homogenous database and digital platform architecture, with the addition of the data source pipeline, effectively combines heterogeneous sleep data from diverse sources into an abstract and dynamic homogeneous representation. These characteristics contribute to the digital platform's flexibility called for by Yoo et al. (2010) [7•]. Due to the adaptable data source pipeline designs, new data sources and data points can be easily added in the future by preprocessing and translating them into an accepted data format. This way, the focal point was to ensure the preservation of essential digital platform design characteristics such as reprogrammability, modularity, and homogenous data representation [5•, 8]. The designed and developed homogeneous database and digital platform design is abstract on a technical level and, therefore, widely applicable in architectural terms to other digital platform ecosystems. Thus, we address the call for novel digital platform architecture design [5•] and offer our main contribution through the conceptualization of a dynamic, homogeneous database and digital platform architecture on the one hand, and the data source pipeline to cope with heterogeneous data on the other hand to the literature. Our paper presents the specific case of sleep data as used in the Sleep Revolution, while we would like to argue that the architecture is generic and can be generalized to different healthcare research and might even fit the purpose of research in general.

Our database and digital platform architecture supports heterogeneous data from diverse sources of sleep data and thus has the potential to support multi-disciplinary research needs by effortlessly bringing them the data they need in a preferred format. We see large benefits of such architecture, especially for multi-disciplinary research topics such as sleep. Due to the ownership of data, different end-users can access relevant data without being limited to specific data sources, which is an additional design quality that encourages collaboration between the research fields. Furthermore, the ownership makes deleting participants data entries effortless. This is an essential feature for the Sleep Revolution and other research projects, since the data belongs to its participants, and they can withdraw their participation and their data at any time [17].

The design is particularly suited for projects that involve multiple stakeholder groups and various data sources and, thus, fulfills the needs for a digital research platform as outlined by Arnardottir et al. [4]. The previously mentioned design qualities are not only sought after in interdisciplinary research projects but are also relevant for other application areas that deal with large amounts of data. In contrast to existing architectures such as the modular layered architecture by Yoo et al. [7•], our design is of a technical nature. This way, we narrow the gap in the literature for more detailed architecture. Moreover, our design is close to practice and easily applicable. This sets it apart from mainly theoretical findings such as those presented by Bygstad et al. [12]. Our approach also differs from other practical solutions such as the research by Bache et al. [11], as it does not operate as part of the communication with the database but instead tackles the issue on an architectural level.

There are limitations to our dynamic, homogeneous database and digital platform design, e.g., it is a less suitable fit for raw signal value entries. In those situations, direct use of raw data requires additional database tables or a file system as a supplement. However, the design fits metadata on each of the raw signal values such as average, min, max, duration, and more. The metadata values can therefore give the end-user a sufficient representation to understand the raw data's amount and diversity.

We did not eliminate the need for preprocessing of the data, and new data sources still require manual work of crystalizing new data source pipelines in order to translate it into the database design. However, the data source pipelines encourage increased automation which minimizes the work needed to add new data points. The data source pipeline's direct communications with our database, that is, the generate-pipeline and insert-pipeline, are independent components that are shared between all data source pipelines. The independent components have a strict evaluation of the data ensuring consistency and homogeneity for all current and future data sources. Thereinafter, the data is presented in the interface, which outlines the front-end of the digital platform. Further research is needed to find suitable visualizations of the data for different end-user groups of the digital platform. The distinct goals for each user group create unique visualization challenges. A high level of customization allows researchers from different research fields to work with data from selected sources to answer their research.

Conclusion

In this paper, we show the complexity of combining various data sources in sleep research and how the researcher's various requirements can be met with a homogeneous database design in a digital platform. First, we contribute with the conceptualization of a simple homogeneous database

and digital platform architecture that uses five tables to represent all data sources with optional additional information that help end-users understand the collected data. Therefore, the complexity of the architecture does not grow with additional data sources. The shared format makes the process of comparing, collecting, and exporting the data effortless for researchers and developers. Furthermore, sharing the data format can connect different research fields by giving researchers a helping hand in using a data source collected outside their field, presenting it in a feasible manner. The design has participant ownership to each data point, which makes the deletion of the participant's data effortless. That is an essential feature in most research projects since the data belongs to the participant who can withdraw its participation, and therefore, their data.

Our action design research findings, derived from Sleep Revolution, provide an illustrative example to show the flexibility and dynamic capacity of the design. However, we argue that this particular digital platform design can be generalized and utilized in other contexts. In addition to that, we contribute with the data source pipeline, which describes all the preprocessing needed for each unique data source from their heterogeneous source into the homogeneous database and digital platform. The data source pipeline with its four key steps, (i) parameter formulation, (ii) forms and entries creation, (iii) process-pipeline creation, and (iv) data insertion, is an obligatory component for the homogenous database to overcome the data sources format constrictions and transform the data into a homogeneous and valid format.

Our design offers a generic design with a high level of customization, and as argued before, it is therefore not limited to sleep research. Instead, it has the potential to fit other fields that require organizing and bringing together large, complex, and diverse datasets in a dynamic manner.

Funding This research is a part of the Sleep Revolution project, with funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 965417.

Data Availability The paper focuses on database and pipeline designs, and no data from this paper can be shared with other researchers.

Declarations

Conflict of Interest Dr. Arnardottir discloses lecture fees from Nox Medical, Philips, ResMed, Jazz Pharmaceuticals, Linde Healthcare, Alcoa—Fjardaral, Visitor (Novo Nordisk), and Wink Sleep. She is also a member of the Philips Sleep Medicine & Innovation Medical Advisory Board. Mr. Sveinbjarnarson discloses fees from Alcoa – Fjardaral and Reykjavikurborg. The other authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any data from studies with human or animal subjects performed. The authors got ethical approval and informed consent before volunteers tested the digital platform.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1.•• Arnardottir ES, Islind AS, Óskarsdóttir M, et al. The Sleep Revolution project: the concept and objectives. J Sleep Res. 2022;31:e13630. **This paper reflects the need for designing digital platforms in sleep research.**
2. Pevernagie DA, Gnidovec-Strazisar B, Grote L, Heinzer R, McNicholas WT, Penzel T, Randerath W, Schiza S, Verbraecken J, Arnardottir ES. On the rise and fall of the apnea–hypopnea index: a historical review and critical appraisal. J Sleep Res. 2020;29: e13066.
3. De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. Med Sci Sports Exerc. 2019;51:1538–57.
4. Arnardottir ES, Islind AS, Óskarsdóttir M. The future of sleep measurements: a review and perspective. Sleep Med Clin. 2021;16:447–64.
5.• de Reuver M, Sørensen C, Basole RC. The digital platform: a research agenda. J Inf Technol. 2018;33:124–35. **This paper outlines the importance of novel architectural design to cope with increasingly diverse data**.
6. Islind AS. Platformization: co-designing digital platforms in practice. Trollhättan: University West. PhD Thesis, University West. 2018;25:123. Available from: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1238297&dswid=-9964.
7.• Yoo Y, Henfridsson O, Lyytinen K. Research commentary—the new organizing logic of digital innovation: an agenda for information systems research. Inf Syst Res. 2010;21:724–35. **This paper outlines the need for strategic thinking in designing digital infrastructures**.
8. Constantinides P, Henfridsson O, Parker GG. Introduction—platforms and infrastructures in the digital age. Inf Syst Res. 2018;29:381–400.
9. Wareham J, Paul FP, Cano Giner JL. Technology ecosystem governance. SSRN Electron J. 2013. https://doi.org/10.2139/ssrn.2201688.
10 Nikayin F, de Reuver M. Opening up the smart home: a classification of smart living service platforms. Int J E-Services Mob Appl. 2013;5:37–53.
11. Bache R, Miles S, Taweel A. An adaptable architecture for patient cohort identification from diverse data sources. J Am Med Inform Assoc. 2013;20:e327–33.
12. Bygstad B. Generative innovation: a comparison of lightweight and heavyweight IT. J Inf Technol. 2017;32:180–93.
13. Sein MK, Henfridsson O, Purao S, Rossi M, Lindgren R. Action design research. MIS Q. 2011;35:37–56.
14. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational

research informatics support. J Biomed Inform. 2009;42(2):377–81. https://doi.org/10.1016/j.jbi.2008.08.010.
15. Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, Morin CM. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep. 2012;35:287–302.
16. Schmitz L, Sveinbjarnarson BF, Gunnarsson GN, Davíðsson ÓA, Davíðsson ÞB, Arnardottir ES, Óskarsdóttir M, Islind AS. Towards a digital sleep diary standard. MCIS 2022 Proceedings. 2022;10. https://aisel.aisnet.org/amcis2022/sig_health/sig_health/10.
17. Radley-Gardner O, Beale H, Zimmermann R, editors. Fundamental Texts on European Private Law. Oxford, United Kingdom: Hart Publishing. 2016. https://doi.org/10.5040/9781782258674.

Appendix C

Publication III

Data Work in Healthcare:

Mediating Data Quality and Data Governance in a Data-Intensive World

Abstract. Data is a critical resource in healthcare research, yet ensuring high data quality remains a persistent challenge. Errors introduced at the point of data entry can propagate through the data lifecycle, affecting usability, integrity, and research outcomes. In response to this challenge, we designed and developed a digital infrastructure over the course of four years capable of accommodating diverse structured health data while enforcing predefined data standards through a validation system. We use that digital infrastructure to examine the root causes of data quality issues at the input stage by analysing the type of data work needed for meaningful data curation. Through a combination of feasibility evaluation, surveys, and semi-structured interviews, we identified recurring mistakes, user behaviour patterns, and underlying reasons for poor data quality. We contribute six key elements of data work necessary for meaningful decision-making and research purposes, grouped under three categories: (i) the data work needed for reaching intrinsic data quality in data governance, (ii) the data work needed for reaching contextual data quality in data governance, and (iii) the data work needed for reaching representational and accessibility data quality in data governance.

Keywords: Data Quality, Data Work in Healthcare, Data Governance, Digital Infrastructure, Information Systems, Healthcare Data.

1 Introduction

Data has become a central resource in the modern world, yet the effort required to make it usable is often overlooked (Abbasi et al., 2016; Bossen et al., 2019). This issue is particularly evident in healthcare, where data-driven decision-making and personalized medicine are increasingly prioritized. As the demand for high-quality, actionable data grows, so does the need to investigate and document the data work involved in achieving these goals (Davenport & Patil, 2012; Parmiggiani et al., 2022). This work, often described as backstage (Parmiggiani et al., 2022) or invisible (Islind et al., 2021), is gaining recognition in the literature as a crucial yet frequently underestimated aspect of data governance. The digitalization of healthcare, supported by digital infrastructures and various healthcare technologies, aims to improve efficiency and quality of care through data. However, data does not simply exist; it must be produced, collected, structured, validated, and interpreted (Cruz, 2023; Gregory et al., 2019; Su et al., 2022). This effort is known as data work (Bossen et al., 2019), encompassing technological, analytical, and contextual tasks required to make data usable within healthcare (Fiske et al., 2019).

Data work in healthcare is inherently complex, requiring not only technical skills but also domain-specific knowledge to maintain data quality and integrity (Jones, 2019). Data work involves continuous human interaction with data curation processes and technologies, making data work prone to errors and inconsistencies. These inaccuracies can significantly impact healthcare decisions, diagnoses, and patient outcomes, alongside research outcomes down the line if faulty data is used for research purposes. Given these potential consequences, examining the intricacies of data work and the factors contributing to errors is crucial. By understanding where and how inaccuracies arise, better informed strategies for improving the data quality of health data can be formed, but for that to happen, data work needs to be put center stage.

Healthcare relies heavily on rigorous research, which in turn depends on high-quality data. As the volume of data flowing through healthcare systems increases, so do the tasks required to manage, process, and ensure its reliability (Janssen et al., 2016). Understanding the data work underpinning healthcare research is becoming increasingly important, particularly as data-driven digital infrastructures expand. While these data-driven digital infrastructures enhance both the quality and quantity of available data, they also generate a substantial amount of behind-the-scenes work. This *backroom data work* is essential for shaping raw data into a usable form, enabling researchers and practitioners to leverage its full potential. With the rise of big data and the ongoing process of datafication, data-driven digital infrastructures have become a cornerstone of digital transformation strategies in healthcare (Galliers et al., 2017; Parmiggiani et al., 2022). As a result, multidisciplinary teams now collaborate to discover, transform, model, and report vast and complex datasets for research purposes (Davenport & Patil, 2012; Parmiggiani et al., 2022). This transformation involves both visible and invisible data work, spanning data collection, data analysis, and securing data quality.

Despite the increasing recognition of the need for high-quality data, much of the existing literature focuses on the benefits and potential of digital

infrastructures—such as efficiency, improved service quality, and enhanced decision-making—while overlooking the extensive data work required to achieve these outcomes (Bertelsen et al., 2024). Researchers play a critical role in this process, yet the effort involved in preparing, curating, and structuring data remains largely behind the scenes, still invisible to a large extent.

As data-driven digital infrastructures become the backbone of healthcare research, understanding how they shape and are shaped by data work is essential. This paper highlights the often-overlooked efforts of researchers in establishing and maintaining the scaffolding for data-driven healthcare, emphasizing the labour-intensive process of transforming raw data into meaningful, high-quality information by examining the data work in detail. Our research focuses on improving data quality in healthcare by identifying the necessary data work linked to meaningful contextual data quality by developing a lean database structure alongside a rigorous validation system embedded within our digital infrastructure. Our empirical context is within sleep research—a context that, despite its profound impact on health, has historically been overlooked as a major pillar of well-being. Sleep influences hormonal regulation, organ function, cognitive and physical performance, and overall health, making it a crucial factor in daily life (Arnardottir et al., 2021).

Sleep disorders, such as insomnia, which affects 10% of adults, and sleep apnea, which impacts nearly 1 billion people worldwide, are prevalent across numerous research fields (Arnardottir et al., 2021). This diversity is reflected in the data collected, spanning multiple disciplines, and presents unique challenges in building and governing heterogeneous data. Our digital infrastructure has accumulated data from various sources, leading to significant data governance complexities. The need to merge smaller studies into larger datasets has further amplified data quality issues, particularly as responsibilities are distributed among researchers in various roles and from multiple disciplines; each with different priorities and approaches. As is common in healthcare research, data collection often centers on individual study objectives rather than the collective and cohesive effort required to ensure high-quality data for future decision-making and research. This paper examines these challenges, highlighting the invisible data work necessary to bridge the gap between data acquisition and the long-term usability of thoughtful and reliable healthcare data.

We designed and developed the aforementioned digital infrastructure on the foundations of a lean database structure, capable of accommodating diverse healthcare data without requiring additional tables, thereby minimizing complexity. It currently houses 63 distinct data sources, encompassing approximately 2,632 parameters from over 2,000 participants, all contributing complex sleep-related data. Over the past four years, researchers working with this data have frequently encountered errors that bypassed the validation system, prompting continuous refinements. To mitigate these challenges, we integrated a robust validation system within our digital infrastructure. In this paper, we examine how effectively researchers can independently upload their collected sleep data using our digital infrastructure and examine the data work connected to acquiring and working with robust research data. Our empirical

data is based on a mixed-methods study involving thirteen participants, combining quantitative data with qualitative insights from semi-structured interviews. Our paper poses the following research questions: *What kind of data work is involved in curating quality healthcare data?*

Our analysis identifies three key elements of data work necessary for curating healthcare data for meaningful decision-making and research purposes: (i) the data work needed for reaching intrinsic data quality in data governance (ii) the data work needed for reaching contextual data quality in data governance, and (iii) the data work needed for reaching representational and accessibility data quality in data governance. Moreover, the validation system embedded within our data-driven digital infrastructure proved its effectiveness in enforcing standards, as no participant was able to upload all data sources without the system identifying at least one mistake, helping the participants along in their data work. Our findings emphasize the need to make data work more visible and integrate it as a fundamental, widely recognized aspect of scientific inquiry, and the need to have both systemic measures in place, as well as a cohesive human understanding across teams. We also argue that decisions about data quality and data governance must move beyond behind-the-scenes work and become a visible, integral part of research practice (Barley & Bechky, 1994) or data science (Parmiggiani et al., 2022). Instead, these discussions must take centre stage, ensuring that data governance and data quality standards are explicitly acknowledged and actively incorporated into research efforts where healthcare data is collected, managed, and analysed.

2 Related Work

A growing discourse in the literature emphasizes adapting data governance to specific contexts to manage the increasing complexity and volume of data. Data governance is a broad concept, generally encompassing data definition, implementation, and monitoring (Alhassan et al., 2016). It also involves establishing policies, standards, and guidelines to define when data quality is considered sufficient (Alhassan et al., 2016). A key aspect of data governance is the assignment of roles, tasks, and responsibilities to ensure compliance with these standards (Benfeldt et al., 2020). Additionally, continuous monitoring and iterative adjustments to roles, guidelines, and processes are essential for improving data quality over time (Petzold et al., 2020).

However, implementing data governance models presents challenges such as defining responsibilities, balancing strictness (Janssen et al., 2020; Petzold et al., 2020), and overcoming resistance to change (Ladley, 2019). Furthermore, errors are common during the adoption of data governance, which can have unintended ripple effects, such as increased data work. Effective error prevention is, therefore, crucial in establishing sustainable and efficient data governance frameworks. For any data governance model to be effective, well-defined data processes are essential. The literature on data governance highlights that a lack of shared understanding of these processes is a major contributor to data governance challenges (Parmiggiani & Grisot, 2020). Additionally, unclear data governance frameworks can erode trust in

data, reducing its perceived value and usability. Moreover, Alhassan et al. (2019) demonstrate how missing or ambiguous data governance structures led to distrust, emphasizing the critical role of well-defined data governance frameworks. In their study, the absence of clear data policies necessitated repeated manual data verification, further increasing data work and creating inefficiencies.

The literature also underscores the need to balance human efforts—through guidelines, shared learning, and collaboration—with strict systemic enforcement to ensure effective data governance. While defining data governance activities has been a longstanding focus, research still lacks practical guidelines for adoption and monitoring (Alhassan et al., 2019). Addressing these gaps could benefit scientific research broadly and improve data curation and data acquisition in healthcare, where trust and consistency in data are paramount. Furthermore, data governance has been framed as a collective action problem, particularly in complex organizations where responsibilities and priorities differ (Benfeldt et al., 2020). Given the gaps in the literature on data governance within such complex settings, data governance can be viewed as a mechanism for aligning organizations through shared data governance structures. While data governance is widely recognized for its potential to create value, this value is shaped by how responsibilities are distributed.

Alhassan et al. (2019) identify six essential elements of data governance: (i) cross-functional effort—collaboration and role assignments crucial for governance; (ii) framework—structured support materials, including diagrams and organizational structures; (iii) data as a strategic asset—treating data as a valuable resource with defined goals; (iv) decision rights—determining authority over data pipelines and governance decisions; (v) data policies, standards, and procedures—establishing rules to support governance objectives; and (vi) compliance monitoring—ensuring adherence to established standards and policies. This highlights the deep interconnection between data governance and the data work needed to make data useful. Effective data governance cannot function without extensive data work, reinforcing the need for structured processes and continuous oversight.

The literature on data governance highlights the importance of sustaining effective data governance over time, as it is critical to long-term success (Ladley, 2019) , and one of the pillars of effective data governance is the notion of data quality. Poor data quality is often outlined as the root cause of the most significant challenges in data governance (Janssen et al., 2020). Given that high-quality data is fundamental to rigorous decision-making and research, the extensive efforts required to establish and maintain good data quality as an embedded element of data governance frameworks should be widely acknowledged. While existing research recognizes the importance of data quality, there remains a gap in understanding the data work involved in curating and maintaining high-quality data, a gap which our paper addresses. Poor data quality can lead to significant social and economic consequences, with ripple effects through decision-making and research over time (Wang & Strong, 1996). While organizations are adopting practical strategies to enhance data quality, these efforts often focus primarily on accuracy (Redman, 1998; Steuperaert et al., 2025). However, there are multiple aspects of good data

quality. *Firstly*, intrinsic data quality refers to the inherent quality of the data itself. *Secondly*, contextual data quality emphasizes that data quality must align with the specific task or use case. *Thirdly*, representational data quality refers to how well data is structured and presented, and *fourth*, accessibility data quality underscores the significance of system-related factors in data quality and the notion of making data understandable and usable for its intended purpose. These elements combined illustrate that high-quality data should be inherently reliable, suitable for the given context, well-represented, and easily accessible to users (Wang & Strong, 1996; Mohammed et al., 2025; Steuperaert et al., 2025).

The literature on data quality has been growing in recent years, brought on by the notion that quality data is needed to make quality decisions through machine learning models (Gröger, 2021; Mohammed et al., 2025) as well as generic decision-making, based on quality data. The current stream of literature is focused on data lifecycle concerns and effective data management, much of which is focused on insightful machine learning. However, data quality does not simply happen. Instead, it takes considerable data work, which is largely hidden and invisible. Furthermore, without data work, the data would neither be usable for machine learning purposes nor would it become insightful for its upstream use. In this paper, we focus on the data work needed to cultivate data governance on the one hand and data quality aspects on the other hand.

3 Research Approach

Our study is based on data collected through a mixed-methods research approach. The Sleep Revolution is a multidisciplinary initiative focused on advancing sleep research by collecting and preparing data for future studies. The project integrates prospective datasets with up to 20 years of retrospective clinical data sourced from hospitals, research centers, and universities across Europe. Its primary objective is to support ongoing research while establishing a solid foundation for future scientific inquiry related to researching sleep. Researchers can securely access and analyse this data through a high-performance cluster, making the data, embedded within our digital infrastructure, the project's most valuable asset.

3.1 The Digital Infrastructure

To manage the project's complex and heterogeneous data, we designed a lean database framework intended to reduce friction in data work across disciplinary boundaries (see Figure 1). The framework organizes over 2,600 variables from diverse datasets into seven unified tables, each linked to individual participants and studies via a shared primary key. This relational model supports a broader digital infrastructure used by 39 research centers, with most being in Europe, integrating a range of data types, including sensor data, mobile app logs, clinical records, and survey inputs.

The design is modular and schema-light, centering on four core entities: (i) Form, which defines the table template; (ii) Entry, representing individual

data fields or parameters; (iii) FormResult, corresponding to rows of submitted data; and (iv) EntryResult, storing cell-level values. This structure allows new data sources to be added without altering the table architecture, supporting extensibility while maintaining structural consistency. Metadata fields further support evolving data needs, enabling documentation and traceability without additional schema complexity.

A key component of this digital infrastructure is a built-in validation system that enforces standards. This validation system evolved into the Data Integrity Assurance System (DIAS), developed iteratively over four years. DIAS was shaped by user feedback and co-design sessions, leading to a system capable of enforcing format and logic rules while remaining usable by non-technical contributors. The goal was to offload quality enforcement from manual review to automated validation without requiring users to write scripts or understand the database internals.

As DIAS matured, it became central to the infrastructure's approach to data governance. Its rules and policies were refined through continuous use, addressing issues such as inconsistent identifiers, formatting errors, and structural mismatches. Improvements focused on enhancing findability, usability, and data consistency under resource constraints. This chapter outlines the architecture and development of this infrastructure as a basis for evaluating how users interact with automated data validation systems and whether such systems can support independent standard application and improved data quality in practice.

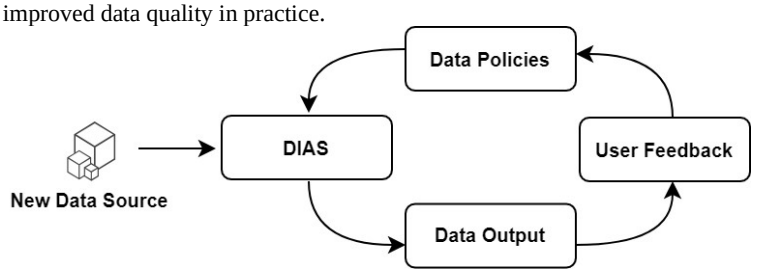


Figure 1. Illustrates the iterative process used to create data policies and rules in the project.

3.2 Study Design, Data Collection & Data Analysis

This paper relies on the digital infrastructure, which we have designed and developed for the past four years, as well as a mixed-methods empirical data collection, combining a feasibility test, survey, and semi-structured interviews to evaluate the data work behind applying data standards independently through automation and to investigate how user behavior, preferences, and data practices impact data quality. The feasibility test focused on whether the DIAS validation system could be used without assistance, relying only on a brief introduction and documentation. The broader aim of our empirical data gathering was to assess whether such validation systems could shift responsibility for standardization and validation from technical personnel to

the individuals who collected and understood the data; broadening the data work.

The participants were engaged in a structured process lasting between 65 and 120 minutes (see Figure 2). All 13 participants who partook in our data gathering had prior experience working with healthcare data. At the beginning of the session, participants completed a consent form explaining the study's purpose, the data being collected, their right to withdraw at any time, and how their information would be handled. Following this, the first author provided an oral introduction to the system, including an overview of the files provided, the task ahead, and how to interpret the documentation. Each participant received a ZIP file containing example input files, a small CSV dataset (12 rows, 8 columns), documentation, and a README file. After the introduction, they were asked to convert the data into six structured CSV files and upload them to the validation system.

Each file corresponded to a specific table in the relational database; participants, studies, study participation, forms, entries, and form results. The complexity of the tasks varied depending on the structure and formatting requirements of each file. More complex files, such as form results, required column-level metadata and stricter formatting, while others involved simpler mappings. The goal was to simulate realistic scenarios in which users prepare structured data independently and rely on documentation and feedback from the validation system to complete the task. The feasibility test allowed us to observe how participants interacted with the system and highlighted challenges in understanding and applying standards.

The participants uploaded their files to a minimalist website connected to the database. After each upload, the validation system provided immediate feedback indicating whether the submission met predefined standards or contained any violations (see example output in Appendix A). In remote sessions, participants shared their screens, allowing the researcher to observe how they engaged with documentation and responded to feedback. The first author answered clarifying questions about data content during the introduction of the system, but after participants began the task, further assistance was avoided unless participants became completely stuck, which did not occur. In such cases, they were simply reminded to refer to the example files and documentation provided.

During the data collection, observational data were collected on task completion rates, time spent, standard violations encountered, participants' use of documentation, and the files uploaded by participants were stored by the system for further review. Following the feasibility test, participants completed a survey capturing their research backgrounds, challenges encountered in data work, time investment in data work, and perspectives on data quality. To deepen our understanding, a semi-structured interview with 17 questions was conducted, exploring participants' experiences with the validation system, its design, and broader views on data governance and data handling.

The qualitative data were analyzed using an abductive thematic approach. We engaged in an interplay between real-world observations through our interview transcripts and insights from existing literature on data work more broadly (Gregory and Muntermann, 2011), going between the two during our analysis. The thematic analysis was conducted in three phases,

through iterative coding of the interview transcripts, starting with open descriptive tags. *The first phase* was conducted by the first author and involved primary coding, assigning initial label segments with descriptive and conceptual tags. This phase was open-ended and exploratory, focused on capturing the essence of the data yet trying to organize and interpret it deeply (Tracy, 2024). *The second phase* was also conducted by the first author and involved secondary coding, revisiting the codes, and beginning to group them, refining them, and identifying the relationship between the codes. These were grouped into broader categories to identify patterns such as difficulties in interpreting data, engagement with documentation, and strategies for addressing validation feedback. *The third phase* was conducted in collaboration between the first and the last author of this paper and included the theme creation. This phase involved axial coding as well as interpretation and abstraction, which ultimately led to the three themes presented in the findings in this paper. The three key elements presented in the findings were drawn from both the frequency of these categories and their relevance to the study's aims, including participants' perceptions of the system and responses to questions related to data work, derived through the third phase. The quantitative survey data were also coded and categorized. Full details of the survey and interview instruments are provided in Appendix A.

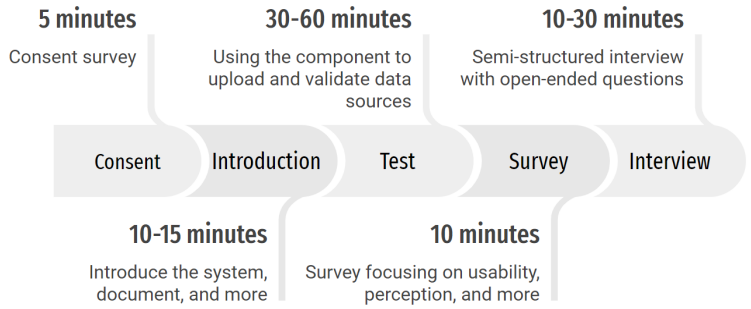


Figure 2. Participant engagement and task sequence.

3.3 Recruitment and Ethical Considerations

The participants were recruited through convenience sampling via professional networks and referrals, with a focus on individuals actively engaged in data work. All had prior experience working with healthcare data and came from diverse fields, including sports science, neuroscience, information systems, and machine learning. Eligibility criteria required prior data experience, English proficiency, no involvement in the development of the validation system, and access to spreadsheet software.

The interviews were audio recorded, transcribed verbatim, and lightly edited for clarity. Non-English interviews were translated into English before analysis. The collected data included survey responses, uploaded task files, interview transcripts, and observational notes. The survey data were cleaned for consistency prior to analysis. Although multiple ethical approvals were

obtained for the broader project, this specific study did not require ethical approval under Icelandic regulations, namely Act No. 90/2018 on Data Protection and the Processing of Personal Data, as no personal or sensitive data were collected. All participants, who were researchers, signed informed consent before participating. They were informed that their anonymized data would be used for research purposes and that they could withdraw their consent at any time, either by contacting the research team or by revisiting the consent form and changing their response from yes to no. This study is part of the Sleep Revolution project and was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 965417.

4 Results

To explore the complexities of data work and its implications for data governance, we identified six key elements essential for curating healthcare data and advancing healthcare research. *First*, data work is prone to frequent errors, leading to a high mistake rate. *Second*, researchers often rely on assumptions when handling data, even when unfamiliar with its context. These first two elements of data work are illustrated in the section called *the data work needed for reaching intrinsic data quality in data governance*. Intrinsic data quality refers to the inherent quality of the data itself, and our paper shows that there is data work needed to unfold intrinsic data quality and that data work involves gaining a deep understanding of data needs and trying to lessen the likelihood of errors and assumptions.

Third, there is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential in balancing systemic measures and human effort despite structured data governance. *Fourth*, perceptions of data quality vary significantly—what one considers crucial, such as metadata or consistent formatting, may be dismissed as unimportant by another. These two elements of data work are illustrated in the section called *the data work needed for reaching contextual data quality in data governance*. Contextual data quality emphasizes that data quality must align with the specific task or use case, and our paper shows that there is data work needed to align the human and systemic measures.

Fifth, data users often lack awareness of the broader needs of others when collecting and processing data for shared use. *Sixth*, many researchers have never engaged with data standards, data principles, or data governance, resulting in significant, often unseen data work later in the research process. These two elements are illustrated in the section called *the data work needed for reaching representational and accessibility data quality in data governance*. Representational data quality refers to how well data is structured and presented, and accessibility data quality underscores the significance of system-related factors in data quality. Our paper shows the importance of doing so through digital infrastructures and cultivating excellence in data work early on, as that, in turn, reduces later workload.

4.1 The Data Work Needed for Reaching Intrinsic Data Quality in Data Governance

During the feasibility evaluation, we analysed the type of data work needed within the digital infrastructure for reaching intrinsic data quality and how the data could be checked for its feasibility for the purpose at hand. Our findings show that a deep understanding of data needs is the essence of data work, and that, in turn, elevates data curation. Our participants were tasked with splitting a provided spreadsheet into six new files, adding metadata columns to three of them to examine their approach to data work. They were given example input sheets for each file type, a short description file, and detailed documentation, which they were encouraged to use. The external tools participants chose varied, including Excel, Google Sheets, LibreOffice Calc, and text editors. While some modified the example files to incorporate the new data, others created entirely new files—both approaches reflecting different forms of data work with potential ripple effects on data governance down the line.

All participants successfully uploaded the six required files as part of our study. However, none completed the task without encountering at least one error flagged by the validation system within the digital infrastructure. Although the provided data was free of initial errors or formatting issues, participants introduced new errors, triggering validation checks that identified problems such as incorrect formatting, missing or additional columns, duplicate entries, incorrect file types, and structural inconsistencies. These errors highlight critical aspects of data governance that must be addressed to ensure sustainable data quality.

Common mistakes included spelling errors, incorrect date formatting, and the accidental deletion of rows perceived as duplicates. Such errors, if not caught early, could necessitate additional data cleaning efforts later, increasing the data work required for future research. Some errors, such as misinterpreted metadata or mislabelled columns, were not flagged by the validation system but were identified through direct observations. These observations underscore the importance of early intervention in preventing faulty data from entering research databases, ultimately safeguarding the integrity of data used for healthcare analysis and scientific discovery.

Participant behaviour exhibited several distinct trends. Initially, many displayed low confidence, frequently seeking clarification rather than engaging with the documentation or system feedback. Over time, they became more efficient and confident in using the digital infrastructure, though frustration was common when uploads were rejected. Notably, many participants approached the task as a challenge to *"get the system to accept their input"* rather than an exercise in meaningful data work. This mindset led them to prioritize satisfying validation requirements over ensuring that the data was properly structured and meaningful for future use.

Participants expressed diverse preferences on how structured and raw data should be shared. Some favoured storing data in a database with query capabilities, while others preferred simpler formats like folder structures or Excel files. These varying perspectives highlight the lack of a universal

approach to data storage and underscore the need for adaptable digital infrastructures that accommodate different workflows.

Overall, the results reveal a pattern of minimal engagement with documentation, a reliance on assumptions, and a reactive rather than proactive approach to data preparation. The high frequency and diversity of errors detected by the digital infrastructure emphasize the critical role of robust validation processes in sustainable data governance. Additionally, these findings highlight the need for improved user guidance and education on the significance of preparing data for long-term reuse. The participants' engagement with documentation revealed another critical challenge. Although documentation was provided and participants were encouraged to use it, only a minority actively relied on it during their tasks. Participant 13 admitted, *"I think they are important, but no one reads them."* Even those who found documentation helpful engaged with it sparingly. Participant 1 noted, *"I didn't work much with the documentation, but I found it helpful for specific tasks, like how to write institution names."* Similarly, Participant 10 stated, *"In general, the documentation is essential, but I did not use it except for one thing I was unsure of."* These responses suggest that relying solely on documentation to ensure proper data handling is insufficient. Instead, cultivating excellence in data work requires a shift in practice—one that emphasizes ongoing discussions, training, and embedding data quality practices into daily research activities. Based on that, we offer two lessons learned. Data work is prone to frequent errors, leading to a high mistake rate, and furthermore, researchers often rely on assumptions when handling data, even when unfamiliar with its context.

4.2 The Data Work Needed for Reaching Contextual Data Quality in Data Governance

Despite the high number of errors detected by the digital infrastructure during testing, survey results indicate that participants generally found the system and the process of creating the required data files to be relatively straightforward, even as first-time users. Most participants rated most of the system functions as 'very easy' to relate to and work with or 'somewhat easy', with only a few selecting 'neutral' in terms of difficulty when working within the data governance structure. Three participants rated the final task as 'somewhat difficult', but no one considered any task 'very difficult.' Overall, nearly all participants agreed or strongly agreed that the digital infrastructure and its embedded data governance mechanisms were beneficial for research projects, particularly within healthcare, with only one participant expressing a neutral stance.

Interestingly, our survey findings revealed significant disagreement among participants regarding what constitutes 'good' data quality. While some emphasized the importance of consistent data formats, unique research identifiers, and the elimination of duplicates, others considered these aspects less critical. This variability underscores the diverse priorities and perspectives researchers bring to data governance, highlighting the need for digital infrastructure that can accommodate differing interpretations of data quality.

However, one area of strong consensus emerged: the importance of identifying which dataset a specific data source belongs to. Most participants considered this essential, particularly in healthcare, where accurately linking patient data to symptoms and associated records is crucial for both research and clinical decision-making.

Participants also highlighted challenges in working with data they had not collected themselves, often revealing underlying trust issues. Many struggled to understand parameters or variables, especially when labelled with abbreviations or general terms they had not defined. These ambiguities directly contributed to distrust in the data. The absence or inaccessibility of metadata further compounded these issues, making interpretation and reuse difficult and reinforcing scepticism about data reliability. Additionally, merging datasets proved challenging due to missing identifiers and inconsistent formats, illustrating the complexities researchers face in ensuring data compatibility as a crucial aspect of data work.

When discussing broader challenges in data handling, participants pointed to several key factors: educational gaps, human error, and technological limitations. Participant 7 identified multiple issues, stating, *"Human errors, not understanding what is important, how to format, not complete… poor documentation, time-consuming."* Education was also seen as a critical factor in mitigating these issues. Participant 3 argued that *"I think with data processing and programming as a basic education from primary school and upwards, this would be much less of a problem."* Participant 5 further noted, *"The learning process is different for different people… [data is] not easy to see where the errors surface."* Participants also expressed concerns about inconsistencies in data collection and reliance on temporary fixes rather than long-term solutions. Participant 8 explained, *"Temporary fixes instead of preventing errors […] We are limited by older systems or solutions that have been used for a long time."*

Furthermore, several participants acknowledged the tendency to prioritize immediate efficiency over long-term usability. Participant 13 observed, *"When people start working on something, they don't think enough about the future […] structure, and all their work so that it could be used in the future."* This sentiment reflects a key challenge in data governance—balancing short-term ease of data handling with sustainable practices that ensure long-term data quality. The lack of adherence to data principles, inconsistencies in formats, and poor conversion tools further complicate data governance. As Participant 11 succinctly put it, *"Not everyone follows the principles. All of those different formats are difficult, and there are formatting issues all over."* These findings emphasize the need for structured but adaptable data governance models that not only enforce quality standards but also integrate best practices for data work in a way that researchers find practical and useful.

Our participants reported spending considerable time on data work tasks such as data conversion, data merging, and data cleaning—processes that should be more streamlined with standardized formats. These inefficiencies in current data governance practices highlight the need for digital infrastructure like the one presented here, which aims to ensure that data is clean, validated, and ready for use from the outset. However, our findings also reveal tensions

and reduced trust when researchers work with data, they did not collect themselves, reinforcing the notion that data work is inherently complex and deeply human. Based on these findings, we offer two additional lessons learned in addition to those already posed in the first subsection of the findings. There is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential for balancing systemic measures and human effort despite structured data governance. Moreover, perceptions of data quality vary significantly; what one considers crucial, such as metadata or consistency in formatting, may be dismissed as unimportant by others.

4.3 The Data Work Needed for Reaching Representational and Accessibility Data Quality in Data Governance

In our interviews, we examined participants' training, readiness, experience in handling data, familiarity with data quality principles, and perspectives on why data is often problematic. Through our analysis, it was clear that data work remains undervalued and is rarely treated as a formal skill. Many participants received little to no formal training in data handling or data quality. Instead, they relied on self-taught methods or guidance from colleagues, learning 'on the fly' rather than through structured education. This lack of formalized training suggests that data work remains largely invisible, assumed rather than taught, and often overlooked as a crucial component of research.

Participant 8 noted: *"No, everything [was] self-taught,"* and Participant 10 echoed: *"No, only brief courses on handling data. Maybe something brief, but I cannot remember."* Similarly, Participant 4 stated, *"Not really, everything I've learned has been through work."* Even those with formal training described a limited scope, typically focusing on technical aspects like database design or data cleaning. Participant 5 shared: *"I did my master's in data science […] We had courses on data quality and cleaning,"* while Participant 6 added: *"Yes, I have had many courses on structuring and creating data in databases and other structures."*

The absence of structured education in data governance points to a systemic challenge that hinders collective action and collaboration in data work. Researchers approach data inconsistently without a shared foundation of best practices, leading to errors and inefficiencies. Our findings suggest that cultivating excellence in data work requires more than individual expertise—it necessitates greater visibility, institutional recognition, and structured training to ensure data quality is not left to assumption or chance.

We also explored the extent to which participants engaged with data principles or formal standards in their data work. Most had little exposure to established data standards, and those who did often relied on ad-hoc practices or conventions specific to individual research projects rather than universally recognized frameworks. Participant 2 stated, *"None I can remember, no guidelines or standards in my current or previous work."* Similarly, Participant 7 noted, *"No, no documentation, but maybe somewhat through adapting others' work."* This absence of structured principles suggests that data

management often lacks uniformity, making it difficult to share and reuse data across research teams.

Some participants attempted to create their own standards to maintain consistency. For instance, Participant 12 shared, *"Yes, but I made it myself. I organized the data we store and the [sleep] data format."* However, the implementation was often incomplete even among those who followed the guidelines. As Participant 4 pointed out, *"Yes, some, but way too few. [We use] naming standards for sleep scoring, for example."* These findings indicate that without widely accepted data governance structures, researchers develop their own fragmented approaches, potentially reinforcing inconsistencies and increasing data work over time.

When asked about implementing the digital infrastructure for validation in their own work, participants largely supported its use, particularly for improving data consistency and strengthening data governance. Participant 1 remarked, *"Yes. The data is so much. Especially good for participants in multiple studies."* Similarly, Participant 9 stated, *"Yes, that would be highly beneficial as long as you have access to the source data for transparency and trust."* Others emphasized its role in unifying data across research projects. As Participant 6 noted, *"I think we need to have it to unify all of the data."* Our findings also show that although it might make sense to implement data governance in general, pressured everyday practice often leads to opting for minimal viable effort due to resource scarcity. It can, however, shift the burden onto others, creating a cascade effect that leads to more work down the line.

Furthermore, some participants raised concerns about flexibility, suggesting that strict validation rules might hinder creativity in certain research contexts. Participant 3 elaborated, *"Having a strict format that your data has been in could become an unnecessary hurdle […] It really depends on what kind of experiment it is."* These findings highlight the balance between enforcing data standards and allowing adaptability in data governance. Based on that, we offer two more lessons learned. Data users often lack awareness of the broader needs of others when collecting and processing data for shared use. Furthermore, many researchers have never engaged with data standards, data principles, or data governance, leading to extensive and often invisible data work later in the research process.

5 Discussion

Knorr-Cetina's (1997) influential research on knowledge work suggests that modern work increasingly mirrors the characteristics of scientific research. This perspective aligns with Barley and Bechky's (1994) notion of the "backrooms of science," which highlights the often-overlooked labour of laboratory technicians in preparing, managing, and curating scientific objects and research data. Building on this foundation, Parmiggiani et al. (2022) emphasize the significant yet hidden efforts required to make data ready for analysis and knowledge curation. Expanding on these insights, our paper explores the complex and nuanced data work undertaken by researchers in

healthcare settings—work that is critical for the design, development, and sustainability of data-driven digital infrastructure and fundamental to advancing healthcare decision-making and research through quality data. Combined, our findings highlight several insights that shed light on data work for mediating data quality and data governance in a data-intensive world. *First*, data work is prone to frequent errors, leading to a high mistake rate. *Second*, researchers often rely on assumptions when handling data, even when unfamiliar with its context. *Third*, there is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential in balancing systemic measures and human effort despite structured data governance. *Fourth*, perceptions of data quality vary significantly—what one considers crucial, such as metadata or consistent formatting, may be dismissed as unimportant by another. *Fifth*, data users often lack awareness of the broader needs of others when collecting and processing data for shared use. *Sixth*, many researchers have never engaged with data standards, data principles, or data governance, leading to extensive and often invisible data work later in the research process.

5.1 Data Quality is Fostered Through a Symbiosis Between Human Effort and Systemic Measures

Data governance has emerged as a key pillar for ensuring data quality in both organizations and research (Benfeldt et al., 2020). However, as our findings show, effective governance requires continuous oversight and dedicated teams managing data throughout its lifecycle. This includes ongoing data work to ensure availability, inform stakeholders, and address challenges. Despite its importance, research grants rarely fund dedicated data governance, focusing instead on outputs like publications. As a result, data work becomes frequently sidelined—even though high-quality research depends on high-quality data (Khong et al., 2023).

The growing literature on data quality reflects its role in enabling sound decisions through artificial intelligence and machine learning (Gröger, 2021; Mohammed et al., 2025). Yet, data governance and quality efforts require personnel and resources that many research projects lack. However, this paper is a call for action; data work in healthcare research must be recognized, prioritized, and resourced appropriately, as data work is pivotal for data quality. Our findings highlight the deep interconnection between data governance and data work. Beyond technical aims, investing in human expertise is essential for leveraging data's full potential in healthcare innovation. Numerous data principles and governance frameworks exist (Janssen et al., 2020), but adherence is often merely partial. While this may reduce their immediate data work, our findings show that it instead frequently shifts the burden onto others, creating a cascade effect. Shortcuts taken to minimize effort at one stage can lead to significant additional data work for others later, which in turn may ultimately undermine data quality and data governance effectiveness.

Our feasibility evaluation showed that—despite being encouraged—only two participants fully utilized the provided documentation, leading to errors

directly linked to their approach. This underscores the challenge of ensuring compliance with data governance models, even when their benefits are well understood (Benfeldt et al., 2020). Our findings emphasize the value of early-stage data work and meaningful human engagement, which can greatly enhance downstream scientific inquiry. They also reveal the vast scope and critical importance of often-overlooked data work in healthcare research. Neglecting data at the point of entry can lead to misinformation, flawed analyses, and unreliable findings, particularly dangerous in healthcare (Adams et al., 2023). Ensuring quality data requires effort in collection, curation, and validation, highlighting the need to acknowledge and invest in the labour behind data-driven healthcare research. Poor data quality can have serious social and economic consequences, rippling through decision-making and research (Wang & Strong, 1996).

Our findings revealed a high error rate. No participant uploaded all files without at least one validation error. Common mistakes included accidental row deletions and incorrect date entries. Despite built-in data governance mechanisms, adhering to guidelines remained difficult. This reinforces the need for both human effort and systemic measures. The errors we observed show that unless proactively prevented, data errors will inevitably occur. This highlights a critical trade-off between relying solely on systemic controls and investing in human awareness of data quality. Meaningful progress requires both: strong technological safeguards and a deeper understanding of the importance of data work. As Jarvenpaa & Essén (2023) argue, data governance must bridge technological and human dimensions. Our paper supports this, showing that sustainable data quality demands both systemic measures and human effort.

A closer analysis of the errors reveals a key insight: participants were tasked with uploading a small dataset, only eight columns and 13 rows. In contrast, our digital infrastructure integrates 63 data sources, covering 2,632 parameters from around 2,000 participants, totaling millions of data entries. Unlike controlled test data or structured financial data, real-world healthcare data is full of inconsistencies—spelling errors, duplications, formatting issues—highlighting the complexity and variability of such data. This increases the likelihood of validation errors and underscores the ongoing need for robust data work. Healthcare data, inherently tied to human variability, is neither static nor simplistic.

These findings underscore that while data governance provides essential guidelines, embedded validation systems are vital for reducing errors (Alhassan et al., 2019). More broadly, data governance requires a careful symbiosis between human effort and systemic measures to ensure data quality. These findings highlight the importance of further research into trust in digital technology and its data (Li et al., 2008; Müller et al., 2024), particularly in the context of data governance. Establishing trust is crucial for fostering user adoption and ensuring that data work—the foundation of effective data governance—becomes an integral and well-supported part of research workflows.

5.2 A Key Element of Data Governance is Making Data Work Visible

Despite all participants being healthcare researchers, their views on data quality, storage, and challenges varied significantly. Some emphasized eliminating duplicates, consistent formatting, and thorough documentation, while others downplayed these aspects. This divergence likely stems from their different backgrounds and environments. For instance, researchers with database experience found duplicate removal easy, whereas others lacked the knowledge to handle such issues. These findings reflect the complexities of multidisciplinary research and highlight the importance of effective data governance (Pansara, 2023) and collaboration across disciplines. While literature stresses data governance, our study shows its practical adoption is challenging and demands both, as stated earlier, systemic measures and a cohesive shared understanding of data principles.

Collaborative data work is hindered by differing priorities and expertise, leading to inconsistent attention to data quality. This was evident in our study, as only two participants who managed metadata effectively also stressed documentation as vital in the interviews. Their views directly shaped their data practices. While all participants agreed on data's centrality in today's research landscape, the data work to make it usable often remains invisible and undervalued (Abbasi et al., 2016; Bossen et al., 2019). To address this, we advocate for open discussions to build shared understanding and actively promote that data work is made visible in healthcare research.

Although most researchers could identify causes of poor data quality, over half had no formal training in data standards or principles. Instead, they relied on intuition and trial-and-error. Minimal standards—such as naming conventions or storage locations—were typically group-specific. This contradiction underscores the invisibility of data work and the need to prioritize training and support, and the importance of making data work increasingly visible.

Quality data requires thoughtful data work, supported by validation systems that bridge human effort and systemic measures. Our findings point to a lack of mutual understanding between disciplines, further complicated in collaborative, multidisciplinary research. This aligns with the seminal work of Schmidt and Bannon (1992) on articulation work, which emphasizes the importance of coordinating and communicating invisible labour. Decisions about data storage, standards, and governance are always made by someone. However, they often fail to account for the needs of all stakeholders (Alhassan et al., 2019). We argue that such decisions should not remain hidden in the background. Instead, they must be visible and recognized as essential elements of data work in healthcare research.

The literature has shown that there are multiple aspects of what can be seen as good data quality, and our paper illustrates the data work connected to each of those, contributing to the literature within information systems on data governance in general and data quality in particular. *Firstly*, intrinsic data quality refers to the inherent quality of the data itself, and our paper shows that data work is needed to reach intrinsic data quality by gaining a deep

understanding of data needs, which is the essence of data work, and that, in turn, elevates data curation. We show that data work is prone to frequent errors, leading to a high mistake rate, and show that researchers often rely on assumptions when handling data, even when unfamiliar with its context. *Secondly*, contextual data quality emphasizes that data quality must align with the specific task or use case, and our paper shows that data work is needed to address trust issues, which in turn can prevent additional data work downstream. Data users often lack awareness of the broader needs of others when collecting and processing data for shared use. Our paper shows that there is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential in balancing systemic measures and human effort despite structured data governance. We also show that perceptions of data quality vary significantly—what one considers crucial, such as metadata or consistent formatting, may be dismissed as unimportant by another. *Thirdly*, representational data quality as well as accessibility data quality underscore the significance of system-related factors in data quality, and our paper shows the importance of doing so through digital infrastructures and cultivating excellence in data work early on, as that, in turn, reduces later workload. We show that many researchers have never engaged with data standards, data principles, or data governance, leading to extensive and often invisible data work later in the research process and a lack of contextual data quality understanding, which could be cultivated through making that element of the data work visible. Ergo, high-quality data should be inherently reliable, suitable for the given context, well-represented, and easily accessible to users, and high-quality data work is needed to reach each of these. *Lastly*, we argue that data work must be made visible and recognized as an essential component of scientific research. Furthermore, decisions regarding data quality standards and data governance frameworks should not be confined to the backrooms of science (Barley & Bechky, 1994) or data science (Parmiggiani et al., 2022). Instead, data work should be treated as visible and integral elements of research, shaping how healthcare data is thoughtfully collected, processed, maintained, and used.

5.3 Limitations and Future Work

Despite the valuable insights from our study, several limitations must be acknowledged. *First*, our participant pool included only researchers working with research data; ergo, our findings may not generalize to other disciplines or settings. *Second*, although our sample size was small (13 participants), we mitigated this by providing rich, detailed descriptions of observed data work, supported by qualitative interview insights. Future research could develop scalable data governance frameworks for multidisciplinary research, particularly in healthcare, with an emphasis on making essential data work visible.

6 Conclusion

In this paper, we examine the data work required to establish data-driven digital infrastructure in healthcare and researchers' efforts to prepare healthcare data for its use. Our findings highlight several insights that shed light on data work for mediating data quality and data governance in a data-intensive world. Our paper illustrates the data work needed for reaching intrinsic data quality in data governance, (ii) the data work needed for reaching contextual data quality in data governance, and (iii) the data work needed for reaching representational and accessibility data quality in data governance.

The paper explores various dimensions of data quality—intrinsic, contextual, representational, and accessibility dimensions—and the essential role of data work in achieving each. We highlight the ways in which data work can be error-prone, assumption-driven, and affected by disregard for standards and regulations. Our paper emphasizes the need for understanding data context, fostering trust, and early engagement with digital infrastructure to reduce future data work. We also argue that data work, though frequently invisible, is vital for scientific research and should be recognized as such. Ultimately, decisions about data quality standards and data governance frameworks must be recognized as central to research practice, shaping how healthcare data is collected, curated, and used, rather than being relegated to invisible, backroom work.

References

Abbasi, A., Sarker, S., & Chiang, R. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17*(2). https://doi.org/10.17705/1jais.00423

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) Is Misinformation a Problem? *Perspectives on Psychological Science*, *18*(6), 1436–1463. https://doi.org/10.1177/17456916221141344

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: An analysis of the literature. *Journal of Decision Systems*, *25*(sup1), 64–75. https://doi.org/10.1080/12460125.2016.1187397

Alhassan, I., Sammon, D., & Daly, M. (2019). Critical Success Factors for Data Governance: A Theory Building Approach. *Information Systems Management*, *36*(2), 98–110. https://doi.org/10.1080/10580530.2019.1589670

Arnardottir, E. S., Islind, A. S., & Óskarsdóttir, M. (2021). The Future of Sleep Measurements: A Review and Perspective. *Sleep Medicine Clinics*, *16*(3), 447–464. https://doi.org/10.1016/j.jsmc.2021.05.004

Barley, S. R., & Bechky, B. A. (1994). In the Backrooms of Science: The Work of Technicians in Science Labs. *Work and Occupations*, *21*(1), 85–126. https://doi.org/10.1177/0730888494021001004

Benfeldt, O., Persson, J. S., & Madsen, S. (2020). Data Governance as a Collective Action Problem. *Information Systems Frontiers*, *22*(2), 299–313. https://doi.org/10.1007/s10796-019-09923-z

Bertelsen, P. S., Bossen, C., Knudsen, C., & Pedersen, A. M. (2024). Data work and practices in healthcare: A scoping review. *International Journal of Medical Informatics,* 184, 105348.

Bossen, C., Chen, Y., & Pine, K. H. (2019). The emergence of new data work occupations in healthcare: The case of medical scribes. *International Journal of Medical Informatics*, *123*, 76–83. https://doi.org/10.1016/j.ijmedinf.2019.01.001

Cruz, T. M. (2023). Data politics on the move: Intimate work from the inside of a data-driven health system. *Information, Communication & Society*, *26*(3), 496–511. https://doi.org/10.1080/1369118X.2021.1954972

Davenport, T. H., & Patil, D. (2012). Data scientist. *Harvard Business Review*, *90*(5), 70–76.

Fiske, A., Prainsack, B., & Buyx, A. (2019). Data Work: Meaning-Making in the Era of Data-Rich Medicine. *Journal of Medical Internet Research*, *21*(7), e11672. https://doi.org/10.2196/11672

Galliers, R. D., Newell, S., Shanks, G., & Topi, H. (2017). Datification and its human, organizational and societal effects. *The Journal of Strategic Information Systems*, *26*(3), 185–190. https://doi.org/10.1016/j.jsis.2017.08.002

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *Journal of the Association for Information Science and Technology*, *70*(5), 419–432. https://doi.org/10.1002/asi.24165

Gregory, R., & Muntermann, J. (2011). Theorizing in design science research: inductive versus deductive approaches. *In 32nd International Conference of Information Systems.*

Gröger, C. (2021). There is no AI without data. *Communications of the ACM,* 64(11), 98-108.

Islind, A., Vallo Hult, H., Johansson, V., Angenete, E., & Gellerstedt, M. (2021). Invisible Work Meets Visible Work: Infrastructuring from the Perspective of Patients and Healthcare Professionals. *Proceedings of the 54th Hawaii International Conference on System Sciences*, https://doi.org/10.24251/HICSS.2021.431.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, *37*(3), 101493. https://doi.org/10.1016/j.giq.2020.101493

Janssen, M., Wallenburg, I., & de Bont, A. (2016). Carving Out a Place for New Health Care Occupations: An Ethnographic Study into Job Crafting. In H. Albach, H. Meffert, A. Pinkwart, R. Reichwald, & W. von Eiff (Eds.), *Boundaryless Hospital: Rethink and Redefine Health Care Management* (pp. 119–141). Springer. https://doi.org/10.1007/978-3-662-49012-9_7

Jarvenpaa, S. L., & Essén, A. (2023). Data sustainability: Data governance in data infrastructures across technological and human generations. *Information and Organization*, *33*(1), 100449. https://doi.org/10.1016/j.infoandorg.2023.100449

Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, *28*(1), 3–16. https://doi.org/10.1016/j.jsis.2018.10.005

Khong, I., Yusuf, N. A., Nuriman, A., & Yadila, A. B. (2023). Exploring the Impact of Data Quality on Decision-Making Processes in Information Intensive Organizations. *APTISI Transactions on Management*, *7*(3), Article 3. https://doi.org/10.33050/atm.v7i3.2138

Knorr-Cetina, K. (1997). Sociality with objects: Social relations in postsocial knowledge societies. *Theory, culture & society,* 14(4), 1-30.

Ladley, J. (2019). *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program*. Academic Press.

Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, *17*(1), 39–71. https://doi.org/10.1016/j.jsis.2008.01.001

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., ... & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems,* 102549.

Müller, L. S., Nohe, C., Reiners, S., Becker, J., & Hertel, G. (2024). Adopting information systems at work: A longitudinal examination of trust dynamics, antecedents, and outcomes. *Behaviour & Information Technology*, *43*(6), 1096–1128. https://doi.org/10.1080/0144929X.2023.2196598

Pansara, R. (2023). Unraveling the Complexities of Data Governance with Strategies, Challenges, and Future Directions. *Transactions on Latest Trends in IoT*, *6*(6), 46–56.

Parmiggiani, E., & Grisot, M. (2020). Data Curation as Governance Practice. *Scandinavian Journal of Information Systems*, *32*(1). https://aisel.aisnet.org/sjis/vol32/iss1/1

Parmiggiani, E., Østerlie, T., & Almklov, P. (2022). In the Backrooms of Data Science. *Journal of the Association for Information Systems*, *23*, 139–164. https://doi.org/10.17705/1jais.00718

Petzold, B., Roggendorf, M., Rowshankish, K., & Sporleder, C. (2020). Designing data governance that delivers value. *McKinsey & Company*, *26*.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79-82.

Schmidt, K., & Bannon, L. (1992). Taking CSCW seriously. *Computer Supported Cooperative Work (CSCW)*, *1*(1), 7–40. https://doi.org/10.1007/BF00752449

Su, Z., He, L., Jariwala, S. P., Zheng, K., & Chen, Y. (2022). “What is Your Envisioned Future?”: Toward Human-AI Enrichment in Data Work of

Asthma Care. *Proc. ACM Hum.-Comput. Interact.*, *6*(CSCW2), 267:1-267:28. https://doi.org/10.1145/3555157

Steuperaert, D., Poels, G., & Devos, J. (2025). A Reference Model for Information Quality in an it Governance Context. *Information Systems Management*, 1-19.

Tracy, S. J. (2024). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact.* John Wiley & Sons.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4), 5-33.

Appendix A

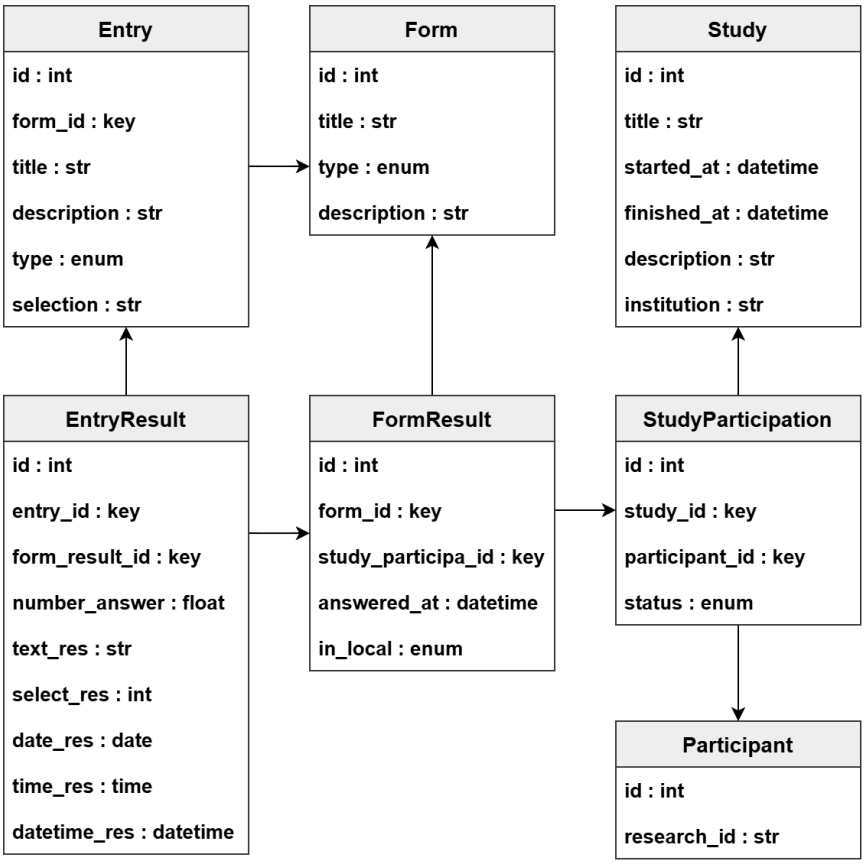


Figure A.1. Lean database structure managing which, by design, does not require additional columns or tables.

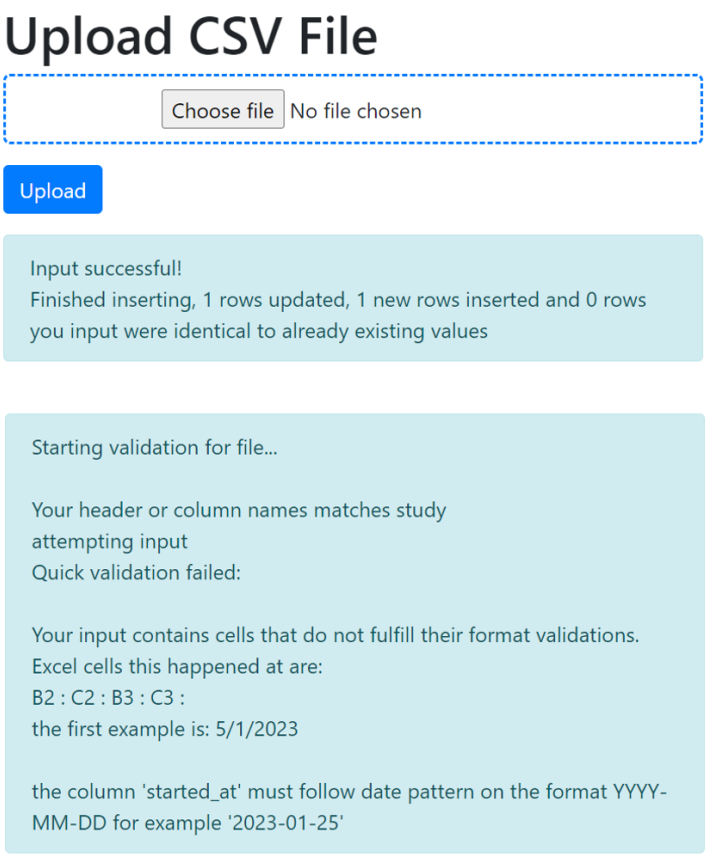


Figure A.2. The interface used in the test for data upload displays two example output scenarios. For space efficiency, both outputs are combined in a single image. The top blue box shows feedback when the uploaded file contains no violations, indicating the extent of database impact. The bottom blue box illustrates feedback when a violation is detected, providing an explanation of the issue to the user.

Interview Questions	
Identifier	**Question**
Q1	In your courses or throughout your work, have you been taught about data quality, usability, or in general, how to make data better?
Q2	In your previous work, have there been data principles or some standards set before? Can you expand on that?
Q3	Do you believe that the validation system you were using now should be used for all result data in the Sleep Revolution?
Q4	Did you think the validation system you were using was too complicated to use? If so, what part of it is complicated?
Q5	Do you believe the documentation was helpful? If so, what parts did the documentation help you with? Which parts were not helpful?
Q6	Did you find the example input CSV helpful?
Q7	How would you describe your experience of processing the data? What aspects of it did you find nice/annoying/frustrating etc?
Q8	How important is it to you that the processes around data are simple and straightforward?
Q9	If you were in control of ensuring data quality, would you prefer a different system for validations?
Q10	Were there any validations in the system that you encountered that you think might be unnecessary?
Q11	Of the files you uploaded, are some files or parts of them still unclear to you?
Q12	How do you believe the result data and the raw data should be structured? That is, after everything has been collected, processed, and validated, how should the final data be presented so it can be used by others, in your opinion?
Q13	What do you believe are important things after data collection has finished?
Q14	Do you believe you could continue uploading data sources after this session?
Q15	Data is troublesome for projects in general when it comes to understanding it, using it, it often requires a lot of processing, contains mistakes, is not complete, and more. Why do you believe that is?
Q16	After a research grant period has finished, do you believe outside researchers who were not part of any of the data collections in that grant should be granted access to the data that was collected? (Can you elaborate on why?)
Q17	Anything to add?

Table A.1. The 17 interview questions we asked participants.

Opinions: Strongly Disagree (1) to Strongly Agree (5)	
Identifier	**Question**
QA1	I believe manual validation (such as agreeing on standards and following them) is better in research than automatic validation (such as software or websites validating the data).
QA2	I believe instead of using automatic validation (such as software or websites validating the data), data users should get the data sources as they were collected, that is, data without any processing.
QA3	I believe I can trust data that has gone through automatic validation (such as software or websites validating the data) is better than data that has not.
QA4	When I receive data that I did not collect, I do not trust it.
QA5	When I receive data I did not collect, I have often trouble understanding what parameters/variable actually is. For example, if they are named with abbreviation or have general names.
QA6	When I receive data I did not collect, I often find descriptions and details (metadata) on the data sources are missing or difficult to find.
QA7	I often find it difficult to join data sources. For example, because of missing identifiers or different formats.
Opinions: Not Important (1) to Very Important (5)	
Identifier	**Question**
QB1	How important is it to you that data has an existing research identifier?
QB2	How important is it to you that data does not have duplicates when you receive it for analysis?
QB3	How important is it to you that data has consistent format? For example if dates are not written as DD/MM/YY and as YYYY-MM-DD or research identifiers are not written sometimes as 'SRI0123' and sometimes 'sri0123'?
QB4	How important is it to you that you know which dataset the data source belongs to? For example, if this data was collected in one study or another?
Time Spent: Percentage	
Identifier	**Question**
QC1	On average how much percentage of your work goes into validating data? This includes making sure your data has correct formats.
QC2	On average how much percentage of your work goes into converting data? By converting, we mean to convert it from one data format (XML, CSV, JSON, TXT) to another in order to, for example, being able to open the data in some environment.
QC3	On average how much percentage of your work goes into cleaning data? By cleaning, we mean to finding errors, make sure formats are the same (for example, all dates written with same format), applying ownership if they are missing, etc.
QC4	On average how much percentage of your work goes into joining data? By joining, we mean to join data sources together so they can be worked on. For example, taking data from a text file and inputting it into an Excel file so the results can be compared, or joining two different excel sheets together.
QC5	On average how much of your work goes into reading, documenting, and writing?
QC6	On average how much percentage of your work goes into meetings?
QC7	On average how much percentage of your work goes into other things? This is the final percentage question so please make sure they add to 100%.

Table A.2. Many of the survey questions we asked.

Appendix D

Publication IV

Let the System Handle It: Simplifying Data Governance Using Automation

Bjarki Freyr Sveinbjarnarson
Reykjavík University
bjarkis@ru.is

April 3, 2025

Abstract

Ensuring high-quality, reusable data requires more than storage and access—it demands sustainable, structured data governance. However, maintaining such governance is especially challenging in resource-limited environments, where data quality often suffers due to inconsistent formatting, missing metadata, and lack of validation. This paper investigates how automated validation systems can support simplified and sustainable data governance by reducing human error, standardizing inputs, and reinforcing best practices at the point of data entry.

We present findings from a mixed-methods study involving healthcare researchers who interacted with a validation system embedded in a broader digital infrastructure. Through system interaction, surveys, and interviews, we observed common challenges in data preparation, including semantic ambiguity, metadata neglect, and user reliance on trial-and-error. Our results demonstrate that while participants requested greater flexibility, automation played a critical role in upholding consistency, enabling trust, and guiding users toward better data stewardship.

By enforcing structural constraints, minimizing format variability, and embedding governance into everyday workflows, our system reduced friction and made participation in data governance easier. We conclude that automated validation systems, when carefully integrated into digital infrastructures, offer a scalable and generalizable path toward sustainable data governance across multidisciplinary research environments.

1 Introduction

Data has become a central resource in the modern world, yet the extensive effort required to make it usable is often overlooked [1, 2]. This issue is particularly evident in healthcare, where data-driven decisions, precision medicine, and the integration of digital infrastructure depend heavily on the quality and usability of data [3, 4]. However, data does not simply exist in usable form; it must be produced, collected, cleaned, curated, interpreted, and integrated across systems [5, 6]. These tasks fall under the broader concept of *data work* [2], which encompasses the technological, analytical, and emotional labor needed to transform raw data into meaningful information.

Despite the promise of digital infrastructures in improving service quality, their benefits are often accompanied by hidden, labor-intensive tasks. The invisible work [7] required to ensure data integrity and usability is increasingly recognized as essential to modern research, particularly in healthcare. As datasets grow larger and more heterogeneous, and the goals of individual studies diverge from collective research priorities, maintaining data quality becomes both more critical and more difficult [8]. The development of robust, scalable digital infrastructures must therefore be accompanied by appropriate data governance mechanisms.

In this paper, we investigate the socio-technical aspects of data governance in a resource-limited research setting. Drawing from a large-scale sleep research project, we explore how automated validation systems and lean database design can improve data quality while minimizing manual labor. We examine the capacity of researchers to engage with such systems, the challenges they encounter, and the human practices that affect data quality.

We position this work within the broader context of dynamic data spaces and automated data governance, which have emerged in response to increasing demands for data interoperability, accessibility, and compliance [9–11]. Our research is guided by the following question:

> *How can automated validation systems support simplified and sustainable data governance?*

2 Related Work

2.1 Data Governance in Research Environments

Data governance encompasses the roles, policies, and processes that ensure data is reliable, secure, and useful throughout its lifecycle [12]. Core activities include assigning responsibilities, defining data standards, and monitoring compliance [13]. In complex organizations, these structures are difficult to implement and sustain [14, 15], particularly when priorities differ across departments and human effort is limited.

Governance frameworks are often seen as abstract or bureaucratic, leading to resistance [16]. Without clearly defined roles or sufficient automation, governance tasks tend to be unevenly distributed, resulting in inconsistent adherence to standards and increased downstream data work [17]. Ambiguities in ownership and unclear policies can diminish trust in the data and amplify inefficiencies [12].

Researchers have emphasized the need for a balanced governance model—one that provides clear responsibilities and rules but also incorporates automation to reduce the human burden [18]. In healthcare, where data is often personal, sensitive, and heterogeneous, this balance is particularly crucial. As Alhassan et al. [12] highlight, successful governance depends on six interconnected components: cross-functional collaboration, structured frameworks, data as a strategic asset, decision rights, defined policies, and monitoring mechanisms.

2.2 Automation and Socio-Technical Challenges

A major trend in recent governance literature is the call for automation in environments where manual oversight is infeasible [19]. Automated systems can enforce compliance, validate input, and reduce the risk of human error. However, these systems also introduce challenges related to rigidity, transparency, and usability. Without careful design, automation can exclude users or encourage workarounds [16, 18].

Automated governance must contend with the socio-technical realities of its context. Individuals often opt for the least burdensome path, even if it results in long-term inefficiencies [14]. A validation system, for example, may prevent errors but can also generate frustration if it limits flexibility or fails to provide clear feedback. Designing systems that enforce standards without alienating users requires understanding not just the technical rules, but also the lived experience of those engaging with the data [17].

2.3 Gaps in Current Research

Although the literature offers rich insights into the principles and pitfalls of data governance, it offers limited guidance for resource-constrained settings. Most frameworks assume the presence of governance officers, dedicated data stewards, and sustained funding—all luxuries often unavailable in academic or healthcare research environments [14, 19]. There is also little empirical work examining how automated validation systems are received by end users, or how user practices shape the effectiveness of these systems.

This paper contributes to this gap by empirically studying how a digital infrastructure with embedded validation rules operates in a real-world research setting. We explore how researchers navigate the system, how errors are introduced and resolved, and what socio-technical practices enable or hinder effective governance. Our findings contribute to the development of governance frameworks that are both scalable and adaptive to limited-resource environments.

3 Methods

3.1 Research Methodology

This study adopts the Action Design Research (ADR) methodology, originally proposed by Sein et al. [20] and further elaborated by Mullarkey et al. [21]. ADR integrates principles from both Action Research (AR), as outlined by Susman and Evered [22], and Design Science Research (DSR), as conceptualized by Hevner et al. [23]. This methodological framework is well-suited for tackling complex socio-technical challenges by iteratively designing, developing, and evaluating artifacts within their real-world contexts.

ADR consists of four stages:

- **Problem Formulation**: Defining the research problem and formulating relevant questions grounded in practical challenges.
- **Building, Intervention, and Evaluation (BIE)**: Iterative creation and testing of artifacts in real-world contexts, refining them through empirical use.
- **Reflection and Learning**: Drawing insights from interventions, examining contextual influences and refining theory.
- **Formalization of Learning**: Generalizing insights into design principles and broader theoretical contributions.

By engaging in iterative cycles of artifact development and evaluation, ADR ensures practical utility and theoretical relevance, particularly in research on information systems, digital infrastructure, and data governance [24].

3.2 The Digital Infrastructure

To manage the project's large-scale and heterogeneous datasets, we designed and iteratively developed a relational database framework aimed at simplifying data work across disciplines. The database consists of seven key tables—`Study`, `Participant`, `StudyParticipation`, `Form`, `Entry`, `FormResult`, and `EntryResult`—with each table structured to ensure consistency and modularity. This schema links all data to individual participants and their respective studies through the `StudyParticipation` table, which serves as a universal primary key.

The modular architecture, integrated into the broader digital infrastructure, supports seamless addition of new data sources without expanding the number of tables. Ancillary fields enable metadata integration, and a built-in validation mechanism enforces data quality standards at the point of entry. This validation system evolved into the Data Integrity Assurance System (DIAS), a core component of the data governance model described in the results section. A figure illustrating the database structure is shown in Figure 2.

3.3 Development of the Data Governance Model

The data governance framework was not designed in a single phase but emerged through trial and error over the course of the project. Initially, few pipelines and a basic database structure were in place. As data collection increased and data users encountered recurring problems and inconsistencies, the system was iteratively refined. Feedback loops from researchers informed continuous improvements to DIAS, enabling the enforcement of evolving data standards through automation.

Data pipelines were created by individuals assuming the role of *data converters*, who developed transformation logic for structured sources. If no pipeline existed for a particular source, *data collectors* coordinated with data converters to create one. Once validated through DIAS, the data was exported to a shared folder system via an automated export functionality. This ecosystem enabled reuse, consistency, and traceable validation across all data sources. Examples of feedback shown to users during validation are included in Figure 3.

3.4 Study Design

We employed a mixed-methods design combining feasibility testing, surveys, and semi-structured interviews. The goal was to evaluate whether researchers with no prior training could perform key data preparation tasks using our validation system, and to understand user behavior, preferences, and the data work required to ensure data quality.

Feasibility testing was used not only as a usability evaluation but also as a test of system applicability—whether participants were capable of executing key steps in the data preparation and upload process, and how they approached challenges. Each participant was given a ZIP file containing: (i) example input CSV files, (ii) a README file, (iii) detailed documentation, and (iv) a small test dataset. They were instructed to split and transform the dataset into six predefined files, which they then uploaded through the validation interface. A schematic of the process participants went through is shown in Figure 4.

The web-based interface featured a minimal design: one upload function and one feedback output, designed to reduce bias. During upload, DIAS either approved or rejected files with clear error messages.

Participant interactions with the validation system were observed. Key performance indicators included task completion time, number of upload attempts, error types, and reliance on documentation. A full list of observed metrics is provided in **Appendix A**.

After the test, participants completed a survey on data practices, preferences, and perceptions of the system. This was followed by a semi-structured interview exploring their experience, views on data governance, and how they define or engage in data work. The interviews were transcribed from audio recordings and translated into English if conducted in another language. Thematic analysis was conducted by the first author. Ambiguous responses (e.g., “I don’t know,” “maybe”) were excluded from coding but retained as context.

3.5 Participants and Recruitment

A total of 13 participants were recruited using convenience sampling via professional networks and researcher group chats. All participants had prior experience working with research data. Their academic backgrounds included health sciences, computer science, neuroscience, exercise physiology, and related disciplines.

Participation was contingent on English proficiency, data experience, and lack of prior involvement in developing the system. Each participant used their own computer with standard spreadsheet software. Sessions were conducted either remotely or in person. Screen sharing was used during remote sessions for observation.

3.6 Ethics and Consent

Although ethical approval was not required under Icelandic regulations (Act No. 90/2018 on Data Protection), all participants provided informed consent. They were informed about the nature of the study, data usage, and their right to withdraw at any point without consequence. Participants were also told they could request that their data be deleted at any time. Consent was collected via a Google Form that explained:

> **Test Overview:** This test will take approximately one hour. Participants download data, process it according to documentation, upload the result, complete a survey, and join a short interview.
> **Data Collection:** Names (for internal tracking only), background information, screen recordings (optional), and survey/interview answers were collected anonymously.
> **Rights:** Participation is voluntary. Participants may skip questions, decline recordings, or withdraw at any time. They may revisit the consent form to update their decision.

4 Results

4.1 Perceptions and Practices of Data Governance

Lack of Shared Standards and Guidelines

Our findings reveal a significant absence of formal data governance education and shared standards among participants. The majority of interviewees reported that they had received little to no training related to data quality, documentation, or usable data structures. Instead, participants frequently relied on self-teaching or informal guidance from colleagues or supervisors (Interview Q1, Q2). When asked about previous experience with standards, only a minority could point to even basic conventions, such as variable naming rules or folder structures. Most participants either followed idiosyncratic personal approaches or reported complete absence of any guidelines.

This absence of shared principles was further reflected in the low consistency of responses regarding what constitutes good data quality. For instance, Interview Q13 highlighted that several participants did not know how to document data appropriately, even if they regularly worked with it. Many also noted that data quality issues often stemmed from "human error," "lack of documentation," or from researchers "not thinking long-term" (Interview Q15). This aligns with the broader literature, which emphasizes that socio-technical complexity and human coordination challenges are central to data governance issues [12, 14, 25].

Disagreement on Data Quality Priorities

The survey data supported this qualitative insight. When participants were asked how important they perceived different dimensions of data quality—such as consistent formats, unique research identifiers, and the absence of duplicates—responses varied considerably (Survey: Data Quality Priorities). For example, while most participants rated research identifiers and dataset provenance (i.e., knowing which dataset a data source belonged to) as highly important, opinions were more mixed on consistent formatting and duplication. One participant even rated consistent formatting as unimportant while others considered it essential.

This divergence illustrates a critical issue: even within a shared project, data governance is hindered by divergent interpretations of what quality means and how it should be achieved. This further underscores the need for enforceable standards and systems that promote shared understanding and practices, particularly when data is expected to be reused across different researchers and disciplines.

4.2 Automation and Validation in Data Governance

Trust in Automated Validation

Survey responses and interviews indicate a generally positive attitude toward the automated validation system embedded in the digital infrastructure. Most participants expressed trust in the system and supported its use for research data validation (Survey: trust in validation system; Interview Q3). Participants emphasized its usefulness in catching formatting errors and duplicates—issues that are often overlooked in manual processes (Interview Q3 – P1, P6, P9, P11). As Participant 6 noted, "I think we need to have it to unify all of the data," and others echoed that it promoted trust and consistency across shared data sources.

However, not all participants were equally enthusiastic. One participant (Interview Q3 – P3) raised a critical concern about the restrictiveness of the system potentially limiting creativity in research, particularly when experimental setups required unconventional formats. They emphasized the need for flexibility in how data is structured and described, arguing that rigid validation rules could become an "unnecessary hurdle" when data collection processes evolve during a project.

Desire for Hybrid Systems

Several participants expressed a preference for hybrid approaches that combine automation with flexibility. In both the survey and interviews, participants suggested the system could be improved by guiding users through corrections or automatically fixing minor issues such as date formatting (Survey: "should the system correct data"; Interview Q4, Q10). While some were skeptical about over-automation, others believed that more intelligent feedback or correction support could reduce user frustration without compromising the integrity of the validation process.

This view was supported by feedback that strict rules were sometimes perceived as arbitrary or overly detailed. For instance, some participants questioned why form titles needed to be unique or why entry descriptions had to follow specific formats (Interview Q10, Q11). Others suggested that the system should allow a more flexible interface for entering structured metadata or automate routine structural checks (Interview Q9).

Observed Use of the System for Passive Validation

During the feasibility testing, we observed a pattern in how participants approached data preparation and validation. Rather than thoroughly reviewing or validating their own data prior to upload,

most participants appeared to rely on the validation system itself to catch and report errors. This behavior—treating the upload process as a trial-and-error loop—was observed in 11 out of 13 participants, with multiple re-upload attempts recorded per file (Feasibility Test Observations).

While this behavior may reduce the likelihood of introducing new errors through manual corrections, it also suggests a reduced level of user engagement with the structure and meaning of the data. Participants were, in effect, outsourcing validation entirely to the system. This raises important questions for discussion: does this behavior indicate trust and efficiency, or does it risk users missing more context-sensitive errors that automated systems cannot catch? This tension between automation and human oversight is central to future development of effective data governance models.

4.3 User Behavior and System Usability

Documentation and Learning Curve

User interaction with the provided documentation revealed a consistent trend: participants showed minimal reliance on written guidance. Out of the 13 participants, only 2 actively and sufficiently used the documentation to guide their actions during the feasibility test (Observation: "Used Documentation"). An additional 2 participants opened the documentation briefly but did not rely on it to complete their tasks. The remaining participants either ignored the documentation entirely or preferred other resources such as example files and system feedback.

Interview responses support this trend. Most participants reported preferring to ask questions rather than consult the documentation (Interview Q5). However, as part of the study protocol, no specific solutions were given to participants when they asked for help. Instead, they were directed to the system's feedback and the materials they had already received. Despite this, many participants attempted to solve issues through guesswork or assumptions—strategies that sometimes led to repeated upload failures before eventual success (Observation: Upload Attempts).

Rather than engaging proactively with the system's guidance mechanisms, participants often relied on trial-and-error, a behavior that underscores a reactive rather than reflective approach to validation. This pattern of minimal engagement with formal instructions may reflect either a general reluctance to use documentation or a perception that the documentation was unnecessary for the task at hand.

Impact of Interface and Example Files

The user interface was intentionally designed to be minimalistic, consisting of only an upload button and a feedback box that returned success or failure messages. Most participants were able to complete their tasks with limited external guidance, relying primarily on the system's immediate feedback to identify and correct errors (Observation: Data Result Attempts, Total Attempts).

Participants consistently praised the example CSV input sheets for their clarity and usefulness (Interview Q6). These example files were cited as critical for understanding both the structure of the data and the formatting expectations of the validation system. The clarity and immediate applicability of these files appeared to reduce confusion and were preferred over the longer-form documentation by most users.

However, this reliance on example files may also indicate a weakness in how documentation is typically used in data validation systems. The small number of participants who engaged with documentation limits the strength of this conclusion, but it suggests that providing working examples may be more impactful than written instructions in similar research environments.

Overall, the findings indicate that users often bypass formal instructions in favor of tangible examples and reactive learning through error correction. While this approach enabled task completion, it also reflects a surface-level interaction with the system—one that may fail to instill deeper understanding of data validation principles or the structure of the digital infrastructure.

4.4 Errors and Challenges in Data Preparation

Prevalence of Common Mistakes

Despite being provided with clean input data, all 13 participants introduced at least one error during the feasibility test, as recorded by the embedded validation system. These errors spanned a wide range of issues, including incorrect date formatting, mislabeled or incomplete metadata, entry mismatches, and improper data types or selection values. Some participants mistakenly deleted rows they perceived as duplicates or entered ambiguous descriptions that were difficult to interpret later.

The number of extra upload attempts needed to complete all of the tasks ranged from 1 to 10, with a median of 4 extra attempts, underscoring the difficulty of entering valid structured data even under guided conditions.

Reactive vs. Proactive Behavior

Observation during the feasibility test and interview responses suggest that most participants approached the task with a reactive mindset, focused more on "getting the system to accept the data" than on ensuring the long-term usability or clarity of their submissions. Rather than carefully reviewing their entries before uploading, participants often relied on trial-and-error, adjusting their files incrementally based on system feedback (Interview Q7).

This behavior was also visible in how users engaged with validation: corrections were often aimed at satisfying the minimum required format, rather than enhancing the overall interpretability or documentation of the data. As a result, while the system helped catch technical errors, deeper semantic issues—such as vague descriptions or non-standard terminology—often remained unresolved.

Feedback and Frustration

While most participants completed the task without major complications, some expressed frustration when feedback messages did not immediately help them identify the source of their errors. For instance, Participant 2 noted a desire for more precise error messages (Interview Q10), while Participant 8 suggested improving the feedback and offering automated correction options (Interview Q4). Others were confused about formatting conventions, such as quote usage in selection values or date input, which led to multiple failed attempts.

However, these concerns were relatively limited, and most participants were ultimately able to use the system successfully. For example, Participant 10 noted that despite having no prior training, the feedback was clear enough to complete the task unaided. These mixed experiences highlight the importance of both effective system feedback and user education in achieving data quality goals.

4.5 Trust, Metadata, and External Data Use

Low Trust in External Data

A recurring theme across both the survey and interviews was the general lack of trust participants had in using data they did not collect themselves. The survey responses showed strongly that participants did not trust data collected by others (Survey Q30). This was supported by several participants in the interviews, who described issues related to missing context, ambiguous formats, and inconsistent documentation (Interview Q15).

Participants identified specific barriers to trust, including unclear variable names, lack of standardization, and uncertainty about how the data was originally structured. One participant noted, "Ambiguous inputs, use of abbreviations, poor documentation..." (Interview Q15), while another added: "Temporary fixes instead of preventing errors... Different formats being used, for dates for example" (Interview Q15). These concerns highlight how difficulties in understanding external data can significantly hinder its reuse.

Metadata Importance

Survey responses emphasized the critical role of metadata. Participants consistently rated "knowing which dataset the data source belongs to" and "having consistent research identifiers" as very important (Survey Q27–30). Without this information, participants reported that interpreting the data or using it collaboratively became difficult or unreliable.

Interview responses reinforced this perspective. Several participants described how missing or unclear metadata made it difficult to understand what each parameter represented or where the data originated from (Interview Q11, Q13). One participant reflected, "Write everything down... what you have done... even the obvious things" (Interview Q13), underlining how proper metadata serves as a necessary bridge for future interpretation and reuse.

Some participants suggested structural improvements, such as enforcing metadata fields by making them required in the upload process. This was seen as a way to improve long-term usability and ensure that future users could rely on standardized, interpretable data. Suggestions like these support the idea that metadata should not only be encouraged but systematically enforced within data governance frameworks.

Implications for Data Governance

The findings indicate that low trust in external data and inadequate metadata are not just occasional problems—they are core obstacles to collaboration and data reuse. Despite the potential of open science and data sharing, participants expressed hesitation when documentation and context were missing. As one participant put it: "When people start working on something, they don't think enough about the future" (Interview Q15).

For data governance, these results emphasize the need to embed metadata requirements into validation workflows. Systems must ensure that key fields like study origin, parameter descriptions, and formatting standards are consistently included. Without such measures, the value of shared data is undermined, and the burden of data interpretation shifts disproportionately to future users—often without the context needed to make sense of it.

4.6 Effort and Time Spent on Data Work

High Cost of Data Cleaning and Joining

The survey responses showed a high proportion of participant time is spent on tasks related to making data usable, with some respondents indicating that up to 80% of their time is allocated to data cleaning, conversion, and joining (Survey: Time Allocation Questions). This was especially common among participants with computational backgrounds who handled raw or complex data formats.

Interview data reinforced these findings. Participants frequently pointed to format inconsistencies, missing identifiers, and challenges in merging data as significant barriers. One participant noted, "There are so many data formats. Every company has their own. The built-in conversion tools are poor... you need to rely on others" (Interview Q15 - P12). Another shared, "It is expensive to make things better for data quality" and highlighted being "limited by older systems" (Interview Q15 - P8).

Desire for Simpler Workflows

Participants expressed a strong desire for streamlined and user-friendly workflows, particularly after data collection. They underscored the value of clear, unified formats and documentation to reduce overhead in preparing data for reuse.

As one participant explained, "The data should be available... validated... with secure storage and backups" (Interview Q13 - P1). Another emphasized the need for documentation that supports downstream use: "Write everything down, what you have done, even the obvious things... no one can ask 10 years from now" (Interview Q13 - P12).

Participants also commented on the burden of complex processes, especially for data they did not collect themselves. One participant stressed, "It is really important [that processes are simple], especially for data you do not collect, and metadata" (Interview Q8 - P1). This points to a clear expectation for systems that reduce manual overhead while ensuring consistency and long-term usability.

Preferences for Future Data Infrastructure

Storage and Format Expectations

Participants expressed a range of preferences for how data should be stored and structured after collection and validation. Several emphasized the importance of having searchable and structured formats that support diverse workflows. A common preference was for databases with filtering capabilities or structured folders with clearly organized files.

One participant shared, "Database, have filters and applications to get the data you want, even possibly with parameters" (Interview Q12 – P1), while another favored "a folder structure with tools to browse the data, using formats like JSON or TXT for easier viewing" (Interview Q12 – P13). This illustrates a desire for infrastructure that balances flexibility with ease of access and readability.

Calls for Structured Open Access

Participants overwhelmingly supported the principles of open science and external researcher access to collected data, provided that appropriate safeguards are in place. Several participants stressed

the importance of comprehensive documentation, anonymization, and proper crediting of original data collectors.

"Redoing analysis is then possible and is becoming necessary. There is a lot of potential since few people can actually analyze everything" (Interview Q16 – P1), explained one participant in support of post-project reuse. Another noted the ethical caveats, stating, "Only if the original researchers have finished their work and the ethical committee is fine with it" (Interview Q16 – P11).

This support for data reuse reflects broader trends in the research community toward data democratization, but also underscores the continued challenges in achieving usable, ethical, and well-documented open datasets.

Final Reflections on Governance Systems

Support for Continued Use

The majority of participants reported feeling confident that they could continue to use the validation system independently after the session, indicating that the learning curve was manageable with brief exposure. Responses such as "Yes. I think so" (Interview Q14 – P13) and "Yeah. Maybe I would need to ask every now and then to be sure" (Interview Q14 – P2) highlight growing user confidence and suggest that the system has potential for adoption in broader research contexts with minimal training.

Need for Human-Centric Systems

Participants reflected on the broader challenges of data governance and highlighted human behavior and education gaps as critical barriers. Many pointed to a lack of training, awareness, and shared standards as major obstacles to reliable data handling and collaboration. As one participant explained, "It is because we do not know how to do the documentation, structure, even if we work a lot with data" (Interview Q15 – P4). Another noted the tendency toward short-term solutions: "Temporary fixes instead of preventing errors. Limited by older systems or solutions that have been used for a long time" (Interview Q15 – P8).

Emphasis on Clarity and Shared Responsibility

There was also strong emphasis on the need for clear systems and shared responsibility in maintaining data quality. Participants consistently described documentation as essential and called for better structures to prevent errors and miscommunication. One participant stated, "Write everything down, what you have done, especially your choices. Write why you made changes. Write the obvious things even, as no one can ask 10 years from now" (Interview Q13 – P12). These reflections suggest that future governance systems must account for not only technical infrastructure but also social, educational, and organizational factors to be successful.

Resulting Data Governance Design

Overview of the Governance Ecosystem

The resulting data governance model, illustrated in Figure 1, reflects an automated and adaptive system designed to streamline data workflows, enforce consistency, and minimize human error. It operates across four main user roles: *Data Collectors*, *Data Converters*, *Data Users*, and *DIAS*

Developers. Each role contributes to and benefits from a closed feedback loop in which data quality improvements propagate across all data sources.

- **Data Collectors** are responsible for collecting new data. They may only input data into the system when a validated pipeline already exists.
- **Data Converters** are tasked with creating conversion pipelines for new structured data sources. These pipelines ensure that inputs are formatted into DIAS-accepted CSV files (formresult format).
- **DIAS Developers** maintain and improve the Data Integrity Assurance System (DIAS) by responding to feedback, adding new validation rules, and generalizing policies to catch previously overlooked errors.
- **Data Users** access the processed and validated data from a shared, read-only folder structure. They can report issues that prompt broader improvements across the system.

Automation, Feedback, and Generalizability

A key feature of the system is its high level of automation. Once converted, data is uploaded to the DIAS system, which applies validation rules and either rejects invalid files or ingests valid ones into the database. An export module then automatically transforms validated data into a hierarchical folder structure.

Errors flagged by DIAS, if confirmed, lead to new or refined policies. The system is thus self-reinforcing: errors identified in one dataset can lead to improvements that benefit all other datasets. This collective improvement mechanism was also observed during the feasibility test (see earlier subsections), where validation feedback was used not just to correct individual files but to generalize better rule enforcement.

Shared Data Environment and Trust Propagation

Validated data is made available to all users in a consistent folder structure. Its hierarchy supports both *raw* and *result* data, organized by study and data source type. This design improves accessibility while reinforcing data integrity. Because all users interact with the same shared version of the dataset, a single detected error—and its subsequent fix—improves the dataset for all users.

Trust in the system is cultivated through transparency and reliability. As shown in the survey and interviews, participants expressed greater trust in data that had passed through the validation system (Interview Q3; Survey QX). This trust supports compliance, since users are encouraged to engage with DIAS in order to benefit from lower downstream workload and more reliable data.

Simplicity and Low Barrier of Entry

Another design principle is the system's simplicity. Users are not required to understand internal rules or standards to engage with it effectively. Just by participating—collecting, converting, using data, or reporting issues—they become actors in the data governance ecosystem. This removes traditional barriers to governance participation and highlights the system's socio-technical adaptability.

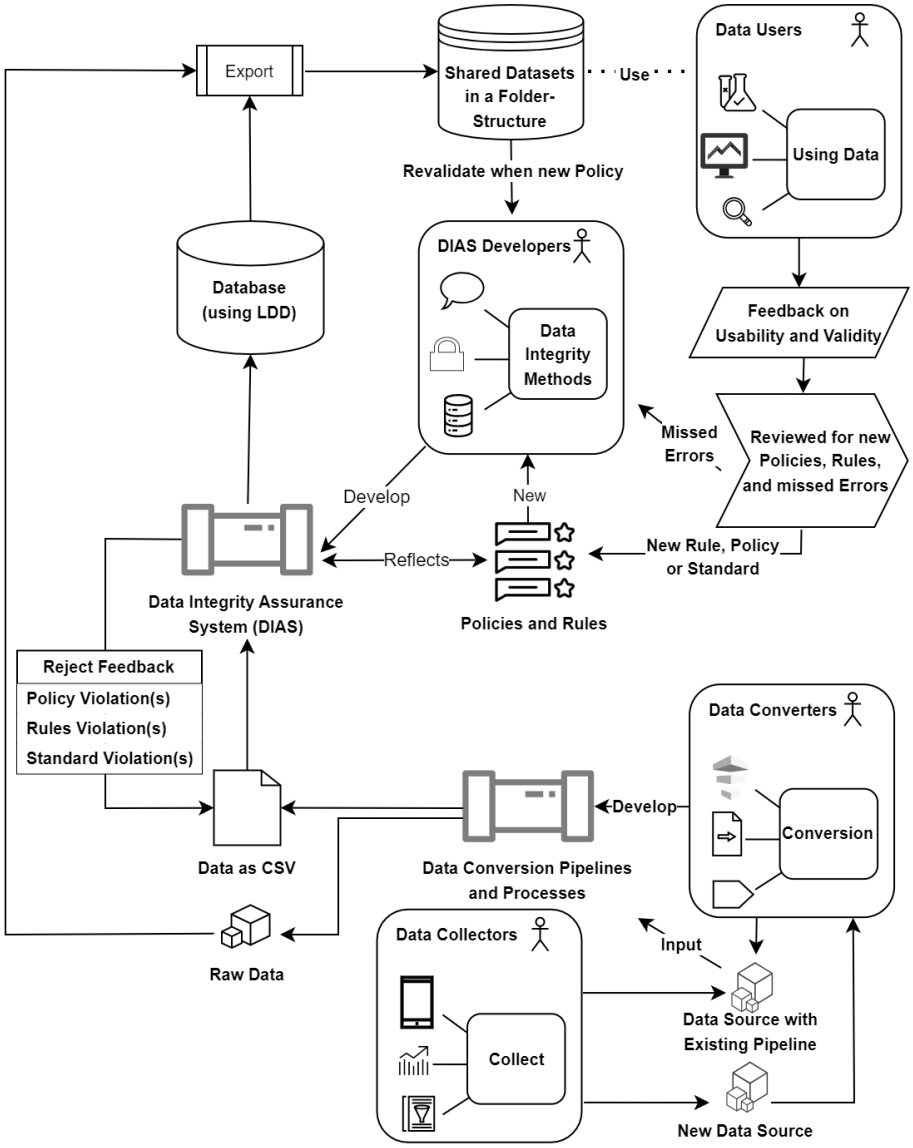


Figure 1: Overview of the resulting data governance model (Sleep Revolution)

5 Discussion

This study set out to address the research question: *How can automated validation systems support simplified and sustainable data governance?* Drawing on the results from feasibility testing, interviews, and surveys, our findings illustrate how automation, simplicity, and structured feedback

mechanisms can support a socio-technical governance model that reduces friction, minimizes errors, and promotes active data stewardship.

Reducing Governance Complexity through Automation

Our digital infrastructure simplified data governance through a homogeneous database schema and standardized input format. By enforcing a single file format (CSV) and embedding validation at the point of entry, we significantly reduced the variability that typically complicates governance processes. This aligns with calls in the literature for standardized input mechanisms to support interoperability and compliance [14, 26, 27].

Importantly, even though the input data used in our study was error-free and simple by design, all participants introduced errors. This underscores a central insight: human error is not the result of system complexity but a natural outcome of manual data work [28]. Rather than criticize the system for these mistakes, this highlights the need to reduce manual engagement where possible. Our findings support a core governance principle: automation should not only enforce standards but also actively reduce opportunities for error.

Balancing Simplicity and Flexibility in Socio-Technical Systems

Participants requested increased flexibility in naming conventions, field constraints, or metadata requirements (Interview Q3, Q9). However, such flexibility could undermine the purpose of validation systems, which is to enforce consistency and avoid ambiguity. As seen in our study, even seemingly minor relaxations—such as allowing duplicate form names—could lead to cascading issues in downstream data use.

These tensions reflect a classic socio-technical challenge: user preferences often clash with system constraints designed to uphold long-term quality [17, 25]. While flexibility may provide short-term satisfaction, it risks sacrificing traceability and trust. Validation systems, therefore, must navigate a fine line—providing support and feedback without ceding control over key standards.

Feedback, Documentation, and Learning-by-Doing

Participants demonstrated low engagement with documentation, preferring to explore the system or request help. Despite this, nearly all completed the tasks with few external prompts, relying on system feedback. This reinforces previous findings that users often disregard documentation unless blocked [29].

The system's feedback loop became a core learning mechanism—what we refer to as *feedback-as-governance*. Rather than treat feedback as corrective alone, we observed that participants used it to guide learning. This real-time correction and reinforcement helped internalize standards without needing formal training, showing how technical design can encode social learning [12, 14].

Trust, Validation, and the Role of Human Oversight

Participants used the validation system reactively—focusing on "getting it accepted" rather than reflecting on data quality. While this behavior suggests surface-level compliance, the system's design ensured that even minimal engagement still resulted in better metadata and formatting. However, this shift toward reactive validation also reveals a risk: users may increasingly offload responsibility to the system, reducing human scrutiny.

Some participants wished the system would "just fix the data" (Survey QX, Interview Q10). This exposes a tension between automation and accountability. Automatic corrections, while convenient, risk introducing semantic errors (e.g., auto-fixing a misspelled name or wrong date). We intentionally avoided automatic edits to preserve data integrity. This design tradeoff emphasizes a key governance principle: systems must preserve the original meaning and traceability of data [28, 30].

Socio-Technical Variation and Governance Alignment

Participants held diverging views on what constitutes "good" data. Some prioritized metadata completeness, others emphasized structural formatting. This divergence reflects the inherently socio-technical nature of governance: while standards can be enforced technically, user assumptions and priorities still vary widely [14, 17, 25].

Our results challenge the notion that governance failures stem primarily from skill gaps. Rather, they suggest that inconsistencies arise from differing priorities and lack of shared understanding. As such, automated validation systems must not only check conformance but also nudge users toward better practices, reinforcing key governance concepts over time [27, 31].

Governance through Interaction and Incentives

Our system demonstrates a sociotechnical design principle: simply engaging with validated data becomes an act of governance. This reframing reduces the burden on users while still promoting standard adherence. As errors are discovered and fed back into the system, improvements cascade—benefiting all users. This cycle of interaction, correction, and reuse turns feedback into evolving policy.

Notably, adoption was not driven by compliance pressure but by perceived convenience. Users realized that using the system saved time, especially when merging data across sources. This aligns with arguments in governance literature that reducing friction increases participation [32, 33].

Toward Simplified, Generalizable Governance

Our governance model enforced structure at input and automated consistency checks through DIAS. While designed for participant-linked data, the underlying architecture can be adapted to support more diverse ownership models. As discussed in the results, minor modifications could allow the same database design to work with different data structures, ownership types, or research domains.

Participants found the system adaptable and helpful despite its rigidity, suggesting that simplicity does not limit scalability. Rather, simplicity in structure and process often facilitates broader adoption and sustainability [34, 35].

Design Recommendations and Future Directions

Our findings suggest several design principles for data governance systems:

- **Enforce metadata completeness**: Require key fields (e.g., description, institution, variable type) to reduce ambiguity.
- **Avoid automatic data corrections**: Preserve user accountability and data integrity.
- **Support learning-by-doing**: Provide clear feedback and minimalistic interfaces rather than rely on formal documentation.
- **Align convenience with governance**: Reduce friction to promote voluntary participation.

Future work should explore how these design principles generalize across disciplines and institutions, particularly in collaborative projects with heterogeneous data sources. Further research could also examine long-term system use, identifying how sustained interaction influences governance culture and data quality.

In addition, several technical and structural areas warrant further investigation:

- **Ownership Model Constraints**: The current schema supports only `Study`, `Participant`, and `StudyParticipation` ownership levels. A more generalizable model (e.g., with `ownership type` and `ownership value` fields) could support more complex data relationships.
- **Performance and Scalability**: Although the lean table structure supports simplicity, performance limitations may arise at scale. A translation layer could generate more traditional table-per-dataset outputs post-validation to support large-scale use cases.
- **Metadata Rigidity vs. Flexibility**: Current metadata is embedded in entries and in the tables themselves. A more modular metadata structure (e.g., `MetadataType`, `MetadataEntry`, `MetadataValue`) could support environments requiring more complex and complete metadata linking.
- **Raw Data Handling**: Raw data (e.g., EDF, images, audio files) are only validated through folder placement. Extending governance to file type validation or integrity checks would further improve system robustness.

Limitations

This study has several limitations. First, our participant pool consisted entirely of researchers working within a healthcare research context. As such, our findings may not be generalizable to other domains or professional settings. Second, the participants were primarily affiliated with the same multidisciplinary project spanning two institutions, which may have influenced their shared assumptions, workflows, or perspectives. Third, the sample size was relatively small (13 participants), though this was offset by a mixed-methods design combining feasibility testing, surveys, interviews, and observation. These complementary data sources allowed for in-depth exploration of user behavior and governance practices.

Additionally, although participants were specialists, their interactions reflected real-world governance challenges, particularly regarding metadata prioritization, assumptions, and format consistency. Future research should broaden the participant pool and investigate governance adoption across different sectors. Longitudinal studies may also offer insights into how system use evolves over time, and whether automated governance frameworks promote lasting improvements in data quality, trust, and user satisfaction.

6 Conclusion

This paper addressed the research question: *How can automated validation systems support simplified and sustainable data governance?* Through a mixed-methods study involving feasibility testing (evaluating whether participants could successfully complete the data tasks), surveys, and semi-structured interviews with researchers actively engaged in data work, we demonstrated that embedding automated validation at the point of data entry—paired with a lean database structure and standardized input mechanisms—offers a viable path toward reducing the complexity of data governance while maintaining compliance, traceability, and user engagement.

Our results show that automation, when thoughtfully integrated, can enforce governance rules without burdening users with additional responsibilities. The system enabled participants to uphold data standards—even without a deep understanding of the underlying governance model—through structured feedback, validation constraints, and a shared environment that improved with use. This illustrates that automation can function not only as a technical enforcement tool but as a participatory mechanism that turns data validation into a governance act.

We contribute to ongoing conversations in data governance, digital infrastructure, and socio-technical systems by providing empirical evidence that governance frameworks can be both simplified and generalized without sacrificing quality or adaptability. The approach outlined in this study highlights a shift away from reactive, manually enforced governance toward proactive, system-embedded practices that encourage broader participation and improve over time through iterative use.

Future research should investigate how these principles translate to other domains, data types, and organizational contexts—particularly in projects with limited resources. The findings offer practical implications for designing scalable governance models that balance automation, accountability, and usability across dynamic research environments.

References

[1] Ahmed Abbasi, Suprateek Sarker, and Roger HL Chiang. Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2):3, 2016.

[2] Claus Bossen, Yunan Chen, and Kathleen H Pine. The emergence of new data work occupations in healthcare: the case of medical scribes. *International journal of medical informatics*, 123:76–83, 2019.

[3] Taylor M Cruz. Data politics on the move: Intimate work from the inside of a data-driven health system. *Information, Communication & Society*, 26(3):496–511, 2023.

[4] Klaus Hoeyer and Sarah Wadmann. 'meaningless work': How the datafication of health reconfigures knowledge about work and erodes professional judgement. *Economy and Society*, 49(3):433–454, 2020.

[5] Amelia Fiske, Barbara Prainsack, and Alena Buyx. Data work: meaning-making in the era of data-rich medicine. *Journal of medical Internet research*, 21(7):e11672, 2019.

[6] Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. Searching data: a review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, 70(5):419–432, 2019.

[7] Anna Sigridur Islind, Helena Vallo Hult, Victoria Johansson, Eva Angenete, and Martin Gellerstedt. Invisible work meets visible work: infrastructuring from the perspective of patients and healthcare professionals. In *54th Hawaii International Conference on System Sciences (HICSS9, Tuesday, January 5, 2021 to Friday, January 8, 2021*, pages 3556–3565. Hawaii International Conference on System Sciences, 2021.

[8] Matthew Jones. What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, 28(1):3–16, 2019.

[9] Lars Nagel, Juan Jose Hierro, Eugenio Perea, Douwe Lycklama, Christoph Mertens, Anne-Sophie Taillandier, Maria Marques, Joshua Gelhaar, Angelo Marguglio, Ulrich Ahle, et al. Design principles for data spaces: position paper. Technical report, E. ON Energy Research Center, 2021.

[10] Edward Curry, Simon Scerri, and Tuomo Tuikka. *Data spaces: design, deployment and future directions*. Springer Nature, 2022.

[11] Rohit A Deshmukh, Diego Collarana, Joshua Gelhaar, Johannes Theissen-Lipp, Christoph Lange, Benedikt T Arnold, Edward Curry, and Stefan Decker. Challenges and opportunities for enabling the next generation of cross-domain dataspaces. In *The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference*, 2024.

[12] Ibrahim Alhassan, David Sammon, and Mary Daly. Critical success factors for data governance: A theory building approach. *Information Systems Management*, 36(2):98–110, 2019.

[13] Ibrahim Alhassan, David Sammon, and Mary Daly. Data governance activities: an analysis of the literature. *Journal of Decision Systems*, 25(sup1):64–75, 2016.

[14] Olivia Benfeldt, John Stouby Persson, and Sabine Madsen. Data governance as a collective action problem. *Information Systems Frontiers*, 22(2):299–313, 2020.

[15] BMRKRCS Petzold, Matthias Roggendorf, Kayvaun Rowshankish, and Christoph Sporleder. Designing data governance that delivers value. *McKinsey Digital*, 26, 2020.

[16] John Ladley. *Data governance: How to design, deploy, and sustain an effective data governance program*. Academic Press, 2019.

[17] Elena Parmiggiani and Miria Grisot. Data curation as governance practice. *Scandinavian Journal of Information Systems*, 2020.

[18] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S Barbosa, and Tomasz Janowski. Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, 37(3):101493, 2020.

[19] Ibrahim Alhassan, David Sammon, and Mary Daly. Data governance activities: A comparison between scientific and practice-oriented literature. *Journal of Enterprise Information Management*, 31(2):300–316, 2018.

[20] Maung K Sein, Ola Henfridsson, Sandeep Purao, Matti Rossi, and Rikard Lindgren. Action design research. *MIS quarterly*, pages 37–56, 2011.

[21] Matthew T Mullarkey and Alan R Hevner. An elaborated action design research process model. *European journal of information systems*, 28(1):6–20, 2019.

[22] Gerald I Susman and Roger D Evered. An assessment of the scientific merits of action research. *Studi organizzativi*, 2022(2), 2023.

[23] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[24] Asif Qumer Gill and Eng Chew. Configuration information system architecture: Insights from applied action design research. *Information & Management*, 56(4):507–525, 2019.

[25] Marina Micheli, Marisa Ponti, Max Craglia, and Anna Berti Suman. Emerging models of data governance in the age of datafication. *Big Data & Society*, 7(2):2053951720948087, 2020.

[26] Boris Otto. Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, 49:424–438, 2019.

[27] Kwanele Ngcobo, Sandiswa Bhengu, Ambani Mudau, Bonginkosi Thango, and Matshaka Lerato. Enterprise data management: Types, sources, and real-time applications to enhance business performance-a systematic review. *Systematic Review— September*, 2024.

[28] John R Talburt, Lisa Ehrlinger, and Justin Magruder. Automated data curation and data governance automation, 2023.

[29] Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An interview study. *IEEE transactions on visualization and computer graphics*, 18(12):2917–2926, 2012.

[30] Sung Une Lee, Liming Zhu, and Ross Jeffery. Data governance for platform ecosystems: Critical factors and the state of practice. In *Pacific Asia Conference on Information Systems (PACIS)*, 2017. Accessed: 2025-02-24.

[31] Paul Brous, Marijn Janssen, and Rutger Krans. Data governance as success factor for data science. In *Conference on e-Business, e-Services and e-Society*, pages 431–442. Springer, 2020.

[32] Gopi Maren and Dataversity. Understanding the potential failures of a data governance program, 2024.

[33] Alexander Procter and Spark. Why data governance keeps holding enterprises back, 2025.

[34] Karol Bližnák, Michal Munk, and Anna Pilková. A systematic review of recent literature on data governance (2017-2023). *IEEE Access*, 2024.

[35] Sara Marcucci, Natalia Gonzalez Alarcon, Stefaan G Verhulst, and Elena Wullhorst. Mapping and comparing data governance frameworks: A benchmarking exercise to inform global data governance deliberations. *arXiv preprint arXiv:2302.13731*, 2023.

Appendix A: Instruments and Observations

Survey Questions (Each of these sections contained multiple specific questions.)

- Position and background
- Familiarity with data formats
- Publication count
- Task difficulty ratings
- Trust in validation
- Preferences on raw vs. validated data
- Data quality priorities

- Time allocation for data work
- Comments on experience and feedback

Interview Questions (Each of these broad questions was followed by prompts.)

- Training on data quality
- Use of data standards
- Opinions on using the system
- Complexity and usability
- Usefulness of materials
- Processing experience
- Expectations for data reuse and documentation
- Reflections on external access and open science

Materials Provided to Participants

- A ZIP archive containing:
  - README file
  - Six example input CSV files
  - One test dataset (CSV)
  - Full documentation PDF

Observational Metrics Collected

- Step-by-step timing and duration
- Upload attempts per task
- Specific validation errors
- Documentation usage
- Total attempts and retries
- Notes on unexpected behaviors

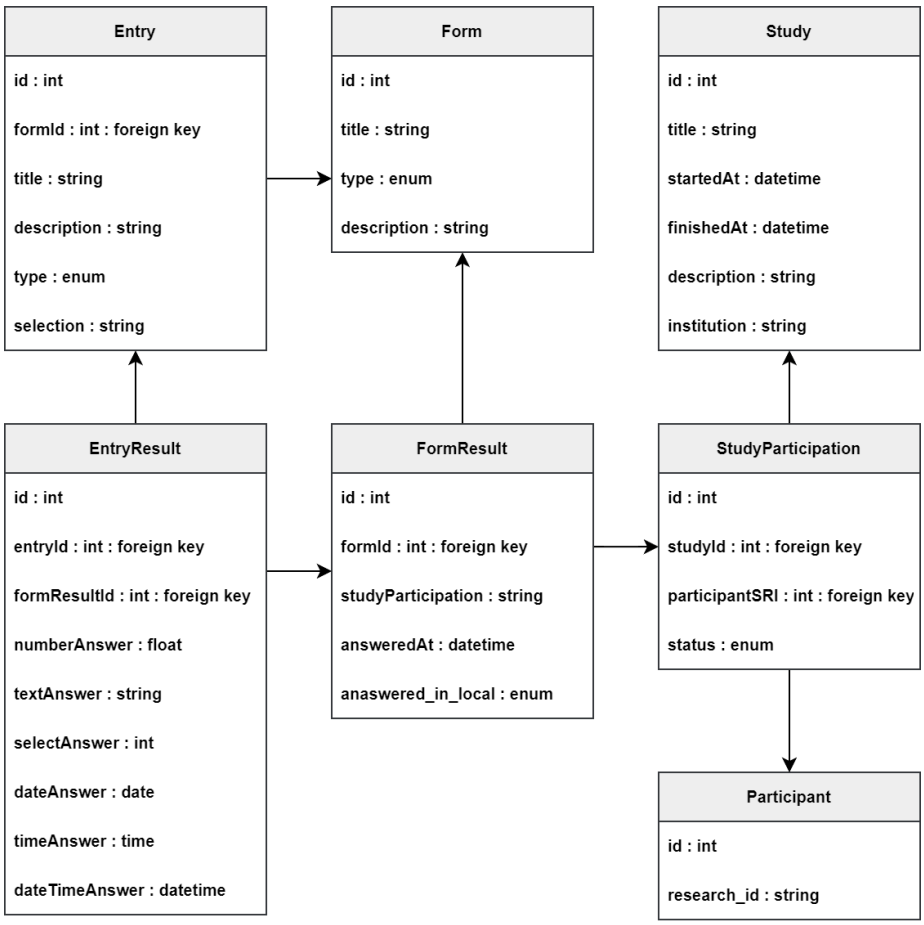


Figure 2: Overview of the relational database design used in the project.

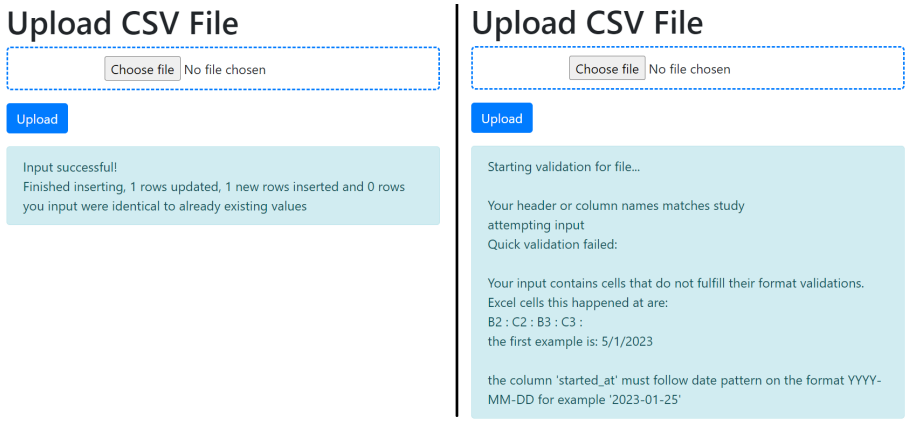


Figure 3: Examples of successful and failed upload feedback shown to users.

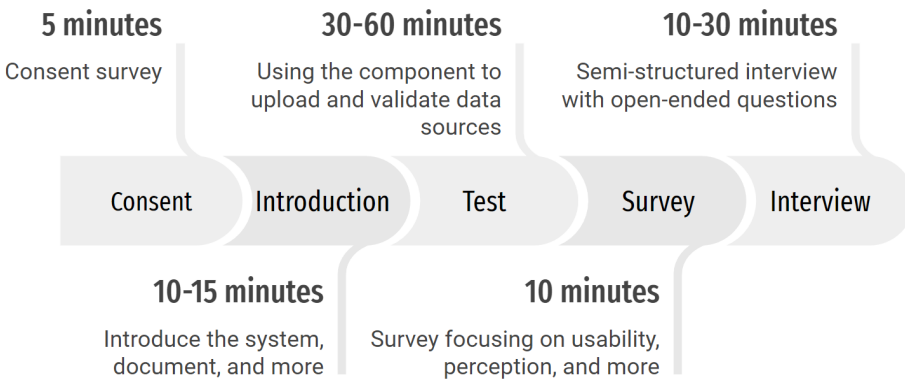


Figure 4: Participant flow through the feasibility study with estimated time allocation.