

The Anatomy of Data Ecosystems: Identifying and Analyzing Archetypes

Completed Research Full Paper

Philipp Kernstock
Technical University of Munich
philipp.kernstock@tum.de

Fabian Michael König
Technical University of Munich
fabian.m.koenig@tum.de

Timo Phillip Böttcher
Technical University of Munich
timo.boettcher@tum.de

Andreas Hein
University of St. Gallen
andreas.hein@unisg.ch

Helmut Krcmar
Technical University of Munich
helmut.krcmar@tum.de

Abstract

Organizations are shifting from traditional business models towards a collaborative, data-centric paradigm, giving rise to data ecosystems. However, while opportunities for leveraging data in these ecosystems are vast, such ecosystems' structure, governance, and operation remain nebulous. Thus, our research delves into the operational intricacies of data ecosystems. Drawing upon a cluster analysis of 142 data ecosystem initiatives and integrating theoretical perspectives on data ecosystems and digital platforms, we identify five distinct archetypes characterized by variations in organizational structure, technical openness, actor interdependence, and governance. Our findings illuminate the core elements of data ecosystems and provide pragmatic insights for their application, fostering an understanding critical for designing multi-sided data platforms and governance mechanisms in public data spaces. This research contributes a comprehensive tool for academia to understand data ecosystems and a guide for industry practitioners to manage organizations within these complex structures.

Keywords

Data ecosystem, archetypes, ecosystem governance, data space.

Introduction

Increasing data-driven value creation between organizations in ecosystems has led to the emergence of data ecosystems. Data ecosystems are networks of autonomous actors that directly or indirectly consume, produce, or provide data and other related resources (Oliveira et al., 2019). They are complex structures where economic value is no longer derived from singular efforts but rather from the synergistic interplay of various actors (Gelhaar et al., 2021). This marks a significant departure from traditional business models to a more collaborative and data-centric approach (Beverungen et al., 2022). Data ecosystems offer novel opportunities for firms to exploit data for cost reduction, customer satisfaction enhancement, and operational efficiency improvements (Gelhaar et al., 2023).

While various initiatives drive the development and standardization of data ecosystems, various structures currently exist to build such ecosystems, leading to different types of data ecosystems in diverse contexts (Gelhaar et al., 2021). Their evolution is vital for fostering trust and transparency, a necessity as exemplified by industry-led initiatives such as the Catena-X Automotive Network or the Open Banking initiative. Participation in these ecosystems is rapidly becoming a strategic imperative for firms aiming to

maintain their competitive edge. However, as data sharing in such ecosystems becomes increasingly relevant and different initiatives emerge in nearly every industry, questions remain about how these ecosystems are structured, governed, and operationalized and what types of data ecosystems might be most promising for specific industries or use cases.

Research investigates mainly the emergent phases of data ecosystems, focusing on challenges and incentives for participation (Gelhaar et al., 2023; Gelhaar & Otto, 2020), roles of participants (Azkan et al., 2022), and design principles for data spaces as multi-sided data platforms. Data platforms enable interactions among diverse user groups, thus creating value through data exchange (Otto & Jarke, 2019). The narrative increasingly emphasizes the transition from private digital platforms to public data spaces, drawing attention to the critical role of governance mechanisms that safeguard data sovereignty and ensure ecosystem sustainability (Beverungen et al., 2022). Despite the growing interest and preliminary frameworks for understanding data ecosystems, there remains a gap in identifying and comprehensively understanding the different structures of their operational functionalities. As different foci of data ecosystem exist, such as engineering, collaboration, or innovation (Heinz et al., 2022), organizations must understand the various aspects and perspectives to join or build a data ecosystem. Further, depending on the problems the organizations aim to tackle through data sharing, different types of data ecosystems might be favorable, e.g., using a platform-centric or a direct data-sharing infrastructure (Jussen et al., 2024). However, various developments of data ecosystems, such as emerging standards like GAIA-X or International Data Spaces, leave ample leeway for actual implementation. Literature has investigated emerging types of data ecosystems and developed taxonomies describing their characteristics and governance mechanisms. However, archetypes are needed to understand, structure, and differentiate between data ecosystem types (Möller et al., 2021). Therefore, we ask: *what are operational archetypes of data ecosystems?*

We answer this research question by following abductive reasoning and cluster analysis to develop a framework of data ecosystem archetypes (van der Valk et al., 2022). We collected data on 142 data ecosystem initiatives and analyzed their characteristics through a coding scheme based on existing literature (Gelhaar et al., 2021). We then used partitioning around medoids clustering and complemented the results with inductive analysis to derive archetypal patterns. This approach allows us to answer the RQ by drawing on existing theory through deductive coding, incorporating specific insights through inductive coding, and deriving patterns across a large number of data ecosystems through cluster analysis.

We develop five archetypes that differ along the dimensions of organizational setup, openness, interdependence of actors, and governance. Two platform-centric archetypes are the open data platform ecosystem and the closed data platform ecosystem. The third archetype is the decentralized data ecosystem, which focuses solely on data exchange. The centralized data ecosystem archetype revolves around one keystone actor, who provides the majority of the data and infrastructure in the ecosystem. Finally, the data trading ecosystem archetype is a marketplace-based data provisioning and selling structure with loosely coupled actors. Our research illuminates the foundational elements of data ecosystems and their implications for operational real-world application. Establishing data ecosystem archetypes gives a comprehensive insight into data ecosystems' structure, governance, and operation. This creates a deeper understanding of these complex structures, which will aid in designing multi-sided data platforms (Otto & Jarke, 2019) and lay a robust foundation for governance in public data spaces (Beverungen et al., 2022). These archetypes are poised to become practical guides for industry practitioners, enhancing organizations' design, management, and strategic alignment in these ecosystems.

Theoretical Background

In the following, we introduce the core concepts of data ecosystems regarding their structure and operational models based on existing research on digital platform ecosystems (DPE). This theoretical background provides the foundation for our data analysis.

Digital Platform Ecosystems

Data ecosystems are rooted in the literature of digital platforms and DPEs (Beverungen et al., 2022; Otto & Jarke, 2019), which provides valuable insights into structural patterns and operational models. Digital platforms can be categorized based on their function as either innovation platforms, transaction

platforms, or a hybrid of both (Cusumano et al., 2019). DPEs are built upon extensible software systems that offer core functionality supplemented by additional modules. They include a platform owner who implements governance to facilitate value creation between themselves and a network of autonomous complementors and consumers (Hein et al., 2020). The generativity of DPEs is a product of the innovative contributions from complementors, leading to the continuous evolution of the ecosystem's offerings (Gawer & Cusumano, 2014). Complementors of DPE can have high autonomy and be loosely coupled to the platform, or they can have low autonomy and be tightly coupled to the platform. The governance of DPEs is a delicate balance between allowing complementors the freedom to innovate and thus opening the platform and maintaining control to ensure quality (Hein et al., 2020). The governance mechanisms set by platform owners orchestrate the DPE, shaping the production and distribution of value through various controls and incentives. Governance mechanisms refer to the rules and procedures put in place by the platform owners to orchestrate the DPE to co-create value (Schrieck et al., 2022). Especially in data-centric platforms, data as a boundary resource is also highly relevant for governing the platform and creating network effects (Otto & Jarke, 2019). The DPE's architecture, which is either centralized around one platform leader (Gawer & Cusumano, 2014), decentralized as a blockchain-based platform, or federated as an alliance-based platform (Otto & Jarke, 2019) has a significant impact on the control over the ecosystem being more centralized or decentralized.

Data Ecosystem

While the research on data ecosystems regularly refers to the fundamental concepts of digital platforms, their continuous evolution and the consequent theoretical insights have led to insights specific to the structure and governance of data ecosystems (Otto & Jarke, 2019). Generally, ecosystems are an agreed structure of actors in multilateral relationships that interact to provide a focal value proposition (Adner, 2017). Drawing on this general concept, data ecosystems are defined as socio-technical complex networks constituted by a plethora of actors who engage in a variety of activities, including the discovery, archiving, publication, consumption, and reuse of data and the fostering of innovation, creation of value, and support for new businesses (Oliveira et al., 2019). The evolution of these ecosystems has been influenced by theories and frameworks from diverse fields, ranging from marketing and service research to business and software ecosystems (Azkan et al., 2022). The terminology has evolved, with concepts such as 'data spaces' and 'data value networks' emerging. Data spaces refer to the technical foundations facilitating data sharing among gathered actors (Gieß et al., 2023). Data marketplaces can operate with a hierarchical structure, where asset-specific goods are traded authoritatively among familiar actors, or with a market structure, where less asset-specific goods are exchanged competitively among actors seeking maximum individual gain (Bergman et al., 2022). Terms like 'collaborative,' 'open,' or 'industrial' add layers to the concept of data ecosystems, indicating the nature of collaboration, the openness of data licensing, and the business-oriented approach, respectively (Gelhaar et al., 2023). The architecture of data ecosystems is underpinned by the interplay of actors, roles, relationships, and resources (Azkan et al., 2022). The structure of data ecosystems is either constructed around one keystone actor significantly contributing to its emergence, around a platform that offers infrastructure and services for data sharing, around a marketplace that provides infrastructure and enables sales of data and related business models, or decentralized without a focal actor (Gelhaar & Otto, 2020; Oliveira et al., 2019). The dynamics within these ecosystems are complex and still under extensive research, presenting novel challenges and opportunities (Oliveira et al., 2019). Roles within the ecosystem, though not yet concretely defined due to the nascent state of research, are crucial. Actors—whether enterprises, institutions, or individuals—engage in the ecosystem fulfilling one or more specific roles and are often driven by a mix of interests, commitment to the ecosystem's goals, and potential incentives (Azkan et al., 2022). One particular benefit is the potential for data sovereignty, promising sovereign data subjects to control, limit, and enforce their rights over their data (Hummel et al., 2018).

Literature on data ecosystems has developed multiple taxonomies related to data ecosystems to improve understanding and start conceptualizing them. Aimed at developing a shared understanding, taxonomies for data ecosystems (Gelhaar et al., 2021), their governance (Lis & Otto, 2021), data spaces as underpinning technological solutions (Gieß et al., 2023), and related concepts such as data marketplaces (Bergman et al., 2022) have been developed. While crucial for fostering an understanding of the different design aspects and options of data ecosystems, these taxonomies often lack the tangible quality of concrete archetypes. Further, some of the taxonomies omit the interaction of actors as a highly relevant

dimension for articulating ecosystem dynamics (Jacobides et al., 2018). Thus, the current conceptual frameworks of data ecosystems lack applicability for how the phenomenon evolves in practice, requiring an improved and structured understanding of practical data ecosystem types.

In the current discourse on data ecosystems, the taxonomies proposed by Gelhaar et al. (2021) and Lis and Otto (2021) represent a substantive integration of the interactional dimension. These frameworks acknowledge the intricate interdependencies that define the interactions among actors within data ecosystems. In particular, Gelhaar, Groß, and Otto's work provides a nuanced classification scheme that delineates the dimensions of actor interrelations with a specificity that allows for the distinction between mutually exclusive and non-exclusive categories (Gelhaar et al., 2021). Upon comparative analysis, their taxonomy emerges as particularly robust, resonating with the core attributes of ecosystem frameworks (Jacobides et al., 2018). In light of its alignment with foundational ecosystem principles, we use this taxonomy as an initial coding scheme for classifying data ecosystems. We develop data ecosystem archetypes that improve understanding, structuring, and differentiating (Möller et al., 2021).

Research Approach

Based on the theoretical foundations of DPE and data ecosystems, we collected data on 142 data ecosystem initiatives. We coded each data ecosystem according to a predefined coding scheme that is based on an existing data ecosystem taxonomy. Similar to recent IS research that has derived archetypes from taxonomy-based data through clustering (Duparc et al., 2022; van der Valk et al., 2022; Weking et al., 2020), we then used cluster analysis to derive patterns across the data ecosystem initiatives (Punj & Stewart, 1983). After deriving an initial set of intermediate archetypes, we used an inductive qualitative approach to derive the final five data ecosystem archetypes (Park et al., 2020). Finally, we conducted six expert interviews to assess the practical relevance and validity of the developed data ecosystem archetypes. The approach allows us to combine abductive reasoning with analyzing a large set of data ecosystem initiatives to produce generalizable insights that advance the current taxonomy-based understanding of data ecosystems (Möller et al., 2021).

Data Sampling, Collection and Preparation

To perform a cluster analysis to derive archetypal patterns, we constructed a dataset of data ecosystem initiatives based on the lists of featured data ecosystem initiatives from the collections of Gaia-X (n=80) and PrepDSpace4Mobility (n=58). We then complemented this dataset through three notable data ecosystem initiatives: Catena-X, Advaneo, and Datarade. To be included, the identified data ecosystems had to have operational character, which means that at least one use case of the initiative had to be usable. We excluded merely conceptual projects without a productive implementation in practice. In total, 142 data ecosystems were included. After sampling the data ecosystems, they were classified across the taxonomy dimensions by Gelhaar et al. (2021). As we aimed to develop archetypes across specific domains and purposes, and ecosystems were often used across multiple domains and purposes, we used the dimensions of organization, infrastructure, openness, interdependence, and control as coding scheme.

Organization captures the different kinds of relationships, interactions, and organization of the actors that form a data ecosystem. Categories are keystone-centric, platform-centric, marketplace-based, or decentralized (Gelhaar & Otto, 2020; Oliveira et al., 2019). *Infrastructure* specifies the main technical infrastructure used to share data within the data ecosystem as centralized or distributed (Otto & Jarke, 2019). *Openness* represents the degree of access to the data ecosystem, which can either be open or closed, where closed means barriers to entry exist that are either technical or legal (Gelhaar et al., 2021). *Interdependence* implies the degree to which ecosystem actors are connected and dependent on each other, where the actors are either tightly or loosely coupled through the data ecosystem (Hein et al., 2020). *Control* relates to the control of the essential data resources in the data ecosystem in the form of how governance mechanisms are structured as either central or decentral (Gelhaar et al., 2021).

The resulting coding distribution relevant to the cluster analysis can be found in Table 1. All labeled use cases had the characteristic *distributed* for the dimension *infrastructure*. This might be the case due to the emerging understanding of data ecosystems as democratic and decentralized phenomena (Beverungen et al., 2022), leading our search results and data points to be skewed. However, the taxonomy category of centralized infrastructure relates to classical proprietary software platforms that

regularly fall under the term digital platforms rather than data ecosystems (Hein et al., 2020). As this characteristic is the same across all use cases, considering the principle of parsimony, we decided not to include the dimension *infrastructure* for the cluster analysis, as the remaining dimensions still allow for a theoretically sound explanation of data ecosystem archetypes based on the characteristics defined in the respective literature (Jacobides et al., 2018).

Dimension	Characteristics			
Organization	Keystone (8 %)	Platform (77 %)	Marketplace (8 %)	Decentralized (6 %)
Infrastructure	Centralized (0%)		Distributed (100 %)	
Openness	Open (49 %)		Closed (51 %)	
Interdependence	Tightly Coupled (35 %)		Loosely Coupled (65 %)	
Control	Central (41 %)		Decentral (59 %)	

Table 1. Distribution of dimensions of the coding scheme for the cluster analysis

Cluster Analysis

We used the programming language R and the “cluster” package to perform the cluster analysis (Maechler, 2019). We first calculated the dissimilarity matrix based on Gower’s distance to perform the cluster analysis on categorical data. Further, it was decided to use the Partitioning Around Medoids (PAM) method for clustering, which is a more robust version of K-means clustering (Schubert & Rousseeuw, 2021). Due to the skewed data set towards platform-centric data ecosystems, we introduced a weighting for the *organization* dimension, leading to more diverse clusters regarding this dimension. This is because the clustering algorithm tries to minimize the distance between data points within a cluster and maximize the distance between clusters. When giving a higher weight to an attribute, the distance matrix will have a higher distance between data points with different characteristics within the organization dimension. We approached the other dimensions with the same systematic approach to address their skewed distribution and enable diverse clusters. This resulted in a weighting matrix of (8,4,2,1) for *organization*, *interdependence*, *control*, and *openness*. The systematic approach to derive the weighting matrix allows us to balance the domination of one attribute, in our case organization, over the other. We used the silhouette width to determine the ideal number of clusters, where a value closer to 1 signifies that observations are very well clustered, around 0 means that observations lie between clusters, and -1 means a wrong cluster assignment, with values over 0.5 being considered reasonable well-assigned observations (Rousseeuw, 1987). For eight clusters, we had a silhouette width score of 0.85, with the silhouette width improving significantly until eight clusters and then becoming asymptotic, e.g. the silhouette width for 5 was 0.61. Thus, the results were promising and were used for further analysis.

Qualitative Adjustments of Clusters and Derivation of Archetypes

After performing the cluster analysis, we drew a random case sample and inspected the analyzed use cases within each cluster to verify their validity. We qualitatively analyzed the eight clusters to ensure that clusters are separable, i.e., that data points between clusters exhibit heterogeneity, and single clusters share common characteristics, data points within the clusters exhibit homogeneity (Weking et al., 2020). We then used an inductive approach (Park et al., 2020) based on the cases assigned to the clusters to adjust the initial clusters and develop the data ecosystem archetypes. For the initial cluster 1, the data ecosystem initiatives within the cluster show high similarity, while neither the dimension *interdependence* nor the dimension *control* was relevant to the ecosystem classification. Additionally, cluster 1 shows a high similarity to the use cases within cluster 4, in which neither the dimension of *interdependence* nor *control* was relevant to the classification. Clusters 2, 3, and 5 also included data ecosystem initiatives of high similarity. All three clusters only differ in the *interdependence* and *control* dimensions. Similarly to clusters 1 and 4, the dimensions of *interdependence* and *control* did not have significant relevance within the ecosystem classification.

After inductively analyzing the data ecosystems with a *platform-centric organization* characteristic, it became apparent that the cases exhibit decentral and central control traits. Data ecosystems in these

clusters, such as Catena-X, build a federated governance model where control is neither fully centralized nor decentralized. By summarizing the above insights, it was decided to combine clusters 1 and 4, and 2, 3, and 5 to form two platform-centric archetypes that leverage federated governance but differ in actor interdependence and openness. For cluster 6, the data ecosystem initiatives focus on connecting different data sources. In the identified data ecosystems, no additional services are offered. Additionally, no further analysis of the data or product or service provisioning happens within these data ecosystems. For cluster 7, a shaping characteristic is that one central company or actor publishes data, and other ecosystem participants depend on the central actor. Cluster 8 is clearly differentiable from all other identified clusters; clear similarities are also recognizable within the cluster.

After the adjustments, we conducted six interviews with experts from organizations involved in building and developing data ecosystems to validate the developed archetypes and to ensure their practical relevance. We sampled the interviewees based on their broad knowledge of various data ecosystem initiatives. The interviews were either working as consultants focusing on data-driven innovation or had a role related to data ecosystems, e.g., a software architect in an automotive company contributing to Catena-X. Next to the general relevance validation, a goal was to validate whether the derived archetypes are mutually exclusive. The experts were interviewed regarding the existence of the archetypes in practice and were asked to name examples. They were asked to consider potentially different archetypes in practice and which characteristics are essential for defining them. We then inductively coded the results of the interviews and matched the results with the developed archetypes. The interviews provided valuable elaborations and examples of archetypes that confirmed our clustering.

Results

In the following, the five identified data ecosystem archetypes are described. The description is based on the theoretical foundations of our research and the analyzed data ecosystem initiatives. The overview of the characteristics within each dimension of the five identified archetypes can be found in Table 2.

#	Archetype Name	Organization	Openness	Interdependence	Control
1	Open data platform ecosystem	Platform-centric	Open	Loosely coupled	Federated
2	Closed data platform ecosystem	Platform-centric	Closed	Tightly coupled	Federated
3	Decentralized data ecosystem	Decentralized	Closed	Loosely coupled	Decentral
4	Centralized data ecosystem	Keystone-centric	Open	Loosely coupled	Central
5	Data trading ecosystem	Marketplace-based	Open	Loosely coupled	Decentral

Table 2. Overview of data ecosystem archetypes

The **open data platform ecosystem** archetype is characterized by being platform-centric, open, loosely coupled, and federated. This means that a central digital platform is offered, which provides an infrastructure and services. These services include the sharing and usage of data in the ecosystem. Access to the platform is considered unrestricted, meaning that everyone can join. While small administrative tasks, like creating an account to access data, are still regarded as open, the sign-up must be free of any additional charges or requirements. Actors within the ecosystem have high autonomy and are only loosely coupled. While control over the data remains mainly decentralized, the platforms establish central governance rules, such as creating meta-data. Exemplary cases mainly consist of open government data platform ecosystems such as the EuRis internet platform, which harmonizes information on river transportation from various sources. The common value proposition is the improvement of safety and transport efficiency by focusing on data exchange between all participants in the ecosystem. The collected data from different sources is offered in a central service that is openly accessible to any user.

The **closed data platform ecosystem** archetype is characterized by being platform-centric, closed, tightly coupled, and federated. A central platform is offered within the archetype, providing infrastructure and services. These services include essential data sharing and usage and more advanced topics like

certification. Contrary to the open data platform archetype, this archetype controls access to its data ecosystem, meaning some type of entry barriers exist. These entry barriers can either be technical or legal. Actors in this archetype tend to be more tightly coupled, for example, through existing business relationships. Governance is often federated between different entities while following a decentral idea. An example of this archetype is the Catena-X data ecosystem. The goal of the ecosystem is to provide an environment in which data chains across the automotive supply chain can be easily created. With these data chains, it is possible to share data across the supply chain while still maintaining data sovereignty and ownership across the supply chain. Another example is Michelin's connected mobility platform, which offers services that connect data from different sources. Some services offered include information on driving events, tire recommendation systems, and a service for trip cost calculation. The ecosystem is considered closed, as users need to buy the service to receive access.

The **decentralized data ecosystem** archetype is characterized by being organizationally decentralized, closed, loosely coupled, and decentrally controlled. Decentralized within the organization dimension means that there is no central actor in place within the data ecosystem. Instead, the actors within the data ecosystem are connected by the common goal of creating joint value. Decentrality within the control dimension refers to data control, indicating that no central actor is in place to have control over a significant part of the data. In addition, the archetype is closed as technical and legal entry barriers exist. Actors are rather loosely coupled within the decentralized data ecosystem. One example is the Gaia-X project, "Decentralized in vehicle MLaaS to EV energy efficiency." This initiative proposes a decentralized learning infrastructure for machine learning models. Currently, companies only train models within their own data and do not publish or sell their data to other competitors. The initiative proposes a decentral platform that organizes the training of machine learning models. The ecosystem participants get a model and only train it locally with their data. As a result, data sovereignty remains within the companies, and no data needs to be published between competitors. The decentralized data control is identifiable, as no central actor controls a significant part of the resources. Only chosen actors can participate within the ecosystem, as contractual agreements between the parties are required.

The **centralized data ecosystem** archetype is characterized by being keystone-centric, open, loosely coupled, and centrally controlled. A data ecosystem is keystone-centric if a central actor is in place that other actors are centered around. The so-called keystone actor is directly or indirectly responsible for a large part of the data in the ecosystem. This keystone actor is responsible for the data, so the control is centralized around him. The archetype is also considered open, as any organization can join the ecosystem to access data from the keystone actor. Consequently, the actors are loosely coupled as they do not have bilateral data exchange relationships with the keystone actor. One example of this is the portal operated by Canal & River Trust. The portal offers various data, ranging from data on aqueducts to data on locks to data on weirs. Within the ecosystem, there are no further services except providing data. As the data is not aggregated by many parties involved but by the Canal & River Trust, it acts as the central party with control and responsibility over the data. Additionally, the data is free to download without registration.

The **data trading ecosystem** archetype is characterized by being marketplace-based, open, loosely coupled, and decentrally controlled. A marketplace-based organization provides the infrastructure for data sharing and different supporting components and functions. These range across specific business models, applications, rules, and services. The archetype is also open, as merely creating an account is necessary to enter the data ecosystem. Additionally, the control of the data is decentralized. In practice, this means that the marketplace operator has little control over most of the data as it is not the main provider of the data on the marketplace. Actors within the ecosystem are loosely coupled, as they have high autonomy regarding their decisions to provide and sell or buy data and mostly do not form tight relationships with other actors. One example of this archetype is the Datarade, which offers data from different industries based on a marketplace-based organization. Actors can search for data without registering on the platform. Data providers can choose for themselves how to offer their data. Typical variants are the provision of an API and a payment per API call or access to the complete dataset for a fixed price. Datarade is not the main data provider and does not control the main part of the data.

In exploring the multifaceted nature of data ecosystems, the expert interviews revealed insightful examples and opinions regarding the various archetypes. In the realm of the open data platform ecosystem archetype expert 6 cited Kaggle as a significant instance where registration leads to broad data availability for machine learning applications, with users able to upload and share datasets widely.

Meanwhile, experts 3, 4, and 6 recognized the closed data platform ecosystem archetype, referencing Catena-X, Mobility Data spaces, or Telekom's Data Intelligence Hub as fitting examples. The concept of the decentralized data ecosystem archetype was touched upon by expert 1, who described the general idea of data spaces as data-sharing concepts. The discussion on the centralized data ecosystem archetype brought forth the notion of intermediaries or keystone actors facilitating data exchange, a structure often seen in practice. Expert 1 mentioned the Data Trustee case, where a central entity collects and aggregates data—such as energy supply data used by city governments for reporting without revealing proprietary supplier details. The experts elaborated on the data trading ecosystem archetype, with Advaneo, Caruso, and Dawex cited as marketplaces for data exchange.

Discussion

Data ecosystems offer many opportunities for firms to exploit data for cost reduction, customer satisfaction enhancement, and new business models (Gelhaar et al., 2023). Research has analyzed the relevance of governance mechanisms for transitioning from private digital platforms to public data spaces. However, questions remained about how these data ecosystems are structured, governed, and operationalized and what types of data ecosystems might be most promising for specific industries or use cases. Therefore, we gathered and analyzed data on 142 data ecosystem initiatives, coding each within the framework of an established taxonomy (Gelhaar et al., 2021). We distilled this data into discernible patterns through cluster analysis using PAM, enabling us to isolate five distinct data ecosystem archetypes. These archetypes—open data platform, closed data platform, decentralized data, centralized data, and data trading ecosystems—are differentiated by their organization, infrastructure, openness, interdependence, and control. The archetypes, validated through expert interviews, reflect a broad range of organizational structures and governance models, from centralized to decentralized control. Our data indicate a prevalent distribution towards platform-centric and open structures.

While our research mainly aligns with the existing literature on data ecosystem taxonomies (Gelhaar et al., 2021), the inductive coding and the validation through expert interviews lead to some characteristics of the archetypes diverging from the earlier conceptualizations. Firstly, the possibility of infrastructure centralization in data ecosystems is challenged as our data sample showed a decentralized infrastructure characteristic. This might indicate an evolving understanding of data ecosystems aligned with decentralized values as propagated through the development of data spaces as the technical foundation of the ecosystems (Otto & Jarke, 2019; Schleimer et al., 2023). Further, most platform-centric data ecosystem initiatives investigated, such as Catena-X, assume a decentralized and democratic ideal where governance and control are supposed to be decentralized or federated. Federation is an increasingly crucial concept in data ecosystems and the enabling data spaces (Schleimer et al., 2023). Our archetypes show that platform-centric data ecosystems frequently rely on federated control. However, governance in data ecosystems involves control over data (Gelhaar, Groß, et al., 2021) and mechanisms to enable the evolution of and innovation in the data ecosystem. Our archetypes enable future research on such governance mechanisms by providing insights into how data ecosystems might be structured by drawing on insights into the governance of centralized or decentralized digital platforms and ecosystems.

Different archetypes are relevant for different environments where organizations try to address economic, technical, or social problems through data sharing in ecosystems (Jussen et al., 2024). The closed data platform ecosystem archetype is suitable in cases where the participating actors are tightly coupled, for instance, through pre-existing business relationships like in the Catena-X ecosystem, where participants are already involved in physical value-creation networks. By joining the data ecosystem, they can create a digital counterpart to their physical value-creation activities and develop data-driven use cases, such as supply chain transparency or product passports. Similarly, the increasingly relevant data trading ecosystem archetype benefits use cases such as providing and selling data to ex-ante unknown ecosystem participants (Bergman et al., 2022). By drawing on the archetypal patterns and the emergence of new and the evolution of existing data ecosystem initiatives, developing a typology of data ecosystems that captures such specific contexts will be possible.

Theoretical and Practical Contributions

Our theoretical contribution is twofold. Firstly, we advance research on the still nascent phenomenon of data ecosystems by providing archetypes that provide a new perspective on data ecosystems and allow

navigating the differentiating research streams (Möller et al., 2021). Secondly, we show that more nuanced approaches to governance emerge for operationalizing data ecosystems, especially as data platform ecosystems (Beverungen et al., 2022). Governance and control approaches often happen on a scale between centralized and decentralized control, with federation being the mainly observable governance mode. From a practical perspective, the developed archetypes could serve as blueprints for designing such platforms, ensuring they are well-structured and capable of fostering beneficial interactions among participants. Consequently, stakeholders can leverage these archetypes to build better data ecosystems, showing them what they must consider for successful implementation. Different archetypes of data ecosystems might be favorable depending on the organizations' goals. Our research gives initial practical guidance on what archetypes might be suitable and how they are structured.

Limitations and Future Research

The research presented acknowledges certain limitations that could be addressed in future studies. The data collection process may carry an inherent bias, as it captures only a snapshot of the current landscape, potentially overlooking the dynamic evolution of data ecosystems. Moreover, the classification into archetypes, though comprehensive, may not encapsulate emerging or evolving ecosystems that deviate from the established frameworks. For future research, there is a significant opportunity to conduct longitudinal studies that track the progression of data ecosystems over time. Such studies would empirically validate the archetypes and may reveal new patterns or structures as these ecosystems develop. Additionally, research could focus on creating a typology for data ecosystems, which could include the role and influence these ecosystems have on fostering circular economies and enhancing supply chain transparency, particularly in relation to Environmental, Social, and Governance (ESG) criteria (Böttcher et al., 2024). This would contribute to the theoretical understanding and provide practical insights for organizations striving to meet sustainability goals. Finally, combining insights from inter-organizational data sharing through data ecosystems and intra-organizational data platforms could form a holistic understanding of how organizations create value through data (Kernstock et al., 2024).

Conclusion

Through our research, we address the remarkable expansion of data generation and the evolution of data ecosystems by identifying and articulating five distinct operational data ecosystem archetypes through cluster analysis, inductive coding, and expert validation. The archetypes echo a departure from traditional business models towards a more synergistic, data-centric approach, recognizing data as a pivotal economic asset. Our findings provide a crucial scaffold for understanding data ecosystems' organization, governance, and interdependencies, offering tangible blueprints for academia and industry alike. The archetypes lay the foundation for a more nuanced understanding of how data can be leveraged within various organizational and technical frameworks to drive innovation, operational efficiency, and economic value. The archetypes presented here are not the end but a beginning, signaling a new era where data's potential can be fully realized in an increasingly connected and data-driven world.

REFERENCES

- Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. *Journal of management*, 43(1), 39-58.
- Azkan, C., Möller, F., Ebel, M., Iqbal, T., Otto, B., & Poeppelbuss, J. (2022). Hunting the Treasure: Modeling Data Ecosystem Value Co-Creation. 43rd International Conference on Information Systems, Copenhagen.
- Bergman, R., Abbas, A. E., Jung, S., Werker, C., & De Reuver, M. (2022). Business model archetypes for data marketplaces in the automotive industry: Contrasting business models of data marketplaces with varying ownership and orientation structures. *Electronic Markets*, 32(2), 747-765.
- Beverungen, D., Hess, T., Köster, A., & Lehrer, C. (2022). From private digital platforms to public data spaces: implications for the digital transformation. *Electronic Markets*, 32(2), 493-501.
- Böttcher, T. P., Empelmann, S., Weking, J., Hein, A., & Krcmar, H. (2024). Digital sustainable business models: Using digital technology to integrate ecological sustainability into the core of business models. *Information Systems Journal*, 34(3), 736-761.

- Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2019). *The business of platforms: Strategy in the age of digital competition, innovation, and power* (Vol. 320). Harper Business New York.
- Duparc, E., Möller, F., Jussen, I., Stachon, M., Algac, S., & Otto, B. (2022). Archetypes of open-source business models. *Electronic Markets*, 32(2), 727-745.
- Gawer, A., & Cusumano, M. A. (2014). Industry Platforms and Ecosystem Innovation. *Journal of Product Innovation Management*, 31(3), 417-433.
- Gelhaar, J., Groß, T., & Otto, B. (2021). A taxonomy for data ecosystems. *Proceedings of the 54th Hawaii International Conference on System Sciences*.
- Gelhaar, J., Müller, P., Bergmann, N., & Dogan, R. (2023). Motives and incentives for data sharing in industrial data ecosystems: an explorative single case study. *Proceedings of the 56th Hawaii International Conference on System Sciences*.
- Gelhaar, J., & Otto, B. (2020). Challenges in the Emergence of Data Ecosystems. *PACIS*.
- Gieß, A., Möller, F., Schoormann, T., & Otto, B. (2023). Design Options for Data Spaces. Thirty-first European Conference on Information Systems (ECIS 2023), Kristiansand.
- Hein, A., Schreieck, M., Riasanow, T., Setzke, D. S., Wiesche, M., Böhm, M., & Krcmar, H. (2020). Digital platform ecosystems. *Electronic Markets*, 30(1), 87-98.
- Heinz, D., Benz, C., Fassnacht, M., & Satzger, G. (2022). Past, present and future of data ecosystems research: A systematic literature review. *PACIS*.
- Hummel, P., Braun, M., Augsburg, S., & Dabrock, P. (2018). Sovereignty and data sharing. *ITU Journal: ICT Discoveries*, 2.
- Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems [Article]. *Strategic Management Journal*, 39(8), 2255-2276.
- Jussen, I., Möller, F., Schweihoff, J., Gieß, A., Giussani, G., & Otto, B. (2024). Issues in inter-organizational data sharing: Findings from practice and research challenges. *Data & Knowledge Engineering*, 150, 102280.
- Kernstock, P., Biermann, K., Sartor, S., Wimbauer, A., Bohnet, J., Böttcher, T. P., & Hein, A. (2024). Data Mesh - A Case Study Perspective on Building Industrial Data Platforms. *ECIS*.
- Lis, D., & Otto, B. (2021). Towards a taxonomy of ecosystem data governance. *Proceedings of the 54th Hawaii International Conference on System Sciences*.
- Maechler, M. (2019). Finding groups in data: Cluster analysis extended Rousseeuw et al. *R package version*, 2(0), 242-248.
- Möller, F., Stachon, M., Azkan, C., Schoormann, T., & Otto, B. (2021). Designing business model taxonomies – synthesis and guidance from information systems research. *Electronic Markets*.
- Oliveira, M. I., Barros Lima, G. d. F., & Farias Lóscio, B. (2019). Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, 61, 589-630.
- Otto, B., & Jarke, M. (2019). Designing a multi-sided data platform: findings from the International Data Spaces case. *Electronic Markets*, 29(4), 561-580.
- Park, Y., Fiss, P. C., & El Sawy, O. A. (2020). Theorizing the multiplicity of digital phenomena: The ecology of configurations, causal recipes, and guidelines for applying QCA. *Management of Information Systems Quarterly*, 44, 1493-1520.
- Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of marketing research*, 20(2), 134-148.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
- Schleimer, A. M., Jahnke, N., & Otto, B. (2023). Architecture Design Options for Federated Data Spaces. *Proceedings of the 56th Hawaii International Conference on System Sciences*.
- Schreieck, M., Wiesche, M., & Krcmar, H. (2022). Governing innovation platforms in multi-business organisations. *European Journal of Information Systems*, 1-22.
- Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms. *Information systems*, 101, 101804.
- van der Valk, H., Haße, H., Möller, F., & Otto, B. (2022). Archetypes of Digital Twins. *Business & Information Systems Engineering*, 64(3), 375-391.
- Weking, J., Mandalenakis, M., Hein, A., Hermes, S., Böhm, M., & Krcmar, H. (2020). The impact of blockchain technology on business models—a taxonomy and archetypal patterns. *Electronic Markets*, 30(2), 285-305.