RESEARCH ARTICLE

Envisioning and proposing Data Mesh for Research Data Management in the Engineering Sciences

Mario Moser [1], Tobias Hamann [1], Anas Abdelrazeq [1], Robert H. Schmitt [1, 2]

1. Laboratory for Machine Tools and Production Engineering (WZL), RWTH Aachen University, Aachen.
2. Fraunhofer Institute for Production Technology (IPT), Fraunhofer Institute, Aachen.

Date Submitted:
2025-04-25

Keywords:
Data Mesh, Research Data Management (RDM), Engineering Data, Engineering Sciences, Decentralised Data Architecture, Data Infrastructure, Data Publishing, Data Reuse

Data availability:

Software availability:

Corresponding Author:
Mario Moser
mario.moser@wzl-iqs.rwth-aachen.de

Abstract. In Research Data Management (RDM), data publishing infrastructures play a crucial role for efficient data provisioning and reusage. Data repositories (generic or discipline-specific) serve for this. Nevertheless, they focus rather on technical aspects without including sociological elements; they struggle to cover the heterogeneous nature of research data (formats, sources); and they are typically centralised, leading to increased complexity in operation and maintenance. In industrial data management, the Data Mesh concept as a decentralised and socio-technical approach has been introduced. Data is handled as products for increased usability, ownership is shifted to the respective domains experts, and a federated governance achieves standardisation while allowing discipline-specific decisions. Based on literature review, the distributed characteristics and further requirements of (engineering) research are mapped with the Data Mesh concept. In this envisioning, Data Mesh and its design principles overall appear appropriate as research data publishing infrastructure. A high level architecture is presented leveraging existing RDM components. Although, as differences in details become apparent, items for further adaptions of Data Mesh for RDM are pointed out.

1 Introduction

The publication and availability of research data is gaining increasing importance. Researchers need to publish their data to make their results reproducible and transparent, and other researchers might reuse existing data to gain new insights. Research data as well as existing data sources, especially in the engineering sciences, are heterogeneous [1] [2]. This requires different types of repositories as specialised solutions for a certain domain or research method. Currently published data is in various data repositories, in data publication journals, or on other web pages. On the other side, these existing sources are often distributed, making it hard to discover them. Once such a dataset is found, its accessibility depends on whether the source is technically open, or isolated – sometimes referred to as 'data silo'. York Sure-Vetter, director of the National Research Data Infrastructure (NFDI), summarised this with his quote[1] that the challenge is less the creation of data, but the findability and accessibility due to missing interconnected and protected research

1. The original quote in German: "„Wir ertrinken in Daten, können sie aber nicht finden." Es fehlten miteinander verknüpfte Datenräume für die Wissenschaft, sagt er, und meint geschützte virtuelle Orte, die den Austausch von Daten über Fachgrenzen hinweg erleichtern." [3].

data spaces for data exchange [3]. Beyond this, data quality and the integration into ecosystems are recognised as as current issues for research data repositories, as well as dynamics in rapid technology development for repositories Schöpfel [4].

In industrial data management, the so-called *Data Mesh* has been developed. In this socio-technical and decentralised solution approach, data sources remain intentionally distributed, to leverage but open up the specialised data silos. Domain Ownership transfers responsibility away from central IT teams closer those responsible for the data creation, following a socio/organisational structure. Product Thinking for data makes data-self-contained and 'productised' for potential reusers. A Self-Serve Platform is the central entry point to find registered datasets and maintain data along its life cycle. A Federated Governance defines global rules while allowing teams to apply additional rules locally to reach standardisation and interoperability.

Objective of this paper is to envision the application of the industrial Data Mesh concept for research data. Based on its conceptualisation, Data Mesh appears to be a fitting solution approach for Research Data Management (RDM), especially in engineering sciences. The heterogeneity of engineering data [1] [2] – ranging from measurement values over simulations and sample provenance tracking to the documentation of technical experiment setups [5] – requires individual specialised repositories instead of a central one. RDM is a 'brownfield' with existing repository solutions, which can be connected and linked in a decentralised way rather than building a new monolith one from scratch; other existing technical services might be leveraged in such a Data Mesh as well. Although these first solution approaches exist, they do not cover organisational and socio aspects. To best knowledge, so far the idea of Data Mesh for RDM has been only mentioned as side note by Diepenbroek et al. [6][2]. The scope of this paper will be specifically on RDM in the engineering sciences in Germany, nevertheless the results might be applicable to other domains or countries. In this paper is concluded that the Data Mesh concept overall fits to RDM in the engineering sciences. Although having similar setting between industry and research, differences in detail will require some transformation. As a result, a conceptual target picture is presented and areas of adaption towards a RDM Data Mesh are identified.

This paper continues with an introduction of current data infrastructures and the Data Mesh approach in Chapter 2. The scientific research landscape as objective for the proposed Data Mesh is presented in Chapter 3, working out the key characteristics. In Chapter 4 the methodological approach used in this vision paper is introduced. Based on the identified characteristics, the application of Data Mesh in RDM of the engineering sciences is envisioned in Chapter 5, concluding with a summary and outlook in Chapter 6.

2 Data Infrastructures

In this chapter, current data infrastructures will be discussed (Section 2.1), before presenting the Data Mesh approach in greater detail (Section 2.2).

2.1 Data Infrastructure and Architectures

In industry, data infrastructures and architectures are used to systematically collect, process, provide, and analyse data. Starting from mid of 1980s, (relational) **Data Warehouse**s (DWH)

2. While there "a networking of RDCs resembling a data mesh" [6] is imagined, with one Research Data Common (RDC) per discipline like e. g. the engineering sciences, this paper here proposes a Data Mesh for the engineering sciences.

have been introduced. Analytical data is separated from operation data, and remodeled for reading performance. Data pipelines are build to extract, transform, and load (ETL) data from (multiple) source systems into the DWH. Transformation steps included quality checks and aligning data structure ('schema on write'). This leads to high-quality data, but increases costs and complexity in maintenance. [7] [8] Those DWHs were not designed to handle non-relational data. Because of these drawbacks and with the raise of 'big data', at the beginning of the 2010s **Data Lake**s (DL) have been introduced. These big data storage systems allows to ingest document-based data in a folder-based structure. Data transformations are applied when data is read ('schema on read'). DLs do not systematically provide the level of quality of DWHs and potentially leading to so-called 'data swamps' if data is only 'dumped' into the DL. [8] Both DWH and DL are centralised architectures. In contrast, the **Data Mesh** as decentralised paradigm is introduced in 2019 (rf. section 2.2 below). At a similar time, **Data Fabric** (DF) has been introduced as a hybrid architecture. Data is integrated from various source systems within a central layer, using data pipelines and standardised connectors, for users and AI applications. [9] [10]

For sharing data within and between organisations, **Data Space**s (DSs) have emerged. They can contain distributed data sources from an organisation. Focus is on provisioning and managing data, while having no or only limited data integration into a common schema. [11] Multiple DSs together can form and support **Data Ecosystem**s (DEs) [12] [13]. From a technical perspective, DEs consists of the components datasets, data operators, metadata, and mappings [14]. DEs are complex network of organisations and individuals, with actors, their roles and relationships, where data is created, managed and shared. Data sharing is done in an interoperable, transparent and self-organised way. [15] [16] [17] [14] DEs can be organisational, distributed, federated, and virtual with regards to level of control over resources and participant interdependence [18] [1]. Scientific data and domain-specific requirements can pose challenges for design of DEs [1]. Examples for industrial DSs and DEs are GAIA-X and International Data Spaces (IDS) [19], in academic the European Open Science Cloud (EOSC) and FAIR Data Spaces[3].

2.2 Data Mesh

While DWHs focus on the ETL pipelines, Data Mesh emphasizes the data itself and lets data remain in its original (decentralised) source systems. According to Serra [8], it aims to solve the four main challenges of centralised systems: missing ownership, low data quality, technical scalability, and organisational scalability [8]. Missing ownership and data quality issues have been mainly recognised in DLs. Centralised (monolithic) systems lack technical scalability and inherent complexity, making it hard to operate and maintain such systems over years [20] [21]. Central teams in data architectures can become bottlenecks, which requires (re)organisation of organisational structures, and therefore introduces sociological components. By doing so, *data democratisation* is aimed to be achieved, i. e. (governed) access to their data within an organisation, along with training the employees respective capabilities [20]. On technical as well as business side, data should be available faster for analytical insights and data-driven products, while automating and therefore reducing the governance effort [10].

The concept under the term "Data Mesh" has been proposed by the IT consultant Zhamak Dehghani, first in two blog posts in 2019 [22] and 2020 [23], and subsequently in 2023 in her book [20]. Main motivation is to make data architectures scalable by reducing organisational and

3. project until 2024-12-31

technical bottlenecks, to increase data quality with business ownership of data, and to remove separation between operational and analytical data. Data Mesh is a decentralised and socio-technical data management approach (even referred to as 'paradigm') for scalable acquisition, access, and management of analytical data in large and complex organisations [20]. Data Mesh is a holistic concept bringing together elements of technology, strategy, and methods, but it does not specify a certain technology; the introduction will require organisational and cultural change [8] [10]. Related concepts have been described by Strengholt [24] as well with a reference to the term 'Data Mesh'. First scientific publications have been made by Machado et al. in [25] and [26]. Due to the origin in industrial data management, Goedegebuure et al. [27] conducted a gray systematic literature review (SLR) on the principles and their relation to service-oriented architecture, and Bode et al. [28] interviewed industry experts regarding challenges and implementation strategies. First use cases are described in scientific literature, e. g. in banking [29], public sector [30], product lifecycle management in automotive [31], and military applications [32].

Data Mesh is summarised in **four principles**, which interact with each other [20] and are in combination more than a decentralised data management soley [8]:

I. **Domain Ownership**: The responsibility for data (ingestion, description, curation etc.) is shifted closer to those responsible for the data generation / data source instead of centralised teams. This causes decentralisation and moves responsibilities and new tasks towards these data owners within their domains. [8] As subject matter experts, they have more domain knowledge to explain their data and to identify potential data quality issues [20], although they need support with new tasks like data ingestion. In companies, lines of business can be chosen as domains [10].

II. **Data Product**: Analytical data itself is not viewed as a byproduct, but as a valuable outcome potentially relevant for other 'consumers' [8]. While domains may have data 'silos', provisioning data as products helps open them up [20]. Data products are a self-contained and independent units providing business value, focused on target users [8]. They are enabled by a *data domain owner* and built by a *data product owner* [10]. To build a data product, raw data is equipped and enriched with metadata, quality-assured, and data rules and policies are added. The data product owner takes care for data profiling, discovery, and transformations to form the data product [10]. Each domain has domain teams, Application Programming Interface (API) code, data, metadata, and infrastructure [8]. Underlying technology can change without affecting the data product, as long as interfaces remain consistent, enabling technological scalability. The combination of existing data products generates new 'aggregated' data products. The idea is described as *data as a product* and implemented as a *data product*.

III. **Self-Serve Platform**: A centralised platform that allows users to maintain their own data along its life cycle, and discover and access existing data [20]. It is created and maintained by a centralised platform team. The platform should provide infrastructure for automated provisioning, maintenance, and monitoring of data products. [8] A data catalog or a data marketplace can be used to make indexed data products findable for others [10]. Within the platform, governance is applied automatically for all data products [20].

IV. **Federated Governance**: A data governance describes rules and procedures in the collaboration between domains, as well as with the central teams. The data governance is federated, i. e. some elements are given organisationally top-down to achieve standardisation within the Data Mesh, while other aspects are left up to the domains for their individual design and local autonomy. Global rules include e. g. standard for interoperability, generic data quality metrics, and data security. Nevertheless, it is expected that having all rules centralised causes slower decision process, inflexible regarding future changes, and missing capability to capture domain-specific nuances and requirements. Therefore domains define their own governance, e. g. on domain-specific standards or specific data quality requirements. Goal is to balance interoperability on the one side with autonomy and agility on the other side. The execution and monitoring of governance should happen in an automated way. [20] [8]

Data Mesh is inspired by **existing related approaches**, particulary applied to data management before. The I. principle, Domain Ownership, is coming from domain-driven design (DDD) by Evans [33] and applied here to data management within a domain-oriented architecture, putting data into contexts [8] [24]. The Product Thinking approach is applied to data within the II. principle for Data Products. The idea of microservices / service-oriented architecture (SOA) is applied to data [27] which requires standardisation, addressed with the Federated Governance (principle IV). Governance has already been introduced in e. g. DLs. Self-serve platforms (principle III) have been used in data management and business intelligence before.

Along with the organisational changes and technical implementations, new **roles** are emerging and existing ones are changing. The domain-orientation shifts tasks from one centralised team to each independent domain team in the respective domains [8]. Main roles for the data product are the *data product owner* from a business perspective, and the *data product developer* from a technical one [20] [8] [27]. On the self-serve platform, the *platform product owner* [20] (or *data platform teams* [27]) are responsible for provided services there. Regarding governance, *federated governance teams* consisting of several representatives from the domains [20] [8] [27].

The previous section on related approaches leads to the first **criticism**: Data Mesh is not a new idea with regards to the data products, working like Data Marts before in a DWH [34]. Serra [8] argues that technological advancements have reduced scalability issues since the 1980s. Interdisciplinary domain teams require business analysts and data engineers to adopt each other's perspectives [34]. Focusing on data over ETL pipelines raises concerns about maintainability of data processing logic compared to model-based development [34]. Governance may not resolve dataset duplication issues effectively [34]. Designed for intra-organisational use, challenges arise in inter-organisational sharing, as seen in clinical trials [35]. The Data Mesh approach is considered as complex in implementation. It is considered as a theoretical concept, with limited technology available, where in practice differences and exceptions from original concept's vision are likely [8]. Dehghani states that the Data Mesh concept described in [20] is still under development and intentionally open designed to be adapted.

3 Environment: Engineering Sciences Research and Research Data Management in Germany

Object of consideration in this work is the German research landscape (Section 3.1) with engineering (Section 3.2) and its RDM (Section 3.3), which are introducted in the following chapters. They are described following 'People, Organisation, Technology' [36]. This structure can be found similarly in Design Science Research (DSR) to characterise the environment [37] [38]; similarly, Donner [39] uses 'structure, task, technology, people', as proposed by Leavitt [40], to evaluate RDM systems and organisation. For building data infrastructure in a field, Schultes and Wittenburg [41] states that this is "in both technical and social domains" [41].

3.1 Research Landscape in Germany

Figure 1 shows a high-level visualisation of the scientific landscape in Germany. Its purpose is to briefly introduce the main groups of actors and their relationship, which are characterised in the following, without claim to completeness.

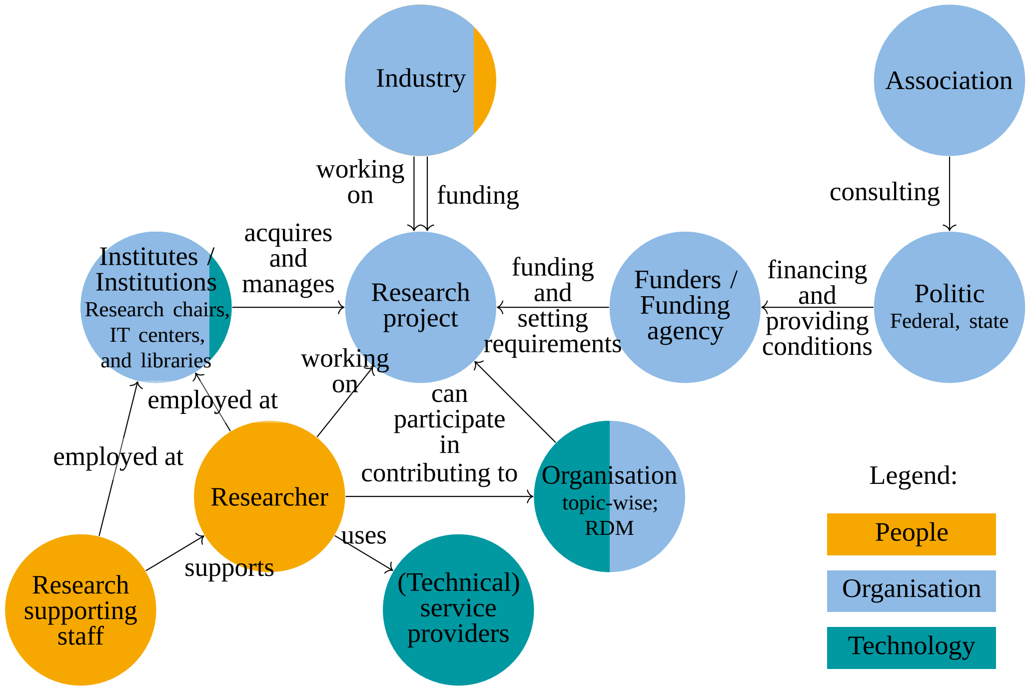

Figure 1: Simplified schematic representation of the German research landscape

'**People**' in research are the individual **researcher**, conducting research in his/her specialised field, next to teaching activities and project acquisitions. Prerequisite is an academic degree, mostly a master's degree or comparable. In Germany researchers typically have to leave academica after latest six years due to the German Academic Fixed-Term Contract Act (*German:* Wissenschaftszeitvertragsgesetz) (WissZeitVG). They are supported by non-scientific staff. [42] Depending on the hierarchical organisation levels, a researcher is supervised by a research group leader, chief engineer, and professor.

Researchers typically work for a **research facility** at an **institution** as **organisations**. Such institutions can be differentiated into mainly three types: In 2023/2024 in Germany there have been 109 universities and 215 universities of applied sciences [43] [44], as well as non-university research institutions (Max Planck institutes, Fraunhofer, Leibniz, and Helmholtz; the German Aerospace Center (*German:* Deutsches Zentrum für Luft- und Raumfahrt) (DLR) is a research institution of the Federal Republic of Germany in the form of a registered association). All these institutions usually have an IT center and a library as central departments.

Based on their interest, researchers within their institutes can participate in **topic-wise associ-**

ation, like the Wissenschaftliche Gesellschaft für Produktionstechnik (WGP) in engineering. Overarching technical **service providers** offer their services to researchers, institutions, or within projects, like the German National Research and Education Network (*German:* Deutsches Forschungsnetz) (DFN), e. g. with their DFN-AAI service for authentication and authorization, the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), repository operators, or identifier services like e.g. Open Researcher and Contributor ID (ORCID).

The political system in Germany is federated, mainly between the Federal Government and the Heads of Government of the States. Research institutions in Germany receive a baseline funding by the **politic**, additional funding can come in partnership with **industry** projects. The funding and financing structure is more complex than depicted and not scope of this article. Public projects are tendered and managed by **funding agencies** (*German:* Projektträger) like German Research Foundation (*German:* Deutsche Forschungsgemeinschaft) (DFG). These funding agencies set financial, organisational, and scientific requirements. The Good Research Practice (*German:* Gute Wissenschaftliche Praxis) (GWP) [45] is a code of conduct for researchers. **Associations** can give recommendations and consulting to politics; in terms of RDM, the German Council for Scientific Information Infrastructures (*German:* Rat für Informations Infrastrukturen) (RfII) sees itself as panel of experts between politic and sciences in questions related to digital sciences.

In summary, research in Germany is decentralised at various institutions, with actors on the level of ‘People’ (researchers, supporting staff), ‘Organisation’ (institutions as well as rules), and ‘Technology’ (tools and service providers).

3.2 Engineering Sciences

Engineering sciences are a group of disciplines, all with technical characteristics [46]. The DFG classifies them into five research areas [47], each divided further, as shown in Figure 2. The evolving nature of engineering sciences is reflected in continuous revisions of this classification.

Especially the engineering sciences are driven by collaboration with **industry** partners. In 2022, the engineering sciences had the highest amount of third-party funding (*German*: ‘Drittmittel’) per university professor among all disciplines [48]. The technology used for data management especially in the engineering sciences is introduced in the next chapter in the context of RDM.

3.3 Research Data and its Management in general and for Engineering Sciences

Research is getting more **data-intensive and data-driven** [49] [50], including the engineering sciences, which Hey [51] refers to as “The Fourth Paradigma data-intensive scientific discovery” [51]. Taking data as foundation to conduct research comes along with the need for structured management of such research data. The term ‘digital research data’ covers all data digitally available that has been created in research processes or is the result of it [52], like measurement data, audio-visual content, texts, surveys, samples, and procedures like software code, simulations, and questionnaires [53]. RDM includes tools and concepts for the systematic management and finally publishing of research data as valuable resource, including “creating, finding, organising, storing, sharing and preserving data within any research process” [54].

Data in the engineering sciences with its various disciplines exhibit the characteristic of heterogeneity, e. g. sensor data, material samples, material models, HPMC data, CAD files, experimental setup documentation, and software code. Data is typically digital or has been

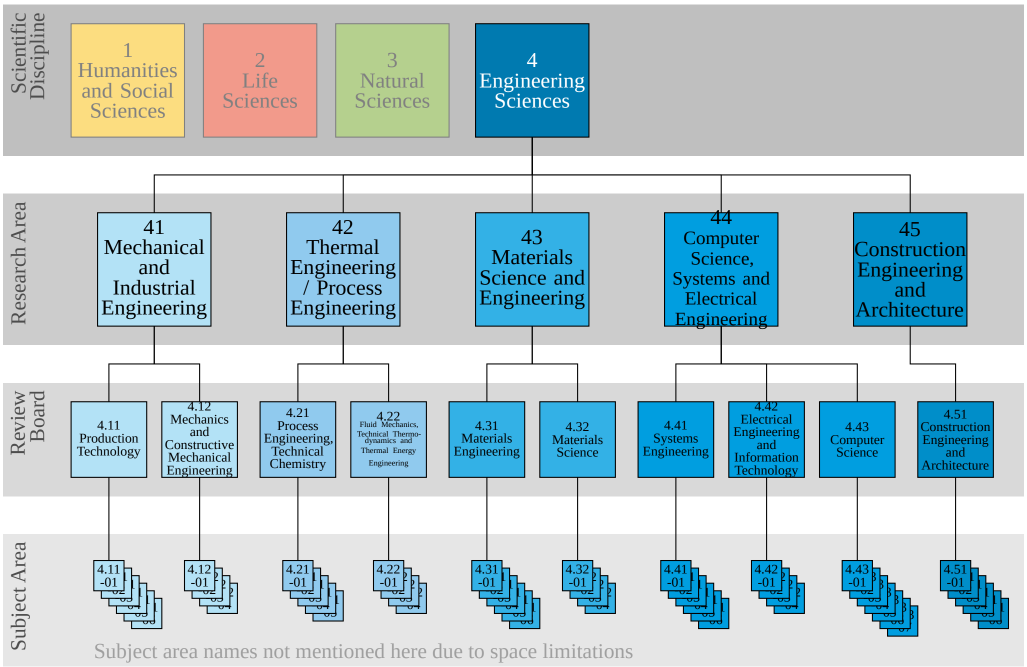

Figure 2: Classification of engineering sciences 2024 – 2028 in Germany (according to DFG [47])

digitalised. These types of data are related to the methods used within engineering sciences, like bespoke experiments, software development, simulation, High Performance Measurement and Computing (HPMC) [5]. Engineering sciences are marked by a high level of interdisciplinarity and collaboration both within its disciplines and with non-engineering fields [55].

Research data is considered along the research **data life cycle (DLC)**, starting from a planning phase, to data collection, analysis, publishing, and archiving phase. Subsequently, another research might reuse the before-mentioned dataset, closing the loop of the DLC. Figure 3 shows a version proposed by Yazdi [56], while various forms with slightly different steps and transitions exists (rf. [57] for an overview). RDM with its DLC contributes to GWP in the research process, including the publishing (guideline 13 / DLC-5) and archiving (guideline 17 / DLC-4).

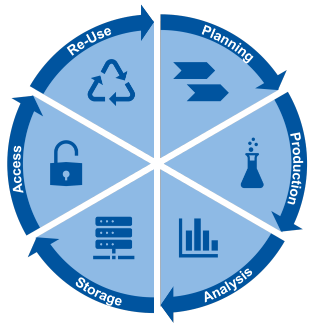

Figure 3: Data life cycle (DLC) phases (adapted from [56])

The **FAIR Principles** by Wilkinson et al. [58] have been established to give guideline on handling research data in a way that it gets ***f****indable*, ***a****ccessible*, ***i****nteroperable*, and finally ***r****eusable* (‘FAIR’) for other researchers. Schultes [59] distinguishes the FAIR principles into technical as well as content- / domain-relevant practices. Different implementations of the FAIR principles in the form of FAIR metrics have been developed. Based on FAIR and GWP, several universities institutionalised FAIR in their RDM guidelines and policies, e. g. RWTH Aachen University [60] and TU Darmstadt [61].

Within **institutions**, RDM as overarching topic is typically originated at IT centers and libraries [62]. Especially libraries come from the preservation of knowledge – formerly in the form of books, now extended with digital data [63]. IT centers provide technical infrastructure for data and information management. In their case study about RDM at the University of Cologne (Germany), Curdt et al. [64] report decentralisation with regards to organisational structures, RDM activities, actors in information infrastructure, and competencies [64]. Several **organisations and associations** are active in the field of RDM in Germany. The Research Data Alliance (RDA) is an international initiative founded in 2013 with a German community, conducting events and creating results in specialised working groups. In Germany, 11 RDM initiatives on the level of federate states are working as regional networks [65]. In 2020 the National Research Data Infrastructure (NFDI) has been founded by Germany's federal and state government. This registered association currently includes 26 discipline-specific consortia for RDM. For the engineering sciences, the National Research Data Infrastructure for the Engineering Sciences (NFDI4ING) and four more specialised consortia exist. NFDI4ING covers all engineering disciplines (according to DFG, rf. Figure 2) with an approach of developing solutions for engineering research *methods* of archetypical researchers ('archetypes'). [5]

From technological perspective, **Digital Objects (DOs)** consisting of data and key metadata, including an identifier (*handle*), which are stored in a network of distributed repositories [66]. For FAIR Digital Objects (FDOs), the DO is extended to improve FAIRness. FDOs are stable units that bundle information for reliable interpretation and processing, using encapsulation and abstraction, to create domain-independent layers around typically domain-specific data. Core is the DO, represented by a digital bit sequence, accessible via a persistent identifier (PID). As layer around the DO and identifier, metadata provide context, while standards such as file formats or operational requests enhance functionality. [67] [68] The FAIR Digital Object Framework applies FAIR principles through a conceptual model [69]. FDOs are central for FAIR data ecosystems [67] and support convergence in distributed data infrastructures [41]. DOs can be identified unique and persisted by applied **identifiers** like Digital Object Identifier (DOI). Beside that, identifiers for researchers (e. g. Open Researcher and Contributor ID (ORCID)) and research organisations (e. g. Research Organization Registry (ROR)) exist.

Best practice in RDM is to publish data in a **repository**. Such repositories are key in RDM [4] and typically provide additional functionalities, e. g. persistence and identifier, compared to uploading a dataset solely on a web page. Repositories can be either generic (e. g. zenodo) or discipline-specific (e. g. for engineering: 4TU.ResearchData, Open Energy Platform and several FIDs[4]). Schöpfel [4] discusses "diversification, not convergence", highlighting the creation of data communities suited for respective practices and needs. There is the tendency towards "diversification, not convergence" [4], highlighting the creation of data communities suited for respective practices and needs. Overviews of existing data repositories are provided in DFG's RIsources portal, by FAIRsharing.org, in the NFDI4ING Data Collections Explorer, or by Registry of Research Data Repositories (re3data) [70].

Contextual information about the published data can be made available in structured form of **metadata schemata and terminologies**. Existing generic metadata schemata are e. g. RADAR and the DataCite Schema for research outputs [71]. Discipline-specific metadata schemata, like for material sciences or for HPMC, can be found via the NFDI4ING Terminology Service, via FAIRsharing, or be build based on existing schemata using the NFDI4ING Metadata Profile.

4. Specialised Information Service (*German:* Fachinformationsdienst) (FID) is a funding program for libraries in Germany, developing and operating repositories for specialised research fields. Regarding the engineering sciences, there are FID BAUdigital (`https://www.fid-bau.de/de/`, DFG-45, years 2020 – 2023), FID Materials Science (`https://www.materials-science.info/`, DFG-43), and FIDmove (`https://fid-move.de`, DFG-44).

4 Methodological approach

So far, based on literature Data Mesh has been presented (rf. Section 2.1) and the main characteristics of Germany's research landscape, esp. for the engineering sciences, have been identified. The environment is described according to "People", "Organizational Systems", and "Technical Systems" (rf. Hevner [38]), as well as problems and opportunities, in Chapters 2 and 3. To envision a Data Mesh for RDM in the engineering sciences, the socio-technical setting of engineering RDM will be reflected within the Data Mesh concept. It is considered along the categories of the overall goal, decentralisation, socio-technical, roles, and the principles I. to IV. in the following. Based on this investigation, similarities are identified, a high-level design is proposed, and needs of adaptions are pointed out.

Scope of this work is RDM in the engineering sciences in Germany, although the results could be generalised to other countries or domains. The engineering sciences appear to be a good fit as forming one community with multiple domains within (e. g. according to the DFG classification in Figure 2, or the NFDI4ING consortium with shared engineering methods). The focus is on data in the narrow sense (esp. databases and files, relational and 'big' data), while software and software repositories are out of scope.

5 Contribution: Envisioning and proposing Data Mesh for RDM in the Engineering Sciences

Long-established (industrial) data management solutions do not meet the complex characteristics mentioned above. Data Warehouses come along with high complexity in integrating heterogeneous data into a maintainable common data model. Data Lakes have been used as research data storage (e. g. [72], [73], [74]), but not primarily of data sharing. As centralised systems, these undergo increased complexity and limited scalability when attempting to be a 'one-fits-all' solution, with regards to data model, providing functionalities, governance structures, etc. Scalability might become even more crucial with further dissemination of RDM and increasing amount of data published. Existing scientific repositories rather serve for isolated data provisioning, but are not designed for interdisciplinary research, leading to 'data silos'. Data Spaces are technical solutions but do not cover sociological aspects. RDM is more than a solely technical topic, containing socio and organisational topics, which needs to be reflected in a respective data architecture. Given these characteristics, Data Mesh appears generally suitable for data provisioning in the engineering sciences. This will be discussed in more detail in the following. Data Fabric is not considered here.

To envision Data Mesh for RDM in the engineering sciences, certain central aspect are used in Chapter 5.1 to compare the Data Mesh concept with the RDM requirements. This chapter concludes with a target picture and implications in Chapter 5.2.

5.1 Category-based Comparison between Data Mesh and RDM

In this section, the quantitative suitability of Data Mesh for RDM will be assessed in the categories of their overall goals, under consideration of decentralisation, as well as socio-technic and roles. Moreover, each of the principles I. to IV. will be discussed for RDM. Finally, it will be evaluated how far previously introduced general critism might impact Data Mesh for RDM.

5.1.1 Overall goal

Data Mesh aims to increase data findability within organisations to accelerate the usage of existing data. This matches with the need for better findability of data in RDM addressed in [3]'s quote in the introduction. For this, interoperability of data is required, both mentioned for industrial organisations via standardisation and governance, as well as in RDM as the "I" in FAIR. In Data Mesh access to data is governed, similarly in RDM FAIR data does not necessarily mean open data. With regards to the DLC (rf. Figure 3), the Data Mesh approach addresses the Access (publication) phase and the Re-Use phase, and beyond this respective data analytics and visualisation functionalities of a Self-Serve Platform would contribute to the Analysis phase. Both aim at analytical data for use cases. This matches with analysing research data [5] in order to gain knowledge, where a research project might be seen as the 'use case'.

5.1.2 Decentralisation and Federation

In Data Mesh the data sources and respective domains are decentralised, while the governance is **federated**. Same applies for RDM: Engineering subjects are heterogeneous, with partially specialised repositories existing, and no central organisation of research institutes in Germany. A federated governance provides the chance to achieve interoperability along this data.

Due to the 'brownfield' nature and heterogeneity of the engineering sciences regarding their methods and data characteristics presented before, heterogeneous data is distributed[6] in specialized repositories. Instead of building a new infrastructure 'from scratch', the Data Mesh approach adopts to this. These various data sources – which might be considered as 'data silos' – can persist in the Data Mesh concept, as long as they follow defined standards and support interoperability. The RfII recommends federated approaches [76] [77]. The idea of decentralisation in RDM has been raised in [78] and [79] before. Lehmann et al. [80] mentions decentralised practices in RDM, but due to missing access authorisations and missing data documentation in various distributed data sources, they decide for a centralized approach [80].

5.1.3 Socio-technical

As socio-technical approach, Data Mesh includes aspects of human and organisation next to IT. This goes along with the findings by Donner [39] for RDM systems, who assessed organisational factors and their interaction for implementing RDM systems. According to the DFG, digital infrastructures for research require new organisation structures and responsibilities [81]. As shown in the scientific landscape (rf. Figure 1), humans and organisations play a vital role in research, which applies more specific to RDM as well.

5.1.4 Roles and Responsibilities

The main characteristics of roles in Data Mesh (rf. page 5) appears to be already fulfilled within research: Instead of centralised teams, researchers itself publish and maintain their data, as they are already close to their collected data. With the introduction of a Data Mesh for RDM, the researcher's responsibility to publish and maintain data remains, but is more formalised. A

5. Borgman and Brand [75] categorises university data into telemetry data, academic administrative data, and research data. In scope of RDM and this paper, it is only referred to the latter, not to other (operational) types of university data.
6. Referring to data sources that are distributed because of their nature and characteristics and organisation; in contrast, data sources distributed e. g. for georeplication are not meant here.

difference is that teams are build on institute or project level, and that researchers leave after a maximum of six years. Data is collected on project level, which appears less consecutive compared to departments in companies continuously collecting business data. As supporting staff for digital data preservation, institutions have (centralised) IT teams and librarians, and may even have data stewards supporting in managing data. Introducing a Data Mesh requires a centralised team responsible to build and operate a self-serve platform and developing a federated governance; associations like NFDI4ING, NFDI or RDA might coordinate this.

5.1.5 Principle I. Domain Ownership

The **I. principle** “**Domain Ownership**” is driven by shifting responsibility towards the one producing data instead of centralised IT teams. For RDM, such decentralisation is already reflected and lived in practice when researchers themselves and not their IT departments are responsible for publishing and maintaining their data. Beyond publishing data and documentation, in Data Mesh this might include answering topic-related questions about the dataset.

A challenge is that researchers who are leaving their research institute might not be available to answer questions about their dataset. The period of maximum six years according to WissZeitVG is way shorter than the usual data retention period of ten years. Due to the project characteristic of research, finished projects might not be handed over once they are finalised. A domain-driven approach might give the chance that even after a researcher left, experts from the same domain might support in understanding the respective dataset, as they might have an understanding of the data (implicit domain knowledge) and domain-specific data quality.

Domains are nowadays already used in sciences to organise data [82] [83]. Similarly, De Smedt et al. [68] sees the organisation in distinct communities as socio-scientific context. According to Borgman and Groth [82], domains are demarcated and share inside a common knowledge, technology, or other forms of grouping. Regarding RDM, the DFG states that this is highly shaped by the respective methods of the scientific disciplines [53]. Nevertheless, the term ‘Domain’ does not necessarily has to have the same meaning between RDM and Data Mesh principle. Concrete domains for a Data Mesh in RDM are still open to be defined yet. A domain structure by organisational hierarchies (university, research institute, etc.) might prove the chance to find a successor as contact person once the initial data owner left the institute. A structure by research discipline might support better in answering domain-specific questions, nevertheless there might be high specialisation as well as interdisciplinary research. The DFG classification of engineering sciences (rf. Figure 2) could be a first approach. Instead of using a hierarchical classification system in Data Mesh, a multi-dimensional tagging approach might be better suited to describe different and interdisciplinary domains.

Although the engineering sciences are heterogeneous with regards to their various disciplines, the consideration of a Data Mesh for the whole engineering sciences seems appropriate, due to their shared methods and data characteristics. Qualitatively spoken, a more narrow scope (e. g. a single engineering sciences discipline) might hinder interdisciplinary interoperability; a wider scope (e. g. one Data Mesh for all sciences) might create difficulties in defining data governance due to different domain-specific characteristics in the individual fields. The design allows datasets to be part of more than one Mesh, as long as they fulfill the respective governances.

The same way the term ‘owner’ and its implications can be discussed. In the context of Data

Mesh, ownership referrers to responsibility to maintain data, but not in terms of a possession. Especially in collaborative research projects, there might be more than one owner. Like per definition all projects, research projects are limited in time. It is not expected that usually a dataset gets updated once a research project ends; nevertheless, the maintenance (‘ownership’) of an existing dataset after project end remains an open point. One element of responsibility is the **data quality**. For Dehghani [20] this is one of the main reasons to give ownership to the domains. In sciences, Iglezakis and Schembera [84] mentions, among others, the need of quality management in repositories for data publishing. It is worth mentioning that – with or without Data Mesh – researchers as data ‘owners’ require respective training, ands incentives.

5.1.6 Principle II. Data as a Product

In the **II. Data Mesh principle**, data is treated as **data products**. The ‘consumers’ in RDM are mainly other researchers reusing data, who’s reuse purpose – beside reproducability of scientific results – might be different and a-priori unknown for the ‘producer’. It is best practice to publish research data in a respective repository. After publishing there, it can be taken as **data source** within a scientific Data Mesh. In contrast to the industrial Data Mesh approach, such repositories as infrastructure are not operated by the researchers (seen here as the ‘domain owners’) themselves, but repository operators like organisations/institutions, research projects, or governmental actors. **Identifiers** like DOI make data findable. The microservice thinking allows to adapt to latest technology without making explicit specifications regarding technology, and prevents the effect of ‘vendor lock-in’. In general every repository could be leveraged and connected to the Data Mesh, as long as it is accessible and complies with the rules of such a Data Mesh, defined in a Federated governance (rf. chapter 5.1.8). Data Mesh claims data products to be **self-contained**, which can be fulfilled for research data practices with regards to metadata, but not regarding the repository as external infrastructure used. From research perspective, data should follow the FAIR principles. Dehghani [20] mentions the FAIR principles as well, and demands data to follow the DATSIS principles: Being Discoverable, Addressable, Trustworthy, Self-describing, Interoperable, and Secure [20] [26]. DATSIS appears to overlap with FAIR in certain aspects, while Trustworthy and Secure go beyond it. Similarities and differences between **FDO** concept and Data Products will be investigated in future publication. Enriching data includes **metadata** as context to the data itself, a respective usage license, and lineage. Respective metadata schema have been developed in RDM; generic metadata enables interoperability across domains, while domain-specific ones express content more precisely. By adding **licenses** (like Creative Commons) rules for data reuse can be described in a common way. In case a dataset is continuously updated within the repository, it requires more than just an **identifier**, in order to refer to a specific version of the data (e. g. by date or a versioning number). Another requirement is **data lineage**, i. e. to comprehend the processing and propagation of data leading to the considered dataset [85]. Data products in Data Mesh which are build on other data products (referred to as “aggregated data products”) represent such a lineage trace. Information about available data products can be provided in e. g. a data catalog.

Within the data product approach, research data can be modelled for different needs: For increased transparency of the preprocessing, researchers could provide raw, preprocessed, and standardised data. Raw data is captured directly, e.g. the sensor; processed data included cleaning, transformation, and enrichment activities; and finally data is provided in a more standardised and interoperable way, e. g. according to *Model in the Middle* [86] [87], or following ontologies like *metadata4ing* (m4i) [88]. The file conversion into a more open file format supports accessibility, but inherents risks of losses during conversion, so for transparency data in both proprietary and open file format could be provided as data products. From a scientific perspective, this would ensure that researchers can comprehend (and potentially reuse) the raw data as well the preprocessing steps until the preprocessed dataset. Processing steps generally inherent the risk of failures or implicit assumptions, therefore requiring the data and processing code. Data Products contain the **software** for data creation and processing as well. In RDM, it is best practice to publish software separately in a software repository. This separates software code from the published data, in contrast to the original Data Mesh concept.

Beyond data from scientific contributions, Data Mesh allows to connect public datasets (‘**Open Data**’) via API and to provide it in the platform, e. g. public weather data, traffic data, geographical information, or material data. Researchers would find such open data and enrich their analytical data with it. However, such open data differs in term of provisioning and usage: Usage rights of the initial publisher need to be clarified. Such a data product would be continuously updated, in contrast to research data finished once the related project ended. With such external data, domain ownership in the sense of the I. principle needs to be clarified.

5.1.7 Principle III. Self-serve Platform

For the management of data products, Dehghani [20] conceptualised a **Self-Serve Platform**. Purpose of this **III. Data Mesh principle** is to have central infrastructure to maintain data along its complete lifecycle, which includes an initial onboarding and reusage until retention. Data products are provided here with domain ownership under a federated governance. A research Data Mesh would address certain phases of the research **DLC** (rf. Figure 3). Data publishing (DLC-5) includes an onboarding and maintenance (until deletion) on the self-serve platform, so that it can be found and reused (DLC-6) from there. Beyond, the self-serve platform might provide some basic visualisation and analysis capabilities (DLC-3), nevertheless data analysis in (engineering) sciences typically requires more specialised individual tools. For the deletion (in the underlying repository, not in the scientific Data Mesh itself) retention periods (often 10 years according to GWP) must be adhered to. According to FAIR principle A2., metadata should remain even if the dataset is deleted. Onboarded datasets are **registered** centrally with their metadata, e. g. in a data catalog or a knowledge graph, making research data from various sources findable. In the RDM landscape, repository indexes have been developed by re3data and FAIRsharing before.

When provisioning a dataset within the RDM Data Mesh, additional **Key Performance Indicators (KPIs)** can be provided. These can be generic or domain-specific, and manually or automatically assessed, e. g. for data quality metrics. Although in an industrial Data Mesh datasets could be rejected due to low data quality, for engineering sciences even a ‘low-quality’ dataset supports transparency and reproducability; nevertheless, domain owners should ensure data quality or probably might add a ‘(warning) indicator’. Usage metrics like the number of reusages (lineage) or citations serve as scientific credit and enhances visibility of the dataset and its owner. Data lineage makes transparent how data is combined to create new datasets. Usage, lineage, and update times might help in the decision if a dataset is kept once the retention period is over. Administrative KPIs like the up-/downtime of a dataset can also be applied.

5.1.8 Principle IV. Federated Governance

Similar to Data Lakes, such a platform requires governance, described in Data Mesh's **IV. principle**. A **Federated Governance** fosters interoperability within the engineering sciences by standardisation. A shared governance is expected to reduce monopolies/oligopolies [4]. Standards are defined globally for all engineering disciplines, while local guidelines leave fields autonomy to maintain standards specific for their field. The balancing, which aspect is either locally or globally to be defined, is subject of future research, and might be challenging for interdisciplinary research. The elements of a governance needs to be defined and aligned with the engineering sciences community. In the intersection of engineering sciences and RDM, NFDI4ING or RDA are places where the engineering community meet to formulate such a federated governance. With RDM as a 'brownfield' of existing tools and services, these might be further established within such a governance. The Creative Commons (CC) data licenses could be chosen. For authentication and authorization, Shibboleth, DFN-AAI, or the upcoming IAM4NFDI as NFDI-AAI are candidates. Various metadata standards as well as data models have been developed that might be leveraged.

5.1.9 Previously mentioned Criticism on Data Mesh with regards to RDM

The general **criticism** on Data Mesh (rf. page 5) is assessed regarding its potential impact on a Data Mesh for RDM. Building upon existing approaches (mentioned before on page 5) offers the benefit to use concepts with a certain maturity and experience available. The room for interpretation, the lack of implementation details [8], and limited scientific literature will diminish over time through ongoing research and documentation. Unlike its industry focus on *intra*-organisational data provision (i. e. within one organisation), Data Mesh for RDM must facilitate *inter*-organisational sharing among multiple research organisations. Initial strategies for this are outlined by Falconi and Plebani [35]. The shift in organisational structures required in industry [8] are less applicable to the inherently decentralised structure of (engineering) sciences. Researchers already perform roles akin to data engineers and analysts, yet decentralisation transfers data security responsibilities to researchers as data providers as well [8]. The issue that domains primarily focus on data products for their domain and their requirements [8] might exist in (engineering) sciences as well; nevertheless, publishing data even with a subsequent reuse purpose unknown in the form of data product potentially fosters reusage. Technical dependency of data products built upon each other ('aggregated data product') and impacts in case of changes [8] remain; nevertheless datasets typically stabilize post-project completion, barring repository changes. Criticisms regarding a lack of technology in place are countered by existing RDM technologies, as shown in Section 3.3 and in Figure 4 below. Overall none of the differences seem to disqualify Data Mesh for RDM in general – rather, it requires slightly adoptions and transformations to scientific data and research.

5.2 Target Picture and Need for Adaption/Transformation

A conceptual target picture is presented in Figure 4, covering the Data Mesh architecture in combination with RDM-specific elements. The Federated Governance, Domain Ownership, Data Products, and Self-Serve Platform as the four Data Mesh principles are depicted, supplemented by available technologies and organisational structures from engineering sciences and RDM.

The Data Mesh approach and RDM can be linked at various points, including repositories, data

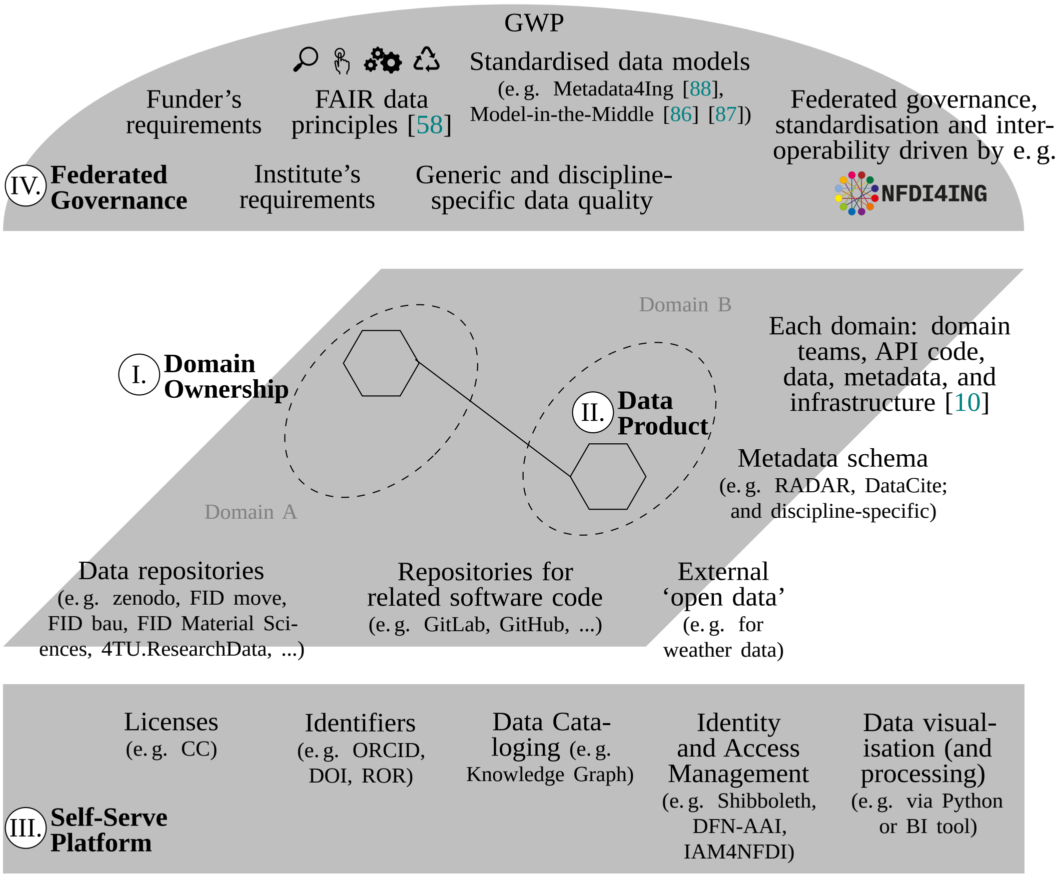

Figure 4: Conceptual target picture of a Data Mesh for RDM in the engineering sciences in Germany

quality, metadata standards, data lineage, integration of knowledge graphs, etc. Nevertheless, the characteristics of (engineering) research and in particular RDM differs to industry, where the Data Mesh concepts is initially originated for. In Table 1 industry on the one hand and research on the other hand are compared regarding the categories from Chapter 5.1. Similarities (symbol =) become clear regarding the overall goal, and the distributed data sources for analytical purpose. The symbol & indiciates elements that can benefit from each other, namely Data Products and FDOs, and DATSIS and FAIR principles. The comparison shows differences where in some aspects Data Mesh cannot be applied 1:1, but transformation/adaption seems required (symbol X), especially regarding intra-/inter-organisational, the openess of data per-se, domain definition, roles, and the head of a federated governance. With having RDM in the engineering sciences technically and organisationally decentralised, this raises the question how to set standards and governance to achieve interoperability. Here, the NFDI and more specifically the NFDI4ING might come into place. They are the central point, where the German community of researchers for RDM in the engineering sciences come together. This enables the chance to develop and establish standards and Data Mesh governance together, reflecting the heterogeneous requirements and characteristics from various engineering domains.

Table 1: Comparison of Data Mesh between industry and science; Comparison: (=) similar or equal / (&) merging concepts or synergies / (X) different, transformation/adaption might be required

category (rf. Chapter 5.1)	industry	research and science	comp.
Overall goal			
Organisational and strategic goal	Provisioning of interoperable data (for data-driven application and data analytics), data democratization	Provisioning of interoperable data (for transparency and reusability), Open Science, and Open Access	(=) / (&)
Principles	Data Mesh principles [20], DATSIS principles [23]	FAIR principles [58]	(&)
Decentralisation and federation			
Organisation	Typically one organisation, i. e. one company	sharing across multiple institutes intended (like in e. g. [35])	(X)
Structure	Decentralised teams working independently		(=)
Roles and responsibilities			
People	Teams within an organisation, e. g. departments	Individual researchers on (collaborative) research projects	(X)
Roles	Several Data Mesh roles in business and IT	Less formal defined roles for researchers, IT not publisher	(X) (&)
Principle I. Domain Ownership			
Data owner	Business units (teams)	Researcher (individual / project team; WissZeitVG), institute	(X)
Domains	The respective business units	tbd, e. g. by discipline or by research method or by institution	(X)
Principle II. Data as a Product			
Data sources	Company data sources, external data sources: ● Distributed ● Typically closed access	Research repositories, Open Data portals: ● Distributed ● Rather open access	 (=) (X)
Data products in domains	Each domain has domain team, data, metadata, API code, and infrastructure	Data, metadata; infrastructure: existing repositories (operated externally) + own infrastructure + software repositories	(=) (X)
Kind of data	Separated: Operational vs. analytical data	project data for analytical purpose (no operational data)	(=)
Data update frequency	Additional data for existing datasets, updated continuously	New datasets in every research project, closed at project end	(X)
Data encapsulation	Data Products	Digital Objects, FAIR Digital Objects	(&)
Data documentation	Self-Describing	in addition: Publication about dataset and research	(&)
Principle III. Self-serve Platform			
Data life cycle (DLC)	Creating, testing, provisioning, saving, managing and sharing of a data product [10]	Research DLC (without retention; focus on research project)	(&)
Principle IV. Federated Governance			
Head of federation	Within company hierarchy	No hierarchy, but potentially organisations (e. g. NFDI4ING)	(X)

6 Conclusion

This paper is motivated by the need to make existing research data more findable, quality-assured, and interconnected compared to how it currently is. Current data infrastructures and the socio-technical approach Data Mesh as presented in the literature have been introduced. A brief overview of the research landscape in Germany in general and a socio-technical consideration of Research Data Management (RDM) in the engineering sciences has been given, showing its decentralisation, main actors, and existing solutions. Based on the identified characteristics, it has been argued how the Data Mesh concept can fit to RDM in the engineering sciences. The decentralised and heterogeneous characteristic of data sources and data types, the data sharing requirement, and the integration of existing tools ('brownfield' landscape) are the main drivers to apply Data Mesh. Domain Ownership formalises the researcher's responsibility for their research data more, and might provide additional options to understand data once a researcher left sciences. Data Products offer a form of standardisation in data provisioning while leveraging existing repositories, under the fact that often a data reuse purpose is not known upfront. Federated Governance balances between local (for domain-individual rules) and global rules (to reach and ensure standardisation and interoperability across the data in the Data Mesh), providing the chance to open up 'siloed' datasets for interconnection. Finally, the Self-Serve Platform should enable researchers to manage their data inside the Data Mesh at a one-stop-shop (data provisioning perspective), as well as to discover interoperable data (data reusage perspective). Data quality will be demanded in data governance, ensured by the owners according to general as well as domain-specific requirements, and will be measured within the platform. A first high level target picture of a Data Mesh in RDM has been designed (rf. Figure 4), taking existing tools and organisations into account. Since Data Mesh has been developed for (industrial) organisations, the approaches cannot be applied 1:1 to research (rf. Table 1), this will require further research on how to adopt and transform the Data Mesh approach for RDM.

Although characteristics are described in detail and a first high-level architecture is presented, this paper proposes the initial idea and more research on the conceptualisation is required. From a methodological perspective, this may including requirement analysis with expert interviews, a systematic socio-technical description of research landscape and RDM, and Design Science Research (DSR). The Data Mesh approach has been considered here isolated without considering combinations with other existing approaches. For the future, e. g. Data Fabric as data management approach (rf. [10]), and data spaces / data ecosystems like EOSC, Gaia-X, FAIR Data Spaces should be taking into account. Data Mesh and such approaches might benefit from each other, and interoperability between each other is desirable. The concrete design of a Data Mesh for research data in the engineering sciences is a task for future research. This includes the before-mentioned adaptions of the industrial Data Mesh approach. Future conceptualisation and implementation will not only serve for a 'engineering sciences Data Mesh', but might be beneficial for other scientific disciplines, and experiences could be fed back to the – relatively young – general/industrial Data Mesh concept.

7 Acknowledgements

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4ING consortium. Funded by the German Research Foundation (DFG) – project number 442146713.

8 Roles and contributions

Mario Moser: Conceptualization, Investigation, Writing – original draft, Visualization, Writing –– review & editing

Tobias Hamann: Writing –– review & editing

Anas Abdelrazeq: Project administration, Supervision, Writing –– review & editing

Robert H. Schmitt: Funding acquisition, Writing –- review & editing

References

[1] E. Ramalli and B. Pernici, "Challenges of a Data Ecosystem for scientific data," *Data & Knowledge Engineering*, vol. 148, p. 102 236, 2023, ISSN: 0169-023X. DOI: 10.1016/j.datak.2023.102236.

[2] S. D. Urban, J. J. Shah, M. Rogers, D. K. Jeon, P. Ravi, and P. Bliznakov, "A heterogeneous, active database architecture for engineering data management," *International Journal of Computer Integrated Manufacturing*, vol. 7, no. 5, pp. 276–293, 1994. DOI: 10.1080/09511929408944616.

[3] C. J. Meier, ""'Wir ertrinken in Daten"'," *Süddeutsche Zeitung*, 2023, https://www.sueddeutsche.de/wissen/nationale-forschungsdateninfrastruktur-daten-forschung-nfdi-1.6016012.

[4] J. Schöpfel, "Issues and Prospects for Research Data Repositories," in *Research Data Sharing and Valorization*. John Wiley & Sons, Ltd, 2022, ch. 12, pp. 211–229, ISBN: 9781394163410. DOI: 10.1002/9781394163410.ch12.

[5] R. H. Schmitt et al., *NFDI4Ing – the National Research Data Infrastructure for Engineering Sciences*, Sep. 2020. DOI: 10.5281/zenodo.4015201.

[6] M. Diepenbroek et al., "Towards a Research Data Commons in the German National Research Data Infrastructure NFDI: Vision, Governance, Architecture," *Proceedings of the Conference on Research Data Infrastructure*, vol. 1, 2023. DOI: 10.52825/cordi.v1i.355.

[7] R. Kimball and M. Ross, *The data warehouse toolkit : the complete guide to dimensional modeling* (Wiley computer publishing), eng, 2nd ed. New York, NY [u.a: Wiley, 2002, ISBN: 0471200247.

[8] J. Serra, *Datenarchitekturen*, ger, 1st ed. O'Reilly Verlag, 2024, ISBN: 9783960108740. [Online]. Available: https://content-select.com/de/portal/media/view/66cc3b99-83f8-4082-94e8-425bac1b0006.

[9] V. Sharma, B. Balusamy, J. J. Thomas, and L. G. Atlas, Eds., *Data Fabric Architectures, Web-Driven Applications*. Berlin, Boston: De Gruyter, 2023, ISBN: 9783111000886. DOI: 10.1515/9783111000886.

[10] E. Hechler, M. Weihrauch, and Y. Wu, Eds., *Data Fabric and Data Mesh Approaches with AI, A Guide to AI-based Data Cataloging, Governance, Integration, Orchestration, and Consumption*. Apress Berkeley, CA, 2023, ISBN: 978-1-4842-9252-5. DOI: 10.1007/978-1-4842-9253-2.

[11] E. Curry, "Dataspaces: Fundamentals, Principles, and Techniques," in *Real-time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems*. Cham: Springer International Publishing, 2020, pp. 45–62, ISBN: 978-3-030-29665-0. DOI: 10.1007/978-3-030-29665-0_3.

[12] F. Möller et al., "Industrial data ecosystems and data spaces," *Electronic Markets*, vol. 34, no. 1, 2024, ISSN: 1422-8890. DOI: 10.1007/s12525-024-00724-0.

[13] B. Otto, "The Evolution of Data Spaces," in *Designing Data Spaces : The Ecosystem Approach to Competitive Advantage*, B. Otto, M. ten Hompel, and S. Wrobel, Eds. Cham: Springer International Publishing, 2022, pp. 3–15, ISBN: 978-3-030-93975-5. DOI: 10.1007/978-3-030-93975-5_1.

[14] C. Cappiello, A. Gal, M. Jarke, and J. Rehof, "Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391)," *Dagstuhl Reports*, vol. 9, no. 9, C. Cappiello, A. Gal, M. Jarke, and J. Rehof, Eds., pp. 66–134, 2020, ISSN: 2192-5283. DOI: 10.4230/DagRep.9.9.66.

[15] M. I. S. Oliveira and B. F. Lóscio, "What is a data ecosystem?" In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, ser. dg.o '18, Delft, The Netherlands: Association for Computing Machinery, 2018, ISBN: 9781450365260. DOI: 10.1145/3209281.3209335.

[16] S. Geisler et al., "Knowledge-Driven Data Ecosystems Toward Data Transparency," *J. Data and Information Quality*, vol. 14, no. 1, 2021, ISSN: 1936-1955. DOI: 10.1145/3467022.

[17] J. Gelhaar and B. Otto, "Challenges in the Emergence of Data Ecosystems," in *PACIS 2020 Proceedings. 175*, 2020. [Online]. Available: https://aisel.aisnet.org/pacis2020/175/.

[18] E. Curry and A. Sheth, "Next-Generation Smart Environments: From System of Systems to Data Ecosystems," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 69–76, 2018. DOI: 10.1109/MIS.2018.033001418.

[19] B. Otto et al., "Reference Architecture Model for the Industrial Data Space," Fraunhofer-Gesellschaft, Tech. Rep., 2017. DOI: 10.24406/publica-fhg-298818.

[20] Z. Dehghani, *Data Mesh*. O'Reilly Verlag, 2023, ISBN: 9783960107248. [Online]. Available: https://content-select.com/de/portal/media/view/62d68bd8-66c4-4aae-918f-0d688677ec64.

[21] A. Loukiala, J.-P. Joutsenlahti, M. Raatikainen, T. Mikkonen, and T. Lehtonen, "Migrating from a Centralized Data Warehouse to a Decentralized Data Platform Architecture," in *Product-Focused Software Process Improvement*, L. Ardito, A. Jedlitschka, M. Morisio, and M. Torchiano, Eds., Cham: Springer International Publishing, 2021, pp. 36–48, ISBN: 978-3-030-91452-3. DOI: 10.1007/978-3-030-91452-3_3.

[22] Z. Dehghani, *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh*, https://martinfowler.com/articles/data-monolith-to-mesh.html, 2019.

[23] Z. Dehghani, *Data Mesh Principles and Logical Architecture*, https://martinfowler.com/articles/data-mesh-principles.html.

[24] P. Strengholt, *Data Management at Scale*, 2nd ed. O'Reilly Media, Inc., 2023, ISBN: 9781098138868. [Online]. Available: https://www.oreilly.com/library/view/data-management-at/9781098138851/.

[25] I. Machado, C. Costa, and M. Y. Santos, *Data-Driven Information Systems: The Data Mesh Paradigm Shift.* E. Insfran et al., Eds., Information Systems Development: Crossing Boundaries between Development and Operations (DevOps) in Information Systems (ISD2021 Proceedings)., Valencia, Spain: Universitat Politècnica de València., 2021. [Online]. Available: https://aisel.aisnet.org/isd2014/proceedings2021/currenttopics/9/.

[26] I. A. Machado, C. Costa, and M. Y. Santos, Procedia Computer Science, 2020. DOI: 10.1016/j.procs.2021.12.013.

[27] A. Goedegebuure et al., "Data Mesh: A Systematic Gray Literature Review," *ACM Comput. Surv.*, vol. 57, no. 1, 2024, ISSN: 0360-0300. DOI: 10.1145/3687301.

[28] J. Bode, N. Kühl, D. Kreuzberger, S. Hirschl, and C. Holtmann, *Data Mesh: Best Practices to Avoid the Data Mess*, version v2, 2023. DOI: 10.48550/arXiv.2302.01713.

[29] D. Joshi, S. Pratik, and M. P. Rao, "Data Governance in Data Mesh Infrastructures: The Saxo Bank Case Study," in *Proceedings of The 21st International Conference on Electronic Business*, Nanjing, China: IECB'21, 2021, pp. 599–604. [Online]. Available: https://aisel.aisnet.org/iceb2021/52/.

[30] K. Vestues, G. K. Hanssen, M. Mikalsen, T. A. Buan, and K. Conboy, "Agile Data Management in NAV: A Case Study," in *Agile Processes in Software Engineering and Extreme Programming*, V. Stray, K.-J. Stol, M. Paasivaara, and P. Kruchten, Eds., Cham: Springer International Publishing, 2022, pp. 220–235, ISBN: 978-3-031-08169-9. DOI: 10.1007/978-3-031-08169-9_14.

[31] Y. Hooshmand, J. Resch, P. Wischnewski, and P. Patil, "From a Monolithic PLM Landscape to a Federated Domain and Data Mesh," *Proceedings of the Design Society*, vol. 2, pp. 713–722, 2022. DOI: 10.1017/pds.2022.73.

[32] S. Dahdal, F. Poltronieri, M. Tortonesi, C. Stefanelli, and N. Suri, "A Data Mesh Approach for Enabling Data-Centric Applications at the Tactical Edge," in *2023 International Conference on Military Communications and Information Systems (ICMCIS)*, Skopje, North Macedonia, 2023, pp. 1–9. DOI: 10.1109/ICMCIS59922.2023.10253568.

[33] E. Evans, *Domain-driven design reference: Definitions and pattern summaries*. Dog Ear Publishing, 2014.

[34] M. DeBellis, L. Pinera, and C. Connor, "Interoperability Frameworks. Data Fabric and Data Mesh Architectures," in *Data Science with Semantic Technologies*, A. Patel and N. C. Debnath, Eds., 1st, CRC Press, 2023, ch. 13, pp. 267–286. DOI: 10.1201/9781003310785-13.

[35] M. Falconi and P. Plebani, "Adopting Data Mesh principles to Boost Data Sharing for Clinical Trials," in *2023 IEEE International Conference on Digital Health (ICDH)*, 2023, pp. 298–306. DOI: 10.1109/ICDH60066.2023.00051.

[36] E. Ulich, *Arbeitspsychologie*, 7th ed. vdf Hochschulverlag AG, 2020, ISBN: 9783728133700. [Online]. Available: https://elibrary.utb.de/doi/book/10.5555/9783728140425.

[37] J. vom Brocke, A. Hevner, and A. Maedche, Eds., *Design Science Research. Cases*. Springer International Publishing, 2020. DOI: 10.1007/978-3-030-46781-4.

[38] A. R. Hevner, "A Three Cycle View of Design Science Research," *Scandinavian Journal of Information Systems*, vol. 19, no. 2, 2007. [Online]. Available: https://aisel.aisnet.org/sjis/vol19/iss2/4/.

[39] E. K. Donner, "Research data management systems and the organization of universities and research institutes: A systematic literature review," *Journal of Librarianship and Information Science*, vol. 55, no. 2, pp. 261–281, 2023. DOI: 10.1177/09610006211070282.

[40] H. J. Leavitt, "Applied Organizational Change in Industry: Structural, Technological and Humanistic Approaches," in *Handbook of Organizations (RLE: Organizations)*, J. G. March, Ed., 1st ed. Routledge, 1965. DOI: 10.4324/9780203629130.

[41] E. Schultes and P. Wittenburg, "FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure," in *Data Analytics and Management in Data Intensive Domains*, Y. Manolopoulos and S. Stupnikov, Eds., Cham: Springer International Publishing, 2019, pp. 3–16, ISBN: 978-3-030-23584-0. DOI: 10.1007/978-3-030-23584-0_1.

[42] RfII, "Digital competencies –– urgently needed! Recommendations on career and training prospects for the scientific labour market," RfII –- Rat für Informationsinfrastrukturen, Göttingen, Germany, Tech. Rep., 2019. [Online]. Available: https://rfii.de/?p=4015.

[43] Statistisches Bundesamt, "Anzahl der Hochschulen in Deutschland in den Wintersemestern 2018/2019 bis 2023/2024 nach Hochschulart," Statista, Tech. Rep., 2024. [Online]. Available: https://de.statista.com/statistik/daten/studie/247238/umfrage/hochschulen-in-deutschland-nach-hochschulart/.

[44] Statistisches Bundesamt, "Hochschulen nach Hochschularten," destatis, Tech. Rep., 2024. [Online]. Available: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/hochschulen-hochschularten.html.

[45] Deutsche Forschungsgemeinschaft, Ed., *Guidelines for Safeguarding Good Research Practice. Code of Conduct*, 2022. DOI: 10.5281/zenodo.6472827.

[46] G. Banse and A. Grunwald, "Coherence and Diversity in the Engineering Sciences," in *Philosophy of Technology and Engineering Sciences*, ser. Handbook of the Philosophy of Science, A. Meijers, Ed., Amsterdam: North-Holland, 2009, pp. 155–184. DOI: 10.1016/B978-0-444-51667-1.50010-0.

[47] DFG Deutsche Forschungsgemeinschaft, Ed., *DFG Classification of Scientific Disciplines, Research Areas, Review Boards and Subject Areas (2024-2028)*. [Online]. Available: https://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/.

[48] Statistisches Bundesamt, "Pressemitteilung Nr. 350 vom 17. September 2024," destatis, Tech. Rep., 2024. [Online]. Available: https://www.destatis.de/DE/Presse/Pressemitteilungen/2024/09/PD24_350_213.html.

[49] S. Büttner, H.-C. Hobohm, and L. Müller, Eds., *Handbuch Forschungsdatenmanagement*. BOCK + HERCHEN Verlag, 2011.

[50] F. J. Montáns, F. Chinesta, R. Gómez-Bombarelli, and J. N. Kutz, "Data-driven modeling and learning in science and engineering," *Comptes Rendus Mécanique*, *Data-Based Engineering Science and Technology*, vol. 347, no. 11, pp. 845–855, 2019, ISSN: 1631-0721. DOI: 10.1016/j.crme.2019.11.009.

[51] T. Hey, "The fourth paradigm – data-intensive scientific discovery," in *E-Science and Information Management*, S. Kurbanoğlu, U. Al, P. L. Erdoğan, Y. Tonta, and N. Uçak, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–1, ISBN: 978-3-642-33299-9. DOI: 10.1007/978-3-642-33299-9_1.

[52] M. Kindling and P. Schirmbacher, "„Die digitale Forschungswelt" als Gegenstand der Forschung / Research on Digital Research / Recherche dans la domaine de la recherche numérique," *Information - Wissenschaft & Praxis*, vol. 64, no. 2-3, pp. 127–136, 2013. DOI: 10.1515/iwp-2013-0017.

[53] Deutsche Forschungsgemeinschaft, Ed., *Leitlinien zum Umgang mit Forschungsdaten*, 2015. [Online]. Available: https://www.dfg.de/resource/blob/172112/4ea861510ea369157afb499e96fb359a/leitlinien-forschungsdaten-data.pdf.

[54] A. Cox and E. Verbaan, *Exploring Research Data Management*. Facet Publishing, 2018, ISBN: 9781783302802. [Online]. Available: https://ebookcentral.proquest.com/lib/rwthaachen-ebooks/detail.action?docID=5419702.

[55] F. J. Ekaputra, M. Sabou, E. Serral, E. Kiesling, and S. Biffl, "Ontology-Based Data Integration in Multi-Disciplinary Engineering Environments: A Review," *Open Journal of Information Systems (OJIS)*, vol. 4, pp. 1–26, 2017. [Online]. Available: https://www.ronpub.com/OJIS_2017v4i1n01_Ekaputra.pdf.

[56] M. A. Yazdi, "Enabling operational support in the research data life cycle," Dec. 2020. [Online]. Available: https://ceur-ws.org/Vol-2432/paper1.pdf.

[57] T. Hamann et al., "Matching data life cycle and research processes in engineering sciences," *ing.grid Preprint*, 2024. [Online]. Available: https://preprints.inggrid.org/repository/view/42/.

[58] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160 018, 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.

[59] E. Schultes, "The FAIR hourglass: A framework for FAIR implementation," *FAIR Connect*, vol. 1, no. 1, pp. 13–17, 2023. DOI: 10.3233/FC-221514.

[60] RWTH Aachen University, *RWTH Aachen Research Data Management Guidelines*, 2024. [Online]. Available: https://www.rwth-aachen.de/cms/root/forschung/forschungsdatenmanagement/~ncfw/leitlinie-zum-forschungsdatenmanagement/?lidx=1.

[61] TU Darmstadt, *Guidelines on Digital Research Data at TU Darmstadt*, 2022. [Online]. Available: https://tuprints.ulb.tu-darmstadt.de/23200/2/Guidelines_Research_Data_2022_en.pdf.

[62] *Praxishandbuch Forschungsdatenmanagement*. De Gruyter Praxishandbuch, 2021. DOI: 10.1515/9783110657807.

[63] C. L. Martin, "Wissenschaftliche Bibliotheken als Akteure im Forschungsdatenmanagement," 2013. [Online]. Available: https://libreas.eu/ausgabe23/03martin/.

[64] C. Curdt, J. Dierkes, and S. Kloppenburg, *RDM in a Decentralised University Ecosystem – A Case Study of the University of Cologne*, 2022. DOI: 10.5334/dsj-2022-020.

[65] foschungsdaten.info, *FDM-Landesinitiativen und regionale Netzwerke*, DE, 2025. [Online]. Available: https://forschungsdaten.info/fdm-im-deutschsprachigen-raum/deutschland/fdm-landesinitiativen-und-regionale-netzwerke/.

[66] R. Kahn and R. Wilensky, "A framework for distributed digital object services," *International Journal on Digital Libraries*, vol. 6, no. 2, pp. 115–123, 2006, ISSN: 1432-1300. DOI: 10.1007/s00799-005-0128-x.

[67] European Commission and Directorate-General for Research and Innovation, *Turning FAIR into reality –- Final report and action plan from the European Commission expert group on FAIR data*. Publications Office, 2018. DOI: doi/10.2777/1524.

[68] K. De Smedt, D. Koureas, and P. Wittenburg, "FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units," *Publications*, vol. 8, no. 2, 2020, ISSN: 2304-6775. DOI: 10.3390/publications8020021.

[69] L. O. Bonino da Silva Santos, T. P. Sales, C. M. Fonseca, and G. Guizzardi, "Towards a Conceptual Model for the FAIR Digital Object Framework," *Formal Ontology in Information Systems*, pp. 227–241, 2023. DOI: 10.3233/FAIA231131.

[70] H. Pampel et al., "re3data – Indexing the Global Research Data Repository Landscape Since 2012," *Scientific Data*, vol. 10, no. 1, p. 571, 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-023-02462-y.

[71] DataCite Metadata Working Group, "DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs. Version 4.6," DataCite e. V., Tech. Rep., 2024. DOI: 10.14454/mzv1-5b55.

[72] H. Mehmood et al., "Implementing big data lake for heterogeneous data sources," in *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, 2019, pp. 37–44. DOI: 10.1109/ICDEW.2019.00-37.

[73] M. Zeng et al., "IESF: Interval Event Streaming Format for the Data Lake of Production," in *2023 Eighth International Conference on Fog and Mobile Edge Computing (FMEC)*, 2023, pp. 159–166. DOI: 10.1109/FMEC59375.2023.10306240.

[74] Y. Zhao, I. Megdiche, F. Ravat, and V.-n. Dang, "A Zone-Based Data Lake Architecture for IoT, Small and Big Data," in *Proceedings of the 25th International Database Engineering & Applications Symposium*, ser. IDEAS '21, Montreal, QC, Canada: Association for Computing Machinery, 2021, pp. 94–102, ISBN: 9781450389914. DOI: 10.1145/3472163.3472185.

[75] C. L. Borgman and A. Brand, "Data blind: Universities lag in capturing and exploiting data," *Science*, vol. 378, no. 6626, pp. 1278–1281, 2022. DOI: 10.1126/science.add2734.

[76] RfII, "Föderierte Dateninfrastrukturen für die wissenschaftliche Nutzung. NFDI, EOSC und Gaia-X: Vergleich und Anregungen für eine engagierte Mitgestaltung des Ausbaus und der Weiterentwicklung," RfII — Rat für Informationsinfrastrukturen, Göttingen, Germany, RfII Berichte No. 4, 2023. [Online]. Available: https://rfii.de/?p=8533.

[77] RfII, “Policy Paper Federated Data Infrastructures for Scientific Use - October 2024,” RfII — German Council for Scientific Information Infrastructures, Göttingen, Germany, RfII Berichte, 2024. [Online]. Available: https://rfii.de/?p=11424.

[78] M. Politze and T. Eifert, “On the Decentralization of IT Infrastructures for Research Data Management,” in *EUNIS 2019 Congress*, Norwegian University of Science and Technology (NTNU) (2019), Trondheim, Norway, 2019. [Online]. Available: https://www.eunis.org/download/2019/EUNIS_2019_paper_57.pdf.

[79] M. Hanke, F. Pestilli, A. S. Wagner, C. J. Markiewicz, J.-B. Poline, and Y. O. Halchenko, “In defense of decentralized research data management,” *Neuroforum*, vol. 27, no. 1, pp. 17–25, 2021. DOI: 10.1515/nf-2020-0037.

[80] J. Lehmann, S. Schorz, A. Rache, T. Häußermann, M. Rädle, and J. Reichwald, “Establishing Reliable Research Data Management by Integrating Measurement Devices Utilizing Intelligent Digital Twins,” *Sensors*, vol. 23, no. 1, 2023, ISSN: 1424-8220. DOI: 10.3390/s23010468.

[81] “Digitaler Wandel in den Wissenschaften,” Tech. Rep., 2020. DOI: 10.5281/zenodo.4191345.

[82] C. L. Borgman and P. T. Groth, *From Data Creator to Data Reuser: Distance Matters*, 2024. DOI: 10.48550/arXiv.2402.07926.

[83] D. Ribes, A. S. Hoffman, S. C. Slota, and G. C. Bowker, “The logic of domains,” *Social Studies of Science*, vol. 49, no. 3, pp. 281–309, 2019, PMID: 31122173. DOI: 10.1177/0306312719849709.

[84] D. Iglezakis and B. Schembera, “Anforderungen der Ingenieurwissenschaften an das Forschungsdatenmanagement der Universität Stuttgart – Ergebnisse der Bedarfsanalyse des Projektes DIPL-ING,” *O-Bib. Das Offene Bibliotheksjournal Herausgeber VDB*, vol. 5, no. 3, pp. 46–60, 2018. DOI: 10.5282/o-bib/2018H3S46-60.

[85] R. Bose, “A conceptual framework for composing and managing scientific data lineage,” in *Proceedings 14th International Conference on Scientific and Statistical Database Management*, 2002, pp. 15–19. DOI: 10.1109/SSDM.2002.1029701.

[86] W. M. P. van der Aalst, “Experiences from the Internet-of-Production: Using “Data-Models-in-the-Middle” to Fight Complexity and Facilitate Reuse,” in *Business Process Management Workshops*, J. De Weerdt and L. Pufahl, Eds., Cham: Springer Nature Switzerland, 2024, pp. 87–91, ISBN: 978-3-031-50974-2. DOI: 10.1007/978-3-031-50974-2_7.

[87] I. Koren et al., “Navigating the Data Model Divide in Smart Manufacturing: An Empirical Investigation for Enhanced AI Integration,” in *Enterprise, Business-Process and Information Systems Modeling*, H. van der Aa, D. Bork, R. Schmidt, and A. Sturm, Eds., Cham: Springer Nature Switzerland, 2024, pp. 275–290, ISBN: 978-3-031-61007-3. DOI: 10.1007/978-3-031-61007-3_21.

[88] D. Iglezakis et al., “Modelling Scientific Processes With the m4i Ontology,” *Proceedings of the Conference on Research Data Infrastructure*, vol. 1, 2023. DOI: 10.52825/cordi.v1i.271.