Metadata Management in Data Mesh: Toward Federated Discovery and Governance*

Rafael H. Suguimoto[1], **Paulo Meirelles**[1], **Kelly Braghetto**[1]

[1]Institute of Mathematics and Statistics – University of São Paulo

rafael_suguimoto@usp.br,{paulormm,kellyrb}@ime.usp.br

Abstract. *Organizations striving for data-driven operations often encounter limitations with monolithic data architectures and centralized data teams, typically in terms of scalability and data governance. The Data Mesh paradigm has emerged as a promising, decentralized alternative, yet practical guidance for implementing its key components, particularly federated metadata management, remains limited. This paper systematically reviews academic contributions on metadata solutions in Data Mesh environments, analyzing architectural patterns, standards, and technologies. We investigate existing approaches, provide a detailed comparison, and pinpoint critical open challenges to enable scalable and interoperable metadata governance across domains.*

1. Introduction

The big data paradigm has driven companies to strive towards becoming data-driven by employing real-time data processing for immediate insights, leveraging predictive analytics to forecast trends and optimize operations, and utilizing machine learning technologies. Despite these efforts, organizations struggle to extract the full potential of their analytical data, as they often opt to adopt monolithic data architectures, such as relational databases, data warehouses, or data lakes, to store and process data from various sources and use centralized data teams to handle requests from throughout the organization. These approaches can hinder their capacity to manage and process data at scale, leading to significant data governance challenges, as the quality, integrity, security, and availability of data within the company are compromised [Goedegebuure et al. 2024].

In this context, Zhamak Deghani proposed Data Mesh [Dehghani 2022], a sociotechnical approach to data management based on the idea of decentralizing data into knowledge domains, allowing organizations to leverage domain-specific insights to improve the quality of data management and transformation processes and solve bottlenecks of central data and IT teams. Beyond data distribution, Data Mesh rests on three core principles: (i) treating data as products to deliver high-quality, reusable data; (ii) a self-serve data platform to lower technical barriers and empower users; and (iii) federated data governance for cross-domain interoperability via quality, access, and security standards. The architecture consists of different planes that enable its core principles: a self-serve plane for building and managing data products, a governance plane that enforces global

*This study was financed, in part, by the São Paulo Research Foundation (FAPESP) and the São Paulo State Data Analysis System Foundation, Brazil, Process Number 2023/18026-8, the São Paulo Research Foundation (FAPESP) Process Number #2023/00779-0, the National Council for Scientific and Technological Development (CNPq) Process Number #420623/2023-0 & the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

policies and interoperability standards, and a product layer composed of the data products built by each domain. Federated metadata management is a vital element of the governance plane, enhancing data quality, discovery, scalability, and enforcing interoperability across domains.

This novel decentralized data architecture is being extensively researched in academia and the industry. [Goedegebuure et al. 2024] provides an in-depth analysis of Data Mesh through a systematic gray literature review, delving into its design principles, organizational roles, inner workings, architectural components, and core concepts. Meanwhile, [Wider et al. 2023] offers a valuable investigation into the technical and practical aspects of implementing Data Mesh architectures, drawing from their own experience in building decentralized data platforms. Several other researchers are applying these concepts to a wide array of contexts.

Although there are many theoretical discussions on Data Mesh, few studies offer practical, actionable guidance for implementation. This results in unresolved questions, particularly concerning vital governance components such as metadata catalogs. These catalogs provide the foundational services for data discovery, comprehension, and access [Oliveira et al. 2024]. Therefore, the primary contribution of this work is to systematize the current body of knowledge on federated metadata management. By cataloging and analyzing architectural patterns, metadata standards, and technical challenges from the literature, we synthesize the state-of-the-art and delineate critical gaps, thereby providing a roadmap for future research into scalable and interoperable metadata governance.

2. Methodology

The primary studies analyzed in this paper were identified through a comprehensive systematic literature review (SLR) on the broader topic of Data Mesh architectures. The SLR was conducted following the guidelines provided by [Kitchenham and Charters 2007], a well-established set of protocols widely adopted in various fields. The main SLR aimed to identify challenges and implementation patterns across all facets of Data Mesh. The search strategy targeted works indexed by renowned academic databases, such as IEEE Xplore, ACM Digital Library, and Scopus, using keywords related to 'Data Mesh' and 'Decentralized data architecture', published between 2018 and 2025.

From the final pool of 55 studies identified in the SLR, a subset was selected for the focused analysis presented in this paper according to the following specific criteria: **(i)** Explicitly discuss the design, architecture, or implementation of a metadata catalog; **(ii)** Propose a technological approach for metadata management. Studies that only mentioned metadata in passing or did not detail a specific implementation approach were excluded from this particular analysis. This focused filtering process yielded a final set of ten primary studies that directly discuss metadata catalog implementations, which form the basis of the analysis in Section 3.

3. Strategies for Metadata Management in Data Mesh Ecosystems

Several researchers examine the role of metadata in large-scale data platforms, emphasizing its importance in preventing data swamps by documenting data models, sensitivity, and lineage [Dolhopolov et al. 2024b]. They highlight current challenges of metadata management. First, there is a mismatch between federated governance and centralized metadata repositories, which are inherently single points of failure. Second, flat formats like XML or JSON are inadequate for storing metadata, as they cannot express complex inter-data relationships. Furthermore, a metadata catalog must fulfill various requirements to operate effectively in a decentralized environment, such as access control, consistency, immutability, auditing, and versioning [Dolhopolov et al. 2024a].

[Driessen et al. 2023b] investigated metadata management in decentralized environments by interviewing IT experts at an automotive manufacturer. The authors identified the challenges faced by the various roles involved in the data exchange, synthesizing five vital requirements that solutions for metadata management in decentralized data architectures must meet. The metadata management should enable data providers to convey the technical and business value of their products within the metadata model, connect data products on both a data and semantic level across domains, automate the creation of metadata, and, finally, allow data consumers to express their needs through the metadata.

The following sections present different approaches in the academic literature to mitigate the aforementioned issues and compare them based on their support for the requirements identified by the reviewed literature [Driessen et al. 2023b, Driessen et al. 2023a, Dolhopolov et al. 2024a]: **(i) Open-Source:** Whether the approach is based on open-source technology; **(ii) Integration:** The ability to connect with different systems and data sources. **(iii) Sensitivity Tagging:** Support for applying globally-defined tags (e.g., Personally identifiable information, financial, sensitive) to data; **(iv) Access Control:** The capability to define and enforce access policies on both data and metadata; **(v) Semantic Connections:** The modeling of business domain concepts and their relationships; **(vi) Data Connections:** Support for schema-level interoperability between different data products; **(vii) Maturity:** The solution's stability, documentation, and adoption level in industry and academia; **(viii) Lineage Tracking:** End-to-end tracking of how data is produced, transformed, and consumed; **(ix) Downstream Propagation:** The automatic inheritance of tags (like sensitivity) across the data lineage.

Metadata models. Further work by [Driessen et al. 2023a] aimed to standardize data product descriptions by proposing a formal metadata model, ProMoTe. The model's design is grounded in established W3C standards, primarily using the Data Catalog Vocabulary (DCAT) as its foundation due to DCAT's focus on describing datasets within a data catalog context. In practice, the ProMoTe model works as a platform-agnostic standardized metadata template. The authors demonstrate its effectiveness in a case study of an organization transitioning into a Data Mesh architecture, where a metadata catalog was implemented in DataHub, an open-source metadata catalog, containing the proposed template within its glossary. Data producers use this template to create structured, consistent metadata for their data products. By providing a rich vocabulary documenting owners, domains, and output ports, this approach establishes strong semantic connections. Moreover, it supports physical data connections by addressing interoperability through traditional techniques, such as foreign keys. Because the model is grounded in the W3C's DCAT standard, it also demonstrates high maturity and simplifies integration with other standard-compliant models. Furthermore, the inclusion of data contracts allows for the explicit definition of both access control policies and sensitivity tagging classifications.

Distributed Ledger-Based Approaches. Emerging from its role as the backbone for cryptocurrencies, blockchain presents a novel architecture for building distributed metadata catalogs. Its properties of decentralization, tamper-proof storage, and auditable transaction history offer a compelling solution to the challenges of managing metadata in a federated environment like a Data Mesh.

[Dolhopolov et al. 2024a] advocated for using blockchain as the foundation for a metadata catalog, arguing it directly addresses the shortcomings of centralized repositories and the various requirements a metadata catalog must fulfill to operate in a decentralized environment. They proposed leveraging blockchain's features to realize the complex requirements of a distributed system. Through the cryptographic hashing of blocks, immutable versioning is provided, while smart contracts enforce global governance policies and enable automated downstream propagation of sensitivity rules. The authors highlighted data interoperability through common semantics as a core requirement; nevertheless, they did not clarify how their blockchain-based approach tackles this issue. Although theoretically robust, this approach currently displays low maturity for enterprise use, with significant performance and scalability concerns remaining open challenges.

Knowledge-Graph-Based Approaches. The literature recommends utilizing knowledge graphs and semantic web technologies to mitigate the shortcomings of current metadata management solutions. Knowledge graphs excel at structuring a metadata catalog by using standardized formal models, such as ontologies, to define the complex concepts and relationships within a domain. This approach creates a global understanding of data products, detailing their content and interconnections in a decentralized manner, whereas semantic web technologies provide standardized, machine-interpretable descriptions.

Several studies outline various models. Some authors proposed linking domain-specific ontologies via a common or main ontology to establish semantic relationships across various domains [González-Velázquez et al. 2024]. To implement this, the authors utilized a hybrid architecture of well-established, open-source technologies, storing the raw metadata in MongoDB while mapping its schema to a GraphDB repository. They also identified a need for access control policies but offered no details on mechanisms for management or enforcement. Conversely, others used standardized semantic blueprints with hierarchical domain structures based on pillar domains, from which subdomains and data products hierarchically descend [Pingos and Andreou 2024]. Each level of this hierarchy was described using Terse RDF Triple Language (TTL) files, a W3C-compliant standard, which can be ingested into a knowledge graph to ensure semantic interoperability. Future work suggests adopting blockchain to manage security mechanisms.

Another strategy, developed for the open-source Open Subsurface Data Universe (OSDU) platform, uses a hybrid model that merges formal ontologies for semantic context with structured JSON files for technical metadata [Abolhassani et al. 2023]. An automated Python pipeline processes this JSON to enrich the ontology model, which is exported as TTL files for integration with graph databases. Their metadata model provides properties for schema versioning, access control, and legal classification, as well as the inheritance of properties from parent classes. Contrasting with ontology-based approaches, a fully decentralized architecture has also been proposed, where each data product publishes its metadata as a self-contained Labeled Property Graph (LPG). In this model, the global catalog is formed by the dynamic union of all independent LPGs, with a graph database, such as Neo4j, as a possible implementation approach [Dolhopolov et al. 2024b]. Nevertheless, their conceptual model does not yet support vital governance aspects, such as access control or sensitivity levels.

Complementing these approaches, [Oliveira et al. 2024] focused on the functional architecture of the catalog itself, idealizing a modular architecture composed of individual components for automated data profiling, data quality assessment, security alerts, and data access control. These modules autonomously generate and manage metadata within a graph data model structured in layers. First, a metadata model preserves the data's semantic meaning, vocabulary, and structural rules as an ontology. Second, a data model is formed as a property graph containing the tangible instances that conform to the rules defined in the ontology.

Vendor-Based Approaches. To reduce implementation overhead and access mature features, metadata catalog development often favors adopting established vendor or open-source solutions over custom builds. Notably, [Wider et al. 2023] documented their practical experiences developing a Data Mesh platform. The authors detail their use of Apache Atlas as a data catalog, noting its critical support for automated governance features, namely, the propagation of sensitivity tags across data lineage. Complementarily, broader landscape surveys [Ashraf et al. 2023] assessed tools against Data Mesh principles, exploring the synergy between Data Mesh and microservice architectures, and identifying which tools best align with both paradigms. These analyses highlight prominent open-source solutions, including Apache Atlas and DataHub, and proprietary options like Collibra and Alation.

Comparing Different Approaches and Technologies. Based on the analysis in the previous sections and the criteria introduced in Section 3, we present a comparative evaluation of various approaches and tools in Tables 1 and 2, respectively, for implementing metadata management solutions. The identified vendor-based data catalog solutions offer robust integration capabilities and support core features, including data sensitivity classification, data interoperability, and data lineage tracking. **Apache Atlas**, **DataHub**, **Collibra**, **AWS Glue Catalog**, and **Alation** are well-rounded solutions, broadly meeting these requirements out-of-the-box or via supporting systems. Notably, **AWS Glue Catalog** relies on additional AWS services (e.g., Lake Formation, DataZone) for comprehensive access control, semantic connections, and downstream propagation. **Neo4j** and **GraphDB**, despite being graph databases rather than metadata catalog solutions, serve as powerful engines for building custom data catalog solutions. Their native ability to model complex relationships makes them highly effective for semantic connections and lineage tracking. Among open-source options, **Amundsen** is capable of data discovery, lineage, and enabling semantic connections through flexible tagging, though data access is not yet supported, and dedicated business glossaries often require customization. **Marquez**, conversely, is a viable specialized option emphasizing technical lineage for ETL jobs and data versioning, rather than semantic documentation.

4. Open Research Challenges

While significant progress has been made in managing metadata for Data Mesh, several challenges and open research questions remain. A key challenge is the architectural inconsistency between the philosophy of decentralization and the practical reliance on centralized catalogs for data discovery [Goedegebuure et al. 2024]. This approach, albeit practical, risks reintroducing bottlenecks and governance conflicts, while decen-

Table 1. Metadata management strategies and tools identified in the literature

Strategy	Integration	Sensitivity Tagging	Access Control	Semantic Connections	Data Connections	Maturity	Lineage Tracking	Downstream Propagation
[Dolhopolov et al. 2024a]		✓	✓				✓	✓
[Driessen et al. 2023a]	✓	✓	✓	✓	✓	✓		
[González-Velázquez et al. 2024]	✓		✓*	✓	✓	✓		
[Pingos and Andreou 2024]	✓	✓*	✓*	✓	✓			
[Abolhassani et al. 2023]	✓	✓	✓	✓	✓	✓	✓	✓
[Dolhopolov et al. 2024b]	✓			✓	✓		✓*	
[Oliveira et al. 2024]	✓	✓*	✓*	✓	✓		✓*	

(*) unaddressed requirements identified by the authors.

Table 2. Metadata management tools identified in the literature

Tool	Open source	Integration	Sensitivity Tagging	Access Control	Semantic Connections	Data Connections	Maturity	Lineage Tracking	Downstream Propagation
Apache Atlas	✓	✓	✓	✓	✓	✓	✓	✓	✓
DataHub	✓	✓	✓	✓	✓	✓	✓	✓	✓
Amundsen	✓	✓	✓*			✓	✓	✓*	
Marquez Project	✓	✓	✓*			✓		✓	✓
Collibra		✓	✓	✓	✓	✓	✓	✓	
Alation		✓	✓	✓	✓	✓	✓	✓	✓
AWS Glue Data Catalog		✓	✓*	✓*	✓*	✓	✓	✓	✓*
Neo4j	✓	✓	✓	✓	✓	✓	✓	✓	✓*
GraphDB		✓	✓	✓	✓	✓	✓	✓	

(*) supported features; may require custom implementation or additional software, as they are not available out of the box.

tralized solutions like blockchain still face significant performance and storage hurdles [Dolhopolov et al. 2024a]. This architectural uncertainty is heightened by immature tooling that struggles with decentralization [Dolhopolov et al. 2024b], whilst vendor solutions introduce their own trade-offs in cost, usability, and often demand specialized technical expertise, which runs counter to the self-serve principles of Data Mesh.

Beyond the technology, significant gaps exist in operational usability. The literature emphasizes a need for systems that enable a consumer feedback cycle and provide flexible templates for producers to balance standardization with custom needs [Driessen et al. 2023b]. Leveraging automation is a highly recommended strategy to bridge these gaps. Automation is vital for ensuring compliance through features like automated data classification and propagation [Wider et al. 2023], and for abstracting operational complexity to create a truly self-serve environment for non-technical users [Goedegebuure et al. 2024].

5. Conclusions & Future Work

We present the partial results of an SLR aimed at uncovering technological approaches and empirically validated implementation guidelines for Data Mesh architectures. Ten papers were analyzed, concentrating on exploring current metadata management solutions and identifying further gaps in the literature to be addressed in future works. Metadata is vital for proper decentralized governance, as it facilitates data discovery, prevents data swamps through standardized documentation, and contributes to security by tagging data with access and sensitivity levels. Moving forward, a broader investigation will be conducted to synthesize current methodologies and tools covering the end-to-end development of a Data Mesh architecture.

References

Abolhassani, N., Tudor, A., and Paul, S. (2023). A data mesh adaptable oil and gas ontology based on open subsurface data universe (osdu). In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 2: KEOD*, pages 29–39. SciTePress.

Ashraf, A., Hassan, A., and Mahdi, H. (2023). Key lessons from microservices for data mesh adoption. In *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 1–8. IEEE.

Dehghani, Z. (2022). *Data Mesh: Delivering Data-driven Value at Scale*. O'Reilly.

Dolhopolov, A., Castelltort, A., and Laurent, A. (2024a). Exploring the benefits of blockchain-powered metadata catalogs in data mesh architecture. In *Management of Digital EcoSystems*, pages 32–40. Springer.

Dolhopolov, A., Castelltort, A., and Laurent, A. (2024b). Trick or treat: Centralized data lake vs decentralized data mesh. In *Management of Digital EcoSystems*, pages 303–316. Springer.

Driessen, S., den Heuvel, W.-J. v., and Monsieur, G. (2023a). Promote: A data product model template for data meshes. In *Conceptual Modeling*, pages 125–142. Springer.

Driessen, S., Monsieur, G., and van den Heuvel, W.-J. (2023b). Data product metadata management: An industrial perspective. In *Service-Oriented Computing – ICSOC 2022 Workshops*, pages 237–248. Springer.

Goedegebuure, A., Kumara, I., Driessen, S., Van Den Heuvel, W.-J., Monsieur, G., Tamburri, D. A., and Nucci, D. D. (2024). Data mesh: A systematic gray literature review. *ACM Computing Surveys*, 57(1):1–36.

González-Velázquez, R., Fernández, I., Ferreira, R., Carballo, C., Álvaro García, Santamaría, B., and González, D. (2024). Smart factory hub – towards a data mesh in smart manufacturing. *Procedia Computer Science*, 232:2709–2719.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering technical report. Technical report, Software Engineering Group, EBSE Technical Report, Keele University and Department of Computer Science University of Durham.

Oliveira, B., Duarte, A., and Óscar Oliveira (2024). Towards a data catalog for data analytics. *Procedia Computer Science*, 237:691–700.

Pingos, M. and Andreou, A. S. (2024). Discovering data domains and products in data meshes using semantic blueprints. *Technologies*, 12(7):105.

Wider, A., Verma, S., and Akhtar, A. (2023). Decentralized data governance as part of a data mesh platform: concepts and approaches. In *2023 IEEE International Conference on Web Services (ICWS)*, pages 746–754. IEEE.