Challenges to Enforce Data Quality in Data Spaces

Claudia P. Ayala[1], Besim Bilalli[1], Cristina Gómez[1], Jose-Norberto Mazón[2,*] and Oscar Romero[1]

[1]Universitat Politècnica de Catalunya, BarcelonaTech
[2]Universitat d'Alacant / Universidad de Alicante

Abstract

Data Spaces must preserve sovereignty and privacy while ensuring FAIR (Findable, Accessible, Interoperable and Reusable) principles. To do so, policy-based strategies have to be developed in order to describe the agreements reached in the Data Space. In this context, two open questions arise: how to define the right Data Space policies, as well as, how to enforce (and monitor) them. Despite the efforts towards defining and enforcing data access and usage policies, there is no solution to operationalize the enforcement of those considering data quality dimensions. However, data quality is becoming a hot topic due to the surge of federated learning and alternative analytical techniques, which require all providers to guarantee a data quality threshold in order to learn robust models. Currently, we have means to describe policies related to data quality rules (e.g., by combining standards such as ODRL and standard vocabularies) but we are missing means to elicit these policies from data providers and enforce them while preserving the data sovereignty. In this paper, we discuss the challenges and open questions that must be addressed in order to operationalize (and eventually, automate) data quality in Data Spaces, which span from requirements elicitation to data validation.

Keywords

Data Spaces, Data Quality, Data Validation, Federated Data Management, Data Sharing

1. Introduction

Data Spaces are federated ecosystems in which *data providers* and *consumers* share data while preserving data sovereignty and privacy. Currently, the *Data Mesh* architecture [1] is at the core of current technological solutions, since it provides a domain-decentralized paradigm that suits the Data Space requirements [2]. Relevantly, the Data Mesh defines the *Data Product* concept, which provides a product-oriented view of the providers' data assets. In short, the data product is a node that encapsulates three structural components required to function: code for enforcing policies (i.e., the Data Space agreements), data (and its metadata) and infrastructure [3]. By definition, the providers' data assets can be heterogeneous both in the infrastructure used and the data provided (in format and semantics).

Behind the idea of Data Spaces is the objective of extracting value from data sharing. This can be achieved in many ways, but data analysis arises as prominent means to achieve so, either by means of descriptive analysis (e.g., dashboarding and OLAP) or predictive analysis (e.g., learning models). However, how to achieve data analysis in federated environments is an open challenge, and federated learning [4] is currently the most widespread privacy-aware data analysis technique. Many efforts have been devoted to develop robust federated learning but little attention has been paid to the role of data. Yet, the impact of the data quality (DQ) from each provider on federated models learnt is huge [5, 6].

One of the biggest open problems in Data Spaces not properly tackled is how the agreements reached (e.g., on DQ) at the Data Space federated layer (i.e., at the federated -unique- view of the data ecosystem) can be enforced at the providers' data assets regardless of their heterogeneity and preserving data ownership and privacy. Note that this problem has been easily tackled in centralized environments by having a central authority extracting, transforming and preparing data for analysis. However, this is not possible in settings where data is not meant to be shared raw. For example, the minimum number of instances and the variances of key attributes might be set as DQ criteria for all data providers and should be automatically and locally validated by executing a software service (specific for the provider infrastructure) provided by the Data Space services catalog. The result of the service execution should be communicated to the Data Space. To our knowledge, there is no architecture, framework or solution tackling this problem, despite the myriad of standards and definitions blooming around the Data Space concept (e.g., [7, 8]).

We focus on how to validate DQ agreements in the Data Space and discuss the open challenges to make DQ happen in Data Spaces to enact trustworthy federated learning.

2. Challenges and Vision

Data Spaces require a governance model for specifying DQ agreements that stakeholders must adhere to in order to participate. Importantly, this governance model must also specify DQ needs agreed among data consumers and providers when developing specific uses cases. Therefore, our view is that the governance model for Data Spaces should distinguish two levels: 1) a Data Space level for agreements among stakeholders of the Data Space authority from data regulations and strategic issues, and 2) a use case level for agreements among data providers and consumers to build specific Data Products. Based on this view and to facilitate the discussion, we propose a visionary framework with a process for the Data Space and use case levels (see Fig. 1). Our framework follows the Open Data Product specification [9], thus splitting each process into two parts: one declarative, at a higher-level of abstraction specifying *what* (analysis phase), and another one at a lower-level specifying *how* (design and implementation phases). The declarative part defines the DQ dimensions and intended level. The ex-

DOLAP 2025: 27th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT 2025, March 25, 2025, Barcelona, Spain

*Corresponding author.
claudia.ayala@upc.edu (C. P. Ayala); besim.bilalli@upc.edu (B. Bilalli); cristina.gomez@upc.edu (C. Gómez); jnmazon@ua.es (J. Mazón); oscar.romero@upc.edu (O. Romero)
https://futur.upc.edu/ClaudiaPatriciaAyalaMartinez (C. P. Ayala); https://futur.upc.edu/BesimBilalli (B. Bilalli); https://futur.upc.edu/CristinaGomezSeoane (C. Gómez); https://s.ua.es/_MuH (J. Mazón); https://futur.upc.edu/OscarRomeroMoral (O. Romero)
0000-0002-6262-3698 (C. P. Ayala); 0000-0002-0575-2389 (B. Bilalli); 0000-0002-3872-0439 (C. Gómez); 0000-0001-7924-0880 (J. Mazón); 0000-0001-6350-8328 (O. Romero)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3931/panel1.pdf

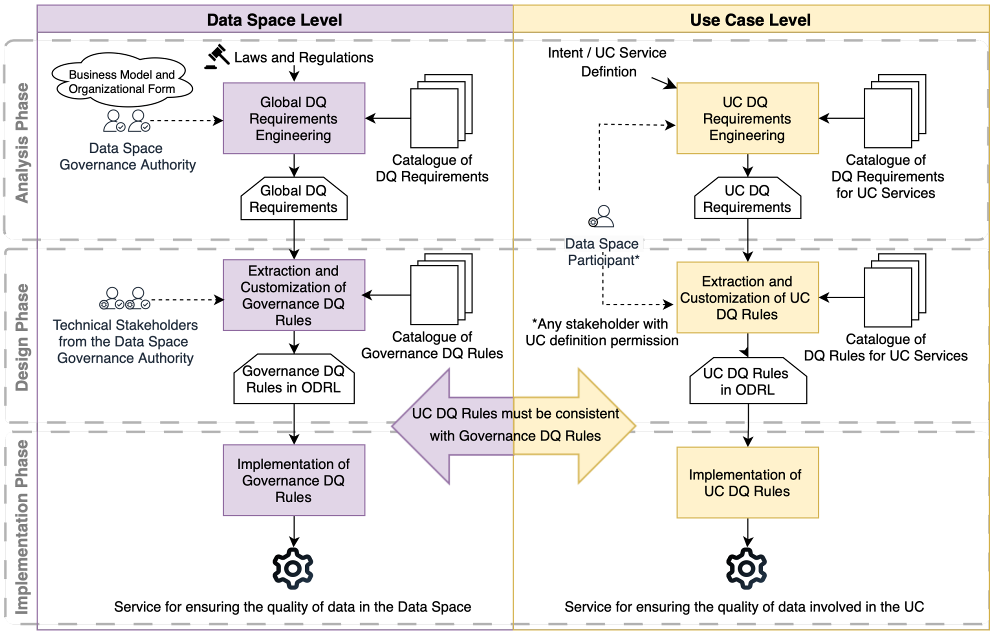
Data Space Level
Use Case Level
Analysis Phase
Design Phase
Implementation Phase
Laws and Regulations
Business Model and Organizational Form
Data Space Governance Authority
Global DQ Requirements Engineering
Catalogue of DQ Requirements
Global DQ Requirements
Intent / UC Service Defintion
UC DQ Requirements Engineering
Catalogue of DQ Requirements for UC Services
UC DQ Requirements
Data Space Participant*
Technical Stakeholders from the Data Space Governance Authority
Extraction and Customization of Governance DQ Rules
Catalogue of Governance DQ Rules
Governance DQ Rules in ODRL
*Any stakeholder with UC definition permission
Extraction and Customization of UC DQ Rules
Catalogue of DQ Rules for UC Services
UC DQ Rules in ODRL
UC DQ Rules must be consistent with Governance DQ Rules
Implementation of Governance DQ Rules
Implementation of UC DQ Rules
Service for ensuring the quality of data in the Data Space
Service for ensuring the quality of data involved in the UC

Figure 1: Visionary framework for considering DQ requirements in Data Spaces.

ecutable part contains the machine-readable "as code" rules, provided as a service, to validate DQ dimensions. Next, we describe both processes and their main challenges.

DQ Requirements Engineering for Data Spaces.

Requirements engineering (RE) for complex systems in open and dynamic environments that extend beyond a single organization is widely recognized as a challenging endeavor [10, 11]. This is particularly true in the context of Data Spaces, where the elicitation and management of requirements must reconcile diverse perspectives, including the strategic business vision, governance, compliance with laws and regulations, infrastructure, scalability demands, and DQ considerations. Our visionary framework proposes applying RE practices to elicit, specify, and manage the Data Space requirements. We advocate for the development and use of a Catalogue of DQ Requirements at two levels: the Data Space level and the use case level. These catalogs promote knowledge sharing and requirements reuse, building a robust repository of experiences and best practices. The proposed process is aimed to: 1) Ensure a common understanding of DQ dimensions by considering established standards; 2) Facilitate the elicitation of diverse DQ requirements from diverse stakeholders to enable effective data sharing; 3) Support the structured specification and management of DQ requirements to ensure compliance and alignment between the Data Space and use case levels for their subsequent operationalization; and 4) Address trade-offs between conflicting DQ requirements. This approach aims to bridge the gap between diverse stakeholder perspectives and the technical requirements for robust DQ management in Data Spaces.

Extraction and Customization of DQ Rules. The complexity of DQ requirements and their textual or semi-structured formalization make their direct operationalization challenging. With the aim of making DQ requirements executable in an operational environment, our visionary framework proposes to transform, in a semi-automated way and using specific catalogues for supporting this transformation, DQ requirements (at Data Space and use case levels) into formalized DQ rules that may be easily implemented. We propose to use a rule language with well-defined semantics (e.g., ODRL), to formalize DQ rules. Several challenges need to be tackled when performing this transformation: 1) the identification of relevant and suitable stakeholders with the specific knowledge for performing this activity in both levels; 2) the definition of specific catalogues with reusable transformation patterns for translating DQ requirements into rules, preserving their semantics; 3) the definition of the artifacts needed (e.g., specialized metamodels or new ODRL profiles), for automating the extraction and customization of DQ rules to the specific domain and level.

Implementation available as a Service of DQ Rules.

The inherent heterogeneity of providers in the context of Data Spaces renders the process of translating formal DQ rules into executable services a significant challenge. The main goal of this activity is to avoid building and maintaining custom solutions that are tightly coupled to specific execution environments or platforms. To address this, we propose an agnostic solution that leverages best practices from software engineering, such as containerized solutions, ensuring portability, scalability, and interoperability. However, the intrinsic characteristics of Data Spaces introduce several challenges that must be addressed: 1) dealing with heterogeneity at the infrastructure level by abstracting the differences while ensuring consistent performance and security across environments; 2) allowing for dynamic and federated execution across multiple distributed nodes, ensuring real-time validation without requiring data centralization.

As conclusion, there is a need for further research to enact DQ in Data Spaces, a must for qualitative federated data analysis. In this sense, we have discussed a visionary framework, its main phases and challenges to be tackled.

Acknowledgments

This work has been partially supported by the EU-HORIZON program under GA.101135513 (CYCLOPS) and by CIAICO/2022/019 project from Generalitat Valenciana.

References

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

[2] M. Bacco, A. Kocian, S. Chessa, A. Crivello, P. Barsocchi, What are data spaces? systematic survey and future outlook, Data in Brief 57 (2024) 110969.

[3] Z. Dehghani, Data Mesh: Delivering Data-driven Value at Scale, O'Reilly, 2022.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: A. Singh, X. J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA, volume 54 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1273–1282.

[5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: I. S. Dhillon, D. S. Papailiopoulos, V. Sze (Eds.), Proceedings of the Third Conference on Machine Learning and Systems, MLSys 2020, Austin, TX, USA, March 2-4, 2020, mlsys.org, 2020.

[6] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, V. Munigala, Overview and importance of data quality for machine learning tasks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, 2020, p. 3561–3562.

[7] Fiware for data spaces: Position paper, https://www.fiware.org/wp-content/uploads/FF_PositionPaper_FIWARE4DataSpaces.pdf, 2024. Accessed: 2024-12-20.

[8] International data spaces association, https://internationaldataspaces.org/why/international-standards/, 2024. Accessed: 2024-12-20.

[9] Data product specification, https://opendataproducts.org/v3.1/#optional-attributes-and-elements, 2024. Accessed: 2024-12-20.

[10] P. Malcher, E. Silva, D. Viana, R. P. dos Santos, What do we know about requirements management in software ecosystems?, Requir. Eng. 28 (2023) 567–593.

[11] P. Hagenhoff, S. Biehs, F. Möller, B. Otto, Designing a reference architecture for collaborative condition monitoring data spaces: Design requirements and views, in: M. Mandviwalla, M. Söllner, T. Tuunanen (Eds.), Design Science Research for a Resilient Future - 19th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2024, Trollhättan, Sweden, June 3-5, 2024, Proceedings, volume 14621 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 355–369.