Aug 10th, 12:00 AM

Data Quality in Data Ecosystems: Towards a Design Theory

Marcel Altendeitering
Fraunhofer ISST, marcel.altendeitering@isst.fraunhofer.de

Stephan Dübler
Fraunhofer ISST, stephan.duebler@isst.fraunhofer.de

Tobias Moritz Guggenberger
TU Dortmund University, tobias.guggenberger@tu-dortmund.de

Data Quality in Data Ecosystems: Towards a Design Theory

Completed Research

Marcel Altendeitering
Fraunhofer ISST, Dortmund, Germany
marcel.altendeitering@isst.fraunhofer.de

Stephan Dübler
Fraunhofer ISST, Dortmund, Germany
stephan.duebler@isst.fraunhofer.de

Tobias Guggenberger
TU Dortmund University, Dortmund, Germany
tobias.guggenberger@tu-dortmund.de

Abstract

Data quality is an important aspect for the success of data ecosystems. Sharing low-quality data causes large data preparation efforts, can disrupt the chain for value co-creation, and can damage the mutual trust among partners in the ecosystem. While there are many data quality tools available in literature and practice, there is limited knowledge on the peculiarities of assessing and managing data quality in data ecosystems. In this study, we present the results of a design science research project that was concerned with the development of design principles for ensuring data quality in data ecosystems. The proposed concept extends an existing ecosystem with an artifact that technically enforces data quality checks on shared data in the manufacturing domain. Our work aims to provide the prescriptive design knowledge needed for such systems. For practitioners, we offer generalized design principles that can inform custom implementations of data quality tools in data ecosystems.

Keywords

Data Quality, Data Ecosystems, Design Science Research, Design Principles

Introduction

The adoption rate of artificial intelligence (AI) and machine learning (ML) initiatives is astonishing. Organizations from all industries are implementing customized use cases to experiment with the potentials of AI to enhance organizational decision-making and innovation (Gröger 2021; Kabalisa and Altmann 2021). To provide these initiatives with the vast amount of data they need, data ecosystems play an important role as they facilitate sharing and reusing data among partners for mutual benefits (Azkan et al. 2020; WEF 2021). By breaking data silos organizations and governments can help facilitate data sharing and leverage data for better outcomes (UNCTAD 2021; WEF 2021). Such cross-institutional data flows enable the cooperative creation of data-driven innovation, which is vital for co-creating impactful solutions, organizational value, and exploiting new business opportunities (Oliveira et al. 2019; WEF 2021). For example, the Mobility Data Space (Mobility Data Space 2022) is a group of numerous stakeholders from the mobility industry in Germany. Hereby, participants share mobility data from different sources to create innovative mobility solutions, such as multimodal mobility.

Despite their popularity, many AI initiatives fall short of their promises in industrial practice (Gröger 2021). The shift from prototypical to production-ready applications is difficult and many initiatives remain insular as organizations face several data-related challenges such as governance, democratization, or quality (Bosch et al. 2021; Gröger 2021). In collaborative data ecosystems, the handling of data quality (DQ) raises new requirements and is considerably important for multiple reasons (Bosch et al. 2021; Gröger 2021; WEF 2021). First, erroneous data at the source causes large efforts in data preparation for partners who often lack data domain knowledge (Bosch et al. 2021; Redman 2020). Second, a lack of DQ can mitigate the required level of trust that partners in a data ecosystem need to share their data (Azkan et al. 2020; WEF

2021). Third, the propagation of low-quality data, such as imbalanced, drifted, or unlabeled data, can deteriorate inter-organizational business processes (Amadori et al. 2020).

A plurality of DQ works and tools have emerged from science and practice that address the aspect of DQ to measure, monitor, or prepare and clean erroneous data sets in different steps of the data lifecycle (Altendeitering and Tomczyk 2022; Ehrlinger et al. 2019). However, these tools do not cover the 'Data Sharing' step of the data lifecycle and neglect the requirements of handling DQ in data ecosystems. Both processes (i.e., DQ and data sharing) are usually disjunct and, in general, DQ tools lack integration with their surrounding infrastructure (Walter et al. 2022). There is a lack of the prescriptive design knowledge, which participants in a data ecosystem need for to share high-quality data. To address this shortcoming and investigate the design of DQ solutions in data ecosystems, we define the following research question:

Research Question: *What are general design principles for DQ tools in data ecosystems?*

In our study, we present the results of a Design Science Research (DSR) project based on Hevner et al. (2004) and Peffers et al. (2007). Based on meta-requirements that we derived from both the literature and expert interviews, we developed a set of generalized design principles for the development of DQ tools in data ecosystems within the manufacturing domain. To demonstrate and evaluate the applicability of the design principles, we implemented these in a DQ tool that performs data quality checks on data-streams. With our results, we advance the scientific body of knowledge on DQ and data ecosystems by offering a theoretical foundation for designing solutions in this domain. This way, we contribute to the nascent design theory on DQ in data ecosystem and pave the way for further research. Organizations can use the generalized design principles to inform custom implementations of tools in their respective contexts and better leverage their data sets.

We structured the remainder of this article as follows. First, we describe the theoretical background of our study. We outline the adopted DSR approach and our course of actions in section 3. In section 4, we present the accumulated design knowledge as generalized design principles. In section 5, we describe the results of demonstrating and evaluating the design knowledge as a software prototype and summarize our study with a conclusion in section 6.

Theoretical Background

Data Ecosystems

The ecosystem concept was introduced by Moore (1993) and is increasingly gaining interest in both strategic management and information systems research (Guggenberger et al. 2020). Relying on the biological analogy, ecosystems describe the interactions amongst actors and their environment, co-evolutionary processes, lifecycle, and value creation (Guggenberger et al. 2020; Moore 1993). An ecosystem comprises different actors, such as the keystone, which typically is a provider of stability and orchestration, and niche players that develop and contribute specialized capabilities (Iansiti and Levien 2004; Oliveira et al. 2019). Regarding the specific focus of the ecosystem, a plethora of ecosystem concepts exist, which are rather fuzzy than clearly delimitable (Guggenberger et al. 2020). Jacobides et al. (2018) for example, differentiate between the "business", the "innovation", and the "platform ecosystem", based on the particular emphasis of a study. More recently, data ecosystems gain interest as a key enabler for the transformation of whole industries towards an integrated digital economy (Capiello et al. 2020). "Data Ecosystems are socio-technical complex networks in which actors interact and collaborate with each other to find, archive, publish, consume, or reuse data, as well as to foster innovation, create value, and support new businesses" (Oliveira et al. 2019, p.589) Within data ecosystems, actors share data to create data-related business innovations, such as new business models (Oliveira et al. 2019; WEF 2021).

Data Quality

DQ is a concept comprising multiple dimensions, which represent different characteristics of data. The importance and definition of these characteristics is context-dependent and can vary between users. Wang and Strong (1996) emphasized this aspect and stated that data must fulfill a "fitness for use" to be of high-quality. There is a broad consensus in science and practice that a strong relationship exists between the quality of data sets and their value for different business processes (Otto 2015). Organizations that can leverage high-quality data are more agile and successful in creating competitive services and products

(Redman 2020; Setia et al. 2013). In particular, data-intensive applications, such as AI or ML, are susceptible to poor DQ, as their accuracy and the confidence in data-driven decision making rely on correct data (Gröger 2021). Without an acceptable level of DQ, organizations struggle to develop data strategies, which they need to sustain in the digital, competitive environment they are nowadays facing (Azkan et al. 2020). Redman (2020, p.1) summarized the importance of DQ by stating "You can't do anything important in your company without high-quality data". Following this, DQ is still one of the most important challenges in data management and many initiatives in research and practice are investigating the topic (Gröger 2021; Setia et al. 2013).

To support the provisioning of high-quality data, a plurality of DQ-centric tools emerged from science and practice. We can distinguish these in three categories: data preparation and cleaning tools (Chu et al. 2016), DQ management and monitoring tools (Ehrlinger et al. 2019), and general-purpose tools, which combine functionalities of the two former types (Altendeitering and Tomczyk 2022). Lately, the functionality of DQ tools is shifting from validating data against manually specified quality rules to more collaborative and automated approaches that can handle multiple DQ dimensions at once (Altendeitering and Guggenberger 2021; Ehrlinger et al. 2019). Having high-quality data is becoming less of an IT-centric task and more of a joint effort involving multiple people across the entire data lifecycle who mutually benefit from the availability of high-quality data (Altendeitering and Tomczyk 2022).

Against this background, our study advances the current body of knowledge on DQ and data ecosystems by providing design knowledge on DQ tools that specifically address the 'Data Sharing' step within a data lifecycle (Otto 2015). Despite the high importance of DQ for this step, there is limited research available and, to the best of our knowledge, no DQ tool specifically addresses this need. Our solution leverages AI and ML techniques for evaluating DQ and, thus, contributes to the trend for automation in DQ tools.

Design Science Research Approach

To develop the artifact and design principles, we follow the established DSR guidelines provided by Hevner et al. (2004) and the DSR methodology suggested by Peffers et al. (2007). For this, we developed design principles, designed and evaluated a useful artifact addressing an important organizational problem, and, therewith, contribute a theory of design and action (Gregor 2006). Over the course of a six-month project, we completed three major design cycles (see Figure 1). Each cycle lasted two months, and we visited the six DSR-phases as introduced by Peffers et al. (2007). We could conduct our research in cooperation with a large organization from the manufacturing industry, which we will call MCo in our study for anonymity. MCo was actively taking part in an International Data Spaces (IDS) based data ecosystem (IDSA 2022b). Concretely, they were sharing several sensor-based data streams from a machine (e.g., temperature, pressure, etc.) with an external consultancy who offered data analysis services. It was the goal of both partners to share high-quality data to ensure correct analytical results and avoid data cleaning efforts. The project team comprised two researchers and two practitioners from MCo who provided feedback and supported the integration of our prototype at MCo.

In the beginning of **design cycle 1,** we conducted two important steps: first, a review of available DQ tools and, second, interviews with data engineering experts at MCo. We started reviewing the tools with two recent survey papers on different DQ tools (Altendeitering and Tomczyk 2022; Ehrlinger et al. 2019), which results in the identification of a lack of design knowledge on DQ in data ecosystem. This is likely because both data ecosystems and inter-organizational DQ management are novel research areas (Altendeitering and Tomczyk 2022; Azkan et al. 2020). For the expert interview, we conducted a two-hour group discussion, including two data-engineering experts at MCo and two researchers. The first expert has 15 years of experience in a system architecture role and has designed and managed multiple data engineering pipelines. The second expert has been working for three years in a data scientist role and is responsible for implementing data analytic solutions. Both experts are required to share data internally and externally on a regular basis and are involved with the IDS-based data ecosystem in place at MCo. The two researchers moderated the discussion using the results of the previous DQ tool review as a basis. Overall, the discussion was fruitful, as we identified two key challenges for sharing high-quality data. First, a separation of the DQ management and data sharing processes, which results in manual data validation and a lack of technically enforced DQ checks prior to sharing data. Second, data cleaning is cumbersome for partners within the ecosystem, as they have no data domain knowledge, which raises a large communication overhead in resolving data issues. These results showed us that there is a need to advance DQ tools in a way that they

incorporate the data sharing aspect and that further developments are necessary. Based on the identified problems, we formulated two meta-requirements that informed the objectives of our intended design knowledge and an initial set of design principles (Peffers et al. 2007). We continued by implementing the design principles as a Minimum Viable Product (MVP) using a prototyping approach (Gregory and Muntermann 2014). This prototype focused on simple, statistical DQ checks and was not yet compatible with the data ecosystem, as this required integration with the IDS Dataspace Connector (IDSA 2022a). We completed the first design cycle by discussing and reflecting on the preliminary results within the project team in an internal review meeting. All participants agreed that the current MVP was promising, and the project team concretized the developments for the second design cycle.

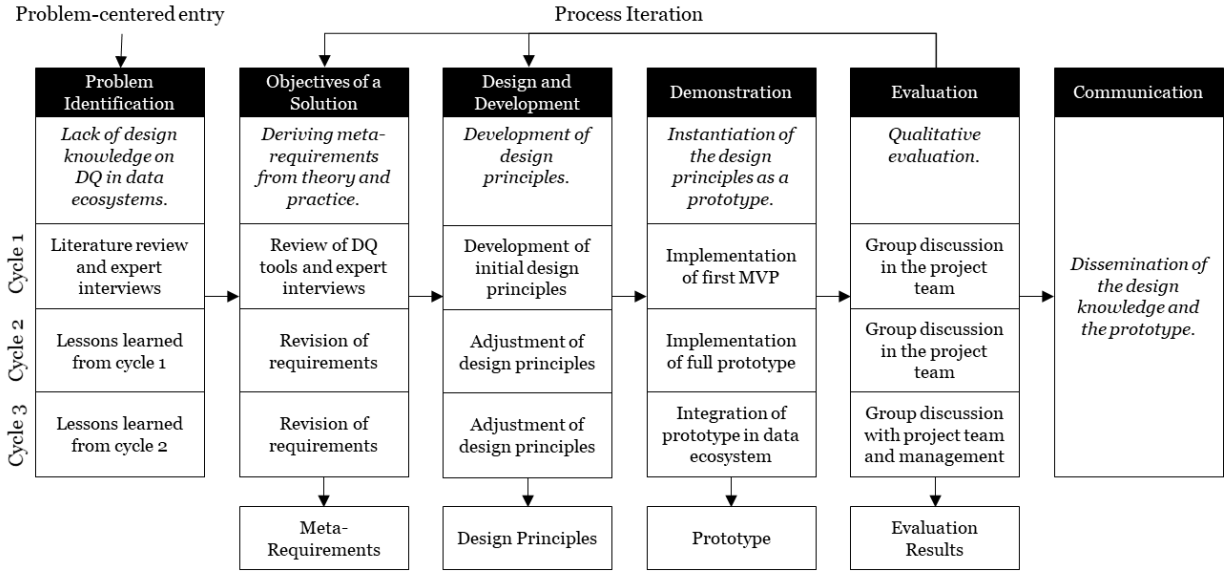

Figure 1. Design Science Research Approach (adapted from Peffers et al. (2007))

By using the analysis of the first cycle as a basis, we refined our solution in **design cycle 2**. We started by revising the initial design principles based on the lessons learned. Specifically, a better understanding of the data sets MCo wants to share, and of the IDS data ecosystem helped us to sharpen our design principles. As a result, we added two design principles for considering multiple DQ aspects and integrating the DQ result as metadata. We only realized that these aspects were necessary after implementing the initial MVP. Like the first cycle, we followed a prototyping approach to implement a refined version of the instantiation (Gregory and Muntermann 2014). At this stage, our prototype was already fully functional and able to conduct four DQ measurements covering different quality dimensions. The prototype also offered a preliminary integration with the IDS dataspace connector. However, this integration was mostly manual and lacked automation, such as an automatically invoked DQ calculation based on changes to the data (IDSA 2022a). Again, we completed the second design cycle by reviewing and discussing the current prototype within the project team. The team agreed that the tool was already well developed, but still lacked some functionalities on integration, which shaped our planning for the third cycle.

In **design cycle 3**, we formulated the final version of our design principles by reflecting on the experiences we made during the first two cycles. As a result, we were able to propose a set of nine design principles that provide a generalizable description of our developments. In the third design cycle, our developments focused on enabling the integration of our prototype with the IDS dataspace connector (IDSA 2022a) and the backend data sources at MCo. For this, it was necessary to transform the prototype to a so called 'Data App', which follows the standards and information models specified by the IDS (IDSA 2022b). At the end of the cycle, we successfully instantiated a DQ tool that validates data before sharing. For the final evaluation, we conducted a qualitative group discussion as described in the 'Demonstration and Evaluation' section.

Artifact Design and Development

For the development of a nascent design theory, we followed the approach of Möller et al. (2020) and derived a set of design principles (DPs) that address the identified meta-requirements (MRs) (see Figure 2). These emerged from the interplay between researchers and practitioners over the three DSR cycles and are grounded in science and practice. Specifically, we used project review meetings at the end of each DSR-cycle to reflect on the lessons learned and the design decisions we made. For formalization, we follow the linguistic template of Kruse et al. (2015).

We separated the design principles into two categories, addressing the two meta-requirements. The first five design principles aim to increase the integrability of DQ in the data sharing process (MR1). The remaining four principles aim to increase the effectiveness of shared data (MR2). With the design principles, we provide a theoretically and empirically grounded answer to common calls for practical research on the integrability and usability of DQ tools and the further advancement of data ecosystems (Altendeitering and Tomczyk 2022; Azkan et al. 2020; Gröger 2021).

Integrate Data Quality in the Sharing Process (MR1)

Trust in the quality of data sets and among participants is an essential aspect in data ecosystems (Azkan et al. 2020). To sustain a high level of trust and ensure using data for mutual benefit, it is necessary to provide all participants with access to DQ tools and avoid a sole use by one participant. In our case, we followed the guidelines for Data Apps by the IDSA and uploaded the app to an ecosystem-wide app store (IDSA 2022b). Accordingly: *Provide the tool in a compatible format for users from different contexts to benefit from DQ validations, given that the data ecosystem has participants from different contexts* ***(DP1)***.

A highly automated and seamless integration of DQ and associated tools (e.g., data catalogs) is vital for connecting isolated data management processes and realize a continuous value delivery (Altendeitering and Guggenberger 2021; Bosch et al. 2021). At MCo, a technical integration of the DQ and data sharing processes was needed to enforce DQ validation prior to data sharing. We realized this requirement by specifying Apache Camel routes as proposed by the IDSA (Apache Camel 2022; IDSA 2022a). Accordingly: *Provide the tool with a technically enforced integration with the dataspace connector for users to receive validated data, given that separated DQ and data sharing processes are in place* ***(DP2)***.

Joint data pipelines often face challenges when using different sources and types of data, as the data might be incompatible or lack normalization (Bosch et al. 2021). To address this issue, a DQ tool should follow standardized information models (e.g., IDSA (2022b), ISO (2022)) to ensure data interoperability (Walter et al. 2022; WEF 2021). Accordingly: *Provide the tool with standardized interfaces and information models for users to receive understandable results, given that different standards are in place* **(DP3)**.

Data-intensive applications are usually implemented in heterogeneous system landscapes that include diverse data sources and DQ definitions (Gröger 2021). To avoid that DQ tools remain insular and are only used for certain data, it is important to provide different functionalities and develop DQ standards that embrace diversity (Altendeitering and Tomczyk 2022; UNCTAD 2021). Accordingly: *Provide the tool with diverse functionalities and DQ standards for users to conduct DQ checks on different data sources, given that multiple data sources and DQ contexts are in place* **(DP4)**.

In distributed data pipelines, it is necessary to maintain a clear trail between a data set and corresponding analytical results (Bosch et al. 2021). For data stream applications, the realization of such a trail can be particularly challenging to not violate the timeliness of data. In our case, we created batches of data from the data stream and assigned the corresponding metadata with an ID and its DQ result. Accordingly: *Provide the tool with methods to store the DQ result as metadata for users to follow the data trail between data set and DQ result, given that the trail can be established* **(DP5)**.

Increase the Effectiveness of Shared Data (MR2)

Real-world data sets often face multiple data errors at the same time (Chu et al. 2016). Therefore, DQ tools should offer functionalities for assessing DQ in multiple dimensions (e.g., accuracy, completeness, timeliness, etc.) to identify different errors at once (Altendeitering and Guggenberger 2021; Wang and Strong 1996). We observed the same issue at MCo, where they had little knowledge about potential errors

and a combination of DQ metrics was necessary to ensure trust in the data. Accordingly: *Provide the tool with functions for analyzing multiple DQ aspects for users to identify errors in different DQ dimensions, given that the errors are not known a priori* **(DP6)**.

The multi-dimensional concept of DQ can be difficult to comprehend by users who are not familiar with the meaning of the different DQ dimensions (Wang and Strong 1996). In cooperation with MCo, we developed a single DQ score that abstracts multiple DQ metrics on a scale from 0 to 1 and is easily interpretable by different users. Accordingly: *Provide the tool with an aggregated DQ score for users to get an easy-to-understand result, given that multiple DQ measurements exist* **(DP7)**.

Explainable AI is an important aspect mentioned in most AI research agendas (Bosch et al. 2021; Gröger 2021). DQ tools pose the same transparency requirement to ensure trust, certifiability, and usability. We offered documentation on the functionality of the algorithms used to address this requirement. Accordingly: *Provide the tool with transparent descriptions of the AI/ML methods used for users to gain trust and understand their functionality, given that there are users without adequate knowledge* **(DP8)**.

An inherent problem of data ecosystems is that actors of different backgrounds have different requirements towards the shared data (Azkan et al. 2020). DQ tools must take this aspect into account and should offer explanations for identified DQ issues. These should be commonly understandable and usable by partners in the ecosystem to resolve and reproduce errors if necessary (Bosch et al. 2021). Accordingly: *Provide the tool with explanations for DQ issues for users to understand and resolve DQ problems, given that there are users without the necessary data domain knowledge* **(DP9)**.

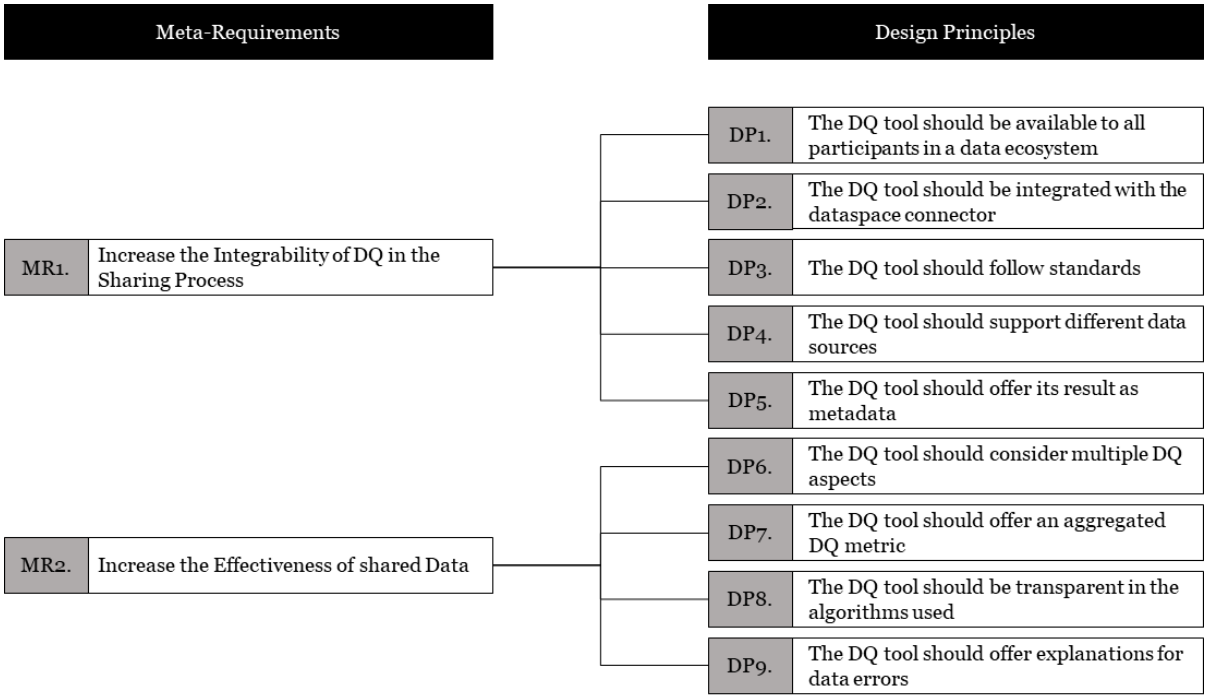

Figure 2. Design Principles for Data Quality Tools in Data Ecosystems

Demonstration and Evaluation

For demonstration and evaluation of generalized design knowledge, instantiation, as a software prototype, is a common practice (Möller et al. 2020). In this study, we present the implementation details of a DQ tool in a data ecosystem that realizes the proposed design knowledge. To demonstrate the concrete technical realization of the design principles, we reference them in the implementation's description. For the evaluation, we conducted a qualitative group discussion involving the DSR project team and members of

the MCo management board. The overall goal of MCo was to extend an existing IDS-based data ecosystem with a DQ component that conducts multiple DQ measurements on sensor-based data streams from machines in their production facilities. This way, data consumers are not only informed about a potential lack of DQ, but they can also identify factors and reasons that have led to low-quality data. This gives them the chance to act upon the root cause and resolve data errors.

The implemented solution comprises two components: an IDS Dataspace Connector (DSC) that is available open-source and includes a docker compose file for fast provisioning (IDSA 2022a), and a DQ application (DQ App) written in Python. To work in different contexts, the DQ App adheres to the specification of Data Apps by the IDS (IDSA 2022b) *(DP1 and 3)*. The DQ App and the DSC are tightly integrated using Apache Camel as an integration framework (Apache Camel 2022) *(DP2)*. This way, the DSC allows the usage of multiple Data Apps if needed.

Data Sharing Process

The overall data sharing process starts with data generation at the source (see Figure 3). After initial data conversions, analytics, and preparation steps, the data provider pushes the file or stream-based data to the DSC, where it is stored as a resource and accessible under user-defined rules *(DP4)*. The DQ App subscribes to changes of this resource and is automatically invoked once new data is available. These steps follow the standardized IDS communication protocols and are REST-based (IDSA 2022a) *(DP3)*. Once the DQ App receives an update of the data, the DQ score is calculated. For traceability between data and the analytical results, a unique ID and the DQ results are added to the metadata of the dataset *(DP5)*. As a results, the DQ extended data is sent back to the DSC where it is saved and made accessible, just like the original dataset. In our solution architecture, we integrated the DQ App closely with the dataspace connector technologies to enforce that DQ is checked prior to data sharing and follow the IDS terminologies for interoperability with other IDS-based data ecosystems (IDSA 2022b).

Once the data is available at the DSC, data consumers within the data ecosystem can access the quality checked data. For this, they negotiate contracts with each other, which include the relevant data sharing terms specified by the participants. Contracts and subscriptions enable automated and sovereign data transfers, so data engineers can focus on using data and avoid the common data governance overhead (Gröger 2021; IDSA 2022a).

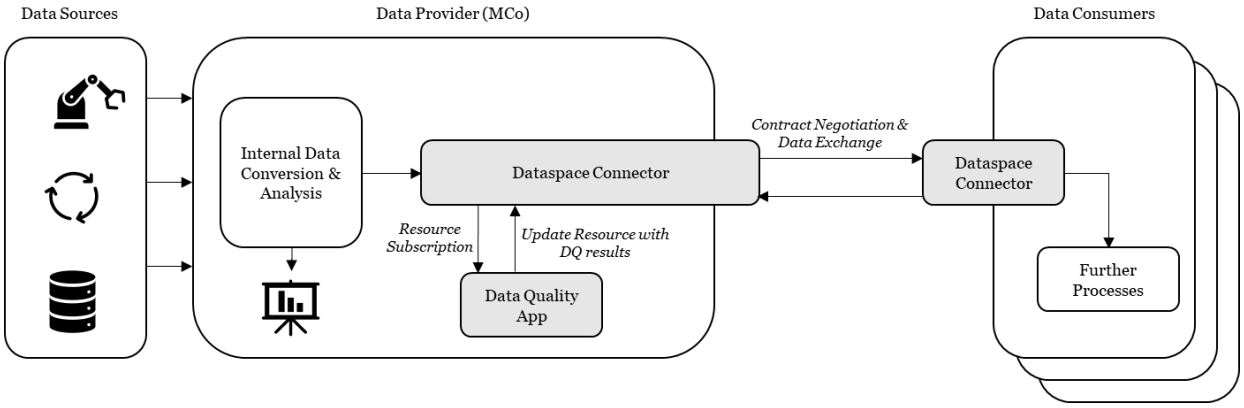

Figure 3. Architecture of the Data Sharing Process

Data Quality App

For implementing the DQ App, we used the programming language Python as it is well suited for data analytics tasks and offers many useful frameworks. However, local Python applications are limited regarding big data. We, thus, decided to use Apache Spark as our processing engine (Apache Spark 2022). Upon creating a Spark RDD (Resilient Distributed Dataset), the data is split up into smaller blocks, which

are then analyzed successively. Through this approach, the DQ App is more scalable and can handle large amounts of data while utilizing widely established data analytics frameworks.

To determine the quality of the provided data set, we combined four different DQ measures that are suitable for sensor-based data streams and cover a variety of DQ dimensions *(DP6 and 8)*. First, we used an Isolation-Forest algorithm to analyze each data block for outliers and assess the validity of data. Based on the commonness of outliers in a data block, we can infer an **outlier measure**. Second, we scan the data set for a potential **concept drift** to determine the accuracy of the data. A concept drift means that a target variable is changing over time in unforeseen ways and is particularly challenging for ML as predictions become increasingly inaccurate. To avoid this, it is important to identify concept drifts early on (Altendeitering and Dübler 2020). For the detection of a concept drift, we assess the approximation of different concepts per data block and store the results in an internal database. In a following step, the concepts of multiple blocks are compared against each other. Through this knowledge, we can detect drifting concepts and deflect a measure for concept drift.

Third, with the **no value measure,** we identify sensors that do not provide values in a whole data block and is used to determine the completeness of a data stream. Finally, the **constant measure** detects sensors staying constant for a long time. All four measures are saved in the database and then combined into an average measure over all blocks in the dataset. Finally, the result in a form of an easy-to-understand ratio between zero and one is added to the metadata of the resource and accessible by other participants within the data ecosystem *(DP7 and 9)*.

Evaluation

For evaluation we conducted a qualitative group discussion on the derived DPs and our developments with the DSR project team and members from the management at MCo. The meeting lasted two hours and we received positive feedback. The derived DPs were considered thorough, complete, and well-suited for ensuring DQ in data sharing. Our instantiated prototype was seen as a good demonstration of the DPs. The easily comprehensible DQ scores were particularly well received, as they help DQ *"to become effective and easy to grasp"* as one participant stated. A data engineer at MCo also liked the consideration of different DQ dimensions as it helps to act upon DQ problems. The participant noted: *"By splitting up the question of the overall quality of a dataset into sub-problems of outliers, concept drift, missing and constant values, we can start concrete actions in the data pipeline"*. Overall, the participants agreed that the developed solution is a useful advancement for data ecosystems as it helps to increase the trust in data sets and supports partners in a data ecosystem to leverage the benefits of data sharing.

Conclusion

In this study, we presented the results of a DSR study on the development of a nascent design theory for DQ tools in data ecosystems. To achieve this goal, we derived meta-requirements from literature and expert interviews and developed design principles that address these meta-requirements. Afterwards, we successfully evaluated our design principles by instantiating a DQ tool in an existing data ecosystem in the manufacturing domain. Specifically, our study offers the following contributions.

The **scientific contribution** of our work lies in the provisioning of the prescriptive knowledge necessary for the design of DQ tools in data ecosystems. Despite the frequently mentioned importance of DQ for data ecosystems (Azkan et al. 2020), and research calls towards more intelligent and collaborative DQ tools (Altendeitering and Tomczyk 2022), there is still limited research on the topic and no study has yet connected these two streams. By following an iterative research approach and evaluating our results in a real-world context, we created an artifact that is theoretically grounded and practically inspired (Peffers et al. 2007). As a result, our study contributes to the current body of knowledge in the domains of both DQ and data ecosystems.

Furthermore, our study provides various **managerial implications**. First, practitioners can use the presented design knowledge to build customized DQ tools in their own contexts (Möller et al. 2020). Specifically, our design principles can act as a framework for organizations and governments to inform the design and development of DQ tools in their ecosystems and environments. This could help break data silos and facilitate the use of data for the mutual benefit (UNCTAD 2021; WEF 2021). Second, by describing a

concrete usage scenario and instantiation of a DQ tool, we can advance the discussions about DQ and raise awareness for the aspect of DQ not only within organizations but also at their boundaries.

However, it is essential to point out that our study is subject to multiple **limitations**. First, we derived the design principles and their implementation as a software artifact in a single case resulting in a limited generalizability of our results. It is likely that the social or organizational context of our study influenced our design decisions. Studies in other contexts or project settings might come to different conclusions and evaluation results that lead to a different artifact at the end of the DSR process. Second, the evaluations we conducted in each of the three cycles are restricted to qualitative data. Further studies that follow quantitative evaluation approaches, such as performance measurements or questionnaires, might be beneficial to confirm our findings. Third, since data ecosystems and DQ are active research areas, it is likely that new requirements arise that need to be reflected in the design knowledge on DQ in data ecosystems.

Based on these limitations, there are several possibilities for **future research**. It would be valuable to implement our solution in other data ecosystems and in cases with different DQ requirements to investigate how the design principles change. Specifically, it would be interesting to extend our solution with standards and information models for more diverse data sources and DQ contexts. Furthermore, future research should examine different architectural designs for implementing DQ applications. For example, a DQ application might be operated at the sensor, connector, or platform level, which poses different organizational and technical requirements. Finally, there is a need for further research on realizing joint DQ efforts across multiple organizations. As a result, cross-organizational DQ initiatives could provide high-quality data products within a data ecosystem (Altendeitering and Tomczyk 2022).

Acknowledgements

This research was partly supported by the EU's Horizon 2020 program and the QU4LITY project (GA no. 825030).

REFERENCES

Altendeitering, M., and Dübler, S. 2020. "Scalable Detection of Concept Drift: A Learning Technique Based on Support Vector Machines," *Procedia Manufacturing* (51), pp. 400-407 (doi: 10.1016/j.promfg.2020.10.057).

Altendeitering, M., and Guggenberger, T. M. 2021. "Designing Data Quality Tools: Findings from an Action Design Research Project at Boehringer Ingelheim," *ECIS 2021 Research Papers*.

Altendeitering, M., and Tomczyk, M. 2022. "A Functional Taxonomy of Data Quality Tools: Insights from Science and Practice," *Wirtschaftsinformatik 2022 Proceedings*.

Amadori, A., Altendeitering, M., and Otto, B. 2020. "Challenges of Data Management in Industry 4.0: A Single Case Study of the Material Retrieval Process," in *International Conference on Business Information Systems*, Springer, Cham, pp. 379-390 (doi: 10.1007/978-3-030-53337-3_28).

Apache Camel. 2022. "Apache Camel," available at https://camel.apache.org/, accessed on Feb 10 2022.

Apache Spark. 2022. "Apache Spark," available at https://spark.apache.org/, accessed on Feb 10 2022.

Azkan, C., Möller, F., Meisel, L., and Otto, B. 2020. "A Service-Dominant Logic Perspective on Data Ecosystems: A Case-Study based Morphology," in *Proceedings of the 28th European Conference on Information Systems*, Marrakech, Morocco.

Bosch, J., Holmström, H. O., and Crnkovic, I. 2021. "Engineering AI Systems: A Research Agenda," in *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*, IGI Global, pp. 1-19 (doi: 10.4018/978-1-7998-5101-1.ch001).

Capiello, C., Gal, A., Jarke, M., and Rehof, J. 2020. "Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391)," (doi: 10.4230/DagRep.9.9.66).

Chu, X., Ilyas, I. F., Krishnan, S., and Wang, J. 2016. "Data Cleaning," in *Proceedings of the 2016 International Conference on Management of Data*, F. Özcan, G. Koutrika and S. Madden (eds.), New York, NY, USA: ACM, pp. 2201-2206 (doi: 10.1145/2882903.2912574).

Ehrlinger, L., Rusz, E., and Wöß, W. 2019. "A survey of data quality measurement and monitoring tools," *arXiv preprint arXiv:1907.08138*.

Gregor, S. 2006. "The Nature of Theory in Information Systems," *MIS Quarterly* (30:3), pp. 611-642 (doi: 10.2307/25148742).

Gregory, R. W., and Muntermann, J. 2014. "Research Note —Heuristic Theorizing: Proactively Generating Design Theories," *Information Systems Research* (25:3), pp. 639-653 (doi: 10.1287/isre.2014.0533).
Gröger, C. 2021. "There is no AI without data," *Communications of the ACM* (64:11), pp. 98-108 (doi: 10.1145/3448247).
Guggenberger, T. M., Möller, F., Haarhaus, T., Gür, I., and Otto, B. 2020. "Ecosystem Types in Information Systems," in *Proceedings of the 28th European Conference on Information Systems*, Marrakech, Morocco.
Hevner, March, Park, and Ram. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105 (doi: 10.2307/25148625).
Iansiti, M., and Levien, R. 2004. "Strategy as Ecology," *Harvard Business Review* (82:3), 68-78, 126.
IDSA. 2022a. "Dataspace Connector," available at https://github.com/International-Data-Spaces-Association/DataspaceConnector, accessed on Jan 27 2022.
IDSA. 2022b. "International Data Spaces," available at https://internationaldataspaces.org/, accessed on Feb 5 2022.
ISO. 2022. "ISO 8000-61:2016(en) Data quality," available at https://www.iso.org/obp/ui/#iso:std:iso:8000:-61:ed-1:v1:en, accessed on Apr 20 2022.
Jacobides, M., Cennamo, C., and Gawer, A. 2018. "Towards a Theory of Ecosystems," *Strategic Management Journal* (39), pp. 2255-2276 (doi: 10.1002/smj.2904).
Kabalisa, R., and Altmann, J. 2021. "AI Technologies and Motives for AI Adoption by Countries and Firms: A Systematic Literature Review," in *Economics of Grids, Clouds, Systems, and Services*, K. Tserpes, J. Altmann, J. Á. Bañares, O. Agmon Ben-Yehuda, K. Djemame, V. Stankovski and B. Tuffin (eds.), Cham: Springer International Publishing, pp. 39-51 (doi: 10.1007/978-3-030-92916-9_4).
Kruse, L. C., Seidel, S., and Gregor, S. 2015. "Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions," in *2015 48th Hawaii International Conference on System Sciences,* HI, USA. 05.01.2015 - 08.01.2015, IEEE, pp. 4039-4048 (doi: 10.1109/HICSS.2015.485).
Mobility Data Space. 2022. "Mobility Data Space," available at https://mobility-dataspace.eu/, accessed on Jan 24 2022.
Möller, F., Guggenberger, T., and Otto, B. 2020. "Towards a Method for Design Principle Development in Information Systems," in *Proceedings of the 15th International Conference on Design Science Research in Information Systems and Technology,* Kristiansand: Norway, pp. 208-220.
Moore, J. F. 1993. "Predators and Prey: A New Ecology of Competition," *Harvard Business Review* (71:3), pp. 75-86.
Oliveira, M. I. S., Barros Lima, G. d. F., and Farias Lóscio, B. 2019. "Investigations into Data Ecosystems: a systematic mapping study," *Knowledge and Information Systems* (61:2), pp. 589-630 (doi: 10.1007/s10115-018-1323-6).
Otto, B. 2015. "Quality and Value of the Data Resource in Large Enterprises," *Information Systems Management* (32:3), pp. 234-251 (doi: 10.1080/10580530.2015.1044344).
Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45-77 (doi: 10.2753/MIS0742-1222240302).
Redman, T. C. 2020. "To Improve Data Quality, Start at the Source," available at https://hbr.org/2020/02/to-improve-data-quality-start-at-the-source, accessed on Jan 24 2022.
Setia, P., Venkatesh, V., and Joglekar, S. 2013. "Leveraging Digital Technologies: How Information Quality Leads to Localized Capabilities and Customer Service Performance," *MIS Quarterly* (37:2), pp. 565-590 (doi: 10.25300/misq/2013/37.2.11).
UNCTAD. 2021. "Digital Economy Report 2021," available at https://unctad.org/system/files/official-document/der2021_en.pdf, accessed on Apr 20 2022.
Walter, V., Gyoery, A., and Legner, C. 2022. "Deploying machine learning based data quality controls– Design principles and insights from the field," *Wirtschaftsinformatik 2022 Proceedings*.
Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5-33 (doi: 10.1080/07421222.1996.11518099).
WEF. 2021. "Data-driven Economies: Foundations for Our Common Future," available at https://www.weforum.org/whitepapers/data-driven-economies-foundations-for-our-common-future, accessed on Apr 20 2022.