D. Demmler, D. Krupka, H. Federrath. (Hrsg.): INFORMATIK 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 1037

TOWARDS A DATA QUALITY INDEX FOR DATA VALUATION IN THE DATA ECONOMY

Dusan Dokic [1], Hannah Stein[2]

Abstract: Data represent a key resource for firm success, being used for strategic decision making and increasing business process efficiency. Despite the large potential of data sharing within data ecosystems or markets, firms are reluctant to do this, due to fear of losing competitive edges, lack of trust and ambiguity regarding data value. According to prior research, data value vastly depends on usage and quality. This paper focuses on data quality, as the lack of methods for quantifying data quality is one main reason for missing comprehensible data valuation approaches. We analyze 15 existing data quality indices (DQI) from theory and practice, identify relevant data quality dimensions and discuss metrics for applicability in data valuation approaches for data ecosystems and markets. Based on a quantitative study, we propose a DQI concept for developing transparent, objective data valuation methods, while providing a better understanding of inter- and intra-organizational data value.

Keywords: data quality index; data quality metrics; data economy

1 Introduction

The importance of data for the success of modern companies is non-debatable as data-driven business models can be identified at the heart of multi-billion dollar companies like Facebook or Telefonica [Fa20]. Data-driven business models are implemented to such an extent in industries like insurance, banking, telecommunication, and tourism, that some already regard them as data monetization industries [HN20]. One example for such business model is the extraction and selling of raw behavioural customer data to third parties [WS22]. The relevance of data-driven business models is expected to increase, as data sharing marketplaces and ecosystems are being increasingly addressed by current research and practitioners [Ab21]. Data sharing within data ecosystems and the resulting increase in data utilization is expected to create up to 30 percent of the world's gross domestic product [JT18] and to increase involved companies' revenue. Engagement in data ecosystems is regarded as necessity for future economic survival of organizations [SHS13]. This increases the pressure on management to leverage existing technologies like artificial intelligence and machine learning, which highly depend on data of high quality [ACS16]. A key challenge in developing and operating such data ecosystems or sharing platforms is the determination of data value and to express it in monetary units, so that participating data providers can be compensated accordingly [Ge21].

[1] German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3, 66123 Saarbruecken, Germany, dusan.dokic@dfki.de
[2] Saarland University, Campus A5.4, 66123 Saarbruecken, Germany, hannah.stein@iss.uni-saarland.de

doi:10.18420/inf2022_86

However, to the best of the authors knowledge, replicable and universally applicable methods to determine monetary data value seem to be missing [ST17]. While many factors are influencing the value of data [MW99], related work points to the fact that data quality (DQ) represents a main value driver for data [OW08, SM22]. This means that the value of data cannot be assessed without assessing its quality first. Since data are the main resource or 'input' in data markets, DQ valuation methods are likely to be at the heart of any business model of such platforms and ecosystems, otherwise the value proposition will not be delivered. Literature is unanimous that DQ represents a multidimensional concept [WS96, PLW02]. Therefore, this paper will focus on the individual dimensions of DQ. To measure the quality of data, it is necessary to measure the single dimensions of it first. Research on data quality dimensions (DQD), their definitions and metrics already exists, and will be discussed in more detail later on in the paper.

The aim of this work is to collect existing attempts of creating a data quality index (DQI) and to analyze them based on the DQD and metrics they use. Based on these results, we propose a first DQI model that that is equipped with guidelines about the structure of such a DQI, what DQD are considered and what metrics are used. Our proposed DQI model represents a sturdy starting point for assessing the quality of data in data ecosystems. The remainder of this paper is structured as follows: in the second chapter, related work in regards to data quality and data ecosystems is being discussed. This includes an overview of 15 current examples of DQI. In the third chapter of this paper, the study design is laid out and the study results are analysed for possible implications regarding our own DQI model. Afterwards, the discovered data quality metrics are compared and filtered for their applicability on data ecosystems in chapter 4. The fifth chapter synthesizes the results of the previous chapters into a first proposal for a DQI for data ecosystems. Finally, a summary of the paper results is given and limitations and future work are discussed.

2 Data Quality for Data Ecosystems

This paper focuses on analysis and differentiation of DQD and their metrics, since these constitute the majority of a data quality index. We will elaborate on the metrics of the single data quality dimensions in a separate chapter, as the metrics make up the most important part for the data quality index.

Literature suggests that information might be more valuable than the actual technology of a company and the discussion on measuring the value of information in a company goes even back to the 1990's [Gl93]. Numerous (research) domains such as mathematics, accounting and logistics developed specific approaches in order to measure data and information value, as well as DQ [Bo14, Wi12]. Other works that focus on the importance of DQ in general and justify developing a measuring method for the same can be found across literature [NS17, ORH05, Re19].

Data Quality Dimension	Definition
Accessibility	The extent to which data are available or easily and quickly retrievable.
Appropriate amount of data	The extent to which the volume of data is appropriate for the task at hand.
Believability	The extent to which data are accepted or regarded as true, real, and credible.
Completeness	The extent to which data are of sufficient breadth, depth, and scope for the task at hand.
Concise Representation	The extent to which data are compactly represented.
Consistent Representation	The extent to which data are presented in the same format.
Value-added	The extent to which data are beneficial and provide advantages from their use.
Accuracy	The extent to which data are correct and reliable.
Interpretability	The extent to which data are in appropriate languages, symbols, and units and the definitions are clear.
Objectivity	The extent to which data are unbiased, unprejudiced, and impartial.
Relevancy	The extent to which data are applicable and helpful for the task at hand.
Reputation	The extent to which data are highly regarded in terms of its source or content.
Security	The extent to which access to data are restricted appropriately to maintain its security.
Timeliness	The extent to which data are sufficiently up-to-date for the task at hand.
Understandability	The extent to which data are easily comprehended.
Portability	Transferability to other usage targets, e.g., sales data, are additionally used for product recommendations.

Tab. 1: Definitions of the DQD according to [PLW02]

DQ is often defined as 'the fitness for use' and is regarded as a multi-dimensional concept [WS96]. Thereby, DQ defines and provides DQD that are being reused throughout literature. For example, a conceptual framework of DQ was developed, including 16 dimensions and group them into four categories: intrinsic, contextual, representational, and accessibility [WS96]. In addition, these authors developed further DQ assessment methods and metrics and address data from a management perspective [KSW02, Le02, PLW02, WML01]. A utility-driven measurement framework for DQD within a specific usage context points to the fact that high DQ has a positive effect on business value[ES07]. It ties the value of data to specific assets in the company and links these to DQD like 'Completeness' or 'Accuracy'. Furthermore, practical examples for the use of these dimension in frameworks and methods for assessing DQ are given. All above-mentioned works are crucial as a basis for the development of a holistic approach for measuring DQ within the context of data valuation in data ecosystems using a DQI. For the development of our own DQI we will use the shortened list of [PLW02] as the basis for developing the index and the associated questionnaire. It includes the dimensions that are being regarded as relevant in the broad literature and for which pre-evaluated questionnaire items already exist [WS96]. Still, not

all dimensions are considered as equally important in literature. Some dimensions are being renamed while keeping the exact definition. For example, 'Completeness', 'Timeliness', and 'Accuracy' are being mentioned more frequently than others [CZ15],[Se16] [Ta16]. Meanwhile, dimensions like 'Believability' are sometimes exchanged with terms like 'Credibility' [CZ15]. The full list of the DQD and their definitions are given in Table 1.

2.1 Data Quality Indices

An index is a model that is being used to capture the information of a multi-dimensional theoretical construct by aggregating the different variables that make up the construct into a single number [DB16]. An index turns highly complex constructs into an understandable and comparable measure. This reduction in complexity comes with a loss of information as the involved indicators are measuring different aspects of the construct that can't be directly compared with each other. While this trade-off between complexity and loss of information is regarded as the main argument against such an index, there are plenty of examples across various disciplines that leverage such indices to convey information in a simple and understandable way to third parties [Un22]. Following the definition of an index provided above, DQ fulfils all the requirements for the use of an index: it is complex in its structure and depends on various dimensions that all measure a different aspect of it. Literature suggested to develop and apply DQI more widely [PLW02]. Since then, a modest number of undertakings can be identified throughout the literature and practice that have attempted to develop a DQI. In the following, we give an overview of these DQIs and analyse them based on their weaknesses and strengths, their comparability, as well as the DQD and their corresponding metrics. Finally, we extract implications for the creation of our own concept of a DQI.

Attempts to create an index for the assessment of DQ are scarce in literature, and those that can be found are mostly context-specific. Table 2 gives an overview of some of these indices. Further Indices can be found in Table 4 and 5 in the appendix. In total, we were able to identify 15 DQIs from literature and practice. For instance, a DQI was used to identify Nursing Quality Indicators for Reporting and Evaluation, using the DQD 'Completeness', 'Timeliness', 'Accuracy', and 'Consistency'[Na20]. They use a data set about prevention of falls and fall injuries to show that an assessment through a DQI was possible and that a specific value could be attached to the data set at hand. In another example, a DQI is used to establish a data cockpit for the Bayer CropScience AG [EHO11]. They aim to harmonize the data infrastructure of the 120 different national subsidiaries of the Bayer CropScience AG by combining the master data of all subsidiaries into a so-called 'golden box' and distributing it from there to the regional systems. Their DQI is kept simple based on a ratio of a faulty data set to the number of data sets in total. Inaccurate data sets are defined as the data sets that violate one of the predefined business rules assigned to validation groups. As a final example for the current use of a DQI, the European Central Bank (ECB) approach in Frankfurt can be identified [SB16]. This DQI is aimed at banks that the ECB is supervising and uses the

dimensions 'Punctuality', 'Accuracy', 'Consistency', and 'Completeness' amongst others. This DQI is supposed to ensure the information submitted by the banks to the ECB is of sufficient quality and to enable a comparison of the organizations.

Origin	Data Type	Use Case	DQD	Strength	Weakness
[EHO11]	Process	DQ cockpit for improving, monitoring, comparing master data across subsidiaries ERP systems.	Unspecified; Business Rules	Comparable measure for corporate DQ monitoring; Measures process relevant data flaws	Company-specific; aggregation on validation group level instead of DQD.
[EBH13]	Customer	DQI-based system for controlling DQ.	Completeness, Accuracy	Example how DQI can improve DQ over time.	Ambiguity how DQD are assessed, measured and corrected.
[SB16]	Finance	ECB-Guidelines for DQ assessment of data from supervised banks.	Punctuality, Completeness, Accuracy, Plausibility	Traffic light system based on rating from 1 (good) to 4 (very bad / missing).	No specific metrics provided; weights unexplained
[Le18]	Corporate	Automatic discovering and review of business rules, incl. a DQI.	Referential integrity, Date, Code, Encryption and Numbering Pattern	Automatic review and discovery of business rules reduces complexity.	Indicators and metrics lack comprehensibility; DQI score is not explained.
[Na20]	Healthcare	DQI assesses DQ for nursing quality, evaluation and reporting	Completeness, Consistency, Accuracy, Timing	DQD and metrics are explained including formulas.	Chosen DQD are non-holistic; context specific weights; partial subjectivity.
[SWL20]	Corporate	DQI-based DQ assessment system for a cross-company batch management system in the tobacco industry.	Integrity, Accuracy, Timeliness, Uniqueness, Consistency, Validity, Stability	Example for the use of a DQI for comparing data between enterprises in the same industry.	Metrics for DQD values are unclear as the values are already 'provided'.

Tab. 2: Overview of the existing DQIs from literature and practice - part 1

In order to be able to determine which DQD should be included in our model, a quantitative study was conducted. The design and execution as well as the results of this study will be discussed in detail in the next section.

3 Study Design

We developed a quantitative survey aiming to rank the individual DQD according to their importance for the companies and to determine possible weights for the later index. The questionnaire was designed to address mainly data professionals such as Chief Data Officer of companies. In the following, an overview of the questionnaire's development as well as a description of the conducted study is given.

3.1 Development of the study questionnaire

We follow the guidelines for creating a standardized questionnaire and which aspects have to be considered [DB16]. As recommended by [DB16], the individual questionnaire items were taken from existing literature [WS96]. A first draft of the questionnaire was tested and discussed in a questionnaire expert conference where a total of 4 people, experienced in the development of questionnaires, were interviewed and asked for comprehensive feedback. The feedback was considered and incorporated until no more suggestions for improvement were made.

The questionnaire is structured as follows. In the beginning, a brief instruction is given on the questions' structure and what the single scale points of the 9-point-Likert scale mean. Thereby we aimed to simplify handling the questionnaire for the participants and to prevent ambiguities. The structure of the questionnaire consists of title, instructions on how the questions should be answered, content-related question blocks, statistical information, a feedback field, and a concluding section. The first questionnaire item asks the participants' perception of the importance of the three data value dimensions, 'Quality,' 'Usage,' 'Timeliness', concerning the value of data. These three dimensions have been identified as potentially important for the valuation of data in a qualitative study [SM22] and the survey at hand aims to verify this assumption. In the questionnaire, the associated definitions of the dimension are given to clarify what is meant by the used terms. The central part of the questionnaire aims to measure the perceived importance regarding different DQD in terms of DQ. For this purpose, the list of already identified data quality dimensions by [PLW02] was adopted for the study. The list contains a total of 16 quality dimensions listed in Table 1. It should be noted that the name of the dimension 'Ease of Manipulation' was changed to 'Portability'. It should also be mentioned here that the dimension 'Free-of-Error' named by [PLW02] is also often titled 'Accuracy' in the literature. No original questionnaire items were developed for the study, instead already validated items, rating scales and ranking methods were adopted [WS96]. The original rating scale from 1 to 9 used by [WS96] is used: 1 stands for 'Extremely important', 5 for 'Important', and 9 for 'Not important at all'. For clarity and streamlining purposes, the 16 DQDs are divided into four blocks of 4 DQDs each. In each block, the question is repeated, and the definitions of the dimensions are given. In the last part of the questionnaire, information on the participant's company is being captured [DB16]. This includes information on the domain, number of employees and

target customer group. Furthermore, information on the participant's years of experience working with data and their current job title are queried.

The survey was conducted via Unipark, an online survey tool provider. The target group were companies operating in the business-to-business sector. We followed numerous acquisition strategies, e.g., through social business networks, our own network, cold calls, etc. In sum, 31 valid questionnaires could be used for analysis. The analysis has shown that 32 percent of the participants are working in companies with less than 50 employees, while 42 percent are working in companies that have employees in the range of 50 to 499 employees and 26 percent of the participants have answered that they are employed in a company with 500 or more employees. 58 percent of these 31 participants declared that they have 5 or fewer years of experience in working with data. The remaining 42 percent declared that they have more than 6 years of experience.

3.2 Study results

The responses were aggregated, and a mean was taken for each of the DQD rated by the subjects (definitions of DQD can be found in Table 1). The dimensions were then ranked in descending order of importance (lowest to highest mean value). The analysis results are given in Table 2 and show that the dimension 'Believability' has the lowest mean value (2.19) and is perceived as the most important by the participants. The upper quarter of the table also includes the dimensions 'Value-added' (2.26), 'Free-of-Error' or 'Accuracy' (2.52), and 'Timeliness' (2.61). In the last quarter are the dimensions 'Reputation' (3.29), 'Understandability' (3.65), 'Concise Representation' (3.81), and 'Portability' (4.19).

Data Quality Dimension	Indicator	Mean
Believability	β_1	2.19
Value-added	β_2	2.26
Accuracy	β_3	2.52
Timeliness	β_4	2.61
Interpretability	β_5	2.65
Relevancy	β_6	2.65
Consistent Representation	β_7	2.87
Accessability	β_8	3.00
Objectivity	β_9	3.06
Security	β_{10}	3.16
Completeness	β_{11}	3.19
Appropriate Amount of Data	β_{12}	3.26
Reputation	β_{13}	3.39
Understandability	β_{14}	3.65
Concise Representation	β_{15}	3,81
Portability	β_{16}	4.19

Tab. 3: Importance of DQD by ascending means

4 Data Quality Metrics

In the following, the different approaches to metrics for the previously described and evaluated DQD are presented and explained. [PLW02] gives three forms that can be used to develop such metrics and provide suggestions on which to use for what dimension. The simple ratio of undesirable outcomes to total outcomes subtracted from 1 can be used for dimensions like 'Accuracy', 'Completeness', 'Concise representation', and 'Relevancy'. Applying this to the dimension of 'Consistency' leads to the following metrics:

Consistency

$$= 1 - \frac{number\ of\ violations\ of\ a\ specific\ consistency\ type}{total\ number\ of\ consistency\ checks} \tag{1}$$

Accuracy

$$= 1 - \frac{Number\ of\ incorrect\ data\ sets}{Total\ number\ of\ data\ units} \tag{2}$$

To be able to use these metrics, a definition of what makes data 'incorrect' has to be introduced beforehand. Business rules can then be used to count the number of data sets that do not adhere to the 'Accuracy'-rule.

Completeness

$$= 1 - \frac{Number\ of\ data\ sets\ with\ missing\ values}{Total\ number\ of\ data\ sets} \tag{3}$$

For metrics like 'Appropriate amount of data', 'Accessibility' or 'Timeliness', a min-max operator is suggested [Le06]. It can handle multi-dimensional input and leaves the option to either be liberal or conservative in the aggregation of the values of the individual indicators. While the authors do not give specific metrics for the dimensions they name, their work is used as the theoretical basis for future studies that provide them [Le06].

Appropriate amount of data

$$= min[\frac{Number\ of\ data\ sets\ provided}{Number\ of\ data\ sets\ needed}, \frac{Number\ of\ data\ sets\ needed}{Number\ of\ data\ sets\ provided}] \tag{4}$$

For this metric, the amount of data has to be specified beforehand. The min operator can then be used to choose the smallest value of the given ratios.

Timeliness

Metric according to [Le06]

$$= \{max[(1 - \frac{Currency}{Volatility}, 0]\}^S \tag{5}$$

with **Currency**

$$= (delivery\ time - input\ time) + age \tag{6}$$

and **Volatility**

$$= length\ of\ time\ over\ which\ the\ data\ remains\ valid \tag{7}$$

Delivery time stands for the time at which the data was delivered to the user, while input time means the time at which the system received the data. Age represents the age at which the system received the data. S denotes the exponent value and parameter that controls the sensitivity of the metrics. If a value of S = 1 stands for no adjustment, a value of S < 1 stands for less sensitivity and makes the metric for timeliness bigger. Following the same logic, a value of S > 1 makes the metric more sensitive and returns a smaller metric value[Le06, Ba98].

Metric according to [Ge18]

$$= 100 * (\frac{Number\ of\ data\ sets\ that\ have\ been\ updated\ since\ a\ chosen\ date}{Total\ Number\ of\ data\ sets}) \tag{8}$$

Accessibility

$$= \{max[(1 - \frac{Request\ time}{Relevance\ time}, 0]\}^S \tag{9}$$

In this metric, 'Request time' represents the current interval of time (in seconds) that it takes from the request by the user to the deliver of the data to the user. 'Relevance time' means the interval of time from request by the user to the time at which the data is no longer of any use to the user. This metric is based around the definition that the earlier data is delivered, the more valuable it is. As the first term would eventually become negative due to the long delivery time, the max operator will limit the minimum to zero. Accessibility can also be based on other approaches besides time depending on the context and should then be calculated by aggregating the different methods [Le06].

Believability

$$\begin{aligned} = min(&Believability\ of\ source;\\ &Believability\ when\ compared\ to\ internal\ commonsense\ standard;\\ &Believability\ based\ on\ the\ age\ of\ data\ ) \end{aligned} \tag{10}$$

The values of the three 'Believability' variables are rated from 0 to 1 by an individual who has background knowledge and can therefore evaluate the credibility of the source of the data. This metric for 'Believability' is based on subjective opinion and personal experience and is therefore open to biases and heavily context-dependent.

For DQD like 'Reputation', 'Interpretability' and others, specific metrics with a defined formula could not be found as they are very context-dependent and often subjective. The general approach for handling these DQD is that data quality rules (DQR) are defined, at best with the help of domain experts. The data sets are then checked for adherence to the specific rules and the number of data sets that violate the DQR are counted. That number is then divided by the total number of data sets verified by the rule and the result is subtracted from 1. A general example for the described approach is shown in equation (11).

$$1 - \frac{Number\ of\ data\ sets\ that\ violate\ DQR}{Number\ of\ data\ sets\ validated\ against\ DQR} \tag{11}$$

Since DQR constitute a separate part of DQ and a detailed discussion of the topic is beyond the scope of this paper, it is assumed in the following that the individual subjective dimensions can be captured via such a DQR in combination with equation 11. This assumption is based on literature that provides methods for the automatic discovery and assessment of such rules [Le06].

5 Data Quality Index

We will now synthesize the results of the previous chapters towards a DQI for data valuation in data economy. When analyzing the DQI, it was noticed that many DQD metrics are based on DQR. This is due to the fact that use cases are usually very specific and no generally applicable metrics exist, in contrast to dimensions like "Completeness", for example. Even with DQDs for which general metrics exist, it is usually assumed that it is defined beforehand what the individual dimensions mean within the application context. E.g., for the DQD 'Believability', it must be defined what makes a data set believable before the metric can be applied. For the DQI of this paper, we therefore conclude that in the first step, rules must be created for each of the DQD. For each DQD, at least one rule should exist. Guidelines and recommendations for the creation of such rules can already be found in the literature [Le18, Fa15].

Regarding DQD, it has been noticed that many existing DQIs consider very few DQD on average or use context-specific labels but refer to the aspects of the DQD from literature. This can be justified by the fact that the existing DQIs were mostly developed for a very narrowly defined use case. This approach is certainly sensible in order to make users comprehend what is assessed through each DQD. Since the goal of this paper is to design a DQI for data valuation in data economy, we use the general DQDs proposed in the literature to ensure the generalizability of the DQI [PLW02]. If more specific DQDs need to be created in individual cases, they must be assigned to one of the general DQDs afterwards to ensure manageability. As for the number of DQDs to be used, it will certainly be difficult to check every record for every DQD. Therefore, it seems advisable to make this decision depending on the data to be considered. In a hypothetical data marketplace scenario, this decision could be made by the buyer of the data as they define what constitutes high quality data in their use case. Since

the previously described study found only a slight tendency about the importance of each DQD, no DQD was excluded. However, a possible exclusion of DQDs would be justified by the fact that many of these DQDs are very subjective or context-dependent and require some domain knowledge to verify the correctness of the rules used.

Many of the existing DQIs provide little or inaccurate information regarding the metrics used. In some cases, these are also very context-specific and difficult to understand for non-experts. This is also reflected in the literature, as there are only a very limited number of publications that precisely develop metrics for specific DQDs or provide guidance on how these should be developed [He18, Fa15, Le06]. Our model uses the metrics of [Le06] in combination with general rule-based metrics for the context-dependent DQD. Weights in an index are a way to incorporate a certain dimension more strongly into the index or to make the index more dependent on one dimension. Within the scope of this work, an attempt was made to develop a possible weighting based on the ranking of the study. Since the differences between the individual DQDs were not sufficiently large to make a meaningful difference in the weighting, this approach was discarded and not pursued further. Reasons for this could be that the number of participants was not sufficient and a more extensive study is needed, or that the weights need to be considered in the context of different use cases. From the analysis of the existing approaches, a multitude of different methods emerges to determine a weighting. However, the basis of these weighting approaches are mostly estimations of domain experts or statistical calculations, which calculate a suitable weighting based on these estimations. For our model, we refrain from specifying a fixed weighting, as this should be determined by those who want to acquire the data. Thus, in the implementation it needs to be ensured that the party acquiring data is given an opportunity to customize the weighting of DQ. In addition, the sum of all weights must equal 1. Otherwise the aggregation equation will not provide the desired output of a value between 0 and 1. For the aggregation of individual values, possible weights are multiplied by the values of the DQD. The resulting weighted values are then summed up. The result is a value between 0 and 1, which represents the quality of the data considered. The higher the value, the higher the quality of the data. It is also possible to multiply this value by 100 to give the DQI as a percentage, if desired.

The resulting model is shown in Figure 1 which illustrates, how the implementation of a DQI can reduce the complexity of multi-dimensional models. As mentioned in the previous chapters, this reduction in dimension comes with an information trade-off. We were able to observe this trade-off in our analysis of already existing DQIs. They lacked comprehensibility in terms of how the DQI was calculated due to missing information either on the used rules, dimensions or metrics. Therefore we recommend to introduce a mandatory description of the data attributes, DQR, DQD and the metrics by the party that provides the data, when implementing this DQI model into valuation approaches for data economy.

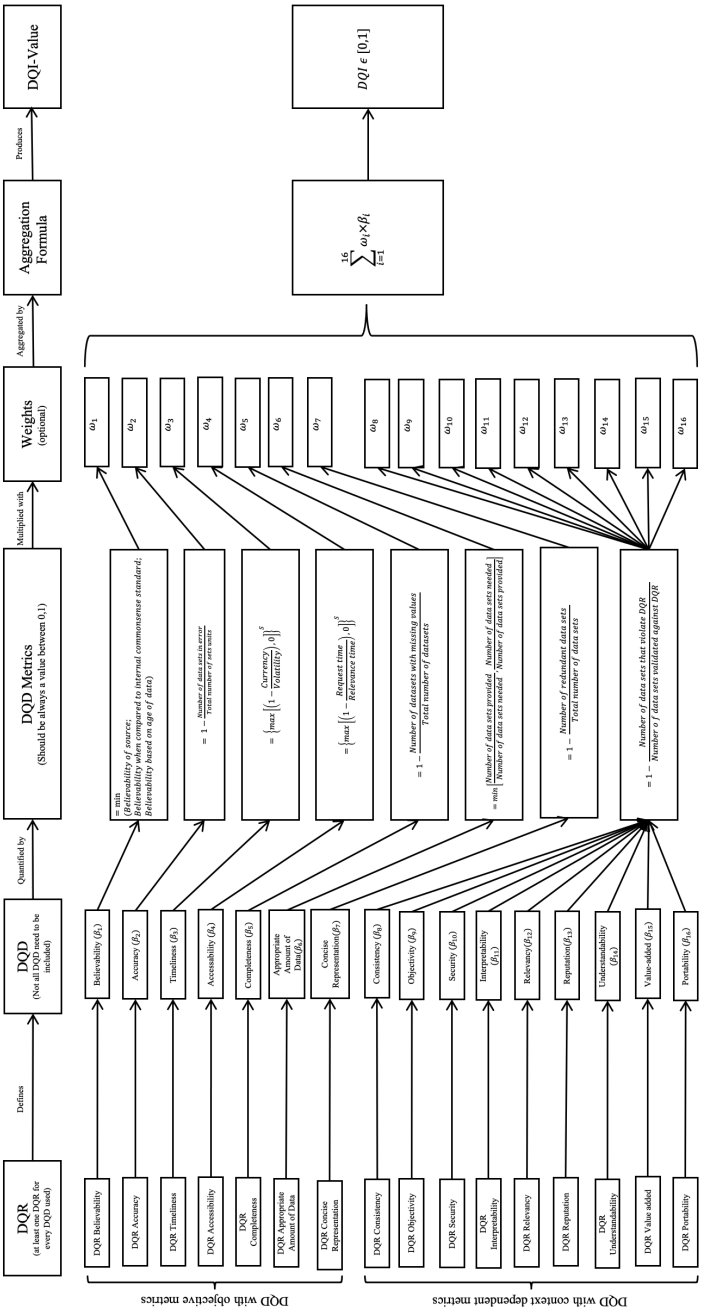

Fig. 1: General framework for developing a DQI

6 Conclusion

Our work does not come without limitations and room for improvement. First, we used the general approach to measure DQD. More precise formulas for the context-dependent DQD could be analyzed and implemented. While the general formula is mainly sufficient, easy to apply and provides a precise definition of DQRs, it cannot be guaranteed that this will do justice to every DQD in every use case. Similarly, this paper could not go further in-depth into the aspect of DQRs, which are a significant building block of the model. Although there are some approaches how these can be created, in the current model, however, the buyer of the data would have to check exactly whether the rules used also measure what they are supposed to. This requires domain knowledge and the availability of detailed descriptions, among other things. Since the study included only a small number of participants, the results are not necessarily representative and were not sufficient to develop a well-founded weighting of the DQD.

Future work should address these points and also examine approaches on how to automate the assessment process, how to incorporate the DQI to data valuation methods, and creating a service around the DQI. These points are especially important with regard to data economies, as the amount of data that will be shared or sold there will certainly not allow for manual valuation. A larger study could contribute to a reduction of the DQD and enable a more generalized weighting. The possible study could also consider different use cases in order to examine a possible difference in importance depending on the context.

In summary, in this paper we have gathered some basic arguments for a DQI and supported them with examples of existing indices from academia and practice. We then analyzed, compared, and evaluated these DQIs. Together with the results of a quantitative study, we have then incorporated these results into our own DQI model, which draws on metrics and suggestions from the literature. We therefore conceptualized a sturdy starting point for further research on DQIs for data valuation in the data economy.

7 Acknowledgement

This work is part-funded by the research project 'Future Data Assets' (grant number: 01MD19010C) funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the 'Smart Data Economy' technology program, managed by the DLR project management agency.

Bibliography

[Ab21] Abbas, Antragama Ewa; Agahari, Wirawan; van de Ven, Montijn; Zuiderwijk, Anneke; de Reuver, Mark: Business Data Sharing through Data Marketplaces: A Systematic Literature Review. J. Theor. Appl. Electron. Commer. Res., 16(7):3321–3339, 2021.

[ACS16] Anand, Abhijith; Coltman, Tim; Sharma, Rajeev: Four steps to realizing business value from digital data streams. MIS Q. Exec., 15(4), 2016.

[Ba98] Ballou, D.; Wang, R.; Pazer, H.; Tayi, G. K.: Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science, 44(4):462–484, 1998.

[Bo14] Borek, Alexander; Parlikad, Ajith Kumar; Woodall, Philip; Tomasella, Maurizio: A Risk Based Model For Quantifying The Impact Of Information Quality. Computers in Industry, 65(2):354–366, February 2014.

[CZ15] Cai, Li; Zhu, Yangyong: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Sci. J., 14:2, 2015.

[DB16] Döring, Nicola; Bortz, Jürgen: Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer, Wiesbaden, 5 edition, 2016.

[EBH13] E.Baghi; B.Otto; H.Osterle: Controlling Customer Master Data Quality: Findings from a Case Study. 05 2013.

[EHO11] Ebner, Verena; Hüner, Kai; Otto, Boris: , Fallstudie Bayer CropScience AG: Entwurf und Implementierung geschäftsorientierter Datenqualitätskennzahlen, 01 2011.

[ES07] Even, Adir; Shankaranarayanan, G.: Utility-Driven Assessment of Data Quality. SIGMIS Database, 38(2):75–93, May 2007.

[Fa15] Fan, W.: Data Quality: From Theory to Practice. SIGMOD Rec., 44(3):7–18, dec 2015.

[Fa20] Faroukhi, Abou Zakaria; El Alaoui, Imane; Youssef, Gahi; Amine, Aouatif: Big data monetization throughout Big Data Value Chain: a comprehensive review. Journal of Big Data, 7:3, 01 2020.

[Ge18] Geuer, M.: , Prozessorientierter Data Quality Index erfolgreich einführen. https://www.business-information-excellence.de/datenqualitaet/91-prozessorientierter-data-quality-index-erfolgreich-einfuehren, 2018. [Accessed 2022-07-23].

[Ge21] Gelhaar, Joshua; Gürpinar, Tan; Henke, Michael; Otto, Boris: Towards a taxonomy of incentive mechanisms for data sharing in data ecosystems. Proc. of 25th PACIS, p. 121, 2021.

[Gl93] Glazer, Rashi: Measuring the Value of Information: The Information-Intensive Organization. IBM Systems Journal, 32:99 – 110, 02 1993.

[He18] Heinrich, B.; Hristova, D.; Klier, M.; Schiller, A.; Szubartowicz, M.: Requirements for Data Quality Metrics. J. Data and Information Quality, 9(2), 2018.

[Hi] Hickey, D.; O'Connor, R.; McCormack, P.; Kearney, P.; Roosa, R.; Brennan, R.: The Data Quality Index: Improving Data Quality in Irish Healthcare Records. [Hi], pp. 625–636.

[HK15] Hönigl, Jürgen; Küng, Josef: Obtaining a data quality index with respect to case bases. Vietnam Journal of Computer Science, 2(1):47–56, 2015.

[HN20] Hanafizadeh, Payam; Nik, Mohammad Reza Harati: Configuration of Data Monetization: A Review of Literature with Thematic Analysis. Global Journal of Flexible Systems Management, 21(1):17–34, 2020.

[In21] International Aid Transparency Iniative: , IATI Data Quality Index. https://prod-iati-website.azureedge.net/prod-iati-website/documents/IATI_Data_Quality_Index_-_Background_Paper_1_.pdf, 2021. [Accessed 2022-07-23].

[Ir21] Ireland, Sheep: , Sheep Ireland Guide and Directory of Breeders. https://www.sheep.ie/wp-content/uploads/2021/08/SHEEP-IRELAND-GUIDE-DIRECTORY-2021-reading-view.pdf, 2021. [Accessed 2022-07-23].

[Jo18] Journal, World Economics: , The Data Quality Index (DQI). https://www.world-economics-journal.com/Pages/Data-Quality-Index.aspx, 2018. [Accessed 2022-07-23].

[JT18] JT Lorenz: . https://www.mckinsey.com/industries/financial-services/our-insights/the-rise-of-ecosystems-and-platforms-what-role-can-insurers-play-and-how-can-they-get-started, 2018. [Accessed 2022-07-23].

[Ka17] Karam, S.: , International trade data quality index. http://researcharchive.vuw.ac.nz/handle/10063/6128, 2017. [Accessed 2022-07-23].

[KSW02] Kahn, Beverly K.; Strong, Diane M.; Wang, Richard Y.: Information Quality Benchmarks: Product and Service Performance. Commun. ACM, 45(4):184–192, April 2002.

[Le02] Lee, Yang W.; Strong, Diane M.; Kahn, Beverly K.; Wang, Richard Y.: AIMQ: a methodology for information quality assessment. Information Management, 40(2):133–146, 2002.

[Le06] Lee, Y. W.; Wang, R. Y.; Funk, J.D.; Pipino, L. L.: Journey to Data Quality. MIT Press, Cambridge, 2006.

[Le18] Lee, E.; Park, Y.; Kim, H.; Jung, B.; Hong, J.: An Efficient Method for Automatic Updating Business Rules. In: Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems. RACS '18, Association for Computing Machinery, New York, NY, USA, p. 331–332, 2018.

[MW99] Moody, Daniel L.; Walsh, Peter: Measuring the Value Of Information - An Asset Valuation Approach. In: ECIS. 1999.

[Na20] Naik, Shanoja; Voong, Stephanie; Bamford, Megan; Smith, Kyle; Joyce, Angela; Grinspun, Doris: Assessment of the Nursing Quality Indicators for Reporting and Evaluation (NQuIRE) database using a data quality index. Journal of the American Medical Informatics Association : JAMIA, 27:776–782, 05 2020.

[NS17] Nagle, Tadhg; Sammon, David: The Data Value Map: a Framework for Developing Shared Understanding on Data Initiatives. In: ECIS. 2017.

[ORH05] Oliveira, Paulo; Rodrigues, Fátima; Henriques, Pedro Rangel: A Formal Definition of Data Quality Problems. In: ICIQ. 2005.

[OW08] Otto, Boris; Weber, Kristin: Data Governance. volume 53. pp. 265–283, 01 2008.

[PA22] Data Quality Index. https://www.paconsulting.com/industries/energy-and-utilities/data-quality-index/, [Accessed 2022-07-23].

[PLW02] Pipino, Leo L.; Lee, Yang W.; Wang, Richard Y.: Data Quality Assessment. Commun. ACM, 45(4):211–218, April 2002.

[Re19] Regneri, Michaela; Georgi, Juliane; Kost, Jurij; Pietsch, Niklas; Stamm, Sabine: Computing the Value of Data: Towards Applied Data Minimalism. ArXiv, abs/1907.12404, 2019.

[SB16] https://www.bankingsupervision.europa.eu/press/conferences/shared/pdf/sup_rep_conf/2017/Data_quality_framework_tools_and_products.pdf, [Accessed 2022-07-23].

[Se16] Serhani, Mohamed Adel; Kassabi, Hadeel T. El; Taleb, Ikbal; Nujum, Al Ramzana: An Hybrid Approach to Quality Evaluation across Big Data Value Chain. 2016 IEEE International Congress on Big Data (BigData Congress), pp. 418–425, 2016.

[SHS13] Selander, L.; Henfridsson, O.; Svahn, F.: Capability Search and Redeem across Digital Ecosystems. Journal of Information Technology, 28(3):183–197, 2013.

[SM22] Stein, Hannah; Maaß, Wolfgang: Requirements for Data Valuation Methods. In: 55th Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences (HICSS-2022), January 4-7, Hawaii, Hawaii, United States. Springer, 2022.

[ST17] Short, James E.; Todd, Steve: , What's Your Data Worth? — sloanreview.mit.edu. https://sloanreview.mit.edu/article/whats-your-data-worth/, 2017. [Accessed 09-Nov-2021].

[SWL20] Shi, Q.; Wang, H.; Lu, H.: Research and Application of AHP-EWM-Based Comprehensive Evaluation of Data Quality. In: Proceedings of the 4th International Conference on Computer Science and Application Engineering. CSAE 2020, Association for Computing Machinery, New York, NY, USA, 2020.

[Ta16] Taleb, Ikbal; Kassabi, Hadeel T. El; Serhani, Mohamed Adel; Dssouli, Rachida; Bouhaddioui, Chafik: Big Data Quality: A Quality Dimensions Evaluation. 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 759–765, 2016.

[Un22] United Nations Development Fund: . https://hdr.undp.org/data-center/human-development-index#/indicies/HDI, 2022. [Accessed 2022-07-23].

[Wi12] Wirtz, Harald: Valuation of Intellectual Property: A Review of Approaches and Methods. International Journal of Biometrics, 7:40, 2012.

[WML01] Wang, Richard Y.; Mostapha, Ziad; Lee, Yang W.: Data quality. Kluwer Academic Publishers, Boston, 2001.

[WS96] Wang, Richard Y.; Strong, Diane M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4):5–33, 1996.

[WS22] Woroch, Robert; Strobel, Gero: Show me the Money: How to monetize data in data-driven business models? 02 2022.

Appendix

Origin	Data Type	Use Case	DQD	Strength	Weakness
[Ir21]	Agriculture	DQI for ranking sheep and evaluating their monetary value based on breeding data and its accuracy.	Accuracy, Timeliness, Completeness	Gives an example for linking data quality to monetary values.	DQD are not separated clearly
[HK15]	Type-independent	DQI for assessing cases in a case-based reasoning (CBR) system	Average solutions per case[3]; Count of similar retained queries[4]; Missing Values[5]	Importance & usefulness of DQI for AI paradigms; Specific DQ criteria for CBR mapped to general DQD.	Same DQD are measured by different criteria
[Ka17]	Trading	DQI to assess consistency in regards to bilateral trade data reporting between countries.	Data availability, Import/ export size ratio and share	Integrate different sources for the same information to detect inconsistencies in data sets.	Focus on "consistency" dimension and provides less insights into other DQD.
[Hi]	Healthcare	DQI for Irish healthcare system records	Correctness of entity specific attributes	Proof that DQI implementation leads to a system wide increase in data quality	Rules merely assess the record for missing values; Unclear if and how other DQD are assessed.

Tab. 4: Overview of the existing DQIs from literature and practice - part 2

[3] reflects the following DQD:appropriate amount of data, concise representation, ease of manipulation, free-of-error, interpretability, timeliness, understandability, value added

[4] reflects the following DQD:appropriate amount of data, free-of-error, interpretability, timeliness, value added

[5] reflects the following DQD:appropriate amount of data, concise representation, completeness, free-of-error, interpretability, relevancy, understandability, value added

Origin	Data Type	Use Case	DQD	Strength	Weakness
[Ir21]	Agriculture	DQI for ranking sheep and evaluating their monetary value based on breeding data.	Accuracy, Timeliness, Completeness	Gives an example for linking data quality to monetary values.	DQD are not separated clearly
[PA22]	Utility Industry	DQI for monitoring and improving DQ. Uses business rules on data assets linked to certain DQD which are weighted and aggregated into a DQI.	Accuracy, Completeness, Consistency, Uniqueness, Validity, Integrity, Timeliness	Provide clear structure for DQI using business rules and connecting them to DQD.	Lacks basics for metrics and their application.
[Ge18]	Process	Monitoring company processes against a desired DQI target.	Timeli-, Unique-, Correct-, Completeness, Accuracy, Consistency, Non-redundancy, Relevancy, Uniformity, Reliability, Understandability	A high number of DQD are included into the DQI; Metrics are explained comprehensively.	Missing literature on how the metrics are being derived.
[Jo18]	Finance	DQI to measure the quality of income data of 154 countries to rank them by their trustworthiness.	Base year, System of national accounts, The informal Economy, Quality of statistics, Corruption	Example of how DQI can be used to establish trust when data is self-reported.	DQI is based on other indices which increases complexity and decreases transparency.
[In21]	Government	DQ Framework based on a DQI to create a DQ standard for transparent data sharing.	Timeliness, Comprehensiveness, Forward-Looking	Continuous feedback rounds for DQI improvement.	Calculation of the DQD values lacks comprehensibility due to domain-specific key figures.

Tab. 5: Overview of the existing DQIs from literature and practice - part 3