Unobtrusive Integration of Data Quality in Interactive Explorative Data Analysis

Michael Behringer[a], Pascal Hirmer[b], Alejandro Villanueva[c], Jannis Rapp and Bernhard Mitschang[d]

Institute of Parallel and Distributed Systems, University of Stuttgart, Universitätsstr. 38, D-70569 Stuttgart, Germany
{firstname.lastname}@ipvs.uni-stuttgart.de

Keywords: Data Quality, Explorative Data Analysis, Human-in-the-Loop, Data Mashups.

Abstract: The volume of data to be analyzed has increased tremendously in recent years. To extract knowledge from this data, domain experts gain new insights using graphical analysis tools for explorative analyses. Hereby, the reliability and trustworthiness of an explorative analysis are determined by the quality of the underlying data. Existing approaches require a manual inspection to ensure data quality. This inspection is frequently neglected, partly because domain experts often lack the necessary technical knowledge. Moreover, they might need many different tools for this purpose. In this paper, we present a novel interactive approach to integrate data quality into explorative data analysis in an unobtrusive manner. Our approach efficiently combines the strength of different experts, which is currently not supported by state-of-the-art tools, thereby allowing domain-specific adaptation. We implemented a fully working prototype to demonstrate the ability of our approach to support domain experts in explorative data analysis.

1 INTRODUCTION

Nowadays, more data are generated than ever before in history (Reinsel et al., 2018), and data are the foundation of almost all business processes and strategic decisions (Grover and Kar, 2017). Oftentimes, at the beginning of the analysis, the exact methodology is still unclear. One speaks of an explorative analysis in which it must first be decided which data sources to use, which data cleaning steps to conduct, and so on (Polyzotis et al., 2018). Here, data analysis processes, such as the KDD process (Fayyad et al., 1996) or CRISP-DM (Shearer, 2000), provide guidance on how to proceed with the analysis and are structured in a highly iterative manner, i.e., with a high number of feedback loops to incorporate new findings and continuously improve the analysis.

In many cases, however, the exploratory analysis is not performed by Data Scientists with in-depth technical knowledge but by domain experts (Behringer et al., 2017). For this purpose, there are, for instance, the approaches Self-Service Business Intelligence (Alpar and Schulz, 2016) or Visual Analytics (Thomas and Cook, 2005). However, the former follows predefined analysis paths while the latter only solves specific challenges (Keim et al., 2010; Stodder, 2015). To provide more freedom for domain experts, graphical data analysis tools are often used. With the help of these tools, data sources are graphically connected with operators (e.g., conducting data preprocessing) in an intuitive way, thus, specifying the analysis workflow. However, a significant challenge here is to assess the underlying data quality since the validity of the analysis results cannot be guaranteed if the data quality is insufficient. In common data analysis tools, the data quality can be evaluated, but this is not intuitive, and not achievable without manual effort.

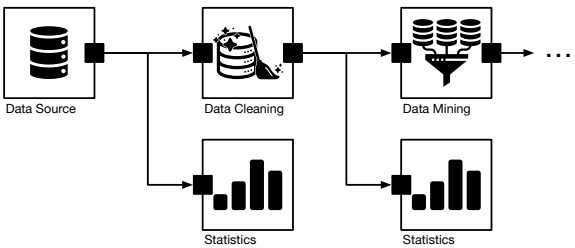

Figure 1: Manual evaluation of data quality in state-of-the-art approaches.

[a] https://orcid.org/0000-0002-0410-5307
[b] https://orcid.org/0000-0002-2656-0095
[c] https://orcid.org/0000-0002-9311-5573
[d] https://orcid.org/0000-0003-0809-9159

Behringer, M., Hirmer, P., Villanueva, A., Rapp, J. and Mitschang, B.
Unobtrusive Integration of Data Quality in Interactive Explorative Data Analysis.
DOI: 10.5220/0011998100003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 2*, pages 276-285
ISBN: 978-989-758-648-4; ISSN: 2184-4992

Figure 1 shows such an analysis workflow enriched by data quality inspection. As can be seen, an additional operator "Statistics" has to be added in each step of the analysis to evaluate the data quality.

Two major problems arise with this approach: (1) the statistics displayed here are not customized to the data, but generic and, thus, interpretation requires appropriate knowledge and (2) a domain expert will tend to be convinced of their own analysis, so this review is rather neglected.

Accordingly, it is essential that data quality is constantly monitored without explicit attention from the domain expert and that, in case of critical low data quality, the domain expert is made aware of quality issues and is able to react to these.

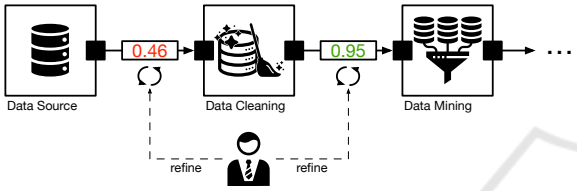

Figure 2: Possible integration of data quality indicators in graphical data analysis tools.

A possible approach is illustrated in Figure 2, showing how data quality can be integrated in an unobtrusive manner. The main contributions of this paper are: (i) a set of requirements for integrating data quality into interactive exploratory analysis, (ii) the results of a comprehensive literature review concerning the implementations of these requirements in different tools available on the market, (iii) our novel approach for unobtrusive integration of data quality into graphical data analysis tools and how different experts can contribute their respective strengths in this context, and (iv) a prototypical implementation of our approach to illustrate how such a tool could look like for domain experts.

The remainder of this paper is structured as follows: In Section 2, we define requirements that are necessary for the appropriate integration of data quality into the interactive analysis process. Next, we evaluate related data analysis tools according to our requirements in Section 3. In Section 4, we introduce our novel approach for unobtrusive integration of data quality into interactive data analysis. Then, we show a brief overview of our implemented prototype in Section 5 and discuss how this prototype fulfills the requirements in Section 6. We present related work in Section 7. Finally, we conclude this work and show future work in Section 8.

2 REQUIREMENTS

In this section, we define several requirements that are necessary to enable domain experts to be aware of data quality at all times during exploratory analysis and react to it when needed. In general, there are 5 requirements for user-centric data analysis processes that should be fulfilled for an optimal involvement of a domain expert in the analysis (Behringer et al., 2017). In this paper, we adapt these requirements to the specific circumstances related to data quality:

(R1) Integration into Entire Data Analysis Process. For a successful analysis, a domain expert must be informed about the data quality at all times. Furthermore, it is indispensable for assessing the performed analysis steps to show their impact on the data quality. This is the prerequisite to understand the quality of the underlying data, to improve it if necessary, and finally to evaluate the reliability of the analysis.

(R2) Feedback at Different Levels of Detail. An important criterion for the involvement of domain experts with regard to interactive data analysis is to avoid information overload. Thus, information about the current data quality has to be presented according to the respective context, i.e., less detailed information considering the entire analysis process and more details on single analysis steps or on request. These levels of detail should be separated into different data quality dimensions, e.g., completeness or timeliness.

(R3) Involvement of Several User Roles. It cannot be assumed that a domain expert has comprehensive knowledge of all available data sources. Thus, metadata related to data quality must already be pre-annotated, e.g., by domain experts working within the respective domain or technical experts residing in the IT department. Hence, several user roles can contribute the respective strengths to the analysis process and support self-service analysis, leading to a clear separation of concerns.

(R4) Automated Background Monitoring. Data quality monitoring should require as less attention from domain experts as possible. Instead, it should take place primarily in the background. Then, if the data quality decreases below a threshold, a domain expert can be alerted, preventing the need to check data quality and ensure appropriate surveillance.

(R5) Assisted Solving of Identified Issues. If the data quality is insufficient, the reasons behind this have to be communicated to the domain expert in a comprehensible manner. Therefore, it is tremendously important that suggestions are made to overcome these deficiencies and to support the domain expert in this analysis process.

Table 1: Coverage of the identified requirements in various tools.

Functionality \ Phase	Data Wrangling	Data Analysis	Hybrid
R1: Integration into the entire data analysis process	◔	◔	◕
R2: Feedback at different levels of detail	●	◔	◔
R3: Involvement of several user roles	◕	◔	◔
R4: Automated background monitoring	◔	◔	◔
R5: Assisted solving of identified issues	◕	◔	◑

◔ None or poor support [0%,25%] ◑ Medium support (25%,50%] ◕ Good support (50%,75%] ● High support (75%,100%]

3 EVALUATION OF RELATED TOOLS AND FOUNDATIONS

In this section, we evaluate related tools according to our requirements and we introduce foundations.

3.1 Evaluation of Related Tools

In order to emphasize the importance of our approach, we analyzed and compared established tools in regard to their consideration of data quality aspects. With regard to the requirements introduced, we evaluated the leading tools based on the respective Magic Quadrants of Gartner in the area of data wrangling/data quality (Gartner Inc., 2021b) and data analysis (Gartner Inc., 2021a). We also added tools that support both data wrangling and data analysis. These tools comprise Rapid Miner, KNIME as well as Tableau[1] including Data Prep and cover all steps of the data analysis process. We evaluated these tools with respect to our introduced requirements. An overview of this comparison is shown in Table 1.

Regarding our first requirement R1, the integration into the entire data analysis process, we conclude that data wrangling and data analysis tools offer very limited support while hybrid tools offer good support. However, in hybrid approaches, data quality needs to be manually integrated into the data analysis processing. In each step of the analysis workflow, domain experts need to define which data quality metric should be evaluated and insert specific steps to do so. In more complex processes, however, this can become a very tedious task. Hence, a full integration without a large amount of user interaction would be desirable.

The second requirement, the feedback at different levels of detail, is supported very well by data wrangling tools, while data analysis and hybrid tools offer nearly no support. In data analysis tools, user feedback regarding data quality is given in a very limited scope, for instance, only on column level and not in separate dimensions. In hybrid tools, it is necessary to insert manual steps into the process which can provide some kind of feedback regarding data quality. However, this also means that feedback is only contained in these inserted steps and not integrated into the overall user interface of the tools.

Regarding the third requirement, the involvement of different user groups, we did not find any higher support in the analyzed data analysis and hybrid tools. With regard to data wrangling tools, it should be noted that these are usually utilized by a different expert than for the subsequent data analysis. This in turn means that although different user roles work together as part of the entire data analysis process, the data quality is predetermined in this case, irrespective of the analysis. Hence, there is a gap regarding R3.

The fourth requirement specifies that a calculation in the background is desirable, without interrupting domain experts while specifying data analysis processes. The analyzed tools do not offer sufficient support for this. While data wrangling tools have another focus, data analysis tools usually do not provide integrated solutions or lack of required metadata to calculate data quality. Hybrid approaches require inserting data quality assessment steps which need to be actively triggered. Hence, there is no sufficient background calculation possible in these tools.

Finally, regarding requirement five, the interactive correction, a good support was provided by data wrangling tools. Here, interaction is possible, however, this is solely focused on the data level and not on contextual level regarding the entire data analysis process. Considering meta data as well would be beneficial. For data analysis and hybrid tools, it is observed that most tools offer some functionality to correct data quality. Regarding analysis tools, some offer limited functionality to interactively remove duplicates or fill in missing values, however, here the tools differ greatly in their powerfulness. In hybrid tools, this interaction has to be manually configured

[1]Tableau: https://tableau.com

by adding an additional step to the data analysis process which requires deeper knowledge.

3.2 Foundations

As our paper aims to measure the underlying data quality, we need to specify metrics that allow assessing whether data quality can be considered as low or high. In the following, assembled from a literature review, metrics are introduced to measure established data quality dimensions.

3.2.1 Accuracy/Correctness

Calculating how syntactically or semantically correct data is, is a very difficult task. We either require correct reference data or plausibility rules for this (Loshin, 2010), e.g., a birth date is not in the future. Hence, the following approaches assume that incorrect data can be identified:

Ratio of incorrect values The easiest way to calculate correctness is by comparing the amount of correct values to the total amount of values (Azeroual et al., 2018; Serhani et al., 2016; Juddoo, 2015).

Distance function A more sophisticated way to calculate correctness is using distance functions, which take into consideration the degree how correct or incorrect data is by calculating the similarity of data. To do so, a distance function is used depending on the type of data. For string values, the Levenshtein distance (Levenshtein, 1966) or Hamming distance (Hamming, 1950) could be used.

3.2.2 Consistency/Integrity

This dimension specifies how consistent the set of data is, i.e., whether a data set has contradictions within itself. So-called consistency rules are used to define what is considered as consistent. An example of a consistency rule is that a zip code has to match the respective city (Azeroual et al., 2018; Batini and Scannapieco, 2016). We can use the following metrics to measure consistency:

Ratio of consistent and inconsistent data: The easiest way to calculate this metric is once again comparing the amount of inconsistent data entries to the total amount of data entries checked for consistency (Lee et al., 2006).

Weighted sum: A more accurate approach is using a weighted sum, which considers the consequences of violating or fulfilling consistency rules. This approach is described in detail by (Alpar and Winkelsträter, 2014; Hipp et al., 2007).

3.2.3 Completeness

The dimension completeness calculates how complete the data is or, in other words, if data is missing. Hereby, one must decide if only values within the dataset (closed-world assumption) or missing but correct values which are not contained in the dataset should be considered (open-world assumption) (Batini and Scannapieco, 2016). For instance, based on this decision, a dataset with 30 states of the U.S. could be seen either as complete or incomplete if we consider the remaining 20 states as well. We can calculate this dimension by the following metrics:

Missing value ratio Here, the ratio of present data to the missing data is calculated to measure completeness (Batini and Scannapieco, 2016; Scannapieco et al., 2005).

NULL tuple ratio Instead of focusing on features, another approach is to compare the ratio of tuples containing at least one NULL value with tuples containing no NULL values (Blake and Mangiameli, 2009). Note that multiple NULL values within a single tuple are not considered.

3.2.4 Timeliness

The dimension timeliness measures the probability that data being processed at a certain point in time still reflects the reality and, hence, is not outdated. This dimension may change over time as it is strongly dependent to the use case (Wang and Strong, 1996).

Probabilistic approach In the probabilistic approach, we assumed that timeliness decreases exponentially. To calculate this dimension, it is necessary to define how fast the timeliness of data decreases in a specified amount of time, e.g., the timeliness dimension decreases by 10% each month (Heinrich and Klier, 2011).

Time-limited approach Another means to measure this dimension is the time-limited approach. Here, it is assumed that data can be considered invalid at a fixed point in time. This approach then calculates the decay of data quality in regard to timeliness from data creation to the point in time of invalidity (Ballou et al., 1998).

Hybrid Approach: Finally, in the hybrid approach, timeliness is calculated using the time-limited or probabilistic approach depending on the current use case. The probabilistic approach is used if it is not possible to define a fixed time limit, i.e., it cannot be foreseen when data become invalid. Otherwise, the time-limited approach is used (Even and Shankaranarayanan, 2005).

4 UNOBTRUSIVE INTEGRATION OF DATA QUALITY IN INTERACTIVE DATA ANALYSIS

In this section, we present our novel approach to overcome the aforementioned limitations of existing approaches and, thus, to enable domain experts to easily maintain an overview of data quality without requiring additional effort while conducting an explorative analysis. To achieve this, (a) different user roles need to collaborate and contribute their respective strengths, (b) generic data quality metrics need to be defined, (c) depending on the analysis context, the data quality metrics have to be adapted, and (d) issues in data quality need to be identified and solutions have to be recommended.

Figure 3 shows an overview of our approach. As described above, a popular approach to involve domain experts in exploratory analysis is the use of graphical data analysis tools. These provide a range of data sources and allow domain experts without deeper technical knowledge to combine these data sources and specify transformations in a graphical manner. To keep the necessary effort for domain experts at a minimum, these data sources should be provided in advance by an IT expert. In our approach, this is done in the preparation phase (Figure 3, 1). Here, an IT expert first adds a new data source to the repository (Figure 3, 1a), e.g., by specifying a connection to the respective database. In a subsequent step, the IT expert has to create a domain-agnostic ground truth for this data source (Figure 3, 1b). This ground truth comprises various data quality metrics that must hold in order to consider the data as qualitative. For instance, if we consider the data quality dimension completeness under the closed-world assumption for one data feature, the IT expert has to specify which values have to be included in the data, e.g., the 50 states of the USA are defined and each differing value or missing value would decrease the quality. When this ground truth is specified for each data feature, the task of the IT expert is done and the ground truth is stored as a data quality artifact in the repository. At this point, the IT expert's task ends for the time being.

This phase is decoupled from the explorative analysis of domain experts to facilitate flexibility without the need to reach out to an IT expert. To support domain experts in their exploratory analysis with regard to data quality, we describe our approach based on graphical data analysis tools. For a domain expert, the first phase is the specification phase (Figure 3, 2), in which the analysis workflow is created.

First, the required data sources are selected (Figure 3, 2a). Since a data quality artifact is provided in the repository for all available data sources in this phase, the data quality can be determined by means of this artifact (Figure 3, 2b). These calculated data quality metrics are then displayed to the domain expert and allow for a direct assessment (Figure 3, 3a) whether this data source is qualitatively sufficient or, if it is not, where possible problems may be located, e.g., if there are data completeness concerns. Subsequently, it can be decided to either change the data source(s) (Figure 3, 2a) or to proceed with the specification of the analysis workflow (Figure 3, 2c), e.g., preprocessing or data mining transformations.

If the latter is chosen, the workflow is being executed (Figure 3, 2d), which is necessary because these transformations affect the data and, thus, influence the data quality. In both cases, the data quality is calculated again based on the data quality artifacts (Figure 3, 2b) and displayed for review (Figure 3, 3a). Up to this point, all data quality metrics are calculated based on the data quality artifact defined by the IT expert in the preparation phase and are, therefore, independent of the context of the domain expert's analysis. Although this is a significant advance over state-of-the-art approaches without automatic data quality monitoring, it is still insufficient in many cases. For instance, it is possible that the data timeliness dimension has been defined in advance in such a way that the data must be up-to-date on a daily basis, but for the current analysis, historical data is required.

In this scenario, the definition of data quality by means of the data quality artifact is no longer suitable and has to be adapted to the intended analysis. This is supported by our approach in the adaptation phase (Figure 3, 4), where the domain-specific quality metrics are defined (Figure 3, 4a). This can be a wide variety of different measures, which is why our architecture is generic and extensible in this respect. Possible adjustments include, for instance, adding or removing quality dimensions, adjusting thresholds, or even weighting the different dimensions according to their importance. With each adjustment, the now domain-specific data quality is immediately calculated (Figure 3, 4b) and again visualized for assessment by the domain expert (Figure 3, 3b).

By this stage of the process, data quality metrics have been predefined by an IT expert and/or adapted to the context by a domain expert. However, if the calculated data quality is still insufficient, or if a more reliable analysis should be performed, the domain expert is required to focus on the data itself and use the improvement phase (Figure 3, 5) to enrich or clean the data until a sufficient data quality is achieved.

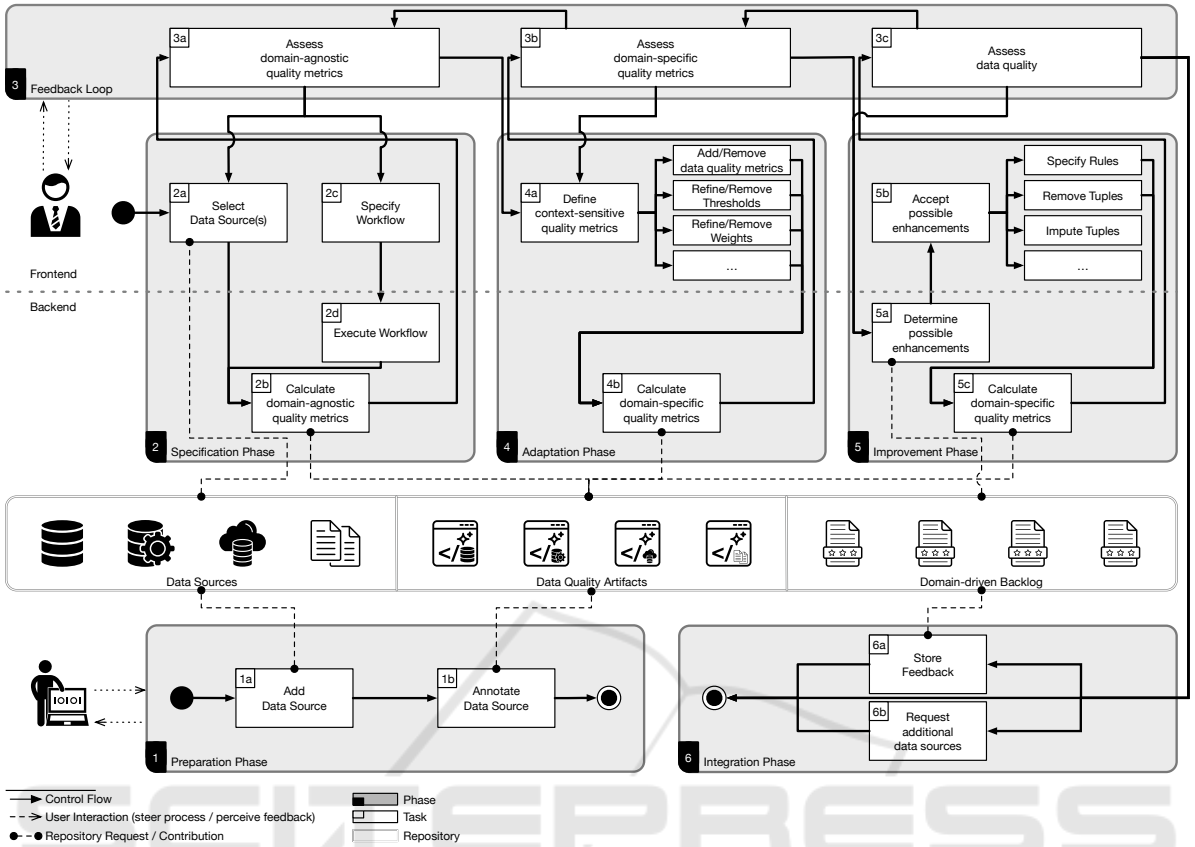

Figure 3: Different phases to unobtrusively integrate data quality in explorative analysis.

Therefore, our approach uses data quality artifacts to automatically identify common data quality problems (Figure 3, 5a) and indicate these problems to the domain expert. At the same time, possible solutions are suggested, such as how to deal with null values or duplicates. A domain expert can now accept these suggestions (Figure 3, 5b) or take more in-depth actions, e.g., specify further rules for data imputation.

Next, there is a further loop consisting of the (re)calculation of data quality metrics (Figure 3, 5c) and the follow-up assessment of whether the data quality is now sufficient (Figure 3, 3c). However, this process is not straightforward. Instead, it can require several iterations depending on the complexity and suitability of the data. Consequently, a domain expert must be able to switch between the phases, e.g., if it turns out during the improvement phase (Figure 3, 5) that the calculation of the domain-specific data quality metrics is still insufficient and the metric calculation needs to be refined. For this reason, our approach includes a feedback loop (Figure 3, 3) to enable the domain expert to leverage knowledge from completed phases and use it to make adjustments in previous phases.

Once these phases are completed, the domain expert's analysis is finished and new insights have been gained. Due to the prior integration of an IT expert, it is expected that the effort for domain experts remains manageable. Nevertheless, it is also feasible that the data quality artifacts are no longer up-to-date, e.g., because the focus of the necessary analyses has shifted since the data quality artifact has been created or the data itself has changed over time. For this reason, our approach comprises one more phase, the Integration Phase (Figure 3, 6), in which feedback is provided back to the IT department (Figure 3, 6a), e.g., domain-specific adjustments made in the adaptation phase (Figure 3, 4) or data cleaning transformations performed in the improvement phase (Figure 3, 5). An IT expert may later adapt and refine the initial data quality artifacts or perform additional data cleaning in advance, provided that this makes sense for the majority of the analyses. Furthermore, additional data sources can be requested if no suitable data sources are available in the repository (Figure 3, 6b). Either way, the domain expert's tasks end at this point, and the possible additional activity of an IT expert is again decoupled in terms of time.

Figure 4: Detail view in the user interface of our approach.

5 PROTOTYPE

We have implemented the presented concepts in a fully functional prototype to provide more insight into what such a system would look like for domain experts This was accomplished by first extending a data mashup tool to provide an overview of the overall data quality during the workflow specification (cf. Figure 2). By clicking on one of the quality indicators for the overall data quality, a small overlay is displayed, which lists the respective data quality dimensions separately and offers the possibility to display more details to the domain expert.

If the domain expert decides to proceed with more details, the user interface depicted in Figure 4 will be opened. In the upper part of this user interface, the overall data quality achieved is shown first of all and is color-coded (Figure 4, a). Thus, it can always be seen at first glance whether the desired data quality has already been achieved. Next to this, this overall data quality is divided into its dimensions, for the data source in this figure, thus, the accuracy, the completeness, the consistency, and the timeliness (Figure 4, b). Now, a domain expert can quickly identify which dimensions are critical at the moment. According to the presented approach, these dimensions are either the domain-agnostic data quality metrics defined by the IT expert or the domain-specific, refined settings of the domain expert. Furthermore, within the top section of the dashboard, a small sample of the dataset is visualized in order to give the domain expert a better feeling about the data characteristics (Figure 4, c). In many situations, however, an assessment at the data source level is not sufficient. For instance, it is not yet possible to tell whether the value of the timeliness dimension is caused by individual features with very low data quality or whether all features have medium data quality instead.

For this reason, the data quality is also calculated at feature level (Figure 4, d) and visualized in color-coded form. Here, once again, it is possible to see at a glance which features exhibit critical data quality, both overall and again subdivided into the individual data quality dimensions (Figure 4, e).

Below this overview, the most common values are indicated for each feature as value distribution (Figure 4, f), as well as additional statistical key indicators (Figure 4, g), e.g., the number of null or unique values, as well as minimum, maximum, or the average. The respective indicators depend on the data type of the feature, i.e., the average is not calculated for categorical data.

In order to increase the data quality, identified issues and potential countermeasures are suggested below these metrics (Figure 4, h). For instance, for the feature date, one finding is that many records are older than the specified policy. In this case, our prototype suggests, among other things, that this outdated data should be removed from the dataset resulting in a higher quality dataset according to the requirement. Here, it is up to the domain expert to judge whether this would corrupt the analysis results. An alternative would be the second suggested action that the policy for the timeliness dimension is adjusted, i.e., older data is also considered as qualitative. For the same feature, an additional problem with the data in this figure is that the date format does not match the specified format. For this, a transformation to the correct date format is suggested, which the domain expert can trigger with just a single click.

A further example can be seen when considering the feature RPM (rotations per minute). Here, either the IT expert or the domain expert has defined a rule according to which the RPMs for a specific machine type must be within a given range. If individual records violate this rule, this is either an indication of underlying issues with the machines or anomalies in the data, which should, therefore, not be considered for the current analysis. As a possible solution in this scenario, our prototype suggests that either these entries are discarded or the corresponding rule is adapted. Any of these adjustments made is immediately applied to the data in the background. The user interface is updated after the calculation has been completed.

Alternatively, the detail view can be closed at any time to proceed with the analysis. Here, the system returns to the workflow specification overview (cf. Figure 2) and the displayed overall data quality value is temporarily replaced with an unobtrusive warning until the calculation of the adjustments has been completed. To enable a wide range of possible adjustments, our prototype is built based on a generic architecture and allows to be extended with additional functionalities to increase data quality in an explorative analysis.

6 DISCUSSION

In order to evaluate the introduced approach to integrate data quality assessment unobtrusively into the data analysis process, we use the five requirements defined earlier in Sect. 2.

(R1) Integration into the Entire Data Analysis Process. The first requirement specified that it is essential to integrate data quality assessment into the entire data analysis process, i.e., a domain expert must be informed about the data quality at all times. By adding a data quality assessment through calculation of domain-agnostic metrics into each step of the data analysis process and by allowing to adjust process steps through feedback loops, we can fulfill the first requirement R1. In our approach, domain experts are always informed about possible data quality issues.

(R2) Feedback at Different Levels of Detail. The second requirement was to avoid information overload of domain experts. Thus, showing less detailed information considering the entire analysis process and more details on individual analysis steps or by specific user request. In our approach, we realized this by providing direct user feedback first in an aggregated overview and then allowing domain experts to request more details on demand in each step. Thus, we follow the popular Information Seeking Mantra (Shneiderman, 1996).

(R3) Involvement of Several User Roles. The third requirement was the consideration of multiple user roles. While domain experts have knowledge about the specific goals of the analysis process and know the meaning of the data very well, they usually lack technical knowledge and IT expertise, e.g., to integrate new data sources into the process. Thus, we clearly extinguish the roles of domain experts and IT experts and we tried to keep the communication overhead between these roles as minimal as possible since the steps of these user roles can be clearly separated, i.e., a clear separation of concerns.

(R4) Automated Background Monitoring. The fourth requirement was that the data quality assessment should be done in the background without being obtrusive for domain experts in workflow specification. As can be seen in Figure 3, our approach fulfills this requirement by calculating data quality metrics fully in the background without blocking workflow specification for domain experts. Once the assessment is finished, the results are shown to the domain experts in the frontend.

(R5) Assisted Solving of Identified Issues. Finally, requirement five was that users should be assisted in improving data quality by applying different measures. This is realized in our approach in two ways: (a) by identifying typical data quality issues when

they arise and proposing possible solutions that can be accepted with one mouse click by a domain expert, and (b) by showing the impact of operations, e.g., data preprocessing tasks, on the resulting data quality for more complex data quality issues.

In conclusion, our approach fulfills the specified requirements. In contrast to existing approaches, the resilience of the data analyses can be predicted better since the domain expert is informed about the data quality in the process at all times.

7 RELATED WORK

Many different process models have been published to describe the methodology for data analysis. The two most popular representatives are the KDD process (Fayyad et al., 1996) and CRISP-DM (Shearer, 2000), which schematically represent the various phases required to transform raw data into knowledge in a structured way. The implementation of these processes is generally performed by technical experts and based on domain-specific knowledge of domain experts. As a consequence, however, these predefined analyses tend to become a black-box and changes can rarely be made independently by a domain expert (Behringer et al., 2017). Therefore, specific domain-knowledge cannot be considered during the analysis process and it is the way to go for well-understood analyses.

Approaches to integrate domain experts are: (i) Visual Analytics (Thomas and Cook, 2005), which extends common visualization through repetitive analysis steps, and (ii) Self-Service Business Intelligence (Alpar and Schulz, 2016), which is designed to enable self-directed analysis. Thereby, Visual Analytics is mainly focused on solving a specific problem (Keim et al., 2010), while Self-Service Business Intelligence offers a generic analysis approach, but follows predefined analysis paths (Stodder, 2015), e.g., choosing the features to use, and lack of advanced data mining tasks.

Another approach is based on graphical data analysis tools (Daniel and Matera, 2014), e.g., KNIME[2] or RapidMiner[3]. Here, a domain expert specifies an analysis workflow step-by-step starting with the selection of data sources, continuing with data preparation, and finally data mining and reporting.

An important issue when it comes to data processing and analysis is data quality. Data quality is oftentimes reduced to accuracy of data, i.e., typos or incorrect data (Firmani et al., 2016). However, data quality should be considered in numerous dimensions. Yet, this understanding is not uniform as a literature review shows, e.g., the recommended dimensions differ between standards (e.g., DIN EN ISO 9001:2015), practice (Askham et al., 2013) and scientific literature (Azeroual et al., 2018; Batini and Scannapieco, 2016; Firmani et al., 2016; Wang and Strong, 1996). Nevertheless, there is a certain consensus with regard to dimensions that occur more frequently. These include, in particular, accuracy/correctness, completeness and consistency (Askham et al., 2013; Azeroual et al., 2018; Batini and Scannapieco, 2016; Firmani et al., 2016; Wang and Strong, 1996), timeliness (Askham et al., 2013; Azeroual et al., 2018; Wang and Strong, 1996) as well as trustworthiness/credibility/reputation (Batini and Scannapieco, 2016; Firmani et al., 2016; Wang and Strong, 1996). In many cases these dimensions are summarized under generic terms.

According to Wang and Strong (Wang and Strong, 1996) these are: (i) intrinsic quality, which by definition is present in the data, e.g., accuracy, objectivity or trustworthiness of the origin, and (ii) data quality which depends on the task at hand and is therefore contextual, e.g., timeliness, relevancy or completeness. Furthermore, data quality can also be evaluated with regard to availability and security (accessibility data quality) or in terms of interpretability (representational data quality). In particular, for interactive data analysis, the first two categories, i.e., intrinsic and contextual data quality, are to be considered.

8 SUMMARY AND CONCLUSION

In this paper, we present a new approach to assess data quality during explorative analysis in an unobtrusive interactive manner. In a first step, we conducted a comprehensive literature review to identify shortcomings of existing tools in regard to detecting data quality issues in the analysis process. By doing so, we compared different data wrangling, data analysis, and hybrid tools based on a set of requirements. After identifying the shortcoming of the existing tools, we introduced an approach that shows how data quality can be assessed during the entire life cycle of a data analysis process while keeping the domain expert in the loop. The goal is to integrate domain-agnostic and optional, domain-specific data quality metrics into each step of the data analysis process.

This approach helps by considering data quality early, i.e., during the specification of the analysis workflow, which enables data quality by design.

[2]KNIME: https://knime.com

[3]RapidMiner: https://rapidminer.com

Hence, a domain expert recognizes timely if the quality of data sources is high enough or if the data sources need to be extended or replaced. Detecting such issues in an early phase of process creation reduces costs and efforts of the entire process. Furthermore, the process contains several feedback loops in each step, which allows returning to an earlier step in the life cycle in case a data quality issue is detected. This could, e.g., lead to an adaptation or enrichment of the used data sources.

Overall, our introduced approach offers a continuous and unobtrusive integration of data quality assessment in the entire life cycle of a data analysis process, easily understandable by domain experts. We evaluated our results through the identified requirements and a prototypical implementation showing its applicability. In future work, we plan to conduct extensive user studies to evaluate our concept and prototype.

REFERENCES

Alpar, P. and Schulz, M. (2016). Self-Service Business Intelligence. *Business & Information Systems Engineering*, 58(2):151–155.

Alpar, P. and Winkelsträter, S. (2014). Assessment of data quality in accounting data with association rules. *Expert Systems with Applications*, 41(5):2259–2268.

Askham, N. et al. (2013). The Six Primary Dimensions for Data Quality Assessment – Defining Data Quality Dimensions. Technical report.

Azeroual, O. et al. (2018). Data measurement in research information systems: metrics for the evaluation of data quality. *Scientometrics*, 115(3):1271–1290.

Ballou, D. et al. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4):462–484.

Batini, C. and Scannapieco, M. (2016). *Data and Information Quality*. Dimensions, Principles and Techniques. Springer, Cham.

Behringer, M. et al. (2017). Towards Interactive Data Processing and Analytics - Putting the Human in the Center of the Loop. In *Proceedings of ICEIS 2017*. SCITEPRESS - Science and Technology Publications.

Blake, R. and Mangiameli, P. (2009). Evaluating the Semantic and Representational Consistency of Interconnected Structured and Unstructured Data. In *Proceedings of AMCIS 2009*.

Daniel, F. and Matera, M. (2014). *Mashups – Concepts, Models and Architectures*. Springer.

Even, A. and Shankaranarayanan, G. (2005). Value-Driven Data Quality Assessment. In *Proc. of ICIQ 2005*.

Fayyad, U. et al. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39(11):27–34.

Firmani, D. et al. (2016). On the Meaningfulness of "Big Data Quality" (Invited Paper). *Data Sci. Eng.*

Gartner Inc. (2021a). Magic Quadrant for Analytics and Business Intelligence Platforms. Technical report.

Gartner Inc. (2021b). Magic Quadrant for Data Quality Solutions. Technical report.

Grover, P. and Kar, A. K. (2017). Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature. *Global Journal of Flexible Systems Management*, 18(3):203–229.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.

Heinrich, B. and Klier, M. (2011). Assessing data currency—a probabilistic approach. *Journal of Information Science*, 37(1):86–100.

Hipp, J. et al. (2007). Rule-based measurement of data quality in nominal data. In *ICIQ*, pages 364–378.

Juddoo, S. (2015). Overview of data quality challenges in the context of big data. In *Proceedings of the ICCCS 2015*, pages 1–9.

Keim, D. A. et al. (2010). Visual Analytics. In *Mastering The Information Age*, pages 7–18. Eurographics Association, Goslar.

Lee, Y. W. et al. (2006). *Journey to data quality*. The MIT Press.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Loshin, D. (2010). *The Practitioner's Guide to Data Quality Improvement*. Elsevier.

Polyzotis, N. et al. (2018). Data Lifecycle Challenges in Production Machine Learning. *ACM SIGMOD Record*, 47(2):17–28.

Reinsel, D., Gantz, J., and Rydning, J. (2018). Data Age 2025: The Digitization of the World. Technical report.

Scannapieco, M. et al. (2005). Data quality at a glance. *Datenbank-Spektrum*, 14:6–14.

Serhani, M. A. et al. (2016). An Hybrid Approach to Quality Evaluation across Big Data Value Chain. In *Proc. of Big Data Congress*, pages 418–425.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Symposium on Visual Languages*, pages 336–343. IEEE Comput. Soc. Press.

Stodder, D. (2015). Visual Analytics for Making Smarter Decisions Faster. Technical report.

Thomas, J. J. and Cook, K. A. (2005). *Illuminating the Path*. The Research and Development Agenda for Visual Analytics. National Visualization and Analytics Center.

Wang, R. Y. and Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4):5–33.