Capturing and Visualizing Provenance from Data Wrangling

Christian Bors, Theresia Gschwandtner, and Silvia Miksch

Abstract—Data quality management and assessment play a vital role for ensuring the trust in the data and its fitness-of-use for subsequent analysis. The transformation history of a data wrangling system is often insufficient for determining the usability of a dataset, lacking information how changes affected the dataset. Capturing workflow provenance along the wrangling process and combining it with descriptive information as data provenance can enable users to comprehend how these changes affected the dataset, and if they benefited data quality. We present *DQProv Explorer*, a system that captures and visualizes provenance from data wrangling operations. It features three visualization components, allowing the user to explore (1) the provenance graph of operations and the data stream, (2) the development of quality over time for a sequence of wrangling operations applied to the dataset, and (3) the distribution of issues across the entirety of the dataset to determine error patterns.

Index Terms—Data Wrangling, Data Cleansing, Data Quality, Quality Metrics, Data Provenance, Sensemaking.

1 Introduction

When analyzing data in any way, the initial step before actual analysis is preparing data and ensuring that it is of adequate quality. Data quality management has developed to be an integral part of almost any data processing workflow, to increase the reliability of analysis results. Data wrangling unites the processes of transforming data for subsequent analysis and cleansing data – ridding the data from quality issues – in order to improve the quality of a dataset. However, the outcome should still be representative of the original dataset. An open challenge in data quality management is that the steps to process a dataset into a usable state are often not documented, and hence are seldom reproducible. When using large datasets and obtaining data from different data sources, it is increasingly difficult to perform quality inspection on the raw data. Data wrangling tools produce transformation histories so users are able to reconcile performed actions. However, these are often not available outside of the system, and thus, the history of data transformations is not available when importing the data in a subsequent data analysis (usually different tools are used for data pre-processing and data analysis). Also, there is a lack of context if these wrangling operations led to the desired outcome so that issues were actually resolved.

Data provenance is captured to allow retracing how it was created, from what data it was derived, and how it was changed. This allows to retain sources of errors and allow re-tracing of previously applied operations. Especially when processing data across multiple systems, provenance enables tracing changes back to their sources. Simmhan et al. [1] described a graph structure to be adequate for storing data provenance, however provenance is mostly captured in scientific workflow applications, and rarely logged in data quality management. Storing the data states in the graph's nodes and the transformation processes in it's edges gives an explorable overview of the provenance structure. The inherent acyclical structure shows the data lineage and allows the identification of process sequences. When capturing provenance during data wrangling, actions can be annotated with contextual information, to give more semantic meaning to the wrangling operations and their impact on the data. So far, existing data wrangling tools and solutions have not embraced data provenance as proficiently as necessary to have analysts benefit from their wrangling attempts. Context information is used in data profiling to recommend data transformations (e.g., Wrangler [2], Trifacta Inc., etc.). Interactive methods for data profiling are often employed to analyze certain characteristics and dimensions of the data, like specific columns of interest, or particular data types, which can be leveraged to facilitate data wrangling. However, we argue that this context information can be used to annotate executed wrangling transformations which can aid analysts in retrospectively analyzing the history of a dataset and the applied operations.

Implementing data quality metrics in a dataset allows for detecting quality issues (cf. Section 3.1). They can serve as a measure of overall quality for a dataset. Also, Bors et al. [3] presented an approach where analysts can interactively explore metrics to assess the prevalence of certain types of errors in the data and estimate the quality of a dataset in detail. We propose that leveraging data quality metrics as data provenance can aid the user in understanding the development of the dataset's qualitative conditions. This builds confidence in the reliability of a dataset.

In this paper we illustrate that by providing an approach for exploring data and workflow provenance captured from data wrangling steps, users are able to build trust in a wrangled dataset. By logging what actions were used alongside the wrangling process (e.g., data profiling, filtering, cleansing), it should be possible to gain understanding of the transformations, and make sense of the entire process. Computing quality metrics continuously for each state of the dataset is supposed to give users the ability to quickly

• *C. Bors, T. Gschwandtner, and S. Miksch are with TU Wien, Vienna, Austria.*

Manuscript received April 8, 2019.

assess the qualitative condition of the data and determine if quality has changed throughout the wrangling process. This should enable the user to draw conclusions if the data is usable in its current state. We found that current approaches for exploring provenance are lacking the ability to annotate the data sufficiently to help users make sense of a data wrangling process. Furthermore, these approaches do not allow interactive exploration of alternative branches that were evaluated before ending up with the current transformation steps.

The contributions of our paper are:

- A model for capturing and incorporating data quality metrics as data provenance, as well as annotating data transformations and data revisions in the provenance graph.
- A visual analytics (VA) approach, called *DQProv Explorer*, that allows users to explore the wrangling provenance and associated quality information.
- A novel quality flow visualization that enables the analysis of changes in data quality over sequences of operations, as part of *DQProv Explorer*.
- A user study indicating that *DQProv Explorer* is well suited for assessing provenance data from data wrangling.

2 Related Work

Over the course of design and development, we have reviewed interdisciplinary works published in the fields of data quality and provenance with particular emphasis on VA and visualization methodologies.

In database research, data quality is a relevant topic that is carrying over into other fields. Redman motivated the importance of dealing with data quality issues in big data and *advanced analysis* [4]. Furthermore, they defined dimensions of data quality that characterized different aspects, which would describe the quality of a dataset. These can be formalized into quality metrics that signal lack of quality in a dataset. A different approach towards characterizing quality issues are various taxonomies of different types of dirty data, while Kim et al.'s [5] characterized generic sources of low data quality, Gschwandtner et al. [6] presented a taxonomy of time-oriented data, which indicates the importance of dealing with domain-specific quality issues. Bors et al. [3] presented visual-interactive methods for applying generic quality metrics to tabular datasets and allowing metrics customization to add context- and domain-specific quality checks.

Kandel et al. [7] identified and motivated research directions in data wrangling and addressed challenges in the field of data quality research. One challenge to be addressed is the lack of extracting provenance from data quality operations and wrangling workflows. Provenance can be integrated into visualization and VA approaches in different ways [8], including extracting provenance from processing workflows [9]. However, capturing provenance from preprocessing data has not been sufficiently addressed. Tan [10] described the challenge for data provenance research to provide a uniform framework for combining data provenance and workflow provenance from data transformations. How provenance is integrated in visualization and VA is tied to the analytic process itself. To assess the appropriateness of a dataset from provenance [1], one approach is utilizing transformation histories to help users with identifying changes, new revisions of a dataset, or forking [8]. In collaborative environments, insights from analytic provenance can be used to retrace actions performed and assess trustworthiness of a dataset or analysis outcome [11].

Interactive data wrangling approaches allow raw data exploration, supported by data profiling elements as low level guidance to facilitate exploration and validation, e.g., Profiler [12], Wrangler [2]. Other wrangling approaches use quality checks to detect quality issues in tabular [13] and time-oriented data for cleansing [14] and rastering purposes [15]. In tabular data, data profiling approaches employ visual encodings to retain the tabular structure, abstracting the raw data, but retaining the location characteristics to identify changes [16]. Along with these different visualization approaches, the models of provenance storage are widely different and seldomly generic to visualize provenance in a more general way.

Schulz et al. [17] classified data descriptors that differentiate types of data extraction. They considered *data space descriptors (DSD)* (gathered during analysis), e.g., data dimensionality, granularity, as useful dimensions for conveying metadata of a dataset. One of the challenges for capturing provenance from data wrangling identified by Herschel et al. [18], is to determine what information is relevant for a retrospective analysis of provenance, for example in a collaborative scenario. Current approaches fall short of giving users sufficient information on the data's condition alongside provenance captured from wrangling workflows. We propose that by providing users with descriptive information allows them to re-trace provenance, put it into context, and gain insights into the analytic process of improving data quality. Moreover, there is a lack of interactive visual methods that facilitate the exploration of this information to enable the user to assess the provenance of a dataset, and thus, its usability for a specific task.

3 Data Wrangling Provenance Model

Without knowledge of the user's domain, the particularity of the data (e.g., structural dependencies, exploitable characteristics), or the pursued task, workflow provenance from data wrangling processes is non-descriptive and can hardly be leveraged for sensemaking. Thus, it is necessary to contextualize these processes. Descriptive information of the data's quality is required to audit wrangling operations and assess if they were applied appropriately. We propose to employ measures of quality throughout each processing step to allow judgment if quality was affected throughout the wrangling process.

Figure 1 shows a generic model of provenance generation for data wrangling. The different entities incorporate different types of provenance (according to [1]). The main entities involved are the data, and correspondingly data revisions, generating data provenance, being generated by transformations, generating workflow provenance. The data can be filtered by a condition into a working dataset. We store the information on each revision, capture which filters

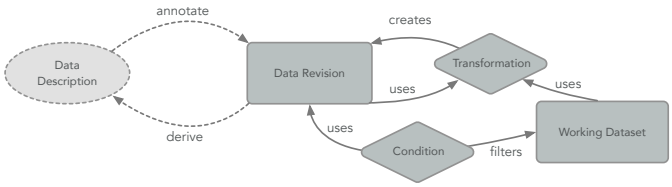

Fig. 1. Model for storing data provenance from data wrangling. The base data is stored as a data revision (i.e., revision 0). A transformation uses a data revision or a filtered working dataset to create a new data revision. Additional data descriptions are derived from every data revision and are used to annotate it subsequently.

were applied, and derive data descriptions to annotate the corresponding revision.

Data Transformations: Information on data wrangling transformations is provided in a log, with the ability to undo/redo. The transformations are stored as workflow provenance, showing the actions taken by the user. Utilizing this logging information, we can construct a provenance graph from these transformations. From each operation we derive parameters and affected rows and columns.

Applied Data Filters: Data filters are employed to process subsets of the data, this can be done to transform a specific selection. Utilizing this information can give users implications whether the analysis was only conducted on a particular subset of the data. This information is stored as row-level data provenance.

Data Descriptions: Interactive profiling of data can be employed during data wrangling to determine data characteristics of the data, e.g., data distributions, anomaly detection. The overall meta-information about the dataset and column characteristics can help to further validate or identify data rows. Descriptive statistical figures of a dataset are often used by data analysts to determine if a dataset is appropriately processed and fit for use. Leveraging these descriptive features for estimating and validating datasets, we can annotate the information extracted from the transformation and filtering operations to make them more meaningful and comprehensible to the user. The data descriptions are stored as row-level or column-level data provenance, depending on the information type.

3.1 Provenance Model Implementation

Within the context of this paper, we will base the definition and use of data quality metrics on the approach provided in Bors et al. [3]. They proposed a method for annotating data with data quality metrics to provide means for visually exploring the quality of tabular datasets. We utilize these metrics and save column- and row-level data provenance to capture contextual information and allow analysts to analyze the development of quality over time.

The implemented definition of a data quality metric is *"the quantified measure of a data quality dimension that gives proportional information about the lack of quality regarding a certain information aspect"*. For each employed metric, we measure the dirtiness of one or multiple columns with respect to a certain quality dimension. The overall measure is the inverted ratio between determined dirty tuples and the number of rows in the dataset. This yields a normalized measure between 0 and 1 for each metric, which can also be interpreted as the percentage of dirty tuples detected by the respective metric. The evaluation of a tuple is done through a validation function $vf_m(\cdot)$, returning a Boolean measure of dirtiness. However, we also retain information on the position of the dirty tuple within the dataset so they can be located. This information will be used as a data descriptor to annotate the data provenance.

The metrics utilized in our approach, for the sake of demonstration, are measures for (1) *column completeness*, (2) *validity*, and (3) *numeric plausibility*. The *column completeness* of a dataset measures the amount of missing values in a particular column, with a missing value described either as an empty entry or equal to an identifier. *Validity* is described as data type compliance to an automatically detected or manually defined data type of a particular column. *Numeric plausibility* is a metric for numeric data, which calculates a statistical distribution to detect outlying values. In addition to these pre-defined metrics, it is possible to define custom quality checks that validate the data with respect to domain-specific characteristics (e.g., numeric constraints, text validation).

The provenance model is implemented as an extension to the open source wrangling tool OpenRefine, to support the implementation of the data quality metrics [3],. OpenRefine is an open source tool that allows data wrangling of multiple types of data in a client-server-style application, with a web-frontend. The two integral extensions of the existing data quality framework are the data quality engine and the provenance model (cf. Figure 2). The data quality engine automatically recommends data quality metrics based on column type. To accomplish this we employ a heuristic validation schema that determines the predominant data type for each column. Custom quality checks can be added in the separately available MetricDoc environment [3] to detect domain-specific issues and hence improve the accuracy of the issue detection. The second feature extending the OpenRefine application is the addition of the provenance annotation model. Data quality metrics are automatically computed and annotated for every data revision and are stored in a provenance graph structure that extends the default data storage. Based on the data quality metric structure, the annotated information stored as data provenance ranges from the overall dirtiness of a particular column and metric, down to the individual indices of dirty tuples. The metrics calculation and provenance annotation is automatically computed on server-internal engines, which reduces the impact of performance during wrangling to a minimum. For a typical wrangling scenario with multiple wrangling branches, the data structure size can be fetched via http access. Since the provenance model extends the default data structure, additional data storage is minimal and only concerns workflow provenance and column- and row-wise data provenance.

4 Requirements Analysis

In Section 2 we gave an overview of research in provenance generation and data quality management, and motivated the opportunities for combining these fields. It can be seen

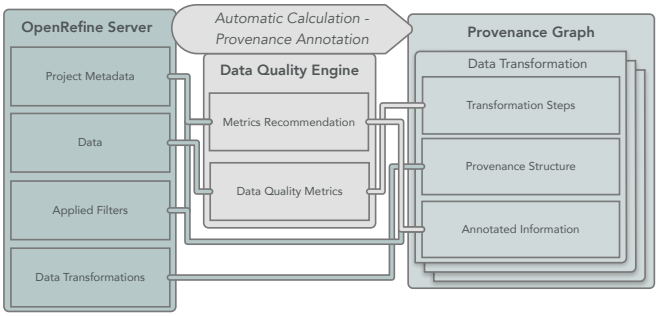

Fig. 2. Overview of our extension for capturing provenance from OpenRefine. Information is propagated from the server to the data quality engine and provenance model. For every data state and wrangling step, the extensions process information from the project to store it as provenance. The result is an annotated provenance graph that can be used for analyzing the outcome of the wrangling process.

that trends have developed towards interactively inspecting data quality based on quality metrics [3], facilitating data wrangling through recommending data transformations [2] and making the effects of such transformations easily comprehensible [12]. To determine the tasks that should be supported by a system that combines provenance and data quality analysis, we performed a requirements analysis of different taxonomies and research directions: Kandel et al. [7] motivated the development of means to (1) diagnosing data problems, (2) editing and auditing transformations, (3) using provenance to track data lineage, and (4) understanding why actions were performed. We elaborated these means further towards the purpose for analyzing provenance, according to Ragan et al. [8].

4.1 Tasks

We deem the following task considerations to be important to effectively support users with analyzing provenance from data wrangling. Prior to defining the tasks, we applied the Data-Users-Tasks design triangle by Miksch and Aigner [19] to first determine the users of our approach – data analysts, software developers, and domain experts concerned with data management –, and the data used – provenance captured during the data wrangling process. We distinguish the following tasks according to Ragan et al.'s [8] characterization of provenance purpose within the scope of assessing data quality.

$\mathbf{T}_{act}$ *Action Recovery*

The analyst wants to see the transformation sequences applied to a dataset and the quality issues retained throughout the process at the level of individual columns. This includes the types of operations, their parameter settings, and the subset of data the operations were applied on.

$\mathbf{T}_{pres}$ *Presentation*

If multiple alternative operation sequences have been created, the analyst wants to visually inspect the differences between different wrangling branches. This includes information if an operation impacted the dataset, what part of the dataset (column- or row-wise changes), and more particularly, if quality was affected. Furthermore, the analyst wants to inspect if subsets of the data exhibit more issues than others (e.g., the sensors of a weather station introduced more measurement artifacts than all others).

$\mathbf{T}_{meta}$ *Meta-Analysis*

When inspecting a sequence of operations, the analyst wants to audit the dataset if it can be trusted for further processing or analysis. To do this, the analyst monitors the development of different quality problems over time to eventually decide on the usability of a dataset. Also, the analyst wants to reconcile what operations the different branches have in common. The analyst wants to use these insights to determine how issues in the dataset were addressed and decide what operations solved these issues most appropriately for downstream analysis.

$\mathbf{T}_{rec}$ *Recall*

The analyst wants to compare the remaining issues in the dataset for two branches (at a time) in order to determine if error patterns were addressed in a similar way, or if different wrangling approaches were employed. By investigating the quality metrics of the dataset over the course of multiple operations, the analyst wants to identify if changes had qualitative impact and trace changes in quality back to the operations that caused them. This includes validation, if either the entire dataset or a particular subset of the data (that has been selected for further analysis) exhibits sufficient quality (e.g., auditing the columns of a dataset).

$\mathbf{T}_{rep}$ *Replication*

The analyst wants to be able to revert the current dataset to previous transformation steps, to either use the dataset for downstream analysis, or as a starting point for further data wrangling. If problems persist in a particular state, the analyst wants to inspect them in detail.

$\mathbf{T}_{coll}$ *Collaborative Communication*

The analyst wants to inspect a sequence of previously applied operations and, in particular, their consequences in terms of quality.

4.2 Design Rationales

Various approaches can be employed for data wrangling, depending on the methods for exploration or evaluation. Individual analysts can have vastly different demands on the quality of a dataset. We conducted a user study to receive feedback on different design alternatives of an early paper version of our prototype, by conducting a usability inspection. The test subjects were all undergraduate computer science students, with basic knowledge of information visualization. The reason we selected these participants is they are similarly trained in methodologically approaching data analysis as our target user group. They were split up into two groups, where the first group (four participants) was interviewed individually on the designs, and the other group (six participants) conducted a focus group usability inspection. The collected positive and negative feedback (please be referred to the supplementary material for additional information) served as a basis to determine the

important design requirements and refine them for the final prototypical implementation.

- $\mathbf{R_1}$ **Allow analysts to navigate through the available quality information from different perspectives.** Enable exploration by investigating details, but also by pursuing a classical overview-first approach. Analysts should be able to navigate towards their specific goals (analyzing a specific branch, data revision or column/row).
- $\mathbf{R_2}$ **The design should emphasize the impact of operations on quality.** This helps users to associate transformation steps with changes of quality.
- $\mathbf{R_3}$ **Cleanness of the dataset should be signaled by cleanness of the visualization** This should emphasize the analyst's perception that no problems can be observed any more.
- $\mathbf{R_4}$ **A graph of operations should show the different wrangling branches.** The branch of the currently selected wrangling operation sequence should be traceable.
- $\mathbf{R_5}$ **The overall size of the dataset and number of quality problems for every data revision should be communicated.** This should help analysts identify what parts of the data are changed during a transformation.
- $\mathbf{R_6}$ **The detail view should give additional information on operations and changes in quality.** This should help to provide insights into how and why an operation influenced the dataset.

During development, the task considerations and design rationales were consulted to prevent inappropriate design or functionality. The upcoming section describes the core features of our prototype, where single or multiple tasks identified were used as design goals for individual components. At the initial design stage, applicability of the design rationales to the components were determined. Throughout design and implementation, the components continuously underwent inspection if design rationales were supported and maintained.

5 Data Quality Provenance Explorer

We present Data Quality Provenance Explorer (*DQProv Explorer*), a VA approach to visualizing provenance that was captured by our data wrangling provenance model. We employed Shneiderman's visual information seeking mantra by giving overview of wrangling provenance in a provenance graph view as well as details on quality in a flow-like visualization. We provide three interactively linked components in our system, the Provenance Graph View (see Fig. 3a), the Quality Flow View (see Fig. 3b), and the Issue Distribution View (see Fig. 4, usually located to the right side of the Quality Flow View). In the Provenance Graph View we can see a graph of all provenance generated around wrangling the current dataset. The Quality Flow View shows a selected wrangling branch in detail, the Issue Distribution View shows how quality issues are distributed across the dataset for the currently selected revision.

5.1 Provenance Graph View

The Provenance Graph View serves as the central (overview first, $\mathbf{R_1}$) navigation element of *DQProv Explorer* (see Figure 3a), showing the captured wrangling provenance ($\mathbf{R_4}$). Inspired by Wu et al.'s [20] uncertainty flow visualization approach, it shows an acyclic graph flow structure, representing transformation operations and data flow between data states. Upon selection of a graph node (i.e., a revision state), the path of transformations is highlighted as a bright yellow path (cf. Figure 3a), and the Quality Flow View is aligned respectively (see Figure 3b). The node heights encode the relative number of rows (compared to the maximum number of rows) in the current data state ($\mathbf{R_5}$). Icons show the operation types for each revision node (e.g. </> indicates a text transformation), and filter icons (▼) along the graph vertices indicate if the dataset was filtered before applying an operation. This overview lets analysts assess which operations were applied at a glance. On demand, detailed information on the applied operations and filters is available (cf. Figure 3b, $\mathbf{R_6}$).

The Provenance Graph View can be used to analyze different aspects of the wrangling provenance model. By following the flow of data along the graph's vertices and the node height, it is possible to see if operations were only applied to subsets of the dataset. Together with the Quality Flow View, branching and branch lengths in the provenance graph shows analysts the history of previous wrangling attempts: short paths or a large number of branches could imply unsuccessful wrangling attempts; Long paths with the same operation icons can indicate small, repetitious operations without significantly changing the dataset (e.g., editing single cells) or impacting quality.

5.2 Quality Flow View

DQProv Explorer's Quality Flow View (see Figure 3c) shows the overall development of quality issues in a dataset over the course of a selected wrangling branch ($\mathbf{R_1}$). The view shows the proportional amount of errors identified in the dataset by stacking bars for each employed data quality metric (for applicable columns) in the data quality engine. This results in a vertical column of quality issues for each data revision. Different colors indicate different types of quality metrics, and correspondingly different types of issues. By showing the development of these quality issues along a selected provenance branch and the corresponding operations, the analyst can assess which wrangling operation changed the dataset and resolved data quality issues ($\mathbf{R_3}$). The stacked bars are connected with a flow-like encoding. The flows are de-saturated for metrics that remain unchanged between revisions and are saturated to highlight a change of a quality metric measure between two revisions. If all issues detected by a particular quality metric are resolved during a wrangling operation, the corresponding flow bundles to zero (cf. Figure 3c: the metric value changes to zero, indicating that the detected *validity* issues of the column *Weight* have been resolved by this operation). Because the Quality Flow View is aligned with the Provenance Graph View, changes in quality can be traced back to the performed transformation operations and the analyst can gain insights if wrangling operations influenced quality ($\mathbf{R_2}$).

Mouseover interaction highlights the entire quality metric's history in the current branch, giving information on the quality metric values ($\mathbf{R_6}$). Figure 3 shows an example where the initial actions did not affect quality. However,

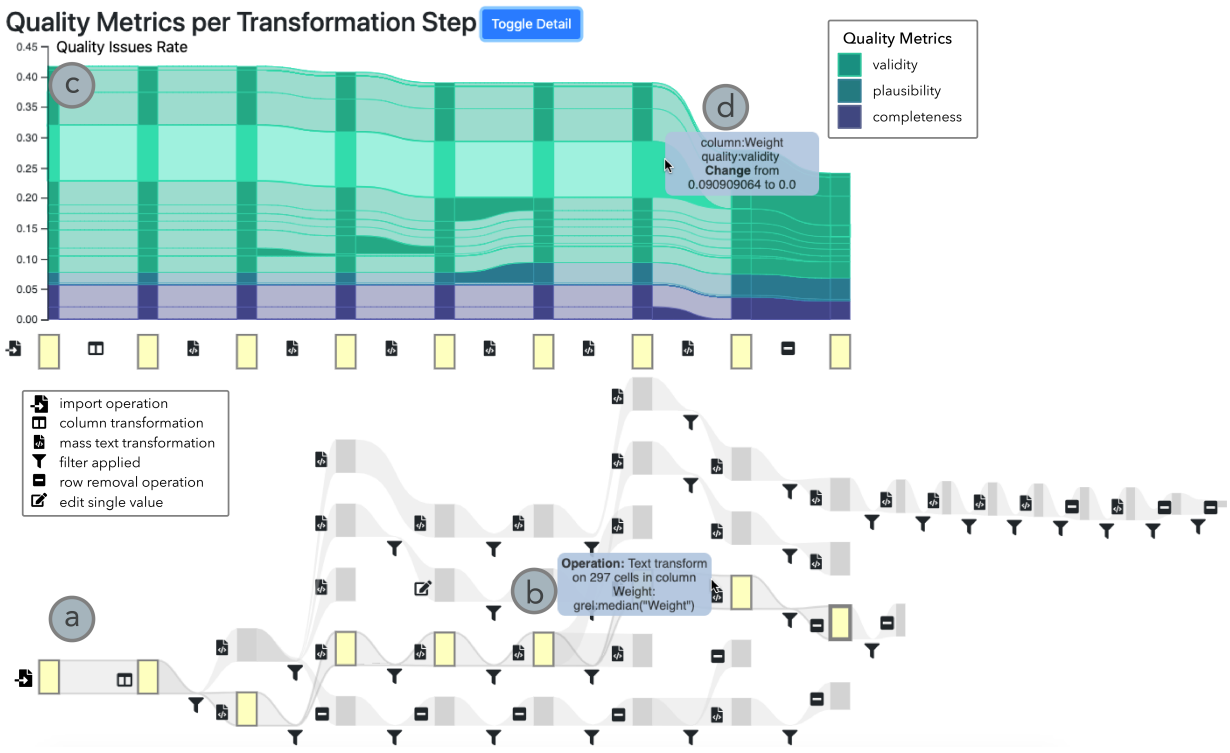

Fig. 3. Two linked views of *DQProv Explorer*: (a) The Provenance graph view allows navigation of the individual data states. The height of the nodes and edges encodes the row size of the data ($\mathbf{R_5}$). Bright yellow graph nodes indicate the currently selected branch, icons indicate the type of operation. (b) On-demand mouseover information on the nodes and vertices shows details on the operations and the dataset size: vertices show information on the employed filtering and transformation parameters, nodes show the number of rows in the dataset. (c) In the Quality Flow View users can observe the development over time for a selected wrangling branch. The bar height indicates the proportional amount of issues detected, color encodes different types of quality metrics. On the horizontal axis, the data revision nodes are duplicated from the selected graph branch to align with the stacked bars to facilitate relating operations to changes in quality. (d) On-demand information on the Quality Flow View highlights the flow of the currently inspected metric (*validity* metric in the *Weight* column) and shows additional provenance information.

after the fourth operation, the number of quality issues continuously decreases. Inspecting the saturated flows with mouseover interactions shows the name of the affected column and metric type(cf. Figure 3d).

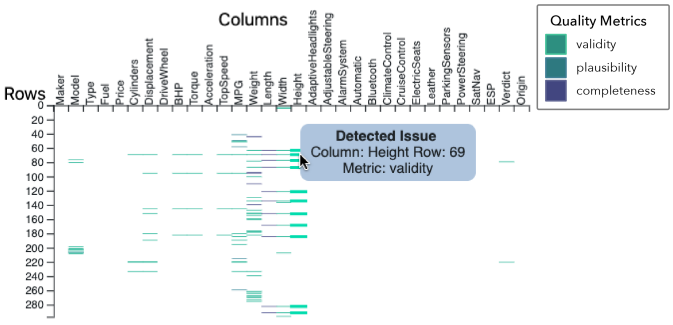

Fig. 4. The Issue Distribution View allows the inspection of issue patterns detected in the current data state. In this particular case it can be observed that (among others) row 69 exhibits multiple errors. The view is linked to the Quality Flow View and mouseover interaction highlights the respective metric flow (cf. Figure 3d).

5.3 Issue Distribution View

The third component in the *DQProv Explorer* is the Issue Distribution View, which can be used for detailed inspection of the distribution of quality issues within the dataset. It shows the relative location of dirty rows within the tabular structure of the dataset. Erroneous entries in the dataset are shown as heat bands, with color encoding the issue type (corresponding to the quality metric identifying the issue). This visualization implies the cleanness of the dataset, with a close to empty view signaling the absence of quality issues ($\mathbf{R_3}$). Inspecting the Issue Distribution View helps discovering error patterns in the dataset (cf. Figure 6, Analysis Step 1). It is an extension of the *schematic error view* presented in the MetricDoc environment [3]. If the number of rows in the dataset exceeds the number of rows available in the visualization, the rows are aggregated to accommodate for insufficient screen space, accumulated errors correspond to higher saturation. That way, it is possible to display datasets exceeding 10,000 entries. When entering into Comparison Mode (see Section 5.4), the difference of the two issue is computed, which allows inspecting the differences in error distribution between these two revisions.

5.4 Comparison Mode

To enable the comparison of the overall quality between two different wrangling branches of the provenance graph, we extended the Quality Flow View to oppose two wrangling branches simultaneously (see Figure 5). The view is displayed when two branches are selected, mirroring the Quality Flow Views, allowing a direct quality comparison of the branches' end points. This view lets analysts compare

the flow of quality over time, but also inspect the difference of employed wrangling operations. For example, if an analyst has to decide to continue analyzing the data, and two wrangling attempts (branches) look similar, it is possible to use the comparison mode to assess which sequence of operations yielded better quality, or used less wrangling steps but was equally as effective. To retain linking to the transformation operations applied to the selected branches ($\mathbf{R_2}$), we duplicate the branches and position them below the Quality Flow View. The selected branches in the Provenance Graph View are highlighted in clearly distinguishable colors, including shared nodes that are bright yellow.

6 Use Case – Cleansing a Car Dataset

We illustrate a use case that shows the Data Quality Provenance Explorer (*DQProv Explorer*) in a concrete wrangling scenario. In this scenario we consider an analyst concerned with the task of wrangling a car dataset (see Supplementary Material for detail information). The analyst investigates the Issue Distribution View showing the automatically computed data quality metrics (cf. Figure 6 Analysis Step 1) for three types of issues (invalid, incomplete, and implausible entries). It shows that 12 of the 33 total columns have issues that need to be taken care of. In the detail view of the initial data state, we see issue patterns which indicate that a few erroneous rows are responsible for multiple detected issues (cf. Figure 6 Analysis Step 1). After identifying the dirty data rows in columns that contain the most errors – namely the *'weight'*, *'width'*, *'height'*, *'displacement'*, and *'miles per gallon' (MPG)* column – and removing them in the data wrangling system, the analyst returns to the *DQProv Explorer* to check how many issues still remain. The analyst finds that most issues have been solved, but the *'MPG'* column still retains implausible values (cf. Figure 6 Analysis Step 2). Upon inspection the analyst determines that these are the result of hybrid cars having better fuel efficiency and reasons that the metric shows false positives.

Upon further inspection of the raw data, the analyst notices that some entries represent electric cars, that should not be removed from the dataset, because otherwise electric cars would be omitted from the dataset. Hence the analyst reverts all operations and restarts the wrangling process. Filtering for `NA` values in the fuel column brings up multiple electric cars. The analyst proceeds to fill in missing cells (*'cylinders'* with `0`, *'displacement'* with `0`, and *'MPG'* with `-1` because the column is not applicable and the numeric value will not create issues in further analysis instead of `NA`) and removes five data rows that exhibit missing values in multiple cells. For the remaining detected issues in columns *'width'*, *'height'*, *'weight'*, and *'displacement'*, the analyst decides to impute missing values with the column's median value instead of removing the entries, like in the first wrangling attempt. The analyst imputes all relevant columns' missing values and returns to *DQProv Explorer* for comparing the overall quality of the second wrangling branch with the first one, where quality was improved mainly by removing data entries (cf. Figure 6 Analysis Step 3). Summary information on the provenance graph's nodes shows that the analyst could retain 293 rows in the second wrangling attempt as opposed to 244 rows in the first wrangling attempt (cf. Figure 6 Analysis Step 4). The analyst continues with selecting two nodes for comparison and inspects the differences in overall quality of the two end states of the branches. It reveals that s/he could successfully remove the similar amounts of errors in the second attempt, but with the benefit of retaining more information by not removing data entries.

7 Evaluation

We conducted a user experience study to determine if *DQProv Explorer* enables users to analyzing provenance generated from data wrangling workflows. We recruited 6 participants (4 male, 2 female; 1 Master Student, 4 Doctoral Students, and 1 Post Doctoral Researcher in Computer Science) with varying degrees of experience in both data quality assessment and visual data analysis. The self-assessed expertise (from (1) = *novice* to (5) = *expert*) of users ranged from intermediate (3) to expert (5) in data wrangling. Expertise in visual data analysis ranged from novice (1) to expert (5).

Within the study we tried to answer if the tasks defined in Section 4.1 are sufficiently supported by our prototype. Specifically we formulated the questions: Can participants determine if quality has changed, and can they decide if the data is usable for subsequent analysis? Are the participants able to compare branches to assess the difference in operations applied to the data, and decide which of the branches poses the most useful dataset for their analysis? Does the prototype allow the users to derive which quality issues were inherent in the dataset and how they were resolved?

7.1 Procedure

Due to limited time with participants, we gave an introduction into the visual encodings and interaction features of the prototype. We then assigned them to complete prepared tasks. We encouraged the participants to think aloud while conducting the tasks. Important actions and comments during the tasks and participant feedback after the session were noted. The sessions took between 75 and 90 minutes and were structured as follows:

Introduction Session (10-15 Minutes): If necessary, the participants were received an introduction into data wrangling and quality metrics to clarify the scope of analysis, specifically because participants had different expectations of a usable dataset. The investigator then exhibited the general functionality and visual encodings of the prototype.

Task Assignment (30-40 Minutes): Participants were instructed to conduct tasks that were oriented around our requirements analysis (cf. Section 4.1). Questions were prepared for each task to guide iterative analysis. If the participant did not provide enough information, the investigator would ask intermittent questions and to suggest possible alternatives to exploring the provenance data. Specifically, questions were intermittently asked to determine what type of provenance participants relied on when conducting analysis.

Interview (10-20 Minutes): In the interview, the investigator asked for feedback about their experiences with the prototype. The participant should reflect on the usability

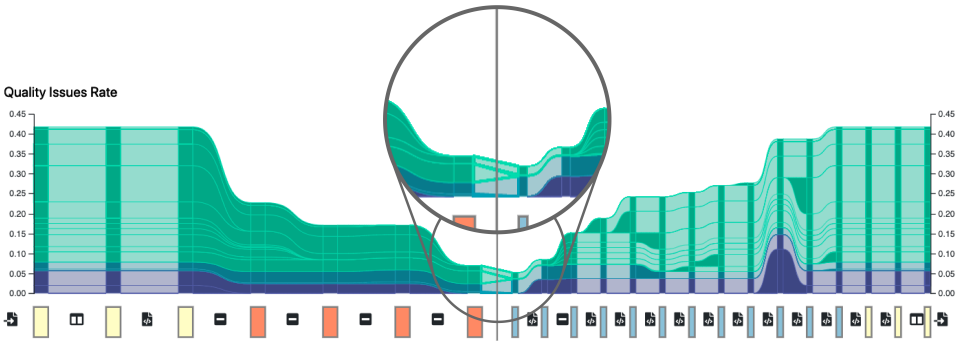

Fig. 5. The comparison mode juxtaposes two wrangling branches. The first branch runs from left to right, while we flip the second branch to run from right to left. This allows for direct matching of the end-states of the dataset of the selected branches. In particular, differences in quality are more easily identifiable. In this image the first three nodes are shared between both branches. It can be observed that different approaches to improve quality have been employed, while in the left branch data elements were removed (multiple ⛔-operations), in the right branch elements were edited or imputed (consecutive </>-operations). However, both approaches led to a reduction of quality problems (height of bars).

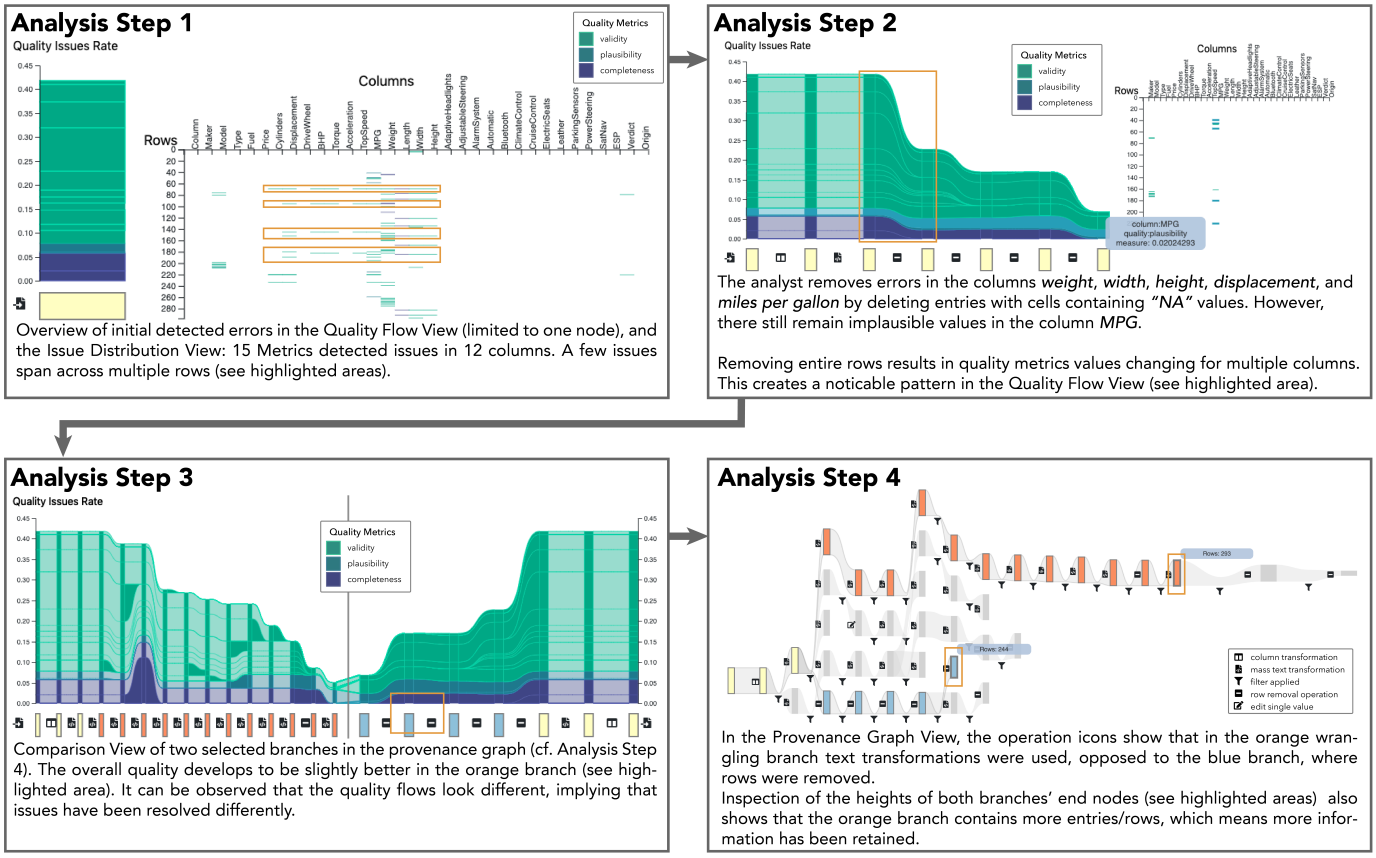

Fig. 6. Visual overview of the wrangling process on a car dataset. The four steps show different stages of the analysis process and how the analyst can use the different views and interactions to determine if the overall quality has improved.

and usefulness of *DQProv Explorer*. This was done to encourage participants to express difficulties they encountered during analysis and to collect suggestions how these could be resolved. The feedback was collected in an unstructured way, participants could express their comments and suggestions in any way they preferred.

Questions: During each separate task participants were asked a series of questions to stimulate iterative exploration and cover the tasks laid out in Section 4.1:

$\mathbf{T}_{act}$ & $\mathbf{T}_{pres}$ - Look at the first state of the dataset and identify the column with the most issues (Column *'weight'*). Now look at the end node of one transformation branch and determine how quality evolved for this column. You can see multiple transformation branches: How different are the two branch end nodes in terms of quality, do similar issues remain? Can you find out what transformation/operation impacted the quality of this column the most?

$\mathbf{T}_{meta}$ - If only the dataset of the second branch was available for analysis, what columns would you use for analysis. If you look at the three different branches and compare remaining quality issues, which one would you choose for analysis, and for what type of analysis? (The *'weight'* column

was affected differently in different branches, cf. Figure 6 Analysis Step 4)

$\mathbf{T}_{rec}$ & $\mathbf{T}_{rep}$ - How did a sequence of actions influence the data? Going back to the Weight column, which of the branches would you use for analysis?

$\mathbf{T}_{coll}$ - Can you determine the user's objective in the sequence of transformations shown in the branch at the bottom of the provenance graph?

7.2 Results

We summarize the results and provide an overview of feedback that was given by multiple participants (a detailed breakdown of the user study and summarized feedback from participants on the different views can be found in the supplementary material). The questions were solved by all participants, with the exception of one participant not being able to solve questions for $\mathbf{T}_{coll}$ (the participant had the lowest self-assessed experience with data wrangling). In summary, we determined two different methodologies of assessing quality issues, based on the participants' patterns of exploration. Two participants iteratively navigated the provenance graph in an detail-first, overview later approach (mainly exploring the Provenance Graph View, using the Quality Flow View for quality inspection). The remaining four participants pursued an overview-first, details on demand methodology (mainly using the Quality Flow View for exploration, and the Provenance Graph View was used only for selecting different branches, and for on-demand context information).

Furthermore, we found implications that the trust in the employed data quality metrics and the trust towards the wrangled dataset depends on the participant's expertise in data wrangling. While two participants simply accepted the metrics as being accurate and subsequently found the Quality Flow View to be sufficient for determining the validity of the dataset, two participants would refuse to make a final statement on the data's quality without exploring the raw data. Specifically participants with higher data wrangling experience demanded for more brushing and linking features, which to us indicated that familiarity with these tools makes users more confident to use complex interaction techniques. Two users suggested to add filtering techniques and toggling techniques to enable more focused exploration on particular types of changes.

Feedback from participants on the different views was mixed. While generally the Quality Flow View and the Provenance Graph View have been well received, the usefulness of the Issue Distribution View was questioned by the majority of participants. This view extended the concept of a *schematic error view* presented by Bors et al. [3], which we adapted to show the distribution of errors across all columns. Participants showed no interest in this view. In future work, it is necessary to determine a more appropriate visualization that supports the analyst in assessing quality issues in detail. Two participants also noted that auditing the data wrangling process of someone else by exploring the provenance graph increased their confidence in the data. This implies the usefulness of *DQProv Explorer* for handoff tasks in collaborative settings. Based on this result we plan to introduce more collaborative features, e.g., adding comments, hiding/disabling provenance revisions.

8 Discussion

The user study could show that *DQProv Explorer* was well received, even though some features were not deemed as necessary by participants. Generalizing this feedback is questionable due to the small number of participants (6). The target audience of our approach are not domain experts, but rather professionals dealing with data analysis who require data pre-processing. This tool is unique in its ability to explore workflow and data provenance from data wrangling, hence we could not use comparable tools in the evaluation of our design. In particular, the Quality Flow and Provenance Graph Views feature custom visualizations to display data provenance specific to wrangling, which is not possible to appropriately encode in general workflow provenance visualizations.

In the introduction we proposed that leveraging data quality metrics aids the user in understanding the quality of the dataset. We can neither confirm nor deny this proposition. One interesting observation from the user study was participants' different perception of quality: While some considered each entry of a dataset as valuable, preferring imputation of values over removal of entries, others solely depended on the quality metrics to signal quality issues and considered the absence of issues as sufficient. Based on the answered questions, specifically questions attempting to validate the understanding task $\mathbf{T}_{coll}$, we can deduce that *DQProv Explorer* supports users in making sense of the wrangling history and in estimating the usefulness of the resulting data based on the user's subjective perception of quality. However, this also implies that employed data quality metrics must be carefully developed and adequately used, because it could also lead to perceiving low/high quality mistakenly. In the future we aim to explore how analysts perceive quality differently, and how this can be used to optimally present wrangling and data quality information.

The presented examples employ quality metrics to detect issues of the types completeness, validity, and plausibility. But our approach is extendable to different types of metrics: Using performance metrics from machine learning algorithms and allowing users to explore the results on different training data could lead to a better understanding of how influential the datasets are on the final algorithmic outcome. Also, measuring introduced uncertainty from wrangling processes could be quantified by quality metrics.

Addressing the concern of scalability of our approach, the Provenance Graph View runs into problems when the graph grows very large. This specifically applies to a large number of consecutive single cell edits, leading to excessively long branches, which skews the provenance graph. Additionally, a large number of wrangling attempts leads to compressed graph nodes, which renders the row size encoding useless. This needs to be addressed by further supporting exploration, by including filtering, zooming, and panning, but also allowing analysts to collapse branches and merge nodes from similar consecutive operations. In contrast, the provenance flow visualization can be scaled well, retaining the development of quality over time also with low space available. However, it could lead to the flow changes not being noticeable, but this will only occur when

there is a very large number of wrangling operations, which can be circumvented by employing a semantic zooming technique.

Our attempt to use context information to annotate provenance, like data quality metrics, or summary information might not cover all changes applied to a dataset. We are interested in further exploring the database and data quality research to find other ways of preserving changes to a dataset.

9 Conclusion

We presented *DQProv Explorer*, a VA approach for capturing provenance data from data wrangling with annotations in the form of data quality metrics and descriptive measures. It enables users to explore the provenance graph of wrangling operations and assess the impact of these operations on the overall quality of a dataset, including the comparison of alternative branches of operations, and detailed issue inspection. In a user study we evaluated the appropriateness of *DQProv Explorer* for different analysis tasks. The results indicate that it enables the users to explore provenance and make sense of the impact of particular operations on quality, as well as to judge usability of the dataset for further analysis purposes.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF), Project No. I 2850-N31, Lead Agency Procedure (DACH) *"Visual Segmentation and Labeling of Multivariate Time Series (VISSECT)"*.

References

[1] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance techniques," *Computer Science Department, Indiana University, Bloomington IN*, vol. 47405, 2005. [Online]. Available: https://www.cs.indiana.edu/ftp/techreports/TR618.pdf

[2] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive Visual Specification of Data Transformation Scripts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 3363–3372. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979444

[3] C. Bors, T. Gschwandtner, S. Kriglstein, S. Miksch, and M. Pohl, "Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics," *J. Data and Information Quality*, vol. 10, no. 1, pp. 3:1–3:26, May 2018. [Online]. Available: http://doi.acm.org/10.1145/3190578

[4] T. C. Redman, "Data Quality Management Past, Present, and Future: Towards a Management System for Data," in *Handbook of Data Quality*, S. Sadiq, Ed. Springer Berlin Heidelberg, 2012, pp. 15–40.

[5] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee, "A taxonomy of dirty data," *Data mining and knowledge discovery*, vol. 7, no. 1, pp. 81–99, 2003. [Online]. Available: http://link.springer.com/article/10.1023/A:1021564703268

[6] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch, "A Taxonomy of Dirty Time-Oriented Data," in *Lecture Notes in Computer Science (LNCS 7465): Multidisciplinary Research and Practice for Information Systems (Proceedings of the CD-ARES 2012)*, G. Quirchmayr, J. Basl, I. You, L. Xu, and E. Weippl, Eds. Prague, Czech Republic: Springer, Berlin / Heidelberg, 2012, pp. 58–72.

[7] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling: Visuatizations and transformations for usable and credible data," *Information Visualization*, vol. 10, no. 4, pp. 271–288, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1177/1473871611415994

[8] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 31–40, Jan. 2016.

[9] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "VisTrails: Visualization Meets Data Management," in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '06. New York, NY, USA: ACM, 2006, pp. 745–747. [Online]. Available: http://doi.acm.org/10.1145/1142473.1142574

[10] W. C. Tan, "Provenance in Databases: Past, Current, and Future," *IEEE Data Eng. Bull.*, vol. 30, pp. 3–12, 2007.

[11] K. Xu, S. Attfield, T. Jankun-Kelly, A. Wheat, P. Nguyen, and N. Selvaraj, "Analytic Provenance for Sensemaking: A Research Agenda," *IEEE Computer Graphics and Applications*, vol. 35, no. 3, pp. 56–64, May 2015.

[12] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ser. AVI '12. New York, NY, USA: ACM, 2012, pp. 547–554. [Online]. Available: http://doi.acm.org/10.1145/2254556.2254659

[13] T. Gschwandtner and O. Erhart, "Know Your Enemy: Identifying Quality Problems of Time Series Data," in *2018 IEEE Pacific Visualization Symposium (PacificVis)*, Apr. 2018, pp. 205–214.

[14] T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, and N. Suchy, "TimeCleanser: A Visual Analytics Approach for Data Cleansing of Time-oriented Data," in *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, ser. i-KNOW '14. New York, NY, USA: ACM, 2014, pp. 18:1–18:8. [Online]. Available: http://doi.acm.org/10.1145/2637748.2638423

[15] C. Bors, M. Bögl, T. Gschwandtner, and S. Miksch, "Visual Support for Rastering of Unequally Spaced Time Series," in *Proceedings of the 10th International Symposium on Visual Information Communication and Interaction*, ser. VINCI '17. New York, NY, USA: ACM, 2017, pp. 53–57. [Online]. Available: http://doi.acm.org/10.1145/3105971.3105984

[16] C. Niederer, H. Stitz, R. Hourieh, F. Grassinger, W. Aigner, and M. Streit, "TACO: Visualizing Changes in Tables Over Time," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 677–686, Jan. 2018.

[17] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann, "A systematic view on data descriptors for the visual analysis of tabular data," *Information Visualization*, vol. 16, no. 3, pp. 232–256, Jul. 2017. [Online]. Available: https://doi.org/10.1177/1473871616667767

[18] M. Herschel, R. Diestelkämper, and H. B. Lahmar, "A survey on provenance: What for? What form? What from?" *The VLDB Journal*, vol. 26, no. 6, pp. 881–906, Dec. 2017. [Online]. Available: https://link.springer.com/article/10.1007/s00778-017-0486-1

[19] S. Miksch and W. Aigner, "A Matter of Time: Applying a Data-Users-Tasks Design Triangle to Visual Analytics of Time-Oriented Data," *Computers & Graphics, Special Section on Visual Analytics*, vol. 38, pp. 286–290, 2014. [Online]. Available: http://www.ifs.tuwien.ac.at/silvia/pub/publications/miksch_cag_design-triangle-2014.pdf

[20] Y. Wu, G.-X. Yuan, and K.-L. Ma, "Visualizing flow of uncertainty through analytical processes," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2526–2535, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6327258

Christian Bors is a PhD candidate in the Institute of Visual Computing and Human-Centered Technology at TU Wien. His main area of research is developing interactive techniques for data wrangling, profiling, and cleansing systems, utilizing quality metrics and data provenance.

Theresia Gschwandtner is Postdoc University Assistant at the Institute of Visual Computing and Human-Centered Technology at TU Wien. Her research focuses on visual and interactive techniques for data analysis, such as data quality management, uncertainty visualization, data provenance, and guidance in visual analytics.

Silvia Miksch is University Professor and head of the Research Division “Visual Analytics” (Centre for Visual Analytics Science and Technology (CVAST)), Institute of Visual Computing and Human-Centered Technology, TU Wien. She served as paper co-chair of several conferences including IEEE VAST 2010 & 2011 and EuroVis 2012 and on the editorial board of several journals including IEEE TVCG and Computer Graphics Forum. She acts in various strategic committees, such as the VAST steering committee and the VIS Executive Committee. Her main research interests are Visualization/Visual Analytics (particularly Focus+Context and Interaction) and Time.