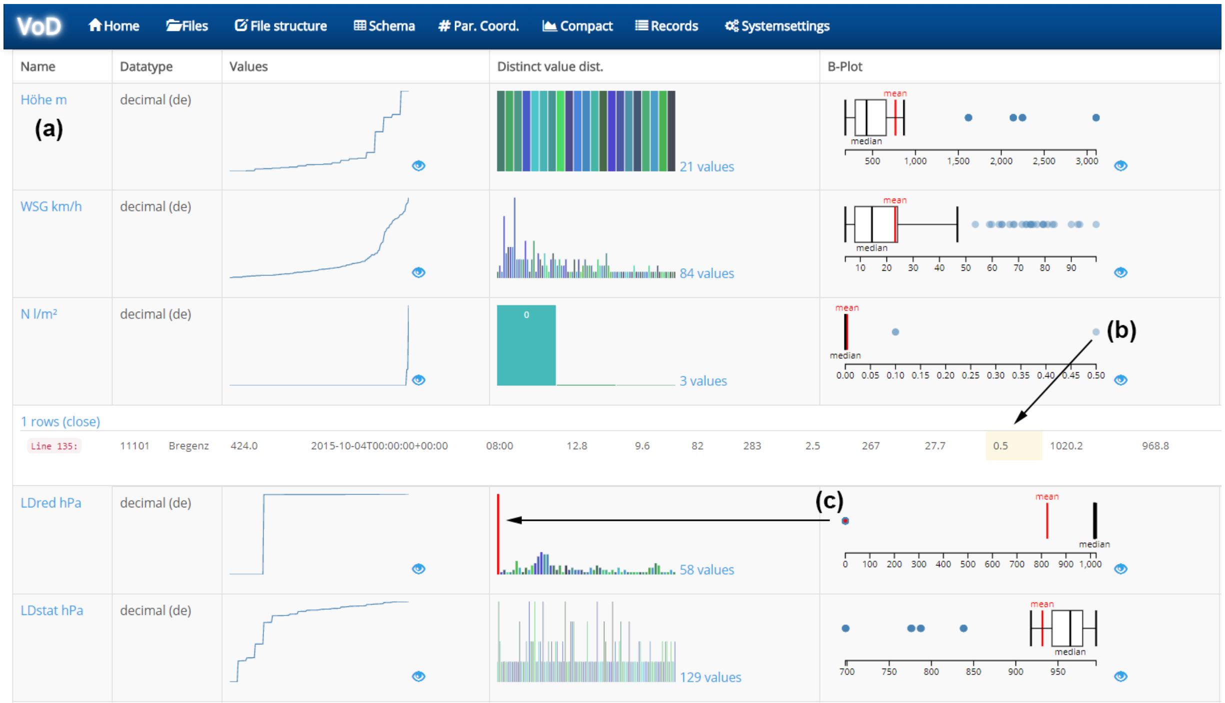
VoD – Understanding Structure, Content, and Quality of a Dataset

Andreas Peterschofsky[*]
TU Wien

Theresia Gschwandtner[†]
TU Wien



Name	Datatype	Values	Distinct value dist.	B-Plot
Höhe m	decimal (de)		21 values	
WSG km/h	decimal (de)		84 values	
N l/m²	decimal (de)		3 values	
LDred hPa	decimal (de)		58 values	
LDstat hPa	decimal (de)		129 values	

1 rows (close)

Line 135: 11101 Bregenz 424.0 2015-10-04T00:00:00+00:00 08:00 12.8 9.6 82 283 2.5 267 27.7 0.5 1020.2 968.8

Figure 1: VoD is a web application that provides summaries of datasets with multiple tabs to show different aspects of the data. The *Compact* page shows different visualizations that in combination help to get a compact overview of the dataset (on the data value level) and possible quality problems. The scroll-able page represents each data attribute in a separate row (here cropped for readability). Row (a) represents the data attribute 'Höhe m' which gives the height of air quality measurement stations. A line chart shows the individual sorted data values in this column. The bar chart showing the quantity of distinct values reveals that there are exactly 21 different heights in this data set corresponding to the 21 measurement stations. The box plot shows that most data elements are measured at stations on lower heights with some exceptions. These data values look plausible. In (b) we see an outlier (far right dot in the box plot) in the amount of precipitation (the data attribute represented in this table row). When selecting this outlier, a row appears below showing the corresponding data entry. As almost all measurements of precipitation were 0 l/m^2 (as can be seen in the bar chart), 0.5 l/m^2 still seems legit and we reason that this is a correct measurement. Hovering the outlier in (c) highlights the corresponding bar in the bar chart in red. This shows that many data entries of '0' correspond to this outlying position. The line chart as well as the bar chart show that there is a noticeable difference between these 0 values and the remaining data. These may be missing values and demand for further investigation.

Abstract

In the age of data science analysts need to handle new data sets on a daily basis. In a first step they need to understand structure, content, and if the dataset is fit-for-use for further processing. However, getting familiar with a dataset by simply scrolling through the data in tabular form is just not feasible for these usually very large sets of data. Thus, we have designed and evaluated a Visual Analytics prototype that provides interactive visual summaries of a dataset on three different levels: the dataset level, the data attribute level, and the data value level. Our results demonstrate the usefulness of our approach and point to further research challenges.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics; Applied computing—Document management and text processing—Document management—Document metadata

1 Introduction

Making sense of huge amounts of data is one of the most important tasks of our time. However, before any meaningful data analysis can be conducted, it is necessary to understand the dataset at hand. At the beginning of any data processing pipeline, the analyst needs to get familiar with the dataset and understand its structure, content, possible anomalies, and quality problems. In a second step it is usually necessary to perform data pre-processing which involves data cleansing, data augmentation, and transformation steps. The amount of data that needs to be handled in this context demands for the computation of effective key figures and summary visualizations. Moreover, the tasks involved are often of interactive nature (e.g., the investigation of data anomalies), which makes interactive visualiza-

[*]e-mail: e9625532@student.tuwien.ac.at
[†]e-mail: theresia.gschwandtner@tuwien.ac.at

Publication forthcoming in Proc. of the Workshop on Visual Summarization and Report Generation: Beyond Scatter-Plots and Bar-Charts (VISREG 2018), IEEE Xplore Digital Library. This is the authors' accepted version (postprint).

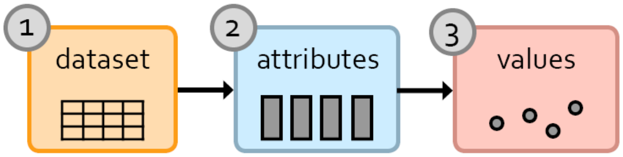


Figure 2: VoD provides summary visualizations on different granularity levels of the dataset, which supports a step wise familiarization – from getting a coarse grained overview to fine grained data value inspection. In a first step we provide summary information about the overall dataset, such as table structure, text delimiters, and amount of missing values. In a second step we provide summary visualizations to dig deeper into the data by investigating the data attributes. This includes value ranges, patterns, and missing values per data attribute, but also the investigation of correlations between these attributes. In a last step we provide visualizations to investigate the data value level and reason about the plausibility of value distributions and outliers.

tions and Visual Analytics (VA) effective means to unburden the user in these tasks.

While there is a number of existing VA approaches tackling different aspects of data quality management, the very first step of getting familiar with an unknown dataset is hardly supported. Thus, we propose a VA prototype to provide visual summaries of a dataset with a focus on understanding content, structure, value distributions, relations, patterns, and possible data quality problems.

2 Related Work

There are a number of VA approaches tackling the problem of data profiling (i.e., the identification of data structures and quality problems). Talend Open Studio [17] is an open-source data profiling tool, which provides mainly statistical information (minimum, maximum, and missing values). Also DataManager [3] provides profiling statistics. Profiler [11] computes suitable representations for different data types. This results in one visualization for each data attribute to give an overview of the data and some automatically identified quality problems. Visplause [1] is a VA system for profiling time series data with a special focus on the investigation of data quality problems and plausibility of the data. KYE [6], on the other hand, provides a number of automatic quality checks in combination with visual exploration to detect additional, hidden quality problems. Other approaches, again, are specialized on specific aspects of data profiling, such as the detection of duplicate data records (e.g., DataMatch [8]). Such data profiling approaches are usually based on taxonomies of different data quality problems, which provide a good overview which tasks and problems should be supported by a data profiling solution (e.g., [7, 12, 13, 15]).

Moreover, there are VA tools that support other data quality management tasks such as data transformations (e.g, Wrangler [10], OpenRefine [18]), or data cleansing (e.g., Potter's Wheel [16], TimeCleanser [5]). More general studies about outlier detection [9], VA methods for Big Data [19], or tabular visualizations [4, 14] cannot be discussed here in detail due to space constraints.

The approaches described above, however, do not focus on providing summary visualizations to get a very first idea of an unknown dataset, but rather tackle subsequent data quality management tasks. Thus, we propose a VA prototype to support this very first step that is necessary before any data-processing can be applied. We guide the user with a step-wise approach in order to ease this task: starting with a very coarse-grained overview and gradually refining it.

3 Our Approach

We designed and implemented a research prototype as web application called Visualizer of Datasets (VoD), with the goal of providing a generic visual summary that allows the user to get a first idea about the structure and quality of different kinds of tabular datasets. Besides structural information, the prototype provides different key figures and visualizations aimed at pointing to potential data quality problems. The web application comprises multiple tabs that communicate different aspects of the data and serve as a step wise guide into the dataset, similar to the idea of a semantic zoom (see Figure 2). In the following we are using an open dataset about weather measurements at Oct. 4, 2015 of different measurement stations in Austria [1] to demonstrate the visual summary pages of VoD.

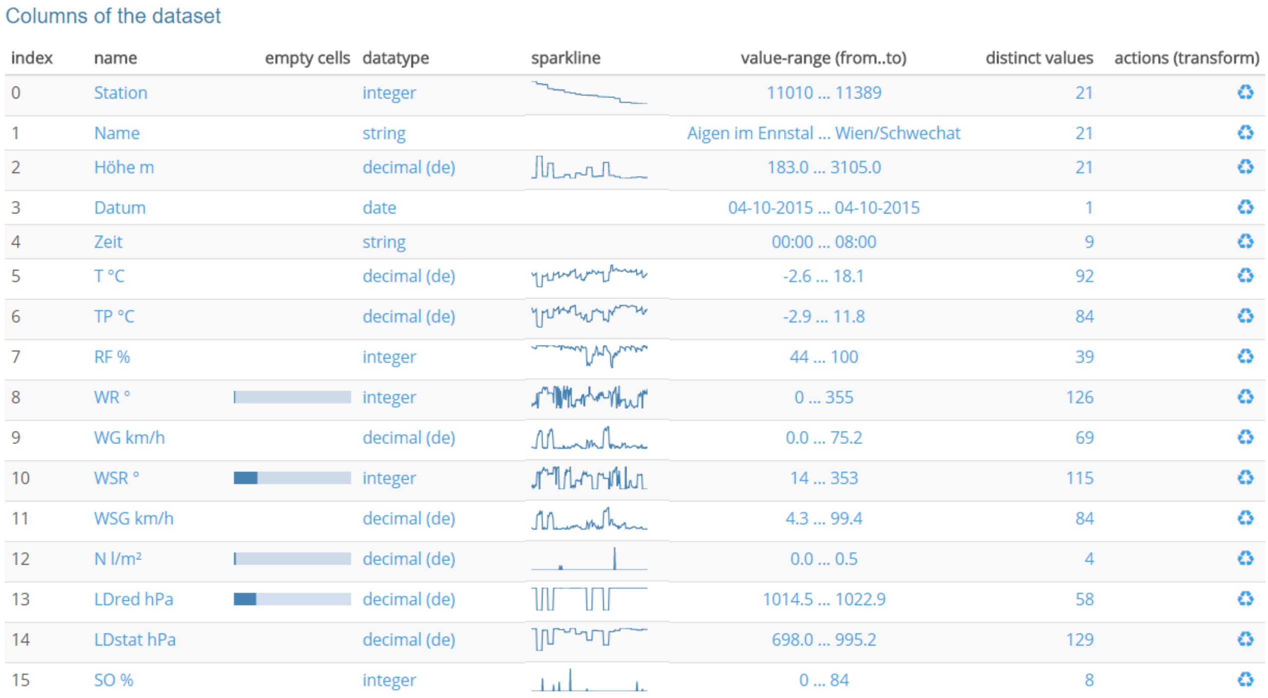
The *File structure* page provides a summary on the general dataset level, giving textual and visual information about the delimiter symbols of text, number of rows and columns, as well as information about the presents of column headers. It also provides information about detected data quality problems of the dataset, such as duplicated rows, incomplete rows, missing values, and data type mismatches.

The *Schema* page provides a summary on the data attribute level, and thus, it provides initial information about the content of the dataset. We provide a table view (see Figure 3), giving key informations about data attributes and means to change the automatically detected data type for each attribute. This table view contains sparkline visualizations that efficiently communicate the amount of empty cells for each attribute as well as line charts showing the values of numerical attributes. Inspecting these line charts in Figure 3 we can see that attributes 5 and 6, 9 and 11, and 13 and 14 seem to correlate and there seems to be a negative correlation between attribute 2 and attribute 14. Of course, reasoning about the validity of the data can only be supported by the provided visualizations and must be decided by a domain expert. We went for a tabular view augmented with sparkline visualizations as this combination is suited to convey compact information about different characteristics (columns) for each data attribute (rows), and thus, it provides a comprehensive, yet manageable overview to get familiar with the data.

The *Paar. Coord.* page allows further investigation of possible correlations between data attributes. Investigating the parallel coordinates plot in Figure 4 affirms the presents of positive and negative correlations between different data attributes. Knowing about the meaning of the data attributes, these correlations are no suprise (e.g., a strong negative correlation between 'station height' (Höhe m) and 'atmospheric pressure data at station level'). However, if a user is not familiar with a dataset, understanding these kinds of correlations between different data attributes is decisive to get a better feeling for the data at hand. We also considered scatterplot matrices to convey correlations, however, they make it hard to identify groups of data items with similar characteristics throughout the dataset, which is needed to reason about the validity of individual data values.

The *Compact* page (see Figure 1) provides a summary on the data value level by means of multiple connected views. These views communicate value distribution and quantities of distinct values per data attribute. We provide three different visualizations that in combination are effective to reason about the plausibility of data values: (1) a line chart of sorted data values, (2) a bar chart showing the quantity of distinct data values, and (3) a box plot visualization showing the data distribution and outliers. Figure 1(a) shows the data attribute 'Höhe m' (the height in meters of different weather measurement stations). The bar chart reveals that there are exactly 21 different height values for this data attribute. These height values exist in equal numbers in the dataset. This seems plausible as there are exactly 21 different measurement stations. The fact that the height values exist in equal numbers shows that the same amount of measurements were taken at each station, which again indicates that there are no missing measurements. The box plot visualization shows that the majority of measurement stations are located at lower heights (with some outliers), and also the line chart

[1] http://at-wetter.tk/index.php?men=raw (accessed: July 12, 2018)



Columns of the dataset

index	name	empty cells	datatype	sparkline	value-range (from..to)	distinct values	actions (transform)
0	Station		integer		11010 ... 11389	21	♻
1	Name		string		Aigen im Ennstal ... Wien/Schwechat	21	♻
2	Höhe m		decimal (de)		183.0 ... 3105.0	21	♻
3	Datum		date		04-10-2015 ... 04-10-2015	1	♻
4	Zeit		string		00:00 ... 08:00	9	♻
5	T °C		decimal (de)		-2.6 ... 18.1	92	♻
6	TP °C		decimal (de)		-2.9 ... 11.8	84	♻
7	RF %		integer		44 ... 100	39	♻
8	WR °		integer		0 ... 355	126	♻
9	WG km/h		decimal (de)		0.0 ... 75.2	69	♻
10	WSR °		integer		14 ... 353	115	♻
11	WSG km/h		decimal (de)		4.3 ... 99.4	84	♻
12	N l/m²		decimal (de)		0.0 ... 0.5	4	♻
13	LDred hPa		decimal (de)		1014.5 ... 1022.9	58	♻
14	LDstat hPa		decimal (de)		698.0 ... 995.2	129	♻
15	SO %		integer		0 ... 84	8	♻

Figure 3: The *Schema* page provides a summary on the data attribute level by means of a table with initial information to get an idea about these attributes. It shows the number and names of attributes and highlights the presence of empty values for each attribute. It displays which data type was automatically detected for which data attribute, and sparklines give a first impression of the behavior of different attributes. Withouth knowing anything about this dataset it can be seen that attributes 5 and 6, 9 and 11, and 13 and 14 seem to correlate and there seems to be a negative correlation between attribute 2 and 14.

of sorted data values does not reveal any suspicious patterns. A more suspicious pattern can be seen in Figure 1(b) where the box plot reveals extreme outliers for the attribute amount of precipitation. Selecting such an outlier in the box plot makes a row appear that shows the corresponding data row from the dataset, giving the raw values and highlighting the selected data value. This reveals that the selected outlier in the box plot is 0.5 l/m^2 of precipitation, while the majority of precipitation values are 0 l/m^2 (as can be seen in the bar chart). This seems to be a plausible value. When hovering the outlier in Figure 1(c), the corresponding bar in the bar chart is highlighted in red. We can see that there is an unusual high amount of 0 values here (also visible in the line chart) which might indicate a data quality problem that demands for further investigation.

3.1 Evaluation

We conducted a qualitative evaluation of our approach with four participants. All participants are technicians and perform tasks with data and information handling within their assignments on a regular basis, but are no visualization experts. They are familiar with general tools such as MS Excel [2].

In individual evaluation sessions we first gave an introduction to the prototype with an explanation of functions and features. Then, we asked each participant to choose one or more of the provided datasets and also to import one of their own files. We further asked them to assume that they have to use the chosen datasets in further analysis steps, so they should use VoD to understand structure, content, and possible quality problems of the dataset. They were autonomously using VoD to explore the data and could ask an instructor in case they encountered any problems during the session. These sessions lasted for about an hour. Subsequently, we asked them to answer a questionnaire about their assessment of the prototype and possible shortcomings.

The study participants graded the overall experience and usage of the prototype as ”good” (second best rating on a 5 point scale). Also the means for data quality assessment and the visualization capabilities were rated to be ”good”. The uniform opinion of all users was that even non technical users would be able to use VoD.

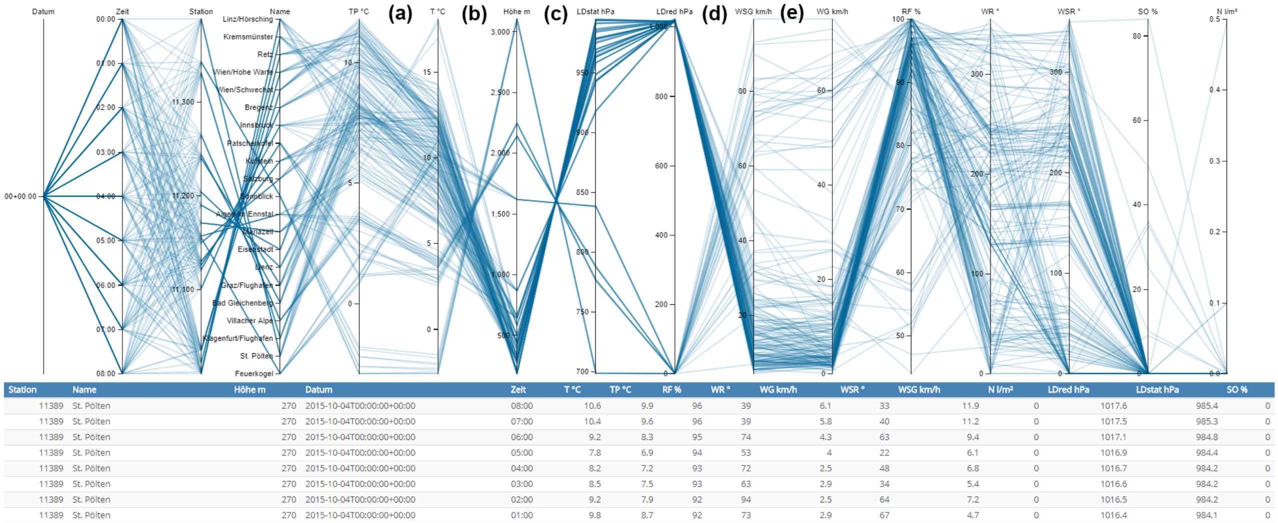
They mentioned criticism about the file import functionality, which is implemented rather rudimentary as this type of usability features was not focus of our work (although the overall rating of the usability was positive). In the following section we will discuss interesting findings of the evaluation results and outline further research challenges.

4 Discussion

When asking the participant what features they would expect from a data summary, all of them reported they expected statistical information about the data, as well as information about missing values. It needs to be considered, however, that all of our study participants have a technical background, so these findings might not be generalizeable to a general public. One participant mentioned that he expected to see correlations between data attributes. We reason that these types of information are important for summarizing a dataset, as they allow users to get an idea of decisive aspects: (1) data structure and attributes, (2) value distributions (preferably in an easy to understand way that does not require statistical knowledge), (3) quality problems such as missing or invalid values, and (4) correlations between data attributes. This list may not be complete but can be used as a starting point for further investigation.

Providing such a summary visualization, however, requires the derivation of a number of data characteristics. The data types of

[2] https://en.wikipedia.org/wiki/Microsoft_Excel (accessed: Aug 17, 2018)



Station	Name	Höhe m	Datum	Zeit	T °C	TP °C	RF %	WR °	WG km/h	WSR °	WSG km/h	N l/m²	LDred hPa	LDstat hPa	SO %
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	08:00	10.6	9.9	96	39	6.1	33	11.9	0	1017.6	985.4	0
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	07:00	10.4	9.6	96	39	5.8	40	11.2	0	1017.5	985.3	0
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	06:00	9.2	8.3	95	74	4.3	63	9.4	0	1017.1	984.8	0
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	05:00	7.8	6.9	94	53	4	22	6.1	0	1016.9	984.4	0
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	04:00	8.2	7.2	93	72	2.5	48	6.8	0	1016.7	984.2	0
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	03:00	8.5	7.5	93	63	2.9	34	5.4	0	1016.6	984.2	0
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	02:00	9.2	7.9	92	94	2.5	64	7.2	0	1016.5	984.2	0
11389	St. Pölten	270	2015-10-04T00:00:00+00:00	01:00	9.8	8.7	92	73	2.9	67	4.7	0	1016.4	984.1	0

Figure 4: A parallel coordinates view is provided to investigate correlations between data attributes and identify interesting patterns. It is linked to the table below by brushing and linking. In (a) we see that the two temperature measures 'dew point' (TPC) and 'temperature' (TC) are not strictly correlated. (b) reveals a strong negative correlation between 'temperature' (TC) and 'height' (Höhe m) and also (c) shows a strong negative correlation between 'station height' (Höhe m) and 'atmospheric pressure data at station level' (LDstat hPa). Another negative correlation can be observed in (d) between 'atmospheric pressure data' (LDred hPa) and 'peek wind speed' (WSG km/h). (e) on the other hand, shows a positive correlation between 'peek wind speed' (WSG km/h) and 'wind speed' (WG km/h) which seems very plausible. Besides investigating correlations, this view also allows for identifying groups of measurements with similar attributes.

different attributes need to be identified to be able to compute, for instance, the number of invalid values. Sometimes this cannot be solved without user intervention. Our study participants reported that they were missing support for specific data types such as currencies, IP addresses, E-Mail addresses, or geographic coordinates. While the number of data types supported by VoD can easily be extended, it might not be feasible to consider all data types possible. To tackle this problem, our participants suggested to include means that allow for an easy and straight-forward way to define new data types.

On a similar note, they wished for a possibility to include expert knowledge to define domain or business specific validation rules (e.g., valid value ranges). However, one of our participants stated that he is not familiar with regular expressions even though he has a technical background. Thus, regular expressions might not be a good choice for a general public either. A combination of a learning-by-example system and sophisticated guidance techniques that help the user to correctly formulate such rule sets might be more promising in this respect.

Our study participants especially praised the usefulness of the provided sparklines, parallel coordinates, bar charts, and box plots. They appreciated sparklines for being compact and informative, and they thought that the parallel coordinates plot, bar charts, and box plots were especially useful for getting a good overview of the data. The line charts of sorted values were neither mentioned positively nor negatively. While this combination of multiple visualizations works for our step wise approach, it would not be compact enough for tasks like comparing multiple datasets or visualizing the provenance of datasets that change over time [2]. These tasks would require very compact summary visualizations that can be used, for instance, as small multiples. Given the wealth of information necessary to get familiar with the structure and content of a dataset, this is a very challenging task which should be tackled in future research.

Another problem that comes with the design of approaches that are not tailored to a specific domain, but instead aim for providing a generic solution which can be applied to a variety of datasets, is that there is usually more work left to the user. In our case, this would be the definition of domain-specific rule sets and data types.

In a final remark we want to point out that most visualizations provided by VoD at its current state support sense making of numeric data attributes. String values and other more complex data types are supported only by the bar chart of distinct values and partly by the parallel coordinates plot. In future work we plan to investigate how these data types can also be visually summarized in an effective way.

5 Conclusion

In this paper we presented the design and evaluation of VoD, a VA research prototype with a special focus on supporting the user in the very first steps of getting familiar with an unknown dataset. To this end, we follow a step-wise procedure giving visual summaries of different granularity leves of the data set: (1) on a overall dataset level for understanding structure and overall amount of detected quality problems, (2) on the data attributes level for understanding data types, value ranges, general patterns, and correlations between data attributes, and (3) on the data value level for reasoning about data value distributions and the plausibility of abnormal values. We qualitatively evaluated VoD with four participants, which demonstrated the usefulness of our approach. From our evaluation results we derived interesting findings that point out open challenges and possible ways of how to proceed with this research. In conclusion, providing expressive summary visualizations of a dataset is an important topic not only for getting familiar with an unknown dataset but also for comparing different datasets or different versions of a dataset. While we support the user with the first task, it is still an unresolved research problems how to provide expressive summary visualizations that are compact enough for the comparison of multiple datasets.

Acknowledgments

This work was supported by the Centre for Visual Analytics Science and Technology CVAST, funded by the Austrian Federal Ministry of Science, Research, and Economy in the exceptional Laura Bassi Centres of Excellence initiative (822746).

References

[1] C. Arbesser, F. Spechtenhauser, T. Mühlbacher, and H. Piringer. Visplause: Visual data quality assessment of many time series using plausibility checks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):641–650, 2017. doi: 10.1109/TVCG.2016.2598592

[2] C. Bors, T. Gschwandtner, and S. Miksch. Visually exploring data provenance and quality of open data. In *Poster Proceedings of the Eurographics / IEEE VGTC Conference on Visualization (EuroVis 2018)*, p. 9–11. The Eurographics Association, The Eurographics Association, 2018. doi: 10.2312/eurp.20181117

[3] Data Manager. Data transformation, cleaning & cleansing. `http://datamanager.com.au/`. Retrieved at Sep 27, 2017.

[4] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresová, M. Ennemoser, A. Lex, and M. Streit. Taggle: Scalable visualization of tabular data through aggregation. *CoRR*, abs/1712.05944, 2017.

[5] 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*, i-KNOW '14, pp. 18:1–18:8. ACM, 2014. doi: 10.1145/2637748.2638423

[6] T. Gschwandtner and O. Erhart. Know your enemy: Identifying quality problems of time series data. In *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis '18)*, p. 10. IEEE, 2018.

[7] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch. A taxonomy of dirty time-oriented data. In *Multidisciplinary Research and Practice for Information Systems*, Lecture Notes in Computer Science (LNCS) 7465, pp. 58–72. Springer, 2012.

[8] D. Hoang. DataLadder - DataMatch 2017. `https://dataladder.com/data-matching-software/`. Retrieved at Sep 27, 2017.

[9] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, Oct 2004. doi: 10.1007/s10462-004-4304-y

[10] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI 2011)*, pp. 3363–3372. ACM, May 2011.

[11] S. Kandel, R. Parikh, A. Paepcke, J. Hellerstein, and J. Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proc. of the International Working Conference on Advanced Visual Interfaces (AVI'12)*, pp. 547–554, May 2012.

[12] W. Kim, B. J. Choi, E. K. Hong, S. K. Kim, and D. Lee. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7:81–99, 2003. doi: 10.1023/A:1021564703268

[13] J. E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, San Francisco, 1st ed., 2002.

[14] C. Perin, P. Dragicevic, and J. Fekete. Revisiting bertin matrices: New interactions for crafting tabular visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2082–2091, Dec 2014. doi: 10.1109/TVCG.2014.2346279

[15] E. Rahm and H. Do. Data cleaning: Problems and current approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 23:3–13, 2000. doi: 10.1145/1317331.1317341

[16] V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pp. 381–390. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

[17] Talend. Open Studio Integration Software Platform. `https://www.talend.com/products/talend-open-studio`. Retrieved at Sep 27, 2017.

[18] R. Verborgh and M. D. Wilde. *Using OpenRefine*. Packt Publishing, 1st ed., 2013.

[19] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. Keim. Visual analytics for the big data era A comparative review of state-of-the-art commercial systems. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 173–182, Oct 2012. doi: 10.1109/VAST.2012.6400554