Developing a formative visual feedback report for data brokering

Delvin Varghese[*] Newcastle University

Jonathan C. Roberts[†] Bangor University.

Panagiotis D. Ritsos[‡] Bangor University.

ABSTRACT

We present the development of a visualisation framework, used to provide formative feedback to clients who engage with data brokering companies. Data brokers receive, clean, store and re-sell data from many clients. However the usage of the data and the brokering process can be improved at source by enhancing the client's data creation and management processes. We propose to achieve this through providing formative feedback, as a visualisation report, to the client. Working closely with a travel agent data broker, we present a three-part framework, where we (1) evaluate data creation and provision processes of the client, (2) develop metrics for quantitative analytics on the data, (3) aggregate the analytics in a visual report.

Index Terms: H.5.2 [User Interfaces]: User Interfaces—Graphical user interfaces (GUI); H.5.m [Information Interfaces and Presentation]: Miscellaneous

1 INTRODUCTION

Travel data brokers are companies that buy and sell information from travel agents (henceforth referred to as clients of the broker). Typically, these brokers do not conduct intensive data analytics or data wrangling processes on these data, choosing instead to spend their resources on collecting, cleaning and generating analytics about the nature of the data.

Travel data is an especially challenging area, because while there are standard formats and guidelines to store the data, due to lack of regulation, many agencies don't adhere to them. Moreover, agencies utilise human operators to input the data, increasing the potential for errors. Finally, outdated systems are rife in this industry and the data that are exported are often disorderly and error-prone, requiring additional pre-processing before they are ready for use by another company. Data brokers are thus forced to spend precious resources to clean that data, and provide feedback to their client companies on how to refine and improve data collection. This is a costly and inefficient process for the data brokers, as they have to perform analytics on all incoming data, assess how fit for purpose the data is, and then get in touch with the sending agency – informing them to, for instance, change their data output formats or update their systems to generate data in an appropriate format, in the first place.

In this paper we discuss the idea of developing a formative feedback visualization as a report card, and consider what analytics and metrics are required, what types of feedback we can provide, explore how to develop the visualization itself through incremental design, and finally outline aspects of behavioural change and aspects of business process re-engineering. Our process for creating and utilizing the formative visual feedback report is shown in Fig. 1, where in three stages the data is analysed, metrics are applied, the feedback is generated, and delivered back to the clients. Building on this metaphor, we envision the relevance of this work for researchers and practitioners working with Data Brokers, who want to explore models for effective feedback among the stakeholders in the data brokering process.

[*]e-mail: D.Varghese2@newcastle.ac.uk

[†]e-mail: j.c.roberts@bangor.ac.uk

[‡]e-mail: p.ritsos@bangor.ac.uk

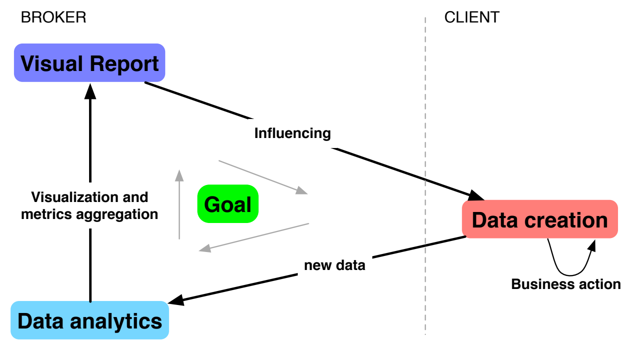

Figure 1: (A) *Data Creation*: Travel agencies generate the data and pass it to the data broker, (B) *Data Analytics*: the broker analyses the data for issues that affect its resale value (C) *Visual Report*: a report is generated which gives formative feedback to the agency on how to improve their business processes to produce better data.

2 DATA SCENARIO

Over the past two years we have been collaborating with The Data Exchange (DX), a company who buys and sells data, specifically for the travel industry. Every month DX receives travel agent data from several travel agents, cleans the data and then sells it on. While data cleaning is an important issue, our focus is to create better data at source. Because much of the data is created and managed by humans, our goal is to create an explanatory visualization that provides the clients with a report which gives formative feedback about aspects of their data processes that could be improved. Depending on where errors occur, the agency may need to change their procedures, policies or ways of storing the information.

To understand the challenges for formative feedback reporting of travel data, we need to explain some of the latter's intricacies. To allow the data to be bought and sold, the data is stored in comma-separated files, that are sent to the data-broker for clearing and storage (and in our case, also for a preliminary quantitative analysis and visualization of the dataset). Each file ranges in size from 5,000 to 40,000 lines of data, and approximately twenty files per month are stored. Because of agent consortia arrangements, one file may contain data from several travel agents. Separate lines of the file are used to store each transaction, change and leg of journey (this is a block of data). Consequently, validity checks need to be achieved across lines. Dates, often in varied format, need to be checked against the booking reference, to check consistency within a booking. The lines therefore represent types of transactions, and each block of data has its own type. Furthermore, a client may make any change to their booking over several months, because the data is saved in files, the booking may be in one file and amendment in another. Each of the blocks of data have associated attributes (ranging from 15 to about 40 values), often with strict requirements (e.g., IATA governed airport codes, which are often not unique). In fact, there are over 80 attribute fields in the dataset. One big challenge is that many of the attributes have default values, therefore they are often left out from the stored files. Fig. 2 shows three lines of typical data, indicating its complexity. The first line shows the field headings, the second line encodes a £30.60 ticket from London Heathrow to Edinburgh airport, and the final line presents a flight from Dubai to Heathrow costing £698.20.

The report card's purpose is provide a visual report of the data, for each client. The end-users (clients of the broker) may not need to interact with the data, but interaction may be required to help clarify some values. Brokers should be able to adapt the reporting parameters as they may have specific requirements, to say focus on one aspect of the travel data, or to focus on specific types of errors. We pair the characteristics of good reporting together, into three core requirements as identified in our collaboration with the data broker: (R1) Circumstances and Provenance, (R2) Absolute and relative quantities, and (R3) Locations and Temporal change. We used these requirements in the visualization design phase, in order to focus on the particular issues that our design needed to address. Good reporting in this domain needs to be able to 1) prioritise and visualize results for different agents in different ways (R1), 2) visualise and include both absolute values, and relative (normative) scores in the report (R2), and 3) provide visualizations that enable exact error positions to be located (R3).

3 Related Work

The concept of data quality in information systems has been discussed since the 70's [16], with subsequent efforts on: data improvement [17], management [22] and the consequences of poor data [29]. Consequently, data cleaning has become a keenly researched area. The survey by Rahm and Do [20] (although published in 2000) still provides a useful review, whereas recently researchers have focused on discrepancy detection [11] and data transformation [13]. Researchers have been calling for a tighter integration of visualization techniques into the data cleaning process itself [12]. Many of these systems use hybrid approaches and integrate editable tables alongside semi-automatic tasks. Example tools include Potter's Wheel [21], Wrangler [13], TimeCleanser [10], and products such as OpenRefine (http://openrefine.org/) or Tableau [15] provide methods to join, split and reshape tabular data.

In this paper we focus on the concept of 'data improvement' [17]. We use the phrase *quantitative analysis*, to refer to the large scale analysis of raw data to check compliance with industry best practices for data storage and handling. While the data we work with is particular to travel agents, we can still draw on principles from current practices in the data analytics community and also from extensive research that has been done in this area. There are different ways to explore and analyse data, and subsequently researchers have explored different dimensions [27] and various aspects of quality that can be measured and analysed. In this work, we have drawn on frameworks suggested by previous data quality practitioners [2, 18, 29] to inform our approach to assessing data quality.

1	AIRTRIP: actualFare, branchlD, conjsickNum, countryCode, currencyCode,BusLei, eTkt, FlightNo, hotelTag, IATA, invDate, LowestLogical Fare, netlnd, numSegs, OandD, 0rigin, Destin, pCarrier, pCarrierNum, resDate, segArriveDate, segCarrier, segClass, segDepartDate, segFare, segFareBasis, segFlightNum, Surcharges, surcharges, taxes, tickNum, tourCode, tripClass, contracted, ancil, ExchangeP/R, ExchangeA/C, AirTrip/Refund, clientldentifier
2	AIRTRIP: 30,6,, GB, GBP, not set, Y, 1442, 645488, 91278040, 01/01/2014,0, not Set, 1, not set, LHR, EDI, BA, 125, 01/01/2014, 02/01/2014,BA, K, 02/01/2014, 30,,144 2,, , 0, 4631677295, not set,Y, not set, not set, not set, not set, AirTrip, 10001092
3	AIRTRIP: 698.2, 6, GB, GBP, not set, Y, 5109, 645488, 91278040, 01/01/2014, 733.11,not set, 1, not set, DXB, LHR, EK, 176, 01/01/2014, 02/01/2014, EK, E, 02/01/2014, 698.2,, 5109, 13.2, 4631677296, not set, Y,not set, not set, not set, not set, AirTrip, 10001092

Figure 2: Example data. Line 1 represents attribute headers. Lines 2 and 3 represent flights (AIRTRIP) from Heathrow to Edinburgh in the UK, and from Dubai to Heathrow.

3.1 Behavioural Change through visualization

One of the challenges in data-error visualizations is that the errors seem minuscule in comparison with the size of the data [9]. This maps onto a practical visual depiction challenge, where with large quantities of data, any visual marks that represent the errors would not be noticeable in the visualization design. Consequently, researchers have investigated data abstraction methods such as sampling, aggregation and clustering [14, 26] to support scalable visualizations for large datasets.

Much work has been done in the field of Business Process Reengineering (BPR) and Business Process Management (BPM) to look at how visualisations can support organisations in modelling and dealing with changes in the behaviour that drives their business decisions. They are both well established fields, and BPM in particular looks at methods, techniques, and tools to support the "management and analysis of business processes" [28]. However, BPM tools often do not offer adequate visualisation support that is personalised to the different stakeholders involved in the process [7]. Little work has been done in how data-focussed organisations can help their partners/clients produce better outcomes. This in turn is beneficial to the broker as they receive better formatted data from the agents. Large scale empirical studies have looked at some of the issues when addressing the topic of behavioural change in businesses including perceived gaps between process design and process execution [3]. Organisations need support in addressing these gaps, and applications that look at analysing processes need to output information that is easily actionable by the organisation.

In relevance to our work, Proviado [5] is a framework for realising flexible and adaptable visualizations of business processes whose data may be scattered over multiple information systems. This project explored the business processes involved and how they are to be mapped in a visualisation application. We take this concept further by developing the idea of formative feedback alongside visual reporting of data processes.

4 Implementing the Data Analytics

As it is evident from the related work, there are many different ways to analyse data within business processes, often dependant on the underlying data. For instance, *accuracy*, *completeness*, *consistency* and *timeliness*, discussed by Ballou and Pazer [2] may be suitable dimensions of data quality of many datasets. Nonetheless, in our use case they are not well-suited. Data *completeness*, for example, is not easy to define in the context of multivariate, intertwining, travel data. Due to this limitation of existing quantitative data analysis characteristics, we have chosen to define a new set of characteristics, especially suited to our case study, along with specific visualization *requirements*. This is in line with the critique of generic visualization tools that are often utilised in business processes mapping and the lack of personalisation that is often possible [23]. We show these requirements diagrammatically in Fig. 3.

4.1 Data metrics

Based on the work of Müller and Freytag [18] we identified three data metrics to utilise in our quantitative analytics, and address our requirements. These metrics have been called *syntactic*, *semantic* and *coverage* errors (see Table 1). The names of the devised metrics were determined after discussion with the broker, and are used to 'identify' each of the error-checking algorithms, rather than being precisely descriptive of a particular technique. Syntactic errors consider problems with each data value. They answer questions like "Is the data in the right format", "Is the entry missing or null"? Semantic errors check the meaning of a row and check if it makes sense. For instance, if a row contains duplicate information to a previous row in a primary key column, then that row is inconsistent with the rules of the table. Coverage errors are statistics generated on special cases. These are not universal rules but rules that apply

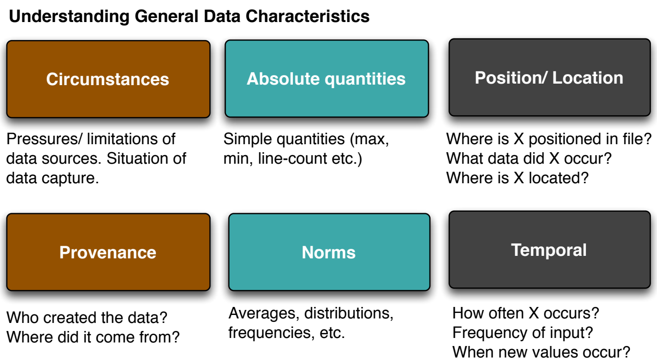

Figure 3: Summary of the key considerations, which describe the considerations for providing effective feedback on travel data. These are categorised in terms of their relation to the data itself, resulting metrics and the specific values in the dataset.

to particular datasets, and are provided as a way to make the error checker extensible, readily for future extensions.

4.2 Parsing & Data mapping

The travel data, originally stored as comma separated values (CSV) is parsed, line-by-line based on the blockType of the row, and the derived information is passed to the error checking component. At each stage error-checking occurs and the data is placed into the database. If errors are presented, they are logged in a relevant database. Major errors or differences to the data-mappings are flagged up to the user, who can manually intervene and map the data appropriately. Statistics on these errors are then calculated. Both the raw and calculated statistical values are passed to the visualization component to build the visual report. Once the mappings have been set, they are saved so that they can be easily reloaded for the next month. The agent identifier is used to automatically switch which mapping is used.

4.3 Data Score

One of the overarching goals of the broker was to define a holistic measure of veracity for a dataset. We name this a "data score".This represents the aggregation of the metrics used in the quantitative analytics. While DX understood potential limitations of such a measure (for instance, being a reductionist approach to data analysis), they were keen to get an overall view of each client so that the brokering team could easily communicate a quick assessment of the dataset's quality with each other. Our goal was therefore to create a heuristic function that would be consistent across different datasets. A weighted average of the error metrics was calculated and used as the final data score. The weighting was determined based on the relative importance assigned to each error by the broker. Despite the subjective nature of this method, it was preferred to alternatives as it gave full control to the broker.The error metric outputs are normalised between 1 and 100 and subtracted from 100 to generate the data score below (Equations 1 – 4).

$$S = 100 - (0.35 \times Mi + 0.13 \times Le + 0.22 \times In + 0.30 \times Fo) \quad (1)$$

$$M = 100 - (0.40 \times Co + 0.15 \times Du + 0.45 \times Se) \quad) \quad (2)$$

$$C = 100 - Sx \quad (3)$$

$$Final = 0.5 \times M + 0.4 \times S + 0.1 \times C \quad (4)$$

5 Report Card Design

Early in our design phase, we used the Five Design Sheet approach (FdS) [24] to collaborate with the broker, exploring various visualization designs. Design inspiration comes from different experiences and walks of life [25]. In our case, we found inspiration from environmental building reports (where the efficiency of a building is

Table 1: Error metrics that informed the analytics

	Symbol	Name	Description
Syntactic (*S*)	Mi	Missing value	Cell value is missing or null
	Fo	Format	Format of the cell does not match the specified attribute format
	Le	Length	Length of the cell exceeds or is below the specified size
	In	Integrity	Cell value is above or below the allowable range of values
Semantic (*M*)	Du	Duplicate	Column(s) are repeated on multiple rows
	Co	Contradictions	Cell value is lesser than the value of the specified column's value
	Se	Semi-empty	More than 50% of the columns in the row are empty
Coverage (*C*)	Ti	Ticket Num[1]	Calculates number of unique ticket numbers as a ratio of the number of rows in the dataset (applicable to AIRTRIP blockType only)
	To	Total Sum[2]	Sum of all values for this column
	Sx	Segment Excess	A particular ticket is repeated more than its allowable max entries (applicable to AIRTRIP blockType only)

[1,2] Ti and To are not used in the coverage (C) calculation, but are depicted on the visualization as supplementary statistics.

visualised by letters and coloured bars) and from progression reports given to students at the end of a teaching period. We split the space into four specific categories, and place individual visualization types in each. Individual reports can then be created for each client (meeting requirement R1) and detailed values are displayed in each segment (meeting requirement R2).

Our prototypes were implemented in Java and Processing as the broker used Java for most of their development. The design consists of five sections which followed closely the classification of data metrics, as described in Sec. 4.1. Referring to the labels on Fig. 4, the five sections are:

(i) Data Overview] (top-left) contains a block of text that aims to provide an overview of the data.

(ii) Coverage Errors] (top-right) shows a single bar chart visualization that depicts a comparison of the number of rows with unique ticket numbers, against rows that have duplicate ticket numbers.

(iii) Semantic Errors] (bottom-left) contains two visualizations showing different dimensions of the same error statistics. On the left side, three columns are used to depicting the position of errors (*contradiction*, *duplication* and *semi-empty* respectively) in the lines of the dataset. The data is binned into 20 blocks of colour (this value can be changed in the configuration file), enabling a quick understanding of where the errors are located in the files. The bar-chart on the right side of the Semantic errors section shows totals for each of the error categories.

(iv) Syntactic Errors] (bottom-right) are shown by a grid of cells. Again bins of 20 parts are used, but this visualization represents a scaled snapshot of the whole data file. The quality of the data is depicted by colour. The four thumbnails represent missing values, length errors, format errors and integrity errors, respectively.

(v) Final score] (middle) depicts the holistic quality measure as a percentage imposed over a pie-chart. This is in line with existing research which highlights that this is a powerful combination and helps the user gain precision in their understanding of the score [1].

Report Card Browser

We added a browser feature to the tool to allow the broker to inspect all quality reports, ordered (by worst or best). Users can explore more detail by brushing over specific visualizations. Individual (or a range of) reports can be printed out, and then sent to respective clients. This browser interface is also used to filter the report cards by the client's data characteristics. The Data Score (see Section 4.3) is prominently displayed for each report card. Comparisons can be

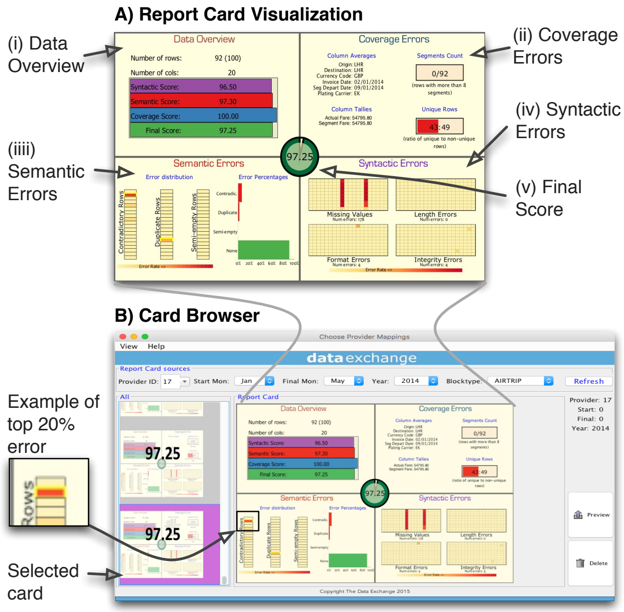

Figure 4: The final implementation of the report card quality visualization. A) the final design, with the updated colour-map. B) depicts the browser interface, where cards can be sorted and ordered.

made between multiple report cards by scrolling through the list of thumbnails, once the filtering has been done.

6 Evaluation

In addition to ongoing testing of our application, we conducted two types of evaluations: (1) tests on the data analysis (including speed testing and error checking) and (2) a usability evaluation of the whole system including the report card. As advocated in Munzner's seminal nested model for Viz validation [19], analyzing the "wall clock time and memory performance" of an implemented algorithm is important when there is a threat that the implementation is suboptimal to the target user's needs.

Thus, one important challenge was to make sure that the tool ran quickly enough. The time taken for the application to parse data and check errors is not only dependent on the number of rows but also on the number of errors encountered in said rows. Our data analysis code evaluates every line and field of the data against multiple criteria. For instance, the key attribute *invDate* has seven characteristics including the allowable range it can contain and the format (in this instance, it is a Date format). These characteristics provide the parameters that enable the *syntactic*, *semantic*, and *coverage* checks. Syntactic errors like *missing values* and *format errors* (incorrect data type) etc. are calculated on a cell-by-cell basis as the program iterates through the data, row-by-row. In the table extracted from the data, for instance, row attributes which contain empty columns (" ") or "not set" will be treated as *missing values*, and syntactic errors will be flagged for all such cases. Semantic errors like contradictions and duplicates are done on a row-by-row basis.

Another check we made was to the veracity of the analysis. We developed a test rig that checks the results against specific test data, comparing the error rate of the generated data quality score against the manually computed error rate of the dataset.

We also evaluated the usability of the Report Card generating application, using the System Usability Scale (SUS) [8]. We note that our evaluation includes the usability of the whole system (including loading data, setting mappings and weightings, deriving the visualization and browsing) and not just the formative feedback visualization part. Early pilot evaluations informed the design of the final implementation. As the travel data comes from a specific domain, we only had access to a few experts who fully understood the nuances of the data and would be able to create appropriate data weightings. We evaluated our final implementation with ten users, each experienced in data-visualization techniques, but with varied expertise with travel data. The resulting SUS score was 72.5, which is both an improvement on the previous evaluation, and is also deemed a 'usable program' [4]. In the post-experiment discussion, our experts commented: "the summary visualization enables you to see where the main problems are, it provides a quick way into the data". and that "while I'd like to see a more comprehensive browser interface, the report card visualization enables me to view the main errors quickly". One expert in particular indicated that "[the tool] is visually appealing and provides a nice summary, but I see myself referring to the generating tool constantly because it also provides me with the most error-prone sections. I may need this to quickly retrieve the error-filled section in question and use it as an example in my discussions with data providers".

7 Discussion & Conclusions

In this paper we present an overview of the design and development of a visualization framework and a corresponding tool for data brokers, intended for providing formative feedback on data veracity to their clients. Our framework for data brokering (Data Creation, Data Analytics, Visual Report) allows the user (broker) to analyse incoming data from travel agents using pre-defined error metrics and then creates a visual report for effective feedback. This feedback is then sent to the travel agents who can refine their Data Creation processes. Our reporting tool addresses our three visualization requirements, set in close collaboration with our expert clients, as follows. First we provided a mechanism to map and adjust the weightings of different error types. Secondly, the browser viewer enables specific reports to be loaded, and then compared with other agents' data. Thirdly, the summary view provides a quick insight for the user into the score of all reports, to depict which agent has the most (or least) erroneous data. Finally, exact error positions (R3) are shown through the greeking technique, and we bin the data into blocks that represent (say) 20 data lines, to give the user an idea of where the information is located. The summary of the number of errors below the syntactic grids, provides context by highlighting the quantity too. The final usability score supports that the visualization is usable.

While at the strategic level of a business an abstract overview or models of the processes may be desired, process actors "need a detailed view of those process parts they are involved in" [23]. This is why personalised approaches to visual reporting are required that help those involved in the daily operations of a data business to perform their role better. We thus included stakeholders throughout design process. A key element of this process is to remove aspects of the business process that are obfuscating to a stakeholder, removing complexity and alowing them to focus on aspects of the business pipeline that they need to be concerned with. Previous attempts have included use of dynamic visualization tools [6] to enable process actors to access relevant aspects of the business processes. We propose instead, an even simpler approach: static reports that are designed well and enable the process actor to quickly and at a snapshot access key information, minimising the need to decipher a process visualization. The final part of the analysis process (Fig. 1) is to *influence* the agents. For that we acknowledge the need for a longitudinal study, despite the fact we have already observed many improvements, through our collaborators.

Acknowledgements

We thank the Data Exchange for their collaboration and encouragement throughout this project. This work was partially funded by Access to Masters (ATM), a pan-Wales higher level skills initiative.

References

[1] Judgment Error in Pie Chart Variations. *Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization (EuroVis)*, pp. 91–95, 2016. doi: 10.2312/eurovisshort.20161167

[2] D. P. Ballou and H. L. Pazer. Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2):150–162, 1985. doi: 10.1287/mnsc.31.2.150

[3] W. Bandara, M. Indulska, S. Chong, and S. Sadiq. Major Issues in Business Process Management: An Expert Perspective. *ECIS 2007 - The 15th European Conference on Information Systems*, 2007:1240–1251, 2007.

[4] A. Bangor, P. T. Kortum, and J. T. Miller. An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6):574–594, 2008. doi: 10.1080/10447310802205776

[5] R. Bobrik, T. Bauer, and M. Reichert. Proviado – Personalized and Configurable Visualizations of Business Processes. *E-Commerce and Web Technologies*, 4082:61–71, 2006. doi: 10.1007/11823865

[6] R. Bobrik, M. Reichert, and T. Bauer. Requirements for the visualization of system-spanning business processes. *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, 2006(August):948–954, 2005. doi: 10.1109/DEXA.2005.165

[7] R. Bobrik, M. Reichert, and T. Bauer. Parameterizable views for process visualization. *Information Systems*, pp. 1–17, 2007.

[8] J. Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[9] Q. Cui, M. O. Ward, E. a. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):709–716, 2006. doi: 10.1109/TVCG.2006.161

[10] 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, New York, NY, USA, 2014. doi: 10.1145/2637748.2638423

[11] V. J. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004. doi: 10.1007/s10462-004-4304-y

[12] 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: Visualizations and Transformations for Usable and Credible Data. *Information Visualization*, 10(4):271–288, Oct. 2011. doi: 10.1177/1473871611415994

[13] 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*, CHI '11, pp. 3363–3372. ACM, New York, NY, USA, 2011. doi: 10.1145/1978942.1979444

[14] 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*, AVI '12, pp. 547–554. ACM, New York, NY, USA, 2012. doi: 10.1145/2254556.2254659

[15] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007. doi: 10.1109/TVCG.2007.70594

[16] J. Marting. *Security, Accuracy, and Privacy in Computer Systems*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1973.

[17] R. C. Morey. Estimating and improving the quality of information in a MIS. *Communications of the ACM*, 25(5):337–342, 1982. doi: 10.1145/358506.358520

[18] H. Müller and J.-C. Freytag. *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.

[19] T. Munzner. A Nested Process Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111

[20] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000.

[21] V. Raman and J. M. Hellerstein. Potter's Wheel: An Interactive Data Cleaning System. *Data Base*, 01:381–390, 2001.

[22] T. C. Redman. *Data quality: management and technology*. Bantam Books, Inc., 1992.

[23] M. Reichert. Visualizing large business process models: Challenges, techniques, applications. *Lecture Notes in Business Information Processing*, 132 LNBIP:725–736, 2013. doi: 10.1007/978-3-642-36285-9-73

[24] J. C. Roberts, C. J. Headleand, and P. D. Ritsos. *Five Design-Sheets – Creative design and sketching in Computing and Visualization*. SpringerNature, 2017. doi: 10.1007/978-3-319-55627-7

[25] J. C. Roberts, J. Yang, O. Kohlbacher, M. O. Ward, and M. X. Zhou. *Novel Visual Metaphors for Multivariate Networks*, vol. 8380 of *Lecture Notes in Computer Science*, pp. 127–150. Springer, 2014. doi: 10.1007/978-3-319-06793-3_7

[26] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, 2004. doi: 10.1109/INFVIS.2003.1249016

[27] G. Tejay, G. Dhillon, and A. G. Chin. Data quality dimensions for information systems security: A theoretical exposition. In P. Dowland, S. Furnell, B. Thuraisingham, and X. S. Wang, eds., *Security Management, Integrity, and Internal Control in Information Systems: IFIP TC-11 WG 11.1*, pp. 21–39. Springer US, Boston, MA, 2006.

[28] W. van der Aalst. Business process management: A personal view. *Business Process Management Journal*, 10:135–139, 2004. doi: 10.1108/bpmj.2004.15710baa.001

[29] R. Y. Wang and D. M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4):5, 1996. doi: 10.1080/07421222.1996.11518099