# Data Quality in Process Mining: A Rule-based Approach

R.M.E. van Cruchten[1]

[1] Tilburg University, Warandelaan 2 5037 AB Tilburg, The Netherlands

**Abstract.** "Garbage in – garbage out", a truism in any data analysis technique. Process mining, a form of data analysis that uses data from event logs, provides some unique data challenges in this respect, including missing events, event granularity, and case heterogeneity. Dealing with these challenges is often regarded as an a priori step and not as an integral part of the process analysis itself. This research proposes a novel and integral approach to data quality in process mining. By investigating existing techniques for data quality rule discovery, a more systematic approach is presented to measure and enhance event data quality. Moreover, a framework for the application of data quality and transformation rules will be investigated to create a more transparent and auditable data preparation approach. Lastly, the extent to which data quality rules can be used to express event data compliance will be investigated.

**Keywords:** Process Mining, Data Quality, Rule-based approach, Data pre-processing.

## 1 Introduction

Process mining is a relatively new technique that fills the gap between data mining at the one hand and business process analysis and modelling on the other. It aims to extract process-related knowledge from event data and enables an organization to discover, monitor and improve its processes (van der Aalst, 2011). The high dependency on information systems in business processes has created a situation in which the digital and the physical world are tightly connected. This connectivity has made it possible to store large amounts of data on the activities that are occurring in business processes, i.e., event data (van der Aalst, 2011). However, as with any data analysis technique the saying "garbage in, garbage out" holds true for process mining. The quality of event data has been recognized as a major challenge in applying process mining in practice (IEEE Task Force on Process Mining, 2012; Bose, Mans, & van der Aalst, 2013; Bose, van der Aalst, Žliobaitė, & Pechenizkiy, 2014; Suriadi, Andrews, ter Hofstede, & Wynn, 2017). Bose et al. (2013) identify four broad categories of event data issues namely: missing, incorrect, imprecise and irrelevant event data. The IEEE Task Force on Process Mining (2012) defines event data quality by two dimensions: (1) the level of abstraction of the events  and (2) the accuracy of the timestamp in terms of (i) its granularity, (ii) directness of registration and (iii) correctness (Andrews, Suriadi, Chun, & Poppe, 2018). While the event data quality frameworks as presented by Bose et al. (2013) and IEEE Task Force on Process Mining (2012) are useful for classifying and describing the impact of various data quality issues, they do not provide any guidance on how to identify or address them. Moreover, methodologies for applying process mining in practice, such as the Process Mining Project Methodology (PM2) by van Eck, Lu, Leemans, & van der Aalst (2015) or the L* life cycle model by IEEE Task Force on Process Mining (2012), only mention the importance of event data quality but do not define addressing data quality as an explicit step in their methodologies. Furthermore, research towards systematically addressing event data quality challenges is scarce (Andrews et al., 2018). Traditionally, the database field has made use of integrity constraints in the form of business rules to enforce data quality. The application of business rules is no new topic in the field of computer science as well but has yet to find its way in the field of process mining. This research will therefore address the following research question: how can event data quality in process mining be systematically addressed using a rule-based approach. The remainder of this paper is structured as follows. In section 2 the research approach will be discussed. In section 3 a rule-based approach will be elaborated on and in section 4 a process mining framework of the proposed rule-based approach is presented. Section 5 presents the conclusion and future work.

## 2 Research Approach

This research will apply a design science approach since it aims to create and evaluate an artifact intended to solve an identified organizational problem (Wieringa, 2014), namely the need to systematically address event data quality. Two type of artifacts will be designed:

1. Method(s) to identify and solve event data quality issues
2. A methodology for applying the designed method(s) to address event data quality.

The requirements for the design artifacts are as follows: the designed artifacts are

- Systematic
- Automated
- General applicable, i.e., domain and system agnostic
- Transparent, so they can be easily audited.

Lab experiments and case studies using real-life event data will be used to develop and validate the designed artifacts. The use of event data from real-life environments to design process mining artifacts is something that has not been done extensively. Moreover, using real-life event data will contribute to the applicability of the designed artifacts in practice.

## 3 A Rule-based Approach to Data Quality

The relationship between event data quality and process mining results is obvious. However, dealing with data quality is often seen as an a priori, laborious activity that requires a lot of manual effort. Research towards a systematic and generalizable approach in addressing the identified quality issues is scarce. Suriadi et al. (2017) and Andrews et al. (2018) are two recent approaches towards systematically identifying and addressing event data quality issues. Both papers however recognize the need for further research towards systematic approaches. This research will address this research gap by focusing on how data quality rules can be systematically applied to repair data to improve data quality.

### 3.1 Rules as a uniform language

Dependency theory is as old as relational databases themselves. Data dependencies, also called integrity rules or data quality rules, provide a uniform logical framework to describe and define data quality rules (Fan & Geerts, 2012). Conditional functional dependencies (CFDs) are an extension of the traditional functional dependencies (FDs) that make use of patterns of semantically related constants to create stricter rules aimed at improving the quality of the data. For example CFD1: ([Country = NL, ZIP =] □ [Street]). In this case CFD1 is an extension of the FD, meaning the combination of Country and ZIP uniquely identify Street, that holds on the subset of records that satisfy the pattern Country = NL. CFDs allow for rules to be more specifically defined, creating stricter rules that are able to discover semantic errors. Moreover, these CFDs can be applied to repair data in a semi-automatic way by discovering dirty records and suggesting the correct value to a user for inspection before a record update (Fan & Geerts, 2012).

### 3.2 Discovering data quality rules

Chiang & Miller (2008) demonstrated a decade ago that conditional functional dependencies (CFDs) can also be mined from a dataset and subsequently be used as data quality rules to measure and improve data quality. While CFDs are much used in practice for data cleaning, research towards the discovery of CFDs has not been conducted extensively (Rammelaere & Geerts, 2019). Defining a set of integrity constraints that reflect an organization's business rules and domain semantics is often a very time consuming effort in which business experts having knowledge of that domain are extensively consulted. Discovery techniques that can (partially) automate this time consuming effort are thus of added value. Furthermore, domain specific rules may exist in the dataset that users are not aware off but can still be useful in enforcing semantic data consistency (Chiang & Miller, 2008). Discovery of rules therefore provide a more unbiased approach in data quality rule definition. However, it must be noted that rule discovery techniques cannot guarantee to produce a set that is complete since it is not possible to absolutely determine the complete spectrum of possible data issues (Suriadi et al., 2017). Therefore, manual validation and refining of the discovered data quality rules will still be important. This research will investigate whether tacit process and domain knowledge can be formalized by mining CFDs, or other forms of integrity rules, from event data in a semi-automatic way, e.g., by mining a set of rules and validating them with domain experts.

### 3.3 Multi-level perspective on rules

While traditional integrity rules are focussed on identifying data issues at a record level (i.e., “intra-record”), event data provides a unique data quality challenges because of the notion of cases (i.e., a subset of records that define a single process instance). For example, identifying missing events in a case requires “inter-record” integrity rules. Such rules could be based on business rules that enforce certain process execution (van Cruchten & Weigand, 2018). Thus, it is proposed that integrity rules can be defined at both the recorded system event and case level. Moreover, if it is required to perform transformations on the recorded system events such as semantic labeling (Alves de Medeiros et al., 2007), aggregating events (Smirnov, Reijers, & Weske, 2012; Montani, Leonardi, Striani, Quaglini, & Cavallini, 2017) or mapping events to a different level of abstraction (Baier, Mendling, & Weske, 2014; Tax, Haakma, Sidorova, & Aalst, 2016), integrity rules could also be defined at the transformed event level. Having intergrity rules at this level enables the measurement of the data quality before and after the transformation effort. Thus, a multi-level perspective on integrity rules is proposed in which rules can be defined at a system event, event, case and model level as shown in Figure 1.

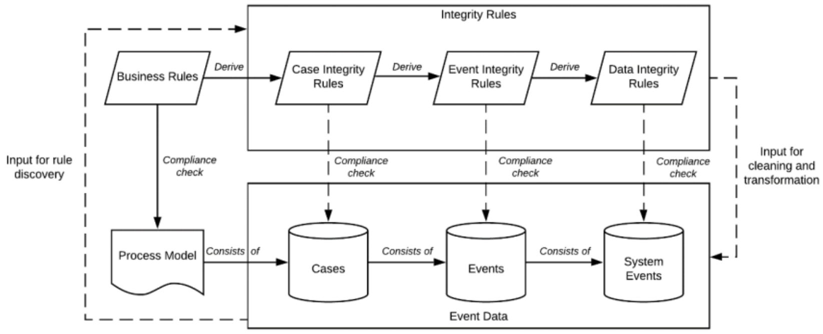

**Fig. 1.** Multi-level perspective on integrity rules

### 3.4 Rule based data transformation

When comparing the event as recorded in an information system to the events in a process model as defined by the organization, one often faces a difference in level of abstraction. Reason for this is that a model is created as an abstract of reality while information systems record events at a detailed and finer level of granularity (Baier et al., 2014). This difference in granularity can lead to misinterpretation of process mining results as the discovered process model is less understandable from a business user perspective. To bridge the difference in level of granularity the recorded system events should be transformed to understandable business events, which will result in more understandable process mining results (Jareevongpiboon & Janecek, 2013). Moreover, it is argued that event data transformations should be rule-based so that the applied domain logic is more transparent and thus auditable, so that the quality and integrity of the transformed data can be guaranteed (van Cruchten & Weigand, 2018). Rule-based data transformation has not been researched extensively (Claes & Poels, 2014; Leonardi, Striani, Quaglini, Cavallini, & Montani, 2018; van Cruchten & Weigand, 2018; Suriadi et al., 2017;). Claes & Poels (2014) apply rules in merging inter-organizational event logs and the notion of rule- and ontology-based transformation is also proposed by (Leonardi et al., 2018). Previous work by van Cruchten & Weigand (2018) has successfully demonstrated rules can be used for both cleaning data as well as transforming process "unaware" data (i.e., data that is not stored with the intentional goal of process logging) to a higher level of abstraction.

## 4 Revised Process Mining Framework

Storing the discovered and/or defined rules in a repository will facilitate that these rules can be used in a systematic approach to various event log preparation activities (e.g., cleaning, transformation, abstraction) regardless of the type of data analysis to be performed (Suriadi et al., 2017). Thus, it is proposed that the well-known process mining positioning framework by van der Aalst (2011) is to be extended with a rule-repository as shown in Figure 2.

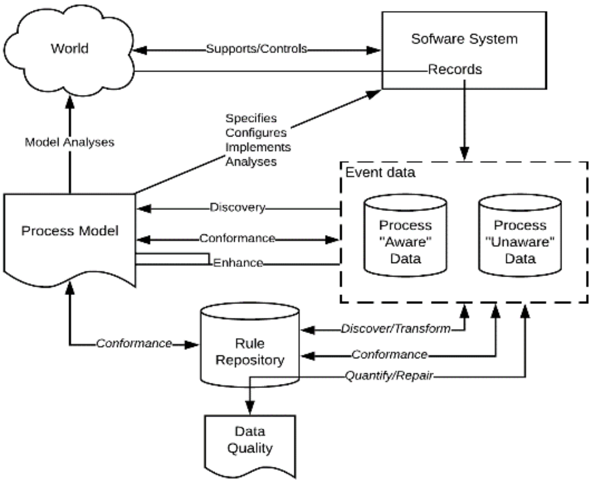

**Fig. 2.** Positioning and extension of the three main types of process mining: discovery, conformance and enhancement. Adapted from Process Mining: Discovery, Conformance and Enhancement of Business Processes, by W.M.P. van der Aalst, 2011 (p.9)

This rule repository could also serve as a conformance checking mechanism that expresses the compliance of the process from a data perspective rather than a control-flow perspective. Or put differently, compliance as a process mining goal should be seen as a multi-level concept, in which the control-flow perspective is the highest level. The added value of compliance checking at the event data level is that no data is left out of the analysis (i.e., data of “bad” quality is also considered in the compliance analysis). It therefore provides more complete and empirical evidence for compliance, which is important if one is to apply process mining in for example auditing (Caron, Vanthienen, & Baesens, 2013). Figure 2 can thus be regarded as the outline of a framework for the application of a systematic, rule-based approach to event data quality improvement and compliance checking. Moreover, the existing PM2 project methodology by van Eck, Lu, Leemans, & van der Aalst (2015) will be revised to incorporate more specific data preparation steps to provide process mining practitioners with more guidance on this topic.

## 5 Conclusion

This research proposes a novel systematic, rule-based approach to address event data quality. Specifically, qualitative data cleaning techniques will be investigated to see if data quality rules can be mined from event data and subsequently be applied to measure and improve event data quality. Furthermore, method(s) to apply rules in data transformation will be designed, to create a more transparent and auditable data preparation approach. The application of rules in expressing compliance from an event data perspective will be investigated as well, along with a framework that defines the activities to apply such a rule-based approach to data quality in practice.

## References

Alves de Medeiros, A. K., Pedrinaci, C., van der Aalst, W. M. P., Domingue, J., Song, M., Rozinat, A., … Cabral, L. (2007). An Outlook on Semantic Business Process. In *OTM Confederated International Conferences “On the Move to Meaningful Internet Systems”* (pp. 1244–1255). Retrieved from http://dx.doi.org/10.1007/978-3-540-76890-6_52

Andrews, R., Suriadi, S., Chun, O., & Poppe, E. (2018). Towards Event Log Querying for Data Quality Let’s Start with Detecting Log Imperfections. In M. R. Panetto H., Debruyne C., Proper H., Ardagna C., Roman D. (Ed.), *On the Move to Meaningful Internet Systems. OTM 2018 Conferences. Lecture Notes in Computer Science, vol 11229* (pp. 116–134). Springer, Cham. https://doi.org/10.1007/978-3-030-02610-3

Baier, T., Mendling, J., & Weske, M. (2014). Bridging abstraction layers in process mining. *Information Systems*, *46*, 123–139. https://doi.org/10.1016/j.is.2014.04.004

Bose, R. P. J. C., Mans, R. S., & van der Aalst, W. M. P. (2013). Wanna Improve Process Mining Results ? It’s High Time We Consider Data Quality Issues Seriously. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2013)* (pp. 127–134). https://doi.org/10.1109/CIDM.2013.6597227

Bose, R. P. J. C., van der Aalst, W., Žliobaitė, I., & Pechenizkiy, M. (2014). Dealing With Concept Drifts In Process Mining Using Event Logs. *IEEE Transactions* *on Neural Networks and Learning Systems*, *25*(1), 154–171. https://doi.org/10.1109/TNNLS.2013.2278313

Caron, F., Vanthienen, J., & Baesens, B. (2013). Comprehensive rule-based compliance checking and risk management with process mining. *Decision Support Systems*, *54*(3), 1357–1369. https://doi.org/10.1016/j.dss.2012.12.012

Chiang, F., & Miller, R. J. (2008). Discovering Data Quality Rules. *Proceedings of the VLDB Endowment*, *1*(1), 1166–1177. https://doi.org/10.14778/1453856.1453980

Claes, J., & Poels, G. (2014). Merging event logs for process mining: A rule based merging method and rule suggestion algorithm. *Expert Systems with Applications*, *41*(16), 7291–7306. https://doi.org/10.1016/j.eswa.2014.06.012

Fan, W., & Geerts, F. (2012). Foundations of Data Quality Management. In *Synthesis Lectures on Data Management* (pp. 1–217). Morgan & Claypool Publishers. https://doi.org/10.2200/S00439ED1V01Y201207DTM030

IEEE Task Force on Process Mining. (2012). *Process mining manifesto*. *Business Process Management Workshop 2011* (Vol. 99). Springer-Verlag. https://doi.org/10.1007/978-3-642-28108-2_19

Jareevongpiboon, W., & Janecek, P. (2013). Ontological approach to enhance results of business process mining and analysis. *Business Process Management Journal*, *19*(3), 459–476. https://doi.org/10.1108/14637151311319905

Leonardi, G., Striani, M., Quaglini, S., Cavallini, A., & Montani, S. (2018). Leveraging semantic labels for multi-level abstraction in medical process mining and trace comparison. *Journal of Biomedical Informatics*, *83*, 10–24. https://doi.org/10.1016/j.jbi.2018.05.012

Montani, S., Leonardi, G., Striani, M., Quaglini, S., & Cavallini, A. (2017). Multi-level abstraction for trace comparison and process discovery. *Expert Systems with Applications*, *81*, 398–409. https://doi.org/10.1016/j.eswa.2017.03.063

Rammelaere, J., & Geerts, F. (2019). Revisiting Conditional Functional Dependency Discovery : Splitting the “C” from the “FD”. In I. G. Berlingerio M., Bonchi F., Gärtner T., Hurley N. (Ed.), *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11052.* Springer, Cham.

Smirnov, S., Reijers, H. A., & Weske, M. (2012). From fine-grained to abstract process models: A semantic approach. *Information Systems*, *37*(8), 784–797. https://doi.org/10.1016/j.is.2012.05.007

Suriadi, S., Andrews, R., ter Hofstede, A. H. M., & Wynn, M. T. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information Systems*, *64*, 132–150. https://doi.org/10.1016/j.is.2016.07.011

Tax, N., Haakma, R., Sidorova, N., & Aalst, W. M. P. Van Der. (2016). Event Abstraction for Process Mining using Supervised Learning Techniques. In B. R. Bi Y., Kapoor S. (Ed.), *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016. IntelliSys 2016. Lecture Notes in Networks and Systems, vol 15.* (pp. 161–170). Springer, Cham. https://doi.org/10.1007/978-3-319-56994-9_18 van Cruchten, R. M. E., & Weigand, H. (2018). Process Mining in Logistics: The need for rule-based data abstraction. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)* (pp. 1–9). https://doi.org/10.1109/RCIS.2018.8406653

van der Aalst, W. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Berlin Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-19345-3

van Eck, M. L., Lu, X., Leemans, S. J. . J., & van der Aalst, W. M. P. (2015). PM2: A process mining project methodology. In *International Conference on Advanced Information Systems Engineering* (pp. 297–313). https://doi.org/10.1007/978-3-319-19069-3_19

Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Berlin Heidelberg: Springer-Verlag. https://doi.org/10.1145/1810295.1810446