RE: CAiSE 2012 Paper Notification [268]

Dear Camille,

In this document we want to explain how we incorporated the comments of the reviewers on our paper #268 „Data Transformation and Semantic Log Purging for Process Mining“ (L.T. Ly, Conrad Indiono, Jürgen Mangler, Stefanie Rinderle-Ma) which has been conditionally accepted for the CaISE 2012 conference.

As summarized in your email, we particularly addressed the following issues:

1- `provide a better comparison with data cleaning literature`
In the related work section (Section 6), a paragraph discussing approaches from the database area on data cleaning utilizing integrity constraints has been added (including the associated references). The main differences between these approaches and the approach presented in this paper are elaborated, i.e., first of all, process-oriented data (represented by process logs) has not been subject to data cleaning approaches so far. Secondly, integrity constraints are merely application-independent. In this paper, we present a data cleaning approach that utilizes knowledge on the application context of the data. In practice, the existence of such knowledge can often be assumed, for example, medical guidelines [12], internal quality controls, or contractual policies.

2- `add a discussion on the impact of the method on overlooking important details about processes that are currently regarded as "noise"`
"A detailed description of possible kinds of noise in the context of process mining has been added to the introduction (page 2), i.e., (a) Incorrect/incomplete log data (in the following addressed as noise), (b) log data contributed by parallel branches, (c) infrequent traces, (d) log data contributed by ad-hoc changed instances. It is further discussed how these cases are tackled by the approach, for example, infrequent cases are not filtered out by our approach, since they do not violate constraints (contrary to existing frequency-based process mining algorithms). A further discussion on how the proposed approach tackles cases such as infrequent cases, parallel executions resulting from late modeling in processes, and ad-hoc changed instances has been added in Section 2.1.

3- `discuss the practicality of finding good constraints`
The results of process mining with semantic log purging depend on the constraints chosen. Often process constraints are already defined within the domain of interest, e.g., internal quality controls, regulatory packages, or medical guidelines, and can be utilized for semantic log purging. Further, constraints can be elaborated together with domain experts as in the case study presented in this paper. This can be particularly useful within an iterative process where the results of process mining with semantic log purging are reported back to the domain experts. In this sense, we also experimented with different sets of constraints. Another issue is that constraints might have a different enforcement level. We have also incorporated this into our experiments.

For a detailed description of our revisions see the following table.

Overall, we addressed the issue of novelty of our approach by extending the related work discussion with particular focus on existing data cleaning approaches. Our approach is of interest for practitioners since it presents a way to improve process mining results. It is realistic to assume the existence of suitable process constraints as we know from many domains and projects conducted with partners from practice, e.g., in the health care domain.

We hope that our revisions are satisfactory.

Thank you very much and best regards,

Thao, Conrad, Jürgen, Stefanie

Reviewer Comments	Our revisions
Reviewer 1 This article presents a technique for data transformation and cleaning. The feasibility of the approach is describes with a case study. The approach applies in principle domain knowledge in form of process constraints to verify process log data. Main conclusion is that the approach will allow process mining to yield better results.	
The paper is well written and clearly presented, however overall I suggest the authors should rather **focus on the conceptual concept than on the technical implementation details.**	To improve the paper, we discussed the findings in 5.1 in more detail and detailed the improvement of the mined models. We further included a discussion on the method to obtain constraints.
Furthermore, and as main limitation of the paper, the approach uses domain knowledge in form of defined constraints to improve and "cleanse" process logs (Log purging). As such, the **novelty of the approach appears to be simplistic** and the results can somehow be expected. The constraints have to be defined by domain experts, in principle a concept long established in database systems. Therefore the **innovativeness of the approach needs to be emphasised and further discussed in the paper.**	In the related work section (Section 6), a paragraph discussing approaches from the database area on data cleaning utilizing integrity constraints has been added (including the associated references). The main differences between these approaches and the approach presented in this paper are elaborated, i.e., first of all, process-oriented data (represented by process logs) has not been subject to data cleaning approaches so far. Secondly, integrity constraints are merely application-independent. In this paper, we present a data cleaning approach that utilizes knowledge on the application context of the data. In practice, the existence of such knowledge can often be assumed, for example, medical guidelines [12], internal quality controls, or contractual policies.
Reviewer 2: The approach is semantic rather than statistical in that it helps the process analysis by starting from a kind of reference models, as	In the related work section (Section 6), a paragraph discussing approaches from the database area on data cleaning utilizing integrity

usual in many business process engineering exercises; **moreover, plausibility constraints similar to the typical integrity constraints in data cleaning are exploited to search for noise to be purge**d. A fundamental issue is that the approach is focussed on **mining for the "normal case behavior", purging exceptions and exception handling**. This can be a strength as well as a weakness, as some important elements of a process might occur only rarely but are crucially important. Consider e.g. the bank process rule that cheques with very high values need to be signed by two independent persons. This might only rarely occur in a trace but would still be essential to be included in the process model. It therefore could be dangerous to consider such trace elements as "noise". **It would be helpful for readers to discuss this issue in the paper**. The technical part of the paper is a detailed description firstly of the extraction programs (using functional programming in their prototype), then of the constraint-based log purging. A case study evaluation shows that the proposed approach results in similar, yet more structured process models than mining the uncleaned process traces. It can be noticed that the related work section **refers only to papers from the very closest scientific community** of the authors, i.e. two or three research groups. A somewhat **broader look at the literature might help the readers in positioning this work**. Overall, I like the paper but believe it could benefit from a more **careful discussion of its underlying assumption and better comparison with related experiences** from the data cleaning literature.	constraints has been added (including the associated references). The main differences between these approaches and the approach presented in this paper are elaborated (see also extended comment on this issue for reviewer 1). A description of challenges for process mining has been added to the introduction (page 2). It contains the cases mentioned by the reviewer, i.e., infrequent cases. It is further discussed how this case is tackled by the approach: infrequent cases are not filtered out by our approach, since they do not violate constraints (contrary to existing frequency-based process mining algorithms). A further discussion on how the proposed approach tackles cases such as infrequent cases, parallel executions resulting from late modeling in processes, and ad-hoc changed instances has been added in Section 2.1. In the related work section (Section 6), a paragraph discussing approaches from the database area on data cleaning utilizing integrity constraints has been added (including the associated references). The main differences between these approaches and the approach presented in this paper are elaborated (see also extended comment on this issue for reviewer 1)..
Reviewer 3: The paper is well written with a few minor problems. Several of the **figures are not readable** which makes it difficult	We resized the figures.

to understand the examples and the tools.	
The manually defined reference model and the moned models **are represented using different formalisms which makes it difficult to compare the results**.	This is due to the process mining tool that does not support BPMN. However, the causal dependencies in the models can still be compared quite well. To make this more intelligible, we included more details on the qualitative analysis in Sect. 5.1.
In Table 2 the Milestone subprocess has 20 nodes in the model generated by the original logs and 21 in the model generated by the purged logs. **How is this possible**?	We revised the results from qualitative analysis and normalized the results (removing start and end nodes). We further describe in more detail the findings of the analysis to make the results more intelligible.
Intuitively, the results from the data purging will **depend heavily on the defined constraints. No discussion on how to choose a good set of constraints** is presented and no investigation on how sensitive the method is of the choice of constraints. **Phases like "fundamental semantic constraints" and "minimal well chosen set of constraints" are used without definition or explanation** on what is meant or methodology on how to formulate such constraints.	The results of process mining in combination with semantic log purging depend on the constraints chosen. Often process constraints are already defined within the domain of interest, e .g., internal quality controls, regulatory packages, or medical guidelines [12], and can be utilized for semantic log purging. Further, constraints can be elaborated together with domain experts as in the case study presented in this paper. This can be particularly useful within an iterative process where the results of process mining with semantic log purging are reported back to the domain experts. In this sense, we also experimented with different sets of constraints. Another issue is that constraints might have a different enforcement level. We have also incorporated this into our experiments To improve the paper, we included a discussion on choosing constraints in Sect. 4 and Sect. 5.2.
The discussion on how the proposed method for log generation compares to **XESame needs to be elaborated** - what exactly are the differences and/or in what ways is the proposed method is better.	We added Section 3.5 to differentiate the XESame tool from the data transformation and querying tool as presented and utilized in this paper.