Data Inaccuracy-aware Design of Business Processes

Yotam Evron[1]

University of Haifa, Mount Carmel, Haifa 3498838, Israel
yevron@is.haifa.ac.il

Abstract. Business processes are designed with the assumption that the data used by the process is an accurate reflection of reality. However, this assumption does not always hold, and situations of data inaccuracy might occur which bear substantial consequences to the process and to business goals. Until now, data inaccuracy has mainly been addressed in the area of business process management as a possible exception at runtime, to be resolved through exception handling mechanisms [9]. Design-time analysis of potential data inaccuracy has been mostly overlooked so far. In this paper we describe a research agenda for developing a method for supporting process modelers in designing more robust processes with respect to data inaccuracy.

Keywords: Business Process Modelling, Data in Business Processes, Data Inaccuracy, Formal Analysis

1 Introduction

Business processes consist of activities, which are executed by humans and by resources with the support of information systems, in order to accomplish specific business goals. An information system holds data which should inform process participants about the current state and serve as a basis for decisions and for selecting the process paths to be taken. An underlying implicit assumption is that the data as recorded and used by the system is an accurate representation of the values in the real world. Based on this assumption, process aware information systems (PAIS) can operate as a closed system and actions can be triggered based on the data values in the system, without a need to actually "sense" the values in the real world. Nevertheless, it is a well-known fact that the data stored in the database of an information system is not always completely reliable [2], and situations may occur when data values do not match the real-world values they should reflect. Those inaccurate data values may affect the ability to reach the process goals.

We use the term data inaccuracy to refer to such situations. Since the data stored in an information system has an impact on business goals, it is important to investigate such situations to improve processes. While several works have addressed the issue of

[1] Supervised by Pnina Soffer and Anna Zamansky.

data quality at process design [3][8][12], to the best of our knowledge, no existing modeling formalism provides explicit support for representing and reasoning about potential data inaccuracy at design time. Until recently, data inaccuracy has mainly been addressed in the area of business process management as a possible exception at runtime [9].

The overreaching goal of this research is to develop a method that will assist process designers to build more robust processes, resilient to data inaccuracy. This will be done by incorporating considerations of data inaccuracy at business process design. In this paper we present a detailed research plan for achieving this goal.

The rest of the document is organized as follows: Section 2 provides background and related work, Section 3 introduces the main research questions. Based on these questions, the approach and research methodology are presented in Section 4. In Section 5, we discuss the evaluation methods. In Section 6, the conclusions and expected contributions are discussed.

2 Related work

Data quality has been extensively addressed outside the context of business processes [2][6]. In business process research, limited work concerning data quality has been done. Rodriguez et al. [8] introduced a data quality-aware model which allows modelers to specify data quality requirements in business process models. Their approach is semi-formal and without a conceptual or analytical method. Gharib and Giorgini [4] introduce a goal-oriented approach to model and analyze information quality (IQ) requirements in business processes from a socio-technical perspective. Their work mainly focuses on human interaction without a comprehensive support including computational methods. Data accuracy is an important dimension of data quality. Soffer [12] was the first to suggest an explicit analysis of data inaccuracy at design time, providing a conceptual formulation of the problem and discussing potential consequences of data inaccuracy in business processes. This provides the main conceptual basis for our research.

3 Research Question

The main objective of this research is to develop an approach to enable assessing and reducing data inaccuracy consequences during process design. Our main research question is:

How can we incorporate considerations of data inaccuracy in business process design?

This question gives rise to several sub questions:

1. How can we formally represent a business process in a way that will enable reasoning about data inaccuracy?
2. How can we utilize this formal representation to identify relevant properties of processes that might be associated with data inaccuracy?

3. How can we utilize this representation and properties as a basis for a method that will support process designers in developing processes that are more robust to data inaccuracy?

4 Approach

4.1 Basic Premises of Our Approach

Following Soffer [12], we take the state-based view of the Generic Process Model (GPM) [11], in which a process takes place in a domain which is typically captured by a set of state variables, whose values at a given moment reflect the domain state at that moment. A subdomain is part of the domain described by a subset of the domain state variables. Note that there are many ways to partition a given domain into subdomains and different partitions can reflect different views of the process domain. A process is viewed as a sequence of state transitions, which are governed by a transformation law. However, not all state variables are relevant (or need to be "sensed") in order to make a transition. Thus, the domain may be decomposed into independent subdomains in some parts of the process.

Observation 1. A process may involve multiple sequences of transitions in several independent subdomains.

As explained in [11], subdomains which operate in parallel or independently, may reach a state where a dependency between them exists. Considering the threads of transitions in these subdomains, we call such states *synchronization points*.

Observation 2. Sequences of transitions which take place concurrently in independent subdomains merge at synchronization points.

In many cases some of these subdomains include only state variables (of the physical world), while others include and rely on data items stored in information systems, under the assumption that they reflect corresponding domain variables[2]. However, as already noted above, this is not always realistic:

Observation 3. A discrepancy between a state variable value x_i and its corresponding data item d_i might occur.

In what follows we refer to such discrepancy as data inaccuracy. As long as a subdomain relying only on a state variable x_i operates independently of a subdomain which includes d_i (and vice versa), the existence of data inaccuracy (discrepancy between x_i and d_i) cannot be recognized. It will only be recognized at a synchronization point between these two subdomains. However, it may be too late and the process might get "stuck", or some compensating action would be needed. Therefore, the exact place of a synchronization point matters for mitigating risks of data inaccuracy.

[2] We make a basic assumption of a well-designed data structure, which means that all relevant domain variables are represented by corresponding data items.

As a running example, consider a process of organizing and executing a business party by a catering company. From the company's perspective, the goal of the process is reaching a state where the business party is completed successfully. Let us assume a customer met the company's representatives, who recorded the agreed upon details (such as date, type of food, price etc.) as data in the company's IS. Now the company can execute the process on its own without a need for any further information in order to proceed. This reflects Observation 1: the company executing a part of the process is an independent subdomain, depending solely on the IS data items, without "sensing" their values outside the IS (e.g., by comparing the date registered in the IS to the date known to the customers). At some point, however, when the party eventually takes place, the company's independent subdomain and the customer's subdomain merge, coming to a synchronization point (Observation 2). Now consider a scenario, in which the planned date has been falsely recorded in the information system and does not reflect the actual agreed upon date. This is a manifestation of data inaccuracy (Observation 3). As noted, it is not recognized as long as each subdomain operates independently, but will necessarily be detected at the synchronization point: the company will have everything ready for the planned date recorded in the IS, while the guests will arrive at the meeting point on the agreed upon date as they know it (see "Party execution" in **Fig. 1**). Since these do not match, the party cannot take place.

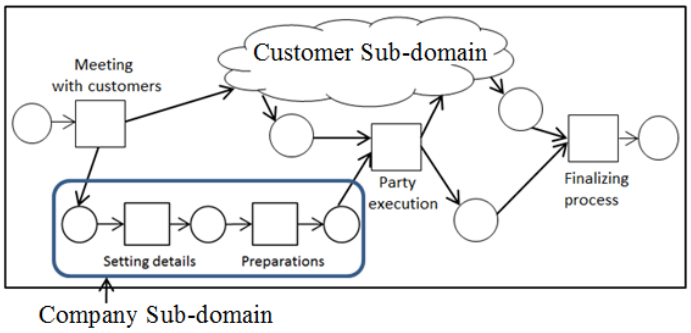

Fig. 1. Conceptual visualization of our example

This example highlights the important role synchronization points play in the detection and handling of data inaccuracy situations.

4.2 Formal Representation of Processes with Data Inaccuracy

Based on GPM, we view a process as a sequence of transformations, in which state variables x_i obtain values and data items d_i are updated to reflect these changes, and as a result data inaccuracy may be introduced in d_i with respect to x_i. In this case, inaccuracy cannot be detected in an independent subdomain relying only on d_i until it reaches a synchronization point with a subdomain containing x_i. This needs to be explicitly captured in formal representations of processes.

While the GPM view is useful for our general conceptual understanding of data inaccuracy, it does not provide computational analysis capabilities needed for analyzing

models of specific processes. For this purpose a model representation which supports design and analysis operations of specific processes is needed. Our first research stage has two goals:

1. To propose a conceptual view of processes which builds on GPM and incorporates the notion of synchronization points (with respect to a given data item) between independent subdomains
2. To propose a more explicit model representation, which can be used for designing specific processes and performing a formal analysis

For achieving the first goal, our starting point is to extend GPM and define in precise terms the general notions of data inaccuracy and synchronization points which emerge from our observations above.

For achieving the second goal, we have posed a number of requirements for a process modelling language to be appropriate for explicit representation of concepts related to data awareness. These included among others explicit representation of data, and the ability to make a distinction between independent subdomains. After a preliminary examination of several well-studied languages (such as Petri nets, YAWL, Workflow nets with data, and BPMN) against these requirements, we decided to choose the Petri-net based formalism of Workflow nets with data (WFD-nets) [10], due to the popularity of Petri-nets and their advanced computational analysis, which can be used as a basis for process analysis with respect to data inaccuracy. WFD-nets are data-based extensions of workflow-nets, which can be naturally adapted for our purposes.

4.3 Identifying Properties of Processes Associated with Data Inaccuracy

Based on the formal model, we intend to investigate two properties of processes related to data inaccuracy. The first is soundness [1], which is an important and well-studied formal property of processes. A process is sound if and only if three requirements are satisfied: Option to complete, Proper completion, and No dead transitions. Current techniques for verifying soundness are mostly restricted to the process control flow and do not consider data. One of our key observations in this research is that data inaccuracy has a direct impact on soundness of processes, namely a process that is sound may in fact not reach proper completion due to the presence of data inaccuracy. Hence, soundness can only be guaranteed if we are *aware* that the data is accurate.

This naturally leads to the second property we intend to investigate: *awareness of (the existence of) data inaccuracy*. As mentioned above, in an independent subdomain operating solely on the basis of data items (without sensing their real values), we may be unaware of data inaccuracy, until a synchronization point is reached, where inaccuracy will be detected, and we become aware of it. Following this, we can decompose a process into aware and unaware parts. Moreover, if a data item $\boldsymbol{d_i}$ is read in an unaware (with respect to $\boldsymbol{d_i}$) part of the process, it might be used based on an inaccurate value, and thus might hamper the process from reaching its goals.

Going back to soundness, in its usual sense it can only be established in case of data inaccuracy awareness (DIA). Data inaccuracy unawareness, on the other hand,

poses a threat to soundness. This leads to the notion of soundness with respect to DIA (DI-soundness).

Our goal at this stage is to formally establish and investigate the properties of DIA and DI-soundness. We intend to provide formal definitions of DI awareness and DI-soundness, and develop algorithms assessing them regarding specific process models.

4.4 A Method for Supporting Process Design

At design time we can anticipate two kinds of data inaccuracy situations that can materialize at runtime: data inaccuracy which we know about, and data inaccuracy which exists and we are not aware of it.

In the first case, correction of the data before its use might be needed and possibly a compensation to avoid negative consequences. In both cases inaccurate data might be used, causing unexpected negative consequences. Our envisioned method seeks to address both situations, based on DI Awareness and DI soundness (see Section 4.3).

Processes have synchronization points which are inherent in their design. When a synchronization point is reached it might be too late to avoid negative consequences, thus, a mechanism that will help avoiding these situations is required. Furthermore, it is possible to artificially introduce additional synchronization points as a control mechanism of data accuracy. To distinguish between these two notions, we refer to the former as “natural” (NSP) and to the latter as “controlled” synchronization points (CSP). NSPs are located at places where independent sub-domains become mutually dependent as part of the process design. The CSP is a synchronization point which can be artificially added to the process for a certain data item, thereby enforcing a ”reality check”. However, such reality checks are costly, and should only be used if necessary. This gives rise to the following questions: inaccuracy of which data items may pose threats to soundness, or require substantial compensation? For such data items, where would be a good place to “plant” CSPs, if at all? Could CSPs for several critical items be put together?

Our third goal is to develop a method for supporting a modeler at the design of processes, allowing him/her to make processes more robust with respect to data inaccuracy. In particular, this will allow the modeler to make informed decisions on the need to insert CSPs, and find the right places for doing so. To this end we plan to progress in the following stages: (i) develop a mechanism for identifying a need for CSPs, (ii) propose an algorithm for identifying critical data items, (iii) based on (ii), propose a method for calculating the most appropriate points in the process to add CSPs, (iv) propose a method for assessing which of the synchronization points need compensating paths to maintain soundness, and (iv) develop a modeling tool with visual support for the proposed method and with process verification capabilities.

We intend to use CSPs as means for avoiding negative consequences of data inaccuracy. They can form a useful solution since they will allow a better control of the data item’s value. An important question arises: where and when to add CSPs in the process, and which are the relevant data items worth examining in order to reduce data inaccuracy consequences. Essentially, CSP will assist us to determine the possibility to encounter a specific data inaccuracy situation at runtime.

To deal with identifying the relevant data items worth examining, we will use data impact analysis. In general, data impact analysis examines the influence of a specific data item on other process elements (e.g. activities, decisions, data items) of a business process. Tsoury et al. [13] provide a method for data impact analysis which can serve as a basis for identifying critical data items. Based on a formal process model definition, relationships among elements of a specific process model (including data) are stored in a database and can be queried for identifying chained dependencies. We will adapt this method by introducing synchronization points into it, identifying critical data items by assessing their impact on the process.

To deal with the questions of where and when to insert CSP in the process, we will develop a time-dependent cost function which will include a probabilistic calculation of cost implications in every point and time in a given process that will assist us in redesigning the process. The decision variable of this function is where and when to add CSPs. Note that there are probably some situations when it will not be worthwhile to add CSP since the insertion of such CSP also has costs. We can view such situation as a tradeoff between improving the process robustness and avoiding high costs (note that "cost" is a general term, and can also refer to other performance indicators such as time). Robustness considerations may point towards handling data inaccuracy as the early as possible, but this may imply higher costs or delays. The developed function will enable evaluating alternative solutions considering additional CSPs.

In addition, we will investigate the impact of each NSP on the DI-soundness property. For example, if we evaluate a specific NSP as crucial for the process to become DI-sound, we will have to incorporate alternative routes for this NSP in the process (to maintain the soundness property of the process).

Last, we will develop a modeling tool (or extend an existing one), with visual support for using our method. It will enable marking the independent subdomains and their synchronization points, assessing costs (or any other user-defined parameter) of CSPs, analyzing a process in terms of DIA areas, and verifying its DI-soundness.

5 Evaluation

As in a design science research, our developed artifacts need to be evaluated. There are many evaluation techniques for design science research in the literature, such as [5][7]. Our evaluation relates to two main criteria: usefulness and usability.

5.1 Usefulness evaluation

The main objective in this phase is validation of the properties of DIA and DI- soundness, and their use for predicting situations of data inaccuracy for a given process. Later on, we can assess our method for proposing improvements in the process.

The above will be performed using a collection of case studies, which are currently being conducted in our research group, such as vehicle industry (small organization, purchasing process), weapon development and manufacturing (large organization, employee training processes), municipality (large organization, water meter manage-

ment process) etc. In these case studies information about data inaccuracy situations is being collected, based on interviews with stakeholders, concrete process models, and event logs.

In our evaluation, we will investigate whether applying our developed method could help (a) predicting existing data inaccuracy situations, and (b) support earlier detection of data inaccuracy at runtime. To this end, we will first transform the process models obtained from the organizations into WFD-based models (enriched with synchronization points). Next we will apply our algorithms for identifying possible data inaccuracy situations based on DIA and DI-soundness, analyze the results and compare them to the baseline of the case study. We expect our method to predict the majority of data inaccuracies which might exist in a given process. Moreover, we may predict a possibility for additional data inaccuracy situations which were not indicated in the case study. These might require additional investigation in the organization.

After exploring all data inaccuracy situations and comparing them to the data collected from the organizations, we will make suggestions for improving the robustness of the processes with respect to data inaccuracy by introducing CSPs in a cost-effective manner. When possible, we will present these suggestions to the process owners and ask them to scrutinize the suggestions and identify new insights (if any). In addition, using the relevant event logs, we will measure the % of cases where data inaccuracy was spotted at a later point in the process as compared to the introduced CSPs. These cases represent improvement achieved by the redesigned process.

5.2 Usability evaluation

A second important evaluation criterion is the usability criterion. Practically, we will examine whether humans can easily use our method, and what difficulties are incurred. We will conduct a controlled experiment to test whether the suggested method, as implemented in a modeling tool, supports modelers in designing more robust processes. The participants of the experiment shall be process designers (if possible) or students. The experiments will rely on a repeatable procedure: with and without the new method (or with different modalities of the modeling tool). The subjects will be asked to design or redesign processes while considering possible data inaccuracy issues. The experiments should provide insight regarding the usability of the method and its perceived usefulness.

6 Conclusion and Expected Contribution

Data has an enormous impact on business processes. Until now, data inaccuracy has mainly been addressed in the area of business process management as a possible exception at runtime, to be resolved through exception handling mechanisms [9]. In this research we focus on analysis of data inaccuracy at design time, introducing the concept of synchronization points and their effect on data inaccuracy awareness (DIA) and soundness. Hence, an innovative aspect of our proposal is the possibility to ex-

plicitly address potential consequences of data inaccuracy at design time. This has the potential to improve the robustness of the processes with respect to data inaccuracy.

The expected contributions of this study are both for research and for practice. For research, this study will open directions to investigate additional process properties that can relate to the accuracy of data. It will also highlight the need to study and quantify related risks. For practice, our research will provide support and guidance for designing robust processes, enabling prediction of data inaccuracy situations so organizations can investigate their existing processes and improve them.

Acknowledgements. The author is supported by the Israel Science Foundation under grant agreement no. 856/13.

References

1. Aalst, W.M., Kees M. van Hee, Arthur HM ter Hofstede, Natalia Sidorova, H. M. W. Verbeek, Marc Voorhoeve, and Moe Thandar Wynn. "Soundness of workflow nets: classification, decidability, and analysis."*Formal Aspects of Computing* 23, no. 3: 333-363, 2011.
2. Agmon, N., & Ahituv, N. Assessing data reliability in an information system. Journal of Management Information Systems, pp. 34-44, 1987.
3. Cappiello, C., Caro, A., Rodriguez, A., and Caballero, I. An Approach to Design Business Processes Addressing Data Quality Issues. In ECIS, pp. 216, 1987.
4. Gharib, M., & Giorgini, P. (2015). Modeling and reasoning about information quality requirements in business processes. In *Enterprise, Business-Process and Information Systems Modeling* (pp. 231-245). Springer International Publishing.
5. Helfert, M. and Donnellan, B. "The case for design science utility - Evaluation of design science artefacts within the IT capability maturity framework", accepted to the International Workshop on IT Artefact Design & Work Practice Intervention, Barcelona, 2012.
6. Orr, K. Data quality and systems theory. Com. of the ACM, 41(2), pp. 66-71, 1998.
7. Pries-Heje, J., Baskerville, R., and Venable, John R. "Strategies for Design Science Research Evaluation". ECIS 2008 Proceedings. Paper 87, 2008.
8. Rodríguez A., Caro A., Cappiello C. and Caballero I., "A BPMN Extension for Including Data Quality Requirements in Business Process Modeling", Business Process Model and Notation (LNBIP 125), pp. 116-125, Springer-Verlag Berlin, 2012.
9. Russell, N., van der Aalst, W.M.P., ter Hofstede, A.H.M.: Workflow exception patterns. In:Dubois, E., Pohl, K. (eds.) Advanced Information Systems Engineering (CAiSE 2006). LNCS, vol. 4001, pp. 288–302. Springer, Heidelberg, 2006.
10. Sidorova, N., Stahl C., and Trčka, N. "Workflow soundness revisited: Checking correctness in the presence of data while staying conceptual." Advanced Information Systems Engineering. Springer Berlin Heidelberg, 2010.
11. Soffer P., Kaner M., Wand Y. Assigning Ontological Meaning to Workflow Nets, Journal of Database Management, 21(3), pp. 1-35, 2010.
12. Soffer, P. Mirror, mirror on the wall, can i count on you at all? Exploring data inaccuracy in business processes. In Enterprise, business-process and information systems modeling (pp. 14-25). Springer Berlin Heidelberg 2010.
13. Tsoury, A., Soffer, P. and Reinhartz-Berger, I. "Towards Impact Analysis of Data in Business Processes". BPMDS, 2016.