ALOHOMORA: UNLOCKING DATA QUALITY CAUSES THROUGH EVENT LOG CONTEXT

Research paper

Fahame Emamjome,
Queensland University of Technology, Brisbane, Australia, f.emamjome@qut.edu.au

Robert Andrews,
Queensland University of Technology, Brisbane, Australia, r.andrews@qut.edu.au

Arthur H.M. ter Hofstede,
Queensland University of Technology, Brisbane, Australia, a.terhofstede@qut.edu.au

Hajo A. Reijers,
Utrecht University, Utrecht, Netherlands, h.a.reijers@uu.nl

Abstract

Big data's rise has amplified the role of information systems in process management. Process mining, a branch of data science, provides analytical tools and methods which can distil insights about process behaviour from big process-related data. Yet challenges remain, including dealing with the quality of big data and the impact of poor quality data on event logs as the input to process mining analyses. We show, through an analysis of 152 case studies, that despite researchers raising concerns about event log data quality, the event log preparation (data pre-processing) phase of process mining case studies is generally handled in a naive manner (as opposed to informed), focusing on fixing symptoms rather than uncovering the root causes of event log data quality issues.This paper considers event log data quality problems from a new angle. We introduce the Odigos (Greek for 'guide') framework, adapted from Mingers and Willcocks (2014), based on semiotics and Peircean abductive reasoning, that explains the notion of process mining context at a conceptual level. From a practical perspective, the Odigos framework facilitates an informed way of dealing with data quality issues in event logs through supporting both prognostic (foreshadowing potential quality issues) and diagnostic (identifying root causes of discovered quality issues) approaches. From a theoretical perspective, the work provides a foundation for the development of a process mining methodology for data pre-processing and for further IS theory development in the area of data analytics.

Keywords: Process mining, Event log context, Event log data quality, Semiotics.

1 Introduction

With the increasing importance of business processes as competitive differentiators for organisations, data analytics and data mining have become the tools to "wring every last drop of value from these processes" (Davenport et al., 2006). Process mining (van der Aalst, 2016), a branch of data science that bridges the gap between data mining and traditional forms of process analysis, provides analytical tools and methods which can deal with the huge volume of process-related data. The rise of Big Data has amplified the role of information systems in process management and has created new avenues for research within the IS discipline. Recent editorials (Abbasi, Sarker, and R. H. Chiang, 2016; Chen, R. H. L. Chiang, and Storey, 2012; Goes, 2014) in IS journals discuss the challenges and opportunities facing IS researchers in the area of big data analytics. Such challenges include dealing with the quality of big data

and the impact of poor quality data on results and on data-driven decision making. Data quality generally is considered as an antecedent for the success of data warehousing initiatives (Wixom and Watson, 2001) and is one of the main success factors for organisational data mining (Nemati and Barko, 2003). Mans et al. (2013) shows that event log quality is a critical success factor for process mining projects. As Marsden and Pingry (2018, p.A1) observe in a paper aimed at starting a wider and deeper discussion of data quality in IS research, "erudite modeling and estimation can yield no value without quality data inputs", i.e. a restatement of the well-known maxim *garbage in - garbage out*.

In general, event data is collected as a by-product of the operation of the systems that support process execution and is often logged for purposes other than process mining (e.g. security auditing). Such event data requires significant manipulation to convert (and clean) to an event log suitable for use in a process mining analysis. Data pre-processing can take up to 60% of the effort invested in a process or data mining project (Cabena et al., 1997; CrowdFlower Inc., 2017) and usually relies on the analyst, possibly informed by some domain knowledge, being able to recognise quality issues and apply appropriate remediation. "Cleaning event logs to address quality issues prior to conducting a process mining analysis is a necessary, but generally tedious and *ad hoc* task" (Suriadi, Andrews, et al., 2017, p.132).

Event log preparation exists as a distinct phase of many process mining methodologies, e.g. PDM (Bozkaya, Gabriels, and van der Werf, 2009), L* (van der Aalst et al., 2011) or PM^2 (van Eck et al., 2015). However, essential elements such as event data quality, the identification of data quality issues, the role of data quality in guiding event data extraction and log construction, and the impact of low data quality on process mining analyses, are generally poorly described (Andrews, M. T. Wynn, et al., 2019). In many process mining projects, researchers limit data pre-processing to merely transforming raw event data to a format that can be consumed by process mining tools, and to uncritically report analysis outcomes, i.e. a *garbage in - gospel out* effect that we refer to as *naive* process mining. As Andrews, M. T. Wynn, et al. (2019) point out, identifying the root causes of quality issues in event logs helps researchers to deal with those quality issues more effectively and get informed insights from their analysis. However, existing approaches to data quality and log cleaning (e.g. (RP Jagadeesh Chandra Bose and van der Aalst, 2010; Cheng and Kumar, 2015)) are more focused on treating data quality symptoms (in a given log) than on recognising the root causes of those issues. Emamjome, Andrews, and Hofstede (2019) proposes the notion of *informed* process mining[1], which involves a consideration of the context in which a process executes as a means of identifying root causes of event log quality issues. We posit that an approach that truly identifies the root causes of event log quality issues serves process mining research better than approaches that deal ex-post with quality issues/symptoms in event logs. Accordingly, the research question that is the focus of this paper is *"How can the root causes of data quality issues in event logs be identified in a systematic way based on a consideration of process mining context?"*

The Odigos framework proposed in this paper provides such a systematic approach to contextualise a process mining project and thus facilitates an informed way of dealing with data quality issues. The Odigos framework is developed based on the research method guidelines proposed in (Danermark et al., 2001) and by adapting the approach of Mingers and Willcocks (2017). The main contributions of this paper are (i) an extensive review of the pre-processing stage in process mining case studies revealing generally naive data cleaning, (ii) a theoretical, semiotics-based framework that frames the process mining context and (iii) illustrations of using the framework as a systematic approach to investigate plausible explanations of the root causes of quality issues in event logs.

2 Background and related work

Process mining is a maturing discipline with an ever-growing suite of tools that builds on process model-driven approaches and data mining to provide fact-based insights into (business) process behaviour and to support process improvements (van der Aalst, 2016). A process mining project begins with a Process

[1] Refers to a high level of maturity (methodological rigour) and a consideration of the organisational context being evident in process mining case studies.

Analyst working with Process Owners to (i) identify the processes to be investigated (or improved) and (ii) specify questions to be answered by the analysis. Central to any process mining project are the records of the execution of individual process steps which are captured (as *event* data) through the interaction between Process Participants with various Information Systems that support the processes. In preparation for analysis, process-related (event) data is identified, extracted, and converted to event log format. It is rare that the extraction is actually performed by the Process Analyst. Rather, there will be an intermediary, usually a Database Administrator whose job it is to manage the information systems that support either the process directly or the organisation's overall information requirements. We refer to the person/role/system responsible for converting source data into the (raw) event log provided to the Process Mining Analyst as the Data Curator. However, the decisions made by the Data Curator, such as which records to include and which to filter out from the log prior to presentation to the Analyst, have the potential to bias/distort analysis results. The Process Mining Analyst will then 'clean' the raw event log in preparation for process mining, conduct the analysis, generate results, and derive insights about the process.

As mentioned in the Introduction, the quality of event logs is critical to deriving useful insights about process behavior (RP Jagadesh Chandra Bose, Mans, and van der Aalst, 2013; Suriadi, Andrews, et al., 2017; van der Aalst et al., 2011). Data pre-processing tools and techniques address data quality issues such as missing data, incorrect data or bringing data to the right or uniform format, etc. There is, however, a lack of attention to methodological identification of quality issues in process mining studies, and there is little awareness of the impact of data quality on the findings of process mining studies (Andrews, M. T. Wynn, et al., 2019). Existing scholarly works on data quality and the pre-processing stage of process mining methodology, represent the researchers' main concerns regarding data quality in process mining research. The focus of these studies can be classified in three main areas:

- **providing a classification of event logs data quality issues to facilitate identification of these problems**. Examples of such works include (RP Jagadesh Chandra Bose, Mans, and van der Aalst, 2013) which identifies 27 distinct event log data quality issues and describes the impact of each on a process mining analysis, (Suriadi, Andrews, et al., 2017) shows that data quality issues can be detected by searching for 'imperfection' patterns in the event log and discusses the impact on a process mining analysis of each pattern, while the framework of Fox et al. (2018) provides a comprehensive list of data quality issues in the healthcare context.
- **approaches to deal with different types of data quality issues and how they impact process mining analysis**. Examples of such works include (Suriadi, Andrews, et al., 2017) which suggests a patterns-based approach to dealing with event log quality issues.Fox et al. (2018) describes the Care Pathways Data Quality Framework (CP-DQF) which uses the quality framework described in (RP Jagadesh Chandra Bose, Mans, and van der Aalst, 2013) to support systematic management (identification, recording, mitigation, reporting) of data quality issues in EHR systems. This framework, helps with identification of data quality issues arising through merging data from different sources, their relation to the research questions and identifying strategies to mitigate the effects of these quality issues on the research.
- **identifying root causes of quality issues in event logs and adopting a proper remedy approach**. This area of focus, even though very critical in relation to data quality, has been addressed by only a few process mining scholars. Examples of works in this area include (Mans et al., 2013) recognising the importance of root cause analysis of data quality issues with a focus on the role of Hospital Information Systems (HISs) in generating data quality issues in the healthcare domain. In (Suriadi, Andrews, et al., 2017), the authors, based on their experience in dealing with multiple event logs, abstract a set of commonly occurring event data quality issues as pattern templates which link the manifestation of each data quality issue to likely underlying causes. Andrews, M. T. Wynn, et al. (2019) proposes a metrics-based approach to assessing data quality and argues that identifying the root causes of quality issues prior to conducting process mining analysis and engagement with stakeholders can provide insights to possible remedies for the quality issues.

This review of existing studies on data quality in process mining, shows that identifying the root causes of quality issues in event logs, although recognised as a critical success factor, requires more attention in order to move towards a systematic, generalisable approach to dealing with data quality issues. The framework proposed in this paper is a step towards addressing this gap.

3 A Survey of process mining case studies

The works mentioned in Section 2 show that some process mining researchers have raised concerns about the need to deal with data quality issues in more depth, to pay more attention to their root causes, and to consider their implications for process mining outcomes. Here we review published process mining case studies from 2007 to 2018 to discern the degree to which researchers' concerns about data quality and pre-processing of event data have been incorporated into applications of process mining methodologies in practice. In order to do this, we define the concept of *informed* data pre-processing. This entails that researchers (i) define *high* quality data in relation to the context and the research questions, and (ii) reflect on how quality issues and data cleaning approaches impact study data and process mining results. In this view, it is *naive* (when pre-processing data) to rely solely on tools to automatically resolve data quality issues and not to be mindful of the underlying reasons as to why these quality issues emerge in the first place. Data quality issues may result from the way systems have been configured (including both operational use and logging) or from organisational rules, e.g. (Andrews, Suriadi, et al., 2018; Ash, Berg, and Coiera, 2004; Suriadi, M. T. Wynn, et al., 2013). From this review, we aimed to answer the following question: *how **informed** are the existing process mining case studies in relation to the pre-processing stage?* To be able to answer this above question, we ranked each case study using a scale of 1 (naive) to 3 (informed), in relation to the pre-processing stage in their methodology. The rankings allowed us to measure the degree of informedness of pre-processing in these case studies across the years and also to get some insights into how papers with higher levels of informedness dealt with data quality issues.

3.1 Literature Review Design

We followed Paré et al. (2015) in conducting a critical literature review to examine process mining case studies in terms of their attention to data quality and the pre-processing stage. Our review approach was also influenced by a number of related guidelines (Paré et al., 2015; Rowe, 2014). Consistent with Paré et al. (2015), in conducting a critical review, we hold each study up against some criteria defined in relation to informedness of the data pre-processing stage. In our approach we (i) extract process mining case studies from the last 18 years (Rowe, 2014), (ii) determine a selection strategy (Paré et al., 2015), (iii) develop coding dimensions and related assessment criteria (Paré et al., 2015) and (iv) perform the coding and the analysis (Balijepally, Mangalaraj, and Iyengar, 2011).
According to (Ghasemi and Amyot, 2016), the combination of Google Scholar and Scopus covers 96% of the published process mining papers in any topic and domain. Consequently, we used these two search engines to locate papers (articles, conference papers and book chapters) containing the phrase "process mining" with a publication date after 1999 (to span the life of the discipline). The search results from the two search engines were combined and duplicate titles removed. The list was filtered to remove obviously irrelevant papers, i.e. general BPM papers and data mining papers (which only mention process mining), 'citation only' references, and articles relating to the process of minerals and ore mining). We then excluded articles: where the principal contribution was a methodology, technique or tool, which was subsequently illustrated with a 'case study'; not written in English; for which the full-text was not freely available to the authors, and; where process mining was only one of several kinds of analyses applied in the case study (for instance, articles where process mining was used to derive an intermediate result which was then used as input to data mining or statistical analysis). We included industry-facing process mining case studies, i.e. process mining case studies which focused on reporting the application of **existing** process mining tools and techniques to a **specific domain** to provide **business value or address**

stakeholders' requirements. We considered papers published by both process mining researchers and domain experts. Since we were doing a critical literature review of process mining case studies, we did not assess the quality of the individual, selected papers (Paré et al., 2015). After initial filtering and subsequent application of inclusion and exclusion criteria, we identified 152 case study papers for analysis.

3.2 Coding Dimensions and Assessment Criteria

As explained before, case studies are coded on a scale from 0 (not reported on), 1 (naive) to 3 (informed) for their data preparation (pre-processing) stage. A paper is coded 1 for the data preparation stage if data quality definition is taken for granted, tools and algorithms are applied to clean data without considering the root causes of data quality issues and how cleaning data can impact the results of the study. To be considered as informed pre-processing (ranked as 3), data quality is defined in relation to the research question and data set, and, changes to a data set, in order to improve quality or to prepare for subsequent analysis, are justified in relation to the organisational context, research questions and limitations/implications of the data cleaning activities. If the paper was between the naive and informed definitions it is coded as 2. To ensure coding reliability, the first 10 papers were coded by two authors, the discrepancies were resolved and the coding criteria revised. Then the whole paper set was coded by one author, reviewed by all authors, the coding criteria were revised for the second time and the papers were again coded by the same author.

3.3 Results of the Review

In this section we present results of our review of case studies. Table 1 presents a frequency distribution of the raw coding of the 152 case studies and the average level of informedness for each phase. It can be observed that (i) 72% of the case studies either did not report on, or are naive in their pre-processing stage (ii) only 5% of the papers are scored as informed in relation to their pre-processing stage. In studies which were scored 1 (84 papers), the common approach to data pre-processing stage is (a) converting data to the right format (e.g. XES) appropriate for process mining tools or, (b) identifying and removing data quality issues such as missing time stamps from the data set without further analysis. 34 of the reviewed case studies were between naive and informed. The common approaches in these studies were (a) using data cleaning tools to clean and filter data quality problems and (b) providing some brief justifications for using specific cleaning tools or filtering criteria.

Methodology Phase	**Informedness Rank**					
	0	**1**	**2**	**3**	**Avg**	**Avg (excl 0)**
Data Pre-processing	26	84	34	8	1.16	1.40

Table 1: Frequency distribution of informedness rankings for data pre-processing methodology phase.

Among 152 case studies, only 8 of them could be considered as informed based on the criteria that we defined. The basic commonality between these papers was their attention to research questions when filtering and cleaning data and also being mindful of the actual causes of identified quality issues. For example, Lemos et al. (2011) identified quality issues in the event logs (such as missing duration of activities, granularity of time stamps and corrupted data entries), explained the causes of those issues (briefly) and, if the quality issue could limit addressing the research questions significantly, proposed an approach to deal with that issue. Lemos et al. (2011) used the formal documentation of the processes as their reference point to remediate the quality issues of their concerns. Suriadi, M. T. Wynn, et al. (2013) suggest that without stakeholders involvement in defining data filtering criteria, it is impossible to arrive on sets of meaningful data which is appropriate for the purpose of analysis and process mining. Andrews, M. Wynn, et al. (2018) identified sets of quality issues in the event logs using (RP Jagadesh Chandra Bose, Mans, and van der Aalst, 2013) framework and analysed the possible causes of these issues. To clean the data, the study authors used multiple available reference sources, including domain experts interviews, to choose the most appropriate filtering

criteria. Another study which scored 3 for the pre-processing stage, (Weber et al., 2018), mentions using qualitative assessments by experts as their approach in data pre-processing and making sense of data. It is evident from these 8 papers that the involvement of stakeholders and domain experts is one of the main approaches in data quality assessment and cleaning. However, further analysis of these 8 papers showed that (i) researchers apply intuitive and ad-hoc approaches that suit their specific data set and context, (ii) there is no systematic approach that specifies how experts can be involved (what are the insights that we can get from domain experts?) and, (iii) there is no systematic approach to identifying other contextual sources potentially useful for better understanding data quality issues. Following our analysis of process mining case studies, we make the following observations regarding the pre-processing stage:

Obs. 1 The high percentage of process mining case studies at the naive side of the scale confirms that there is a definite lack of attention paid to (and perhaps awareness of) quality issues in the data pre-processing stage of process mining projects.

Obs. 2 There is no systematic approach that helps researchers discover the root causes of quality issues in context. The existing approaches are ad-hoc and based on researchers' experience, the particular study's context and data set.

In the next section we propose a theoretical framework that directly supports Obs. 2 (and, by raising the profile of the issue and providing some guidance, may indirectly improve Obs. 1).

4 Theoretical Approach

Quality issues in event logs arise for a variety of reasons - some simple (e.g. incorrect construction of a format mask for a datetime column during ETL) and some complex (e.g. different task completion behaviours across resources - task-by-task completion during the day vs batch completion at the end of the day). Thus in order for process mining researchers to be able to recognize the root causes of these issues in a systematic way (Obs. 2), a frame of interpretation or a theoretical framework that guides process mining researchers in their investigation of the plausible explanations of the root causes of quality issues in event logs is required [2]. Accordingly, we propose a framework that can help to **diagnose** the root causes of identified data quality problems in a systematic manner. The proposed framework can also be used **prognostically** to anticipate quality issues in the event logs (which may or may not be discoverable through the usual syntactic quality symptoms) based on a systematic understanding of the context of the project. Thus, this theoretical framework helps process mining researchers to move towards an understanding of data quality issues beyond merely the symptoms showing in the event logs.

In seeking a theoretical framework as the reference point for investigation of data quality issues, we need to consider some of the characteristics of a process mining project and the specific nature of the event logs. Event logs, usually considered as the starting point for a process mining project, are created as a result of interactions between process participants, automation pieces (e.g. bots), data curators, the information systems, all embedded and influenced by the organisational rules, procedures, norms and, culture. This understanding of event logs implies that quality issues observed in event logs are also caused as a result of interactions between these different actors (process participants, bots, data curators, etc.), systems and the context, and thus, if analysed beyond their form of representation, can provide some insight for a process mining project.

Semiotics is a discipline that seeks to look behind the manifest appearance of data/text. Semiotics is the study of signs, their creation and how they generate meaning. Almost everything that we interact with and is capable of generating some meaning can be a sign. Accordingly, we can consider processes, event data and event logs as signs defined in semiotic studies (Price and Shanks, 2016). The most relevant branch of semiotics in relation to data quality is Peircean semiotics. Peircean semiotics was used in (Price and

[2] According to (Danermark et al., 2001) to be able to guide the explanatory research agenda the nature of the phenomenon and the entities involved in analysis of the phenomenon should be first foregrounded. The theoretical framework proposed in this study is providing this ontological foundation to guide researchers in analysing data quality in event logs.

Shanks, 2016) to determine information/data quality categories and criteria. Process mining researchers usually only have access to event logs and identify data quality issues (symptoms) by statistical, syntactic and semantic (Price and Shanks, 2016) analysis of event log attributes (RP Jagadesh Chandra Bose, Mans, and van der Aalst, 2013). In this paper, we use semiotics to discover the root causes of quality issues in the process mining/event log context.

In IS and ICT research, a considerable body of research has been developed around Peircean semiotics (Peirce, 1974). Mingers and Willcocks (2014) argue that semiotics is at the heart of studying information systems and communication and they propose an analytical framework based on Peircean semiotics. Mingers and Willcocks‘ framework (see Figure 1) can be used to study the relation between signs (data) and the personal, social and material worlds in a communication context. Since, in this paper, we aim to propose a systematic way to explain the relation between quality issues in event logs and the process of creation of event logs (including individual actors, IT systems and the organizational context) we adapt Mingers and Willcocks (2014)'s framework to the context of process mining analysis. To be able to do that we followed (Mingers and Willcocks, 2017) approach in developing a methodology for IS research based on the semiotic framework in (Mingers and Willcocks, 2014)

The framework of Mingers and Willcocks (2014), defines (i) three analytically separable worlds in relation to information system studies: the personal world; the material world; and the social world, and (ii) the interactions between these three worlds — "sociation", "embodiment" and "socio/materiality" (Mingers and Willcocks, 2014, p.61). At the centre of this framework, they define the concept of semiosis to refer to any content created through the interaction of these three worlds. Semiosis is the combination of signs and symbols that represents a meaning in a certain context. In a process mining context, the actual processes, the event data and event logs created for the purpose of analysis are all semiosis content (see Figure 2) generated as a result of interactions between the personal, social and material structures (see Figure 2). We now explain how the Mingers and Willcocks (2014) framework has been adapted to develop a theoretical framework (Odigos[3] framework) for understanding root causes of data quality issues in the context of process mining (see Figure 2).

4.1 The Worlds

Personal world: According to Mingers and Willcocks (2014), the personal world refers to the actors who are involved in the semiotic process of creation of content (semiosis) and meaning, their beliefs, values, motivations and expectations. In this paper and in relation to the concept of data quality, we recognize two main actors in the context of process mining: process participants and data curators who can be understood in terms of their psychological/behavioral structures. Figure 2 shows the process participant and data curator roles in relation to creation of semiosis (content). The role of process participants is to perform the processes and to create event data, while the data curator's role is to create event logs from event data for the purpose of analysis[4].

Social World: Mingers and Willcocks (2014) defines the social world as an "ensemble" of social structures,culture and norms, practices and conventions realised in the form of "position-practices" — role positions and social practices. Social structures, influence the creation of semiotic content not only through interaction with the personal and material worlds, but also directly through established connotation systems. "[...] connotative aspects of sign systems are social rather than individual – they exist before and beyond the individual's use of signs" (Mingers and Willcocks, 2014, p.62). Connotation here refers to pre-existing agreements about the meaning of semiosis content (signs which make that content). Creation of event data is not only influenced by the process participants' intentions (Figure 2, *create*) but also by the connotation systems established in their social context, such as the terminologies they use when

[3] Greek for 'guide' [4] Note that, for the internal arrows in Figure 2, we have considered only the interactions *towards* the semiosis content, since in this study we are interested in understanding the creation and root causes of quality issues in the event logs.

recording data (Figure 2, *connote*). For a process analyst, the event logs are defined based on specific connotations (events, cases, time stamps). Data curators also use their own connotation system to create event logs from event data (Figure 2, *create*). The differences between the data curator's connotation system and a process analyst's connotations can result in data quality issues in the event logs.

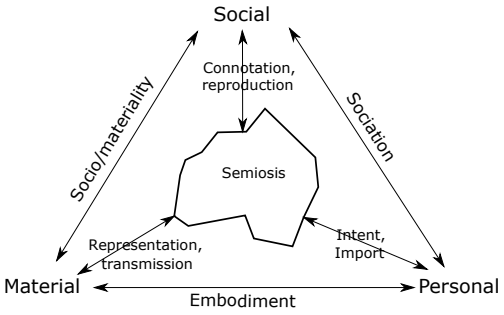

Figure 1: Relations between semiosis and the three worlds from (Mingers and Willcocks, 2014)

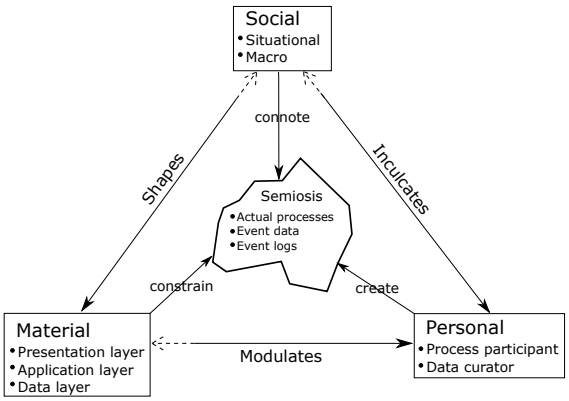

Figure 2: Odigos - Semiotic framework for process mining contextualisation

Material World: Mingers and Willcocks (2014) defines the material world as the physical structure of medium of communication, whether it be technological or not. All means of communication (such as sound, sight) can be considered as an instantiation of a communication medium or as part of the material world. The material world makes the signs accessible and gives them physical embodiment. The interfaces provided by information systems, software logic, the storage and transmission mechanisms are part of the material world and can constrain (affordances and liabilities) the creation of event data [5] (Figure 2, *constrain*). Similarly, the tools used by the data curator to create event logs from event data are also part of the material world. The constraints imposed by these tools can also impact on data quality issues in the event logs. Process participants, performing the actual processes, are interacting with both social structures and the material world (see Figure 2). We now explain in more detail how the interactions between these three worlds also can create forces which influence the quality of event logs.

4.2 Interactions between the Worlds

Interactions Between Social and Personal Worlds (Inculcates): Here, relying on existing theories, we characterise (conceptualise) the interactions between social structures and actors (process participants and data curators) in a process mining context and how they can influence process data, event data, and event logs. We refer to this interaction as *inculcation* (social structures inculcate the individual). Social structures can influence process participants' intentions, attitudes, and behaviours, how they perform their tasks (actual processes) and, their use of information systems (leading to the creation of event data). Social structures embody the requirement for justifiable behaviour, and constrain an individual's justifications and rationality through social norms and expectations (Salancik and Pfeffer, 1978). To better explain social influences on process participants and data curators we characterise social structures in two main categories: *Macro social structures* and *Situational social structures*. Situational social structures consist of norms, power structures, and practices in the immediate context (such as the organisation) against which individual behaviour will be judged (Habermas, 1984). Macro social structures include the wider social context (economy, history, culture, gender and so on) which influences actors' behaviours (Layder, 1998). Event log quality issues can emerge as a result of inculcation of process participants and their social

[5] Different information systems, with different levels of automation are included in this definition. In a fully automated environment the role of process participant changes but is never diminished.

context (*macro* and *situational*) while performing the actual processes or recording the event data. Data quality issues in event logs can also emerge from the inculcation of Data Curators and their social context (*Macro* and *Situational*) while preparing the event logs from recorded event data. While inculcation of actors by the social structure can have immediate effects on the quality issues in the event logs, the transformative effects of the personal world on social structures only has effects on the creation of event logs in the long term. Herein, we mainly focus on the immediate causes of quality issues or the solid headed arrow labelled Inculcates in Figure 2.

Example 1: As an example, let us consider how the interaction between data curator and situational social structures can create data quality issues. In the context of a process mining exercise, the Data Curator has extraordinary power and influence over the analysis. For instance, the Data Curator can anonymise confidential/private information, or can greatly assist the Analyst by grouping/summarising "like" values. Massa and Testa (2005) describe that the specific role of data curators or data administrators (*situational structures*) may provide them with some privilege and power (*inculcates*) which they wish to retain by keeping the ownership of data and providing limited views for process analysts (*creates*). They may also be affected by their own understanding of the goals of the process analyst and the impact of process analysis on themselves and their co-workers (*inculcates*). These situational power structures may affect how and what data they provide to the process analyst. For instance, filtering out cases the data curator perceives to be irrelevant, or worse, wishes to hide from scrutiny by the analyst.

Interactions Between Material and Personal Worlds (Modulates): Data quality issues can also emerge from interactions between the material world (technology) (Hutchby, 2013) and the personal world (process participants and data curators). In the context of process mining, information systems within organisations can be seen as a concrete instantiation of material structures (D'Adderio, 2004). To be able to understand the interactions between actors and technology involved in a process mining context we characterize (conceptualise) information systems in three layers: presentation layer, application layer, and data layer (Mutch, 2010). Each of these layers potentially *modulates* the actions and practices of process participants and data curators. We define the presentation layer to include physical structures (such as personal computers or other devices) and interfaces (such as forms, query interfaces and report generators). The application layer consists of program code supporting business rules and transactions. The data layer, in the context of process mining, consists of data warehouse technologies which support intensive data analysis (Mutch, 2010). Consistent with (Mingers and Willcocks, 2014), we define the interactions between actors and the material world in two ways; the first relates to the presentation layer of an information system and how it *modulates* the process participants' actions (see Figure 2), when a process participant executes a process and records process activities using a device through the interfaces available for the related processes and transactions (Dourish, 2004; O'Neill, 2008). Process participants' errors in entering data and recording the processes can be the result of these interactions.

The second form of interaction between actors and the material world is more related to process participants' interactions with the application layer or data layer. This form of interaction is about the constraints that a system imposes on users through business rules embedded in program code or data structures (*modulates*). Where users can avoid or work around such constraints (Boudreau and Robey, 2005), there is a likelihood that event data which does not reflect the actual processes will be recorded, thus creating quality issues (which may or may not be recognised by a process analyst) in process mining analysis.

The interaction between data curators and material structures can be predominantly defined in relation to the data layer or data warehouses and impacts the creation of event logs. According to Mutch (2010), a data warehouse can be decomposed into software, hardware, and data structures. The data structures imposed by the data curator can constrain and affect the process analysts' views of the data and the result of the analysis (Mutch, 2010). The complexity of software in relation to the data layer, and the privileged access, also provides more power (*modulates*) for data curators within and outside the organisation (Massa and Testa, 2005). The selection of different tools and types of data by data curators is also constraining

event logs provided for the process analyst. **Note** that process participants can individualise the use of information systems. Through time, different patterns of use can modify the design of presentation, application and data layers. These modifying interactions between actors and the systems do not have immediate effects on the creation of event logs and data quality issues. In Figure 2, the dashed head of the arrow from personal world to material world presents these sorts of interactions.

Example 2: In this healthcare example, it is important to note how the embodiment between medical staff and their tasks, the physical devices, and interfaces (presentation layer) can influence the use of the electronic recording systems and the generation of event data. As opposed to at-the-bedside paper charts, electronic recording devices may require clinicians to navigate/search prior to updating the patient's records. The appropriateness of physical devices and their interfaces used by healthcare workers influences (*modulates*) the rate of human errors when these tasks are reflected in a system (Ash, Berg, and Coiera, 2004). These errors in the use of the system can create data quality issues in event logs (duplicate events, inconsistent granularity in event names, or even recording of wrong events). Other aspects of data entry interfaces that may cause errors can be related to the way that data has to be entered e.g. forcing data to be complete upon entry (Ash, Berg, and Coiera, 2004). Such interface issues lead to an increased tendency to prefer paper-based recording over electronic recording. Further, if any clinician's device is for some reason not able to access or immediately update the patient's electronic chart, the overall chart is incomplete.

Interactions Between Social and Material worlds (Shapes): Actors' decisions and behaviours are not only formed through their direct interaction with social and material structures, but also by the way social structures shape technological structures and how the technology is perceived within the social context (Mutch, 2010; Olga Volkoff, Diane M Strong, and Elmes, 2007). To understand the interactions between social and material worlds, researchers have differentiated between two stages of technology construction (Feenberg, 2012). In the first stage, system designers abstract certain features of the social structures (e.g. business rules and processes) to shape the technological artefacts (social-material). Process mining is predicated on the assumption that systems used by process participants faithfully represent roles and practices in the social context. However, according to Volkoff and D. Strong (2013), that is not the case most of the time. Different systems have different capabilities in terms of representing business rules, practices and roles. For example the system may not capture the actual order of tasks, role responsibilities, or fail to record certain exceptions. These inconsistencies will be reflected in the event data in a way that can be misleading. Without knowing about this matter, process analysts do not have a great chance to discover the actual processes from the event logs. The second stage of technology construction can be broken down into two main aspects: 1) how the technology is perceived within the immediate social context (Feenberg, 2012), and 2) how, through time, technology embeds (re-structures) norms, routines, roles and practices into social structure (Olga Volkoff, Diane M Strong, and Elmes, 2007). The former refers to how hardware and the software roles are socially constructed (*shapes*). Therefore, the actors' behaviour is not only related to how they interact with technology and information systems but also how the system is perceived/positioned in their social context (D. Strong and Volkoff, 2010). While the former interaction has immediate effects on the actors and on creation of event logs, the re-structuring effects of technology on social structures has indirect effects which happen over time, e.g. systems' design can eventually change roles and even organisational structure. Consideration of these effects is important if a process mining project includes event log data captured over a long period of time. In Figure 2, these interactions between the material world and social structures are depicted by a dashed arrow head (*shapes*).

Example 3: Features such as "executive dashboards" were initially a manifestation of the focus on performance measurement in the Anglo-American organisational context (*shapes*), but these features changed and established many assumptions about performance management in other contexts as well (Mutch, 2010). These assumptions then may result in power struggles between employees and managers (which

can be presented in the way they perform a task (*inculcates*) or use a system and create event data) striving for their status and rewards (Armstrong, 1986).

5 Illustrative examples of the application of the Odigos framework

We argued that dealing with data quality issues in process mining case studies should be approached by reasoning about the root causes of those issues. Then we proposed a theoretical framework (Odigos framework) which conceptualises the process mining context in order to guide the analysts to understand data quality issues in a systematic way. Here, using some examples, we demonstrate how the Odigos framework can be used for two purposes; *prognosis* or *diagnosis* of data quality issues. The former refers to the role of the framework in identifying potential quality issues in event logs. The latter is about applying the framework to identify the root causes of observed quality issues in event logs.
The example below shows the role of the Odigos framework as a prognosis tool. The example depicts how changes in social structures can impact on event log data. In a prognosis approach, we start with identifying the elements of the framework positioned at the apexes and on the sides of the triangle in Figure 2. Consider the following case scenario: In a process mining project which aimed to discover patient flows in a hospital emergency department (Andrews, Suriadi, et al., 2018), the initial contextualisation revealed that, under an agreement signed by all Australian States[6], financial incentives were associated with public hospitals meeting targets for an agreed percentage of patients physically leaving the emergency department (ED) within four hours of their arrival[7]. Related performance measures were devised and reflected in the design of Hospital Information Systems (*shapes*). However, the targets proved difficult to meet due to the nature of emergency department patient presentations and resulted in pressure on individuals working in an emergency department to improve throughput. Further investigation revealed that, after operating under this agreement for some time, many hospitals introduced a Short-Stay Unit (SSU), logically distinct from the ED in the HIS, but physically co-located/attached to the ED. The short-stay unit allows patients who require monitoring for up to twenty-four hours to be discharged from the ED and admitted to the SSU thus maintaining continuity of care while limiting patients' length of stay in the ED. The following steps are taken to map this scenario to the Odigos framework in Figure 2.

1. **Social world:** *What are the **macro** structures i.e. political/governmental/cultural forces and changes influencing the context of the study*: All states signed up to the National Health Reform Agreement with financial incentives for hospitals meeting LoS targets as per the NEAT.
2. **Social world:** *What are the **situational** structures i.e. organisational rules, norms, culture, business model etc.*: Change in KPIs in the hospital, however, the nature of emergency department processes and the complexity and sensitivity of tasks performed remains unchanged.
3. **Material world:** *What IT systems are used to support the process i.e. presentation, application, and database layers*: HISs are used to record the tasks. Performance measures are embedded in the system application layer. SSU is added to the database as logically distinct unit from ED.
4. **Personal world:** *Who are the process participants (roles, resources)*: Nurses, Paramedics, Doctors and Admin staff working in the emergency department.
5. **Personal world:** *What is the role of the data curator?*: N/A for this scenario.
6. **Shapes:** *a) How do the macro and situational social structures (in 1 and 2) impact on the IT systems specified in 3?:* Performance measures are embodied in HIS systems. *b) How do the systems identified in 3 impact on the organisational Situational structure?:* Implementing the SSU makes it possible to meet the performance targets.
7. **Inculcates:** *How do the macro and situational social structures (in 1 and 2) impact on process participants and data curators?*: Results in pressure on process participants to meet the new performance criteria. Simply changing KPIs does not in itself provide a mechanism for performance improvements.

[6] National Health Reform Agreement 2011 [7] National Emergency Access Target (NEAT)

8. **Modulates:** *a) How are IT systems used by process participants?:* For relevant cases, IT systems are used to discharge patients from ED and transfer them to SSU. *b) How are the IT systems (data base level) used by data curators?:* N/A in this scenario.

Investigate impacts on *Semiosis*: After going through the above steps and finding out about the relevant concepts and their interactions, in the next 3 stages, we move towards inside the triangle in Figure 2 to develop hypotheses about possible data quality issues in the event log.

9. *How are the actual processes performed by process participants affected by the above identified interactions?* In the first stage after introducing the NEAT performance measures in EDs there may be some changes in the performance of processes. We may expect some processes are performed faster or some (not critical) patient care processes may be skipped. Following the introduction of the SSU, the actual performance of the processes may change as the option to discharge from ED in under 4 hours becomes available to process participants.
10. *How is the process data affected by the above identified interactions?* It is anticipated that after introducing the NEAT performance measures we may see small changes in the performance of the processes to get closer to 4 hours LOS. After introduction of the SSU, we anticipate a marked increase in the number of cases with length of stay in the ED being (just) less than the 4 hour target.
11. *What event data quality issues could be expected from 9 and 10 above?* In the first stage, we do not expect to see specific patterns, some cases (with the same level of severity) may take shorter than before (but not significantly) and we may see some missing events in some of the cases. In the second stage, after introducing SSU, we will see distinct process changes and new events such as "Transfer to SSU" i.e. *concept drift*.

The anticipated changes in process behaviour were actually observed and are illustrated in Figure 3 (prior to NEAT and introduction of SSU) and Figure 4 (post NEAT and introduction of SSU) (Queensland Audit Office, 2015).

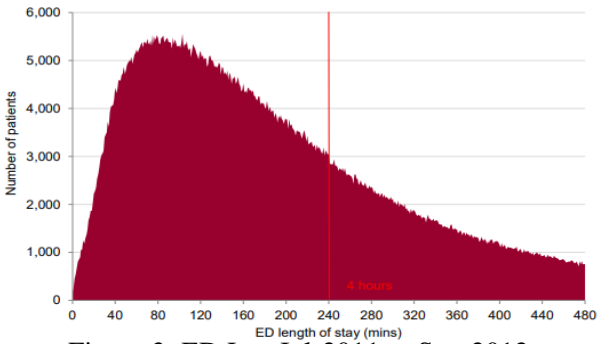

Figure 3: ED Los Jul-2011 to Sep-2012

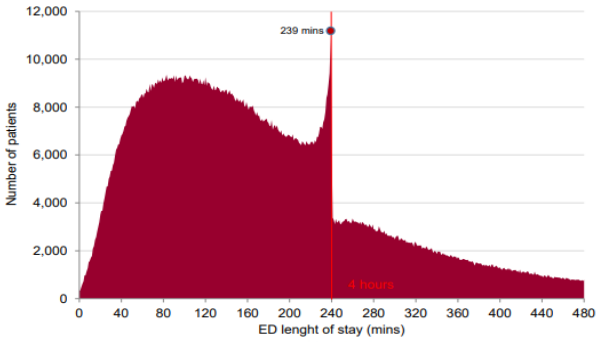

Figure 4: ED LoS Oct-2012 to Jun-2014

The next example demonstrates how the Odigos framework can be applied as a diagnosis tool to understand the root causes of quality issues. Missing cases (i.e. where actual executions of a process do not appear in an event log) were identified in RP Jagadesh Chandra Bose, Mans, and van der Aalst (2013) as a quality issue that can distort the process mining results and hinder discovering critical paths in the processes. For this hypothetical example, let us assume that the case frequency (patient episodes) in an event log intended for use in an analysis of patient flows in a hospital ED does not match actual case frequency, i.e. there are missing cases in the log. Rather than compensating for the effect of the missing cases on the analysis (by generalisation of the behaviours in the event log), we use the missing cases quality issue as the starting point in a deeper investigation of the processes and the process mining context using the framework in Figure 2 through the following steps:

1. **Investigate if the observed data quality issues could be created by Process Participants.** Could the Process Participants have a) actually skipped, or b) executed, but not recorded, those cases?
2. **Investigate if data quality issues could be caused by decisions made by D ata Curators.** Did the Data Curators decide to filter some of the cases from the event data when preparing the event log?

3. **Investigate if the IT systems have imposed some constraints on recording that led to the quality issues.** Are some cases marked as 'confidential' or automatically archived? Are multiple, different systems in use, and do they have different rules for recording event log data elements? For the example above on missing cases in HIS records, we know that, generally, HIS do not impose any constraints on recording of cases or events.
4. **Investigate if the data quality issues are the result of differences in the connotations i.e. the terminologies used to record different tasks by Process Participants and the Data Curators' understanding of those terminologies, or, the Data Curators' understanding of event log structure and process analysis**: In the above example, we know that the concept of case is defined and understood by both data curators and process participants so conflicting connotations could not be the cause of missing cases in the event log.

Since the missing cases in this example are most likely not related to the IT systems (Material constraints) or differences in terminologies (social connotations), we can hypothesise that either the process participants' intentions in recording the cases, or the data curator's intentions while creating event logs from event data could result in missing cases in the event log. Further investigation revealed that the hospital imposes some privacy policies on **releasing** data related to specific groups of patients admitted to the hospital *(situational structures)*. Even though the process participants do record all the cases and related events *(modulates)*, data curators are not allowed to reveal event data related to specific cases without permissions *(inculcates)*. By realising the reason behind the missing cases, the process analyst is able to apply actions to avoid ramification of missing cases in his/her analysis.

6 Conclusion

In this paper we have argued that dealing with pervasive data quality issues requires a deep understanding of the context in which the data was created. We suggested a theoretical approach to the problem and, by building on work by Mingers and Willcocks (2014), we developed the Odigos framework that characterises process mining context and can help with unearthing fundamental issues with data quality. We showed how the work can be applied to deal with data quality issues in process event logs, in both a prognostic (foreshadowing potential quality issues) and a diagnostic (identifying root causes of quality issues) manner. Through a survey on process mining case studies, we demonstrated that the current approaches in dealing with symptoms of data quality problems have been limiting the impact of process mining in practice. Thus, this work has practical significance. Consequently, the proposed Odigos framework can help practitioners conducting process mining case studies to deal with data quality issues in an informed manner. For process mining researchers, the Odigos framework provides the foundation for methodological data pre-processing. By proposing a framework to facilitate identifying root causes of data quality issues in event data, we help researchers to discover the human and social side of data creation rather than treating data as being independent of the people and processes that created it. Identifying the root causes of quality issues highlights the social, material and individual factors which contribute to low quality data which would be overlooked by existing data cleaning methods that focus on symptoms rather than root causes.

References

Abbasi, Ahmed, Sarker, Suprateek, and Chiang, Roger HL 2016. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems* 17 (2), i–xxxii. ( ).

Andrews, Robert, Suriadi, Suriadi, Wynn, Moe, ter Hofstede, Arthur HM, and Rothwell, Sean 2018. "Improving Patient Flows at St. Andrew's War Memorial Hospital's Emergency Department Through Process Mining," in *Business Process Management Cases,* Springer, pp. 311–333. ( ).

Andrews, Robert, Wynn, Moe T, Vallmuur, Kirsten, ter Hofstede, Arthur HM, Bosley, Emma, Elcock, Mark, and Rashford, Stephen 2019. "Leveraging Data Quality to Better Prepare for Process Mining: An Approach Illustrated Through Analysing Road Trauma Pre-Hospital Retrieval and Transport Processes in Queensland," *International Journal of Environmental Research and Public Health* 16 (7), 1138. ( ).

Andrews, Robert, Wynn, Moe, Hofstede, Arthur HM ter, Xu, Jingxin, Horton, Kylie, Taylor, Paul, and Plunkett-Cole, Sue 2018. "Exposing Impediments to Insurance Claims Processing," in *Business Process Management Cases,* Springer, pp. 275–290. ( ).

Armstrong, P 1986. "Management Control Strategies and Inter-Professional Competition; the Cases of Accountancy and Personnel Management," in *Managing the Labour Process,* D Knights and HC Willmott (eds.). Aldershot Crower. ( ).

Ash, Joan S, Berg, Marc, and Coiera, Enrico 2004. "Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors," *Journal of the American Medical Informatics Association* 11 (2), 104–112. ( ).

Balijepally, Venugopal, Mangalaraj, George, and Iyengar, Kishen 2011. "Are We Wielding this Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Systems Research," *Journal of the Association for Information Systems* 12 (5), 375–413. ( ).

Bose, RP Jagadeesh Chandra and van der Aalst, Wil M. P. 2010. "Trace Alignment in Process Mining: Opportunities for Process Diagnostics," in *Int. Conf. BPM,* Springer, pp. 227–242. ( ).

Bose, RP Jagadesh Chandra, Mans, Ronny S, and van der Aalst, Wil M. P. 2013. "Wanna Improve Process Mining Results?" in *IEEE Symposium on Computational Intelligence and Data Mining,* IEEE, pp. 127–134. ( ).

Boudreau, MC and Robey, D 2005. "Enacting Integrated Information Technology: A Human Agency Perspective," *Organization Science* 16 (1) Feb. 2005, 3–18 Feb. 2005. ( ).

Bozkaya, Melike, Gabriels, Joost, and van der Werf, Jan Martijn 2009. "Process Diagnostics: A Method Based on Process Mining," in *Int. Conf. on Information, Process, and Knowledge Management,* IEEE, pp. 22–27. ( ).

Cabena, Peter, Hadjinian, Pablo, Stadler, Rolf, Verhees, Jaap, and Zanasi, Alessandro 1997. *Discovering Data Mining: From Concept to Implementation,* Prentice Hall PTR New Jersey. ( ).

Chen, Hsinchun, Chiang, Roger H. L., and Storey, Veda C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* 36 (4), 1165–1188. ( ).

Cheng, Hsin-Jung and Kumar, Akhil 2015. "Process Mining on Noisy Logs–Can Log Sanitization Help to Improve Performance?" *Decision Support Systems* 79, 138–149. ( ).

CrowdFlower Inc., 2017. *2017 Data Scientist Report,* `https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf` (last visited 23/04/2020). ( ).

D'Adderio, L 2004. *Inside the Virtual Product: How Organizations Create Knowledge Through Software,* Edward Elgar Publishing. ( ).

Danermark, Berth, Ekstrom, Mats, Jakobsen, Liselotte, and Karlsson, Jan 2001. *Explaining Society: An Introduction to Critical Realism in the Social Sciences,* Routledge. ( ).

Davenport, Thomas H et al. 2006. "Competing on Analytics," *Harvard Business Review* 84 (1), 98–107. ( ).

Dourish, P 2004. *Where the Action Is: The Foundations of Embodied Interaction,* MIT Press. ( ).

Emamjome, Fahame, Andrews, Robert, and Hofstede, Arthur HM ter 2019. "A Case Study Lens on Process Mining in Practice," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems",* Springer, pp. 127–145. ( ).

Feenberg, Andrew 2012. *Questioning Technology,* Routledge. ( ).

Fox, Frank, Aggarwal, Vishal R, Whelton, Helen, and Johnson, Owen 2018. "A Data Quality Framework for Process Mining of Electronic Health Record Data," in *IEEE Int. Conf. on Healthcare Informatics,* IEEE, pp. 12–21. ( ).

Ghasemi, Mahdi and Amyot, Daniel 2016. "Process Mining in Healthcare: A Systematised Literature Review," *International Journal of Electronic Healthcare* 9 (1), 60–88. ( ).

Goes, Paulo B 2014. "Editor's Comments: Big Data and IS Research," *MIS Quarterly* 38 (3), iii–viii. ( ).

Habermas, Jurgen 1984. *The Theory of Communicative Action: Jurgen Habermas; Trans. by Thomas McCarthy,* Heinemann. ( ).

Hutchby, I 2013. *Conversation and Technology: From the Telephone to the Internet,* John Wiley & Sons. ( ).

Layder, Derek 1998. *Sociological Practice: Linking Theory and Social Research,* Sage. ( ).

Lemos, Artini M, Sabino, Caio C, Lima, Ricardo Massa Ferreira, and Oliveira, Cesar AL 2011. "Conformance Checking of Software Development Processes Through Process Mining." in *SEKE,* pp. 654–659. ( ).

Mans, Ronny S, van der Aalst, Wil M. P., Vanwersch, Rob JB, and Moleman, Arnold J 2013. "Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions," in *Process Support and Knowledge Representation in Health Care,* Springer, pp. 140–153. ( ).

Marsden, James R and Pingry, David E 2018. "Numerical Data Quality in IS Research and the Implications for Replication," *Decision Support Systems* 115, A1–A7. ( ).

Massa, S and Testa, S 2005. "Data Warehouse-in-Practice: Exploring the Function of Expectations in Organizational Outcomes," *Information and Management* 42 (5) July 2005, 709–718 July 2005. ( ).

Mingers, John and Willcocks, Leslie 2014. "An Integrative Semiotic Framework for Information Systems: The Social, Personal and Material worlds," *Information and Organization* 24 (1), 48–70. ( ).

— 2017. "An integrative semiotic methodology for IS research," *Information and Organization* 27 (1), 17–36. ( ).

Mutch, Alistair 2010. "Technology, Organization, and Structure: A Morphogenetic Approach," *Organization Science* 21 (2), 507–520. ( ).

Nemati, Hamid R and Barko, Christopher D 2003. "Key Factors for Achieving Organizational Data-mining Success," *Industrial Management & Data Systems* 103 (4), 282–292. ( ).

O'Neill, S 2008. *Interactive Media: The Semiotics of Embodied Interaction,* Springer Science & Business Media. ( ).

Paré, Guy, Trudel, Marie-Claude, Jaana, Mirou, and Kitsiou, Spyros 2015. "Synthesizing Information Systems Knowledge: A Typology of Literature Reviews," *Information & Management* 52 (2), 183–199. ( ).

Peirce, Charles Sanders 1974. *Collected Papers of Charles Sanders Peirce,* vol. 2. Harvard University Press. ( ).

Price, Rosanne and Shanks, Graeme 2016. "A Semiotic Information Quality Framework: Development and Comparative Analysis," in *Enacting Research Methods in Information Systems,* Springer, pp. 219–250. ( ).

Queensland Audit Office 2015. *Emergency Department Performance Reporting 3:2014–15,* ( ).

Rowe, Frantz 2014. "What Literature Review is Not: Diversity, Boundaries and Recommendations," *European Journal of Information Systems* 23 (3), 241–255. ( ).

Salancik, Gerald R and Pfeffer, Jeffrey 1978. "A Social Information Processing Approach to Job Attitudes and Task Design," *Administrative Science Quarterly* 23 (2), 224–253. ( ).

Strong, DM and Volkoff, O 2010. "Understanding Organization-Enterprise System Fit: A Path to Theorizing the Information Technology Artifact," *MIS Quarterly* 34 (4), 731–756. ( ).

Suriadi, Suriadi, Andrews, Robert, ter Hofstede, Arthur HM, and Wynn, Moe Thandar 2017. "Event Log Imperfection Patterns for Process Mining: Towards a Systematic Approach to Cleaning Event Logs," *Information Systems* 64, 132–150. ( ).

Suriadi, Suriadi, Wynn, Moe T, Ouyang, Chun, ter Hofstede, Arthur HM, and van Dijk, Nienke J 2013. "Understanding Process Behaviours in a Large Insurance Company in Australia: A Case Study," in *International Conference on Advanced Information Systems Engineering,* Springer, pp. 449–464. ( ).

van der Aalst, Wil M. P. et al. 2011. "Process Mining Manifesto," in *Int. Conf. BPM,* Springer, pp. 169–194. ( ).

van der Aalst, Wil M. P. 2016. *Process Mining: Data Science in Action,* Springer. ( ).

van Eck, Maikel L, Lu, Xixi, Leemans, Sander JJ, and van der Aalst, Wil M. P. 2015. "PM^2: A Process Mining Project Methodology," in *International Conference on Advanced Information Systems Engineering,* Springer, pp. 297–313. ( ).

Volkoff, O and Strong, DM 2013. "Critical Realism and Affordances: Theorizing IT-associated Organizational Change Processes," *MIS Quarterly* 37 (3), 819–834. ( ).

Volkoff, Olga, Strong, Diane M, and Elmes, Michael B 2007. "Technological Embeddedness and Organizational Change," *Organization Science* 18 (5), 832–848. ( ).

Weber, Philip, Backman, Ruth, Litchfield, Ian, and Lee, Mark 2018. "A Process Mining and Text Analysis Approach to Analyse the Extent of Polypharmacy in Medical Prescribing," in *2018 IEEE International Conference on Healthcare Informatics (ICHI),* IEEE, pp. 1–11. ( ).

Wixom, Barbara H. and Watson, Hugh J. 2001. "An empirical investigation of the factors affecting data warehousing success," *MIS Quarterly* 25 (1), 17–41. ( ).