INVITED RESEARCH

Root-cause Analysis of Process-data Quality Problems

Robert Andrews[a] and Fahame Emamjome[a] and Arthur H.M. ter Hofstede[a] and Hajo A. Reijers[b]

[a]Queensland University of Technology; [b]Utrecht University, Utrecht, Netherlands

ARTICLE HISTORY
Compiled June 21, 2021

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. In previous work, we have shown 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 mechanistically, focusing on fixing symptoms and getting the log to a state where it can be consumed by process mining tools, 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. 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. We examine in depth how the framework supports a detailed root-cause analysis of a well-known collection of event log imperfection patterns.

KEYWORDS
Process mining; Organisational context; Semiotics; Data quality; Root-cause analysis

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, 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. Editorials (Chen et al., 2012; Goes, 2014; Abbasi et al., 2016) 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

CONTACT Robert Andrews. Email: r.andrews@qut.edu.au

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 & Watson, 2001) and is one of the main success factors for organisational data mining (Nemati & Barko, 2003). Mans et al. (2013) show 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 et al., 2017, p.132).

Event log preparation exists as a distinct phase of many process mining methodologies, e.g. PDM (Bozkaya et al., 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 et al., 2019). In many process mining projects, analysts 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 *mechanistic* process mining. As Andrews et al. (2019) point out, identifying the root causes of quality issues in event logs helps analysts 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. (Cheng & Kumar, 2015; Bose & van der Aalst, 2010)) are more focused on treating data quality symptoms (in a given log) than on recognising the root causes of those issues. Emamjome et al. (2019) propose 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 identifies the root causes of event log quality issues serves process mining research and practice 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?"* Emamjome et al. (2019) critically review 152 process mining case study papers and conclude that only a small minority of these case studies deal with data quality issues in an informed way.

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

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

method guidelines proposed by Danermark et al. (2001) and by adapting the approach of Mingers and Willcocks (2017). Further, the version of the framework presented in this paper is a modification of an earlier version proposed by Emamjome et al. (2020a) following an evaluation conducted with eleven process mining and data quality experts (Andrews, Emamjome, et al., 2020). The main contributions of this paper are (i) an extensive review of the pre-processing stage in process mining case studies revealing generally mechanistic data cleaning, (ii) a theoretical, semiotics-based framework that frames the process mining context, and (iii) a detailed method for root-cause analysis for a collection of well-known so-called event log imperfection patterns, which capture a range of process-data quality issues. Odigos is intended for use by analysts who will benefit through having a tool that supports a structured approach to identifying data quality issues, and the direct and root causes of these issues. For researchers developing methodological data pre-processing guidelines, Odigos highlights the human and social side of data creation rather than treating data as being independent of the people and processes that created it. The paper is also written with the organisational context in mind. The organisation will benefit from improved understanding of the types and underlying causes of data quality issues, and hence, opportunities to improve systems and operational practices to remediate these issues.

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 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 (Bose et al., 2013; van der Aalst et al., 2011; Suriadi et al., 2017). 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 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	
(Bose et al., 2013)	Identifies 27 distinct event log data quality issues and describes the impact of each on a process mining analysis.
(Suriadi 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.
(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	
(Suriadi et al., 2017)	Suggests a patterns-based approach to dealing with event log quality issues.
(Fox et al., 2018)	Describes the Care Pathways Data Quality Framework (CPDQF) which uses the quality framework described in Bose et al. (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	
(Mans et al., 2013)	Recognises 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.
(Suriadi et al., 2017)	Abstracts 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 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.

Table 1.: Summary of existing scholarly works on event log data quality and the researchers' main concerns regarding data quality in process mining research.

Furthermore, Emamjome et al. (2019) in a critical review of 152 process mining case studies showed that there is only a small percentage of existing studies who deal with data quality issues in an informed manner and 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. This review (Emamjome et al., 2019) also highlighted that there is no systematic approach that guides process mining analysts in dealing with data quality issues. The framework proposed in this paper is a step towards addressing this gap.

3. 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 analysts to be able to recognise the root causes of these issues in a systematic way, a frame of interpretation or a theoretical framework that guides process mining analysts 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 analysts 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 & Shanks, 2016). The most relevant branch of semiotics in relation to data quality is Peircean semiotics. Price and Shanks (2016) uses Peircean semiotics 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 & Shanks, 2016) analysis of event log attributes (Bose et al., 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 organisational context) we adapt Mingers and Willcocks (2014)'s framework to the context of process mining analysis. To be able to do that we followed the approach of Mingers and Willcocks (2017) in developing a methodology for IS research based

[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.

on the semiotic framework in Mingers and Willcocks (2014).

The framework in 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 & Willcocks, 2014, p.61).

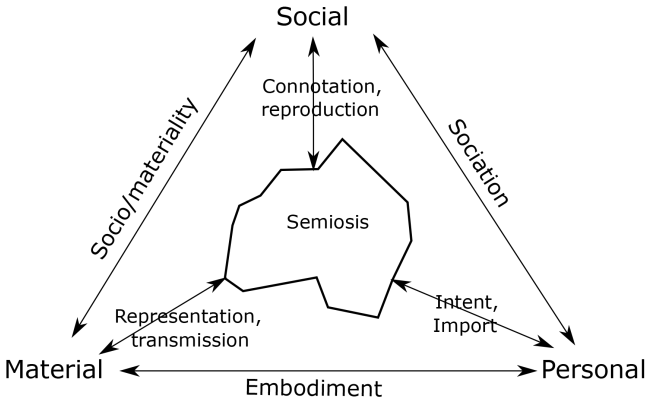

Figure 1.: Relations between semiosis and the three worlds from (Mingers & Willcocks, 2014)

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).

Our theoretical framework is referred to as the *Odigos*[3] framework and was developed by first adapting the framework of Mingers and Willcocks (2014). The resulting framework (Emamjome et al., 2020a) was validated and extended through interviews with process mining experts (Andrews, Emamjome, et al., 2020).

4. 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 the initial conceptualisation of the Odigos framework, two actors in the context of process mining were identified: process participants and data curators (Emamjome et al., 2020a). Through interviews with process mining experts (Andrews, Emamjome, et al., 2020), further roles have been added to the Personal world including process designers and data administrators. 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. The definition of the process participant role is not only limited to the employees of a company but includes any role (such as customers) involved in recording data and completing tasks using IT systems. The data curator's role is to create event logs from event data for the purpose of analysis[4]. This definition also applies to a data curator team or "chain of data

[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.

curators" if applicable. The process designer role has indirect impact on the creation of event data through the development of formal processes embedded in IT systems. The role of data administrators who have access to and control over the data structures of IT systems are another addition to the original version (Emamjome et al., 2020a) of the Odigos framework. This role too only indirectly impacts the creation of event data, by changing the way data is recorded in databases.

Social World: Mingers and Willcocks (2014) define 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 & Willcocks, 2014, p.62). In the initial conceptualisation of the Odigos framework (Emamjome et al., 2020a) we characterised the social world in two categories of *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, legislation, history, culture, language, gender and so on) which influences actors' behaviours (Layder, 1998). 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 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. Andrews, Emamjome, et al. (2020) further specified factors within the social world which influence the creation of event data through interaction between personal and material worlds. These factors include organisational culture, social agreements, management style and organisational structure, performance criteria, legal requirements, and organisational changes.

Material World: Mingers and Willcocks (2014) define 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, and storage and transmission mechanisms used to record process execution 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. In the initial conceptualisation of the Odigos framework (Emamjome et al., 2020a), consistent with (Mutch, 2010), we characterised the material world as IT systems in three layers — presentation, application, and data — as this model provides granularity and generality sufficient to capture data quality

[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.

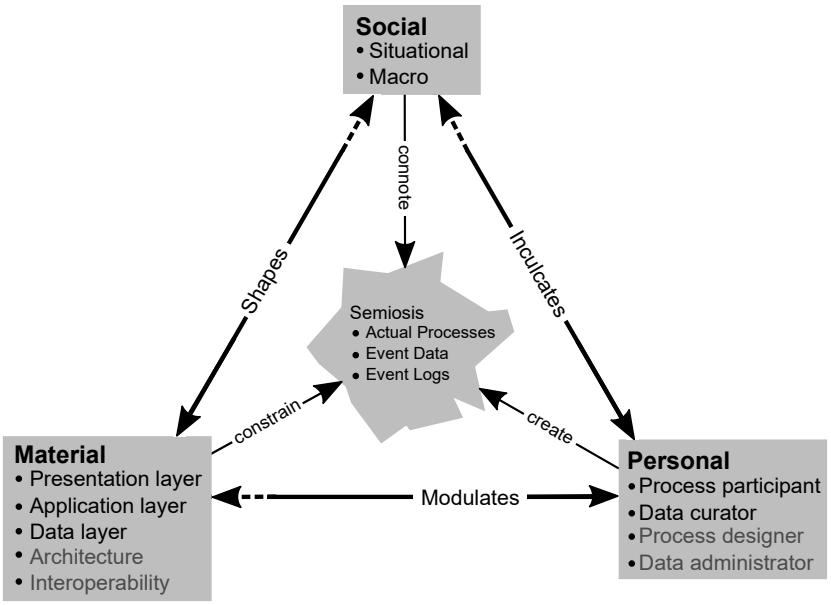

Figure 2.: Odigos – Semiotic framework for process mining contextualisation

issues related to the material world. Interviews with process mining experts (Andrews, Emamjome, et al., 2020) revealed further specific aspects of the material world such as interoperability between different IT systems, information architecture and the design of applications (e.g. the meaning of null values), mixed granularity in recording events and time stamps, and the compatibility of IT systems in recording event data with the requirements of process mining.

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.

5. 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 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 & Pfeffer, 1978). Event log quality issues can emerge as a result of inculcation of process participants and their social context (*macro* and *situational*) while performing the actual processes or recording the event data. Several examples of data quality problems were given by process mining experts (Andrews, Emamjome, et al., 2020) of how organisational culture or performance criteria can influence pro-

cess participants in the way they are conducting their activities or in the way they are using IT systems to record those activities (Andrews, Emamjome, et al., 2020). 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. As evidenced by process mining experts (Andrews, Emamjome, et al., 2020), the quality of extracted event data is also influenced by data curators' knowledge of the domain and privacy regulations and concerns in the context of data collection. While inculcation of actors by the social structure can have direct effect on quality issues in event logs, the transformative effect of the personal world on social structures only influences the creation of event logs in the long term. Herein, we mainly focus on the direct 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 information, or can greatly assist the analyst by grouping or summarising "like" values. Massa and Testa (2005) describe that the specific role of data curators (*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 (e.g. 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 characterise (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. For example, a user friendly interface design can help avoid data quality entries by process participants (Andrews, Emamjome, et al., 2020).

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 & 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. One of the main reasons mentioned by process mining experts in relation to this path was that IT systems are not supporting the processes either through the design of databases, lack of automation or using terminologies which are not consistent with the ones used by process participants (Andrews, Emamjome, et al., 2020).

Consistent with findings from Andrews, Emamjome, et al. (2020), the interaction between data curators and material structures can be predominantly defined in relation to the extraction of event logs. This includes managing (i) data integration between different IT systems, (ii) interoperability challenges between IT systems, and (iii) constraints around the use of data extraction tools. According to Andrews, Emamjome, et al. (2020) the data administrator role can influence the creation of event logs through their interaction with the data layer or data warehouses. According to Mutch (2010), a data warehouse can be decomposed into software, hardware, and data structures. The data structures imposed by the data administrator 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 administrators and data curators within and outside the organisation (Massa & Testa, 2005). The selection of different tools and types of data by data curators is also constraining event logs to be 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 systems do not have a direct effect 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 et al., 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 et al., 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; Volkoff et al., 2007). To understand the interactions between social and material worlds, re-

searchers have differentiated between two stages of technology construction (Feenberg, 2012). In the first stage, systems and process 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 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. In Andrews, Emamjome, et al. (2020) process mining experts mentioned several other factors influencing the design of IT systems, such as the impact of performance criteria, lack of attention to workflow management, and siloed organisational culture. 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 (Volkoff et al., 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 (Strong & Volkoff, 2010). While the former interaction has immediate effects on the actors and on creation of event logs, the restructuring 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).

6. 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.

In a prognostic application of the framework, we adopt an "out to in" approach (see Figure 3). That is, we start by identifying organisational context factors in the three worlds that are relevant to the process. Then, we examine interactions between the worlds (the *Find Root Cause Effects in [World]*, i.e. the *Shapes*, *Modulates*, *Inculcate* arcs in Figure 3), to reason about influences of these factors on the other worlds. Next, we consider how these may affect the semiosis (the *Find Direct Cause*, i.e. the *connote*, *constrain*, *create* arcs in Figure 3).

The example depicts how changes in social structures can impact on event log data. Consider the following case scenario: In a process mining project which aimed to discover patient flows in a hospital emergency department (Andrews 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* in Figure 2). 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 3.

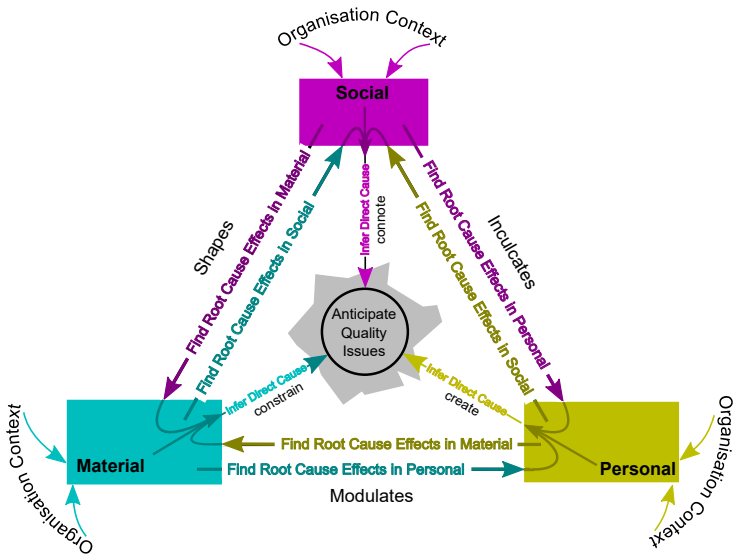

Figure 3.: Prognostic application of the Odigos framework showing the "out to in" traversal of the framework from Organisational Context, through the Worlds, and ultimately, to anticipated quality issues.

(1) **Social world:** *a) 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. *b) What are the **situational** structures i.e. organisational rules, norms, culture, business*

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

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.

(2) **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.

(3) **Personal world:** *a) Who are the process participants (roles, resources)*: Nurses, Paramedics, Doctors, and Admin staff working in the emergency department. *b) What is the role of the data curator?*: N/A for this scenario.

(4) **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.

(5) **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) **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 (database 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 three stages, we move towards the inside of 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 to be performed faster or some (not critical) patient care processes to 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 four 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 four 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 four 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 4 (prior to NEAT and introduction of SSU) and Figure 5 (post NEAT and introduction of SSU) (Queensland Audit Office, 2015).

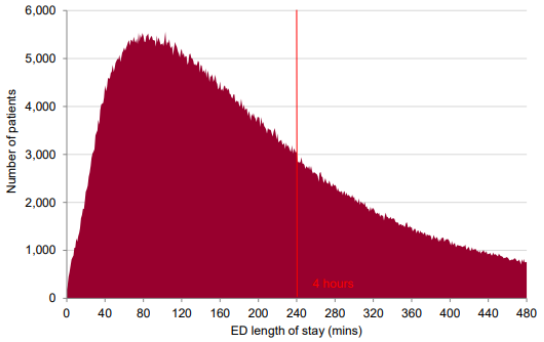

Figure 4.: ED Los Jul-2011 to Sep-2012 *(source: (Queensland Audit Office, 2015))*

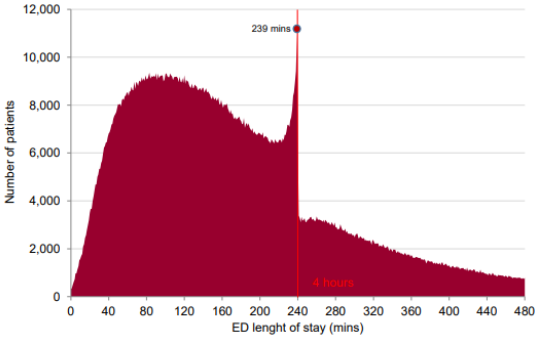

Figure 5.: ED LoS Oct-2012 to Jun-2014 *(source: (Queensland Audit Office, 2015))*

The next example demonstrates how the Odigos framework can be applied as a diagnosis tool to understand the root causes of quality issues. In a diagnostic application of the framework, we adopt an "in to out" approach (see Figure 6).

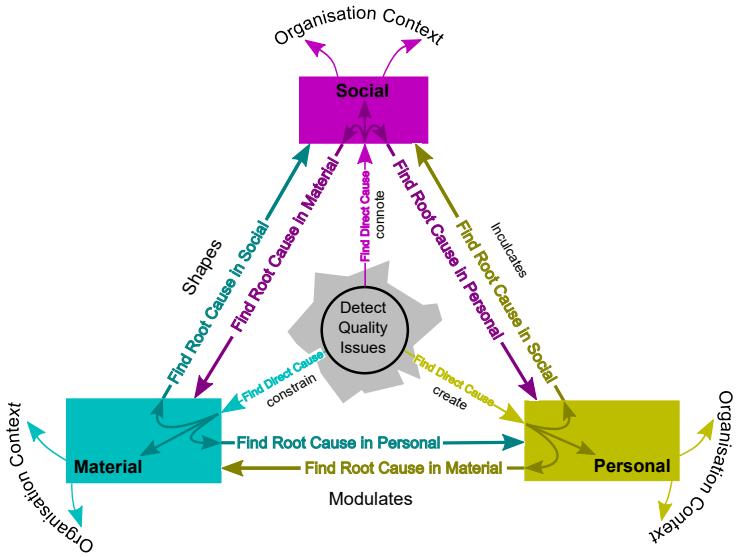

Figure 6.: Diagnostic application of the Odigos framework showing the "in to out" traversal of the framework from detected quality issues, through the Worlds, and ultimately, to root causes (in the Organisational Context).

That is, we start by identifying quality issues in the semiosis, then determine the world (and dimension of that world) directly responsible for introducing the quality issue. Lastly, we examine interactions between the worlds to identify organisational context factors ultimately responsible for the quality issues in the semiosis. Missing cases (i.e. where actual executions of a process do not appear in an event log) were identified in Bose et al. (2013) as a quality issue that can distort the process mining results and hinder the discovery of 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 6 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 data 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 data curators' 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 the ramifications of missing cases in his/her analysis.

7. Odigos and root cause analysis

In the previous section we demonstrated how, by using the Odigos framework, data quality problems can be approached in both prognostic and diagnostic manners. In diagnostic mode, identified data quality issues are related to the semiotic content affected (in the centre of the model). Then, by tracing arcs from the centre of the model outwards, the immediate cause of the data quality issue can be identified (in one of the worlds). Lastly, by considering the arcs on the periphery of the framework (representing interactions between worlds) that lead to the identified world, the root cause of the data quality issue can be identified. In prognostic mode, interactions between worlds can be investigated to determine likely effects in individual worlds, and ultimately, potential data quality issues in the semiotic content.

In the second part of this paper, our aim is to provide more specific definitions and steps in the application of the Odigos framework to discovering the root causes of existing and anticipated data quality issues, and to providing insights into context-sensitive mitigation and data repair actions.

In order to that, we focus on the event log imperfection patterns described in (Suriadi et al., 2017) as this is the only work that not only describes commonly encountered event log quality issues, but considers (with examples) direct causes of the quality issues. Suriadi et al. (2017) show that, in many instances, data quality issues in event logs, while differing in the detailed data values, have certain common (structural) features allowing them to be generalised as (log imperfection) *patterns*. A pattern based approach helps to deal with chaotic domains and make sense of complexity and abundance (Alexander, 1977). By applying the Odigos framework to the log imperfection patterns, we will be able to identify solutions and approaches to recurring data quality problems.

In the following, we explain each pattern and the related immediate and root cause paths. Then in Section 7.2 we introduce the Odigos-RC reference framework and show how this framework can be applied to deal with data quality problems in a real-world context.

7.1. Root cause analysis of the log imperfection patterns

For each of the eleven patterns described by Suriadi et al. (2017), we describe the pattern, then adopt the following approach to using the framework to identify the root causes:

(1) Identify the semiotic content affected (coded as **SCn** – see Table 3 for descriptions).
(2) Identify the potential immediate causes that lead to the data quality problem/pattern (coded as **ICn.n** – see Table 4 for descriptions). In the Odigos framework the immediate causes of data quality problems are process participants, data curators, or IT systems.
(3) Use the Odigos framework to identify potential causal paths that lead to the immediate cause (identified above) to be enacted (coded as **RCn.n** – see Table 5 for descriptions).

To identify the immediate and root causes of each data quality pattern we used brain storming and thought experience (Brown & Fehige, 2019) sessions drawing on the Odigos framework and the authors' experience with process mining. The thought experience was aimed to answer questions such as "how a specific pattern of data quality can be created in event data and how it relates to people (personal world) and IT system (material world)?", "how the immediate cause could come into place? How it could be explained based on the interactions between social, material and personal worlds?"

Form-based Event Capture: The Form-based Event Capture pattern is a direct result of system interface design. The user interface comprises a set of forms with each form being a container for a number of data fields. Once data has been entered, the user clicks 'Save' or 'Submit' and the system logs the form field values, all timestamped at the time the user clicked 'Save' or 'Submit' thus giving the appearance that all activities recorded on the form happened at the same time.

Semiotic Content

SC1 This pattern will affect event data. The signature of this pattern in an event log is groups of events, in the same case, recorded with the same timestamp. Further, these events can be grouped into sets of activities. Such groups of activities occur in multiple cases.

Immediate Cause	*Related World*
IC1.1→SC1	Material
IT system(s) *constraining* the recording of event logs by logging at **too fine** a level of granularity (relative to analysis). That is, the system is not logging the process activity, but is logging either data elements, or activities at too fine a level. For example, all the steps an X-ray machine goes through in warming up to take an image, rather than the business process step of "take an X-ray image". This will result in event level abstraction issues.	
IC1.2→SC1	Material
IT system(s) *constraining* the recording of event logs by logging at **too coarse** a level of granularity (relative to analysis). That is, the logging of timestamps is not specific enough to accurately capture the actual time of the event. This can come about where the field being captured in a form is a date and an activity is derived from the date (as per *Event Constructor* approach in (Andrews, van Dun, et al., 2020)). Where activities are recorded on the same day, they will appear as simultaneous at the level of granularity of the timestamp. This will result in event ordering issues.	
IC3.2→SC1	Material
IT system(s) *constraining* the recording of event logs by capturing all fields on a form when only some have changed. This will give the appearance that more activities than actually occurred have taken place.	

Root Cause	*Related Worlds*
RC1.1→IC1.2→SC1	Social–Material
System design (based on the business requirements) *Shaping* the user interface – Process and system designers have decided on a 'forms' style interface where groups of logically related data elements are captured on a set of (one or more) screens with logging being triggered by user 'saving' or 'submitting' each form. Form fields are written to the log, as separate entries, with each entry having the same timestamp (i.e. when the user 'saved' the form.	
RC2.1→IC1.1, IC1.2→SC1	Social–Material
System design *Shaping* the (non-process aware) logging level – The system is designed/constructed/configured to use a traditional audit/change log as opposed to being process-aware (and writing an event log). Depending on the logging level design, activity timestamps may be either too fine grained or too coarse grained.	
RC2.2→IC3.2→SC1	Social–Material
System design *Shaping* the design of the logging level – Users update some of the fields on a form. The system designed logging mechanism is such that all of the fields on the form are written to the log. This gives the appearance of events having occurred when actually they did not.	
RC5.1→IC3.2→SC1	Personal–Material
Manual data-entry (process participant) *Modulating* the use of the system – User frequently accesses the form (partial completion); Back button of the system to reopen the form. On exit, all the fields on the form are logged, giving the impression that events have occurred when actually they did not.	

Inadvertent Time Travel: This pattern refers to a situation where, due to anomalies in timestamp values, two events that actually occurred in a particular order in real-life, appear to have occurred in a different order in the event log. An example is

the so-called 'midnight problem' where event A happens just before midnight on a particular date, and event B happens just after midnight (i.e. the next day). If the process participant correctly records the time part of each event, but forgets to change the date part, event B will appear to have happened before event A.

Semiotic Content

SC2 This pattern will affect event data. Infrequent (small number of cases) exhibit incorrect (unexpected) temporal event ordering which arises from an incorrect timestamp attribute value.

Immediate Cause	*Related World*
IC1.1, IC1.2→SC2	Material
IT System *constraining* event data through being configured to record at a granularity which is either ill-suited for the purpose of analysis, or inconsistent with other systems from which event data will be extracted for inclusion in the event log.	
IC3.3→SC2	Material
Technical issues *constraining* accurate recording of timestamp values such as faults with sensors on devices intended to automatically record events. These can include faulty devices or loss of connectivity.	
IC5.1→SC2	Material
Technical issues *constraining* accurate recording of timestamp values. For instance, system clocks across multiple systems are not synchronised, or system clocks across multiple systems using date conventions (e.g. Gregorian and Julian calendars) or formats, or faulty sensors not recording correct time.	
IC7.1→SC2	Personal
Process participant *creating* event data through manual data entry either selects incorrect date/time or types an incorrect date/time.	
IC8.1→SC2	Personal
Data curator in *creating* event logs from recorded event data (Extract-Transform-Load) uses an incorrect date/time "mask" while extracting the event log from event data. For instance, instead of using 'yyyy-mm-dd HH:MM:s.sss' (year-month-day 24hour:**minute**:second.ms) uses 'yyyy-mm-dd HH:mm:s.sss' (year-month-day 24hour:**month**:second.ms).	

Root Cause	*Related Worlds*
RC2.5→IC5.1→SC2	Social-Material
IT system design/implementation lacks suitable data validation to prompt/warn about technical fault and automatic recording of events.	
RC3.1→IC1.1, IC1.2→SC2	Social–Material
System configuration – Multiple, different requirements for recording/reporting are reflected in the system design (as it affects data type or recording level of date/time attributes) *shapes* system configuration. This may result in the system not supporting logging at granularity required to provide proper event separation.	
RC3.2→IC1.1, IC1.2→SC2	Social-–Material
System configuration – Different date conventions or different time zones across multiple, international systems from which data is to be extracted *shapes* system configuration such that individual systems record only local time, not universal time.	

Root Cause	*Related Worlds*
RC6.1→IC7.1→SC2	Material–Personal
Manual data-entry (process participant) – System design/implementation permits/validates date/time value without taking into account temporal constraints of the process, i.e. that event A must happen before event B. For instance, the so-called "midnight" problem where event A happens just before midnight and event B happens just after midnight. In recording the date/time for the two events, the process participant correctly records the time component of each event, but forgets that the date has changed (at the midnight) boundary and records the same date for both events. Both date/times are valid, however, it appears as though event B happened before event A, thus violating the partial ordering constraint.	
RC7.1→IC8.1→SC2	Material–Personal
Event log extraction (data curator) – Existing tools and techniques available to data curator, *modulate* data curator in creating the correct date/time format for ETL. For instance, converting from mm/dd/yyyy format to dd/mm/yyyy is intended but some tools, if it is not possible to put a date in the selected format, will convert it into a format that makes sense for the data. If this is not recognised by the data curator, some events, following ETL, may have an incorrect timestamp.	
RC10.1→IC7.1→SC2	Social—Personal
Manual data entry (process participant) – Recording of events "after the fact", e.g. doctor recording clinical notes at end of shift rather than at the bedside. This can result in manually entered/selected dates being inaccurate.	

Unanchored Event: This pattern refers to a situation where date/time values are accurate but are recorded in a format different from the format expected by the tool being used to analyse the data. Such format variations include the confusion between month-day vs day-month format, use of the colon (':') symbol vs the dot ('.') symbol as a separator between hour, minute, and second information, and differences in the way in which timezone information is encoded.

Semiotic Content
SC3 This pattern will affect the event log introducing ambiguity into timestamps and can result in event ordering problems.

Immediate Cause	*Related World*
IC4.1→SC3	Material
Clocks across different IT system(s) are configured to use different date/time conventions thus *constraining* the format of the timestamps in event data.	
IC4.2→SC3	Material
IT system(s) *constraining* date representation in a concrete format (thus forcing some programmatic conversion to a designated format).	
IC7.1→SC3	Personal
Process participants making data entry errors and thus *creating* timestamps with incorrect format in the recorded event data.	
IC7.4→SC3	Personal
Process participants use different data conventions, e.g. American vs British date format, thus *creating* timestamps in event data which are are not suitable for the purpose of process mining (where the date entered is valid in both formats but is to be interpreted in the alternate representation).	

Immediate Cause	*Related World*
IC8.1→SC3	Personal
Data curator, in extracting date/time values uses an incorrect format and *creates* the inappropriate timestamp format in event logs.	

Root Cause	*Related Worlds*
RC3.1→IC4.1→SC3	Social–Material
Systems configuration – Different requirements for time/data records and reporting across different units of the organisation *shapes* systems configuration and the time format used across different systems.	
RC3.1→IC4.2→SC3	Social–Material
Different requirements for recording or reporting across different departments *shapes* the system design in logging level to record date and time in concrete format which needs to be changed to a proper format for process mining analysis.	
RC3.2→IC4.1→SC3	Social–Material
International companies across different locales may use different date conventions *shaping* the design of IT systems in terms of systems clocks and date conventions, such that systems record only local time not universal time.	
RC5.4→IC7.1→SC3	Material–Personal
IT system(s) in the presentation layer lack suitable validation mechanisms to prompt, warn or prohibit (*modulate*) process participants from making errors and entering incorrect values.	
RC7.1→IC8.1→SC3	Material–Personal
IT systems and existing tools are *modulating* the data curator's choice in selecting the format for ETL, e.g., converting from mm/dd/yyyy format to dd/mm/yyyy. (For some tools, if it is not possible to put a date in the selected format, the tool will put it into a format that makes sense for the data. For instance, 03/05/2000, when converted would be a legitimate date 05/03/2020 so the system would make the conversion. However, 03/14/2020 would not convert to a sensible date and so may be left as 03/14/2020 -– in this case, the system thinks conversion has already been done).	
RC9.1→IC8.1→SC3	Social–Personal
Social world is *inculcating* data curator through providing training or experience in using different tools or understanding process mining requirements. The data curator's knowledge (or lack thereof) in the selection and (proper) use of tools and creation of the correct format for process mining analysis can be one of the root causes of unanchored events.	
RC10.2→IC7.4→SC3	Social–Personal
Different users may use different date/time conventions when entering data. In this case, social conventions *inculcate* process participants to create event data with unsuitable timestamps.	

Scattered Event: This pattern refers to events in an event log which have attributes that contain information that can be used to derive new events. In other words, there is information contained within an existing event log that can be exploited to construct additional events, but, the information is hidden in attribute values of several events (often attributes that are textual description or notes fields). The pattern may also arise where the attributes of a single event are recorded as separate entries in the source event log.

Semiotic Content

SC4 This pattern affects the event data and the log in that certain events that did happen in real life are not directly recorded as events in the log meaning that certain process steps that were executed in real-life are missing in the event log.
N.B. There are some similarities between the Scattered Event and Scattered Case patterns, except Scattered Event is finer grained → some causal paths and mitigation may be the same as Scattered Case.

Immediate Cause	*Related World*
IC2.1→SC4	Material
IT system user interface designed such that individual event attributes (case identifier, date/time, activity, resource, ...) are captured in separate form fields and logging writes these individual data fields as separate log records.	
IC2.2→SC4	Material
IT system user interface designed such that multiple event attributes are captured in a single "notes" or "description" form field.	

Root Cause	*Related Worlds*
RC2.1→IC2.1, IC2.2→SC4	Social–Material
System design *shaped* logging without consideration of the potential use/analysis of the data, i.e. not process-aware.	

Elusive Case: This pattern refers to a log in which events are not specifically linked to a case identifier and will often be seen in a log extracted from a non-process aware system or is not designed to support process analysis. An example would be a web server log which simply deals with a stream of incoming requests for pages. Even where multiple requests are received from the same source (IP address), there is no way to link these requests into distinct 'cases'.

Semiotic Content

SC5 No attribute(s) uniquely link events to a case. In other word, the concept of case cannot be identified by using the data in the event log.

Immediate Cause	*Related World*
IC3.1→SC5	Material
IT systems are designed in a way which *constrains* the recorded event data by not logging a case ID or an event ID as part of attributes of events.	
IC7.2→SC5	Personal
Manual data entry (process participants) *creates* potential for inaccurate values recorded for case ID or event ID in the event data.	
IC8.2→SC5	Personal
Data curator incorrectly extracts values for case ID or event ID in *creating* event logs.	

Root Cause	*Related Worlds*
RC2.1→IC3.1→SC5	Social-–Material
A notion of process has not been a concern in the business requirements, thus *shaping* the design of IT systems without consideration of logging the notion of "case".	

Root Cause	*Related Worlds*
RC4.1→IC3.1→SC5	Social–Material
Process mining project's analysis/question of interest imposes a "process" where none is formalised. It implies that the business requirements which *shaped* the design of the IT systems in the first place are logging event data in a way that may not have an attribute that can act as a case identifier.	
RC4.2→IC3.1→SC5	Social—Material
A siloed organisation culture has *shaped* the design of IT systems resulting in implementation of various systems without established record linkages between them; or, individual systems are process-aware but difficulties in linking across systems remain.	
RC6.2→IC7.2→SC5	Material-–Personal
IT system design/implementation (presentation layer) does not provide adequate search/validation to identify existing (or non-matching) case identifiers. Hence *modulation* of process participant's behaviour when entering identifiers manually (with typos) is marked by an inability to link these to related identifiers or event data.	
RC9.1→IC8.2→SC5	Social–Personal
data curators are *inculcated* by their training and their experiences within the social context. Even when they attempt to merge data from different systems they may lack technical, domain, or process mining knowledge and thus can not correctly match *all* entities/episodes across multiple record sets.	

Scattered Case: This pattern manifests where an event log is constructed from multiple source data sets each of which used different (case) identifiers for the same real-world entity. The record sets were amalgamated into a single event log (by the data curator) without proper linking.

Semiotic Content
SC6 This pattern affect the event log and results in multiple (partial) cases, i.e. common sets of activities which are subsets of the total set of activities, but no cases in the log comprised of all their activities.

Immediate Cause	*Related World*
IC6.1→SC6	Material
Integration of data across multiple IT systems *constrains* the quality of the event log. In this case, event data are recorded in multiple, disparate IT systems with no unified "case" view across different systems.	
IC7.2→SC6	Personal
Manual entry of data (process participants) in at least one of the multiple, disparate systems *creates* new, or applies incorrect existing identifier resulting in a "split" case (which will need to be reconstituted later). For example, a patient presenting at a medical centre receives a prescription at an unrelated pharmacy. The patient will have a unique patient number at the medical centre and a unique customer number at the pharmacy. As the underlying systems are distinct, the patient number and customer number will be different, thus creating two distinct (partial) cases.	
IC8.3→SC6:	Personal
Data curator has to merge data across different systems and even though a unique identifier exist for linking data, for various reasons (root causes) the extracted data include different identifiers for the same entity.	

Root Cause	*Related Worlds*
RC4.1→IC6.1→SC6	Social–Material
Process mining project scope is across multiple disjoint processes and IT systems. This could be the case for inter or intra organisations studies. In this scenario, the requirement of the process mining project (having a unified case ID) is not supported by IT systems.	
RC4.2→IC6.1→SC6	Social–Material
The siloed organisation culture has *shaped* the design of IT architecture resulting in multiple, disparate IT systems. Record linkage between systems is not established properly.	
RC6.2→IC7.2→SC6	Material–Personal
IT systems design both in application and presentation layer does not provide adequate search/validation to identify existing (or non-matching) case identifier. This *modulates* process participants' data entry behaviour when choosing incorrect case identifiers.	
RC10.3→IC7.2→SC6	Social–Personal
The nature of the job and the pressure of the work environment *inculcate* process participants' behaviour by making it more convenient for them to create a new case rather than search for a possibly existing and matching case.	
RC8.1→IC8.3→SC6	Material–Personal
Data curator does not take an overarching process/case perspective when extracting event log(s) from source event data. Instead, the data curator simply combines the logs from the contributing systems and merely ensures the event log can be read by process mining tools. For example, a patient will be given a registration number in ED and then in hospital will be given an admission number. Simply combining the two databases (without record matching/linkage) will not capture the notion of a *case*, in the process mining context, as the patients' end-to-end hospital encounter. Rather, patients whose hospital encounter included both an ED presentation and a hospital admission will have their end-to-end hospital encounter scattered across two distinct cases in the event log (one for ED, and one for hospital admission).	
RC9.1→IC8.3→SC6	Social–Personal
Data curators are *inculcated* by their training and their experiences within the social context. Even when they attempt to merge data from different systems they may lack technical, domain, or process mining knowledge and thus can not correctly match *all* entities/episodes across multiple record sets. The unmatched entities result in partial cases.	
RC11.1→IC8.3→SC6	Social—Personal
Lack of communication of analysis requirements *inculcates* data curator in extraction of event logs. Organisational boundaries (lack of communication, lack of transparency – siloed structure) limit the analyst's knowledge of what data is available (in the worst case, the analyst does not know that linking data exists) leading to poor specifications being handed over to the data curator (for event log extraction).	
RC12.1→IC8.3→SC6	Social—Personal
Legal and privacy concerns and regulations *inculcate* data curator in extracting event logs. Some data elements are on purpose hidden (through privacy concerns). The hidden elements (not available to the data curator) are critical to linking records, thus hindering the reconstruction of cases.	

Collateral Events: Collateral events are multiple events in a log that refer to a single process step in a case. This could result from the event log being constructed from

multiple systems, each of which records a reference to the same process step, or where the system's log records detailed, low-level events (such as opening and closing of a form) instead of only the relevant process step.

Semiotic Content

SC7 This pattern creates unnecessary noise in the event data with often trivial, low level activities which in themselves do not represent process steps, and do not contribute to deriving meaningful insights from the event data.

Immediate Cause	*Related World*
IC2.1→SC7	Material
IT systems are *constraining* the recording of event data by logging events in too fine grained levels such that each event relates to the same activity, none, in themselves, represent a process step relevant to the analysis.	
IC2.3→SC7	Material
IT systems application layer *constrains* recording of event data by logging activities with a mixed level of granularity, meaning that there are some events which are too fine grained and not appropriate for the purpose of analysis.	
IC6.1→SC7	Material
Event data is recorded in multiple, disparate systems. Accordingly, integrating data from multiple databases is *constraining* the accuracy of events in event data with multiple events (from different systems) referring to the same process step.	

Root Cause	*Related Worlds*
RC2.1→IC2.1→SC7	Social–Material
The design of IT systems is *shaped* by business requirements or historical conventions resulting in activities being logged in too fine-grained levels.	
RC3.1→IC2.3→SC7	Social–Material
The design of the application layer is *shaped* by business requirements across different units of an organisation. Different requirements for recording/reporting shapes the configuration of different applications, and the level of granularity in which activities are recorded.	
RC4.2→IC6.1→SC7	Social–Material
The siloed culture within an organisation *shapes* the design of IT systems and, as a result, multiple disparate systems are being used across different departments. Integrating logs across these systems can create multiple events representing the same process step.	

Polluted Label: This pattern refers to groups of events in the log having attribute values that are structurally the same, yet the individual values of the attribute are different from each other. This could arise where the attribute is constructed from a mix of boiler plate text and values entered into a form. For example, the system will record case id in the same field used to record activity labels.

Semiotic Content

SC8 This pattern affects the event data and the event log, and will result in discovered process models over-fitting the event log due to the multiple versions of the attribute not being abstracted out. The pattern may affect event attributes such as case identifiers, activity labels or resource identifiers that are critical for process mining.

Immediate Cause	*Related World*
IC2.2→SC8	Material
IT systems *constrain* the recording of event data by logging multiple values in a same field. This could be a default setting of the systems or a setting created by users to address some business requirements.	
IC7.4→SC8	Personal
Process participants *create* activity labels, adopting different conventions when manually recording event data. This results in different syntax used to represent the same concept in event data.	

Root Cause	*Related Worlds*
RC2.1→IC2.2→SC8	Social–Material
The design of IT systems and the way activities are logged to event data is *shaped* by business requirement or traditional auditing requirements. In this case, recording multiple values to one field is (or was) useful for some proposes but not for process mining analysis.	
RC10.2→IC7.4→SC8	Social–Personal
Different social settings or business requirements *inculcate* process participants and thus they may use different conventions for recording event attributes and activity labels.	

Distorted Label: This pattern affects event attributes such as the activity label and is characterised by existence of two or more values of the attribute that are not an exact match with each other but have strong similarities syntactically and semantically. The pattern may be introduced through incorrect data entry (e.g. 'typos').

Semiotic Content

SC9 The pattern, where it applies to the event activity label, means that the label does not accurately reflect the process step that generated the event data. Each variation of the label will be treated as a separate activity by process mining tools, thus negatively impacting the discovered process model.

Immediate Cause	*Related World*
IC7.3→SC9	Personal
Process participants *create* event data (manual recording) and make keyboard proximity errors, general keying errors, etc.	
IC7.4→SC9	Personal
Process participants *create* event data by manually recording activity labels and attributes. In this case, process participants use their own terminologies and language based on their experiences or cultural norms. This can result in activity labels which syntactically are slightly different. For example, due to the use of different spellings for the same word.	

Root Cause	*Related Worlds*
RC5.2→IC7.3→SC9	Material–Personal
The business or task requirements *shape* the presentation layer of IT systems in a way to allow users to overwrite system generated values for activity labels. Even though this feature may be useful to address some user and tasks requirements, it can potentially result in recording of incorrect values for activity labels.	
RC5.4→IC7.3→SC9	Material–Personal
IT system (presentation layer) design is *modulating* process participants' behaviour by lacking suitable data validation to prompt/warn/prohibit users from entering incorrect values. Thus, if process participants make an error, the incorrect value will be recorded without generating any warning.	
RC10.2→IC7.4→SC9	Social–Personal
Different languages, cultures used across different organisation units *inculcate* process participants' behaviour in choosing the values for different data fields in the system.	

Synonymous Labels: There is a group of values (of certain attributes in an event log) that are syntactically different but semantically similar. This pattern may arise where an event log is constructed from multiple systems, each of which refers to the same process step by a different name.

Semiotic Content

SC10 Where the log exhibits this pattern, the event data will have multiple names for the same real-world attribute thus creating ambiguity in the event log.

Immediate Cause	*Related World*
IC6.1→SC10	Material
IT systems (integration of databases) is *constraining* the extraction of event logs. Event data is recorded in multiple, disparate systems with different configurations and this can result in different values for the same concept. This could be the case in cross-organisational analysis involving multiple systems that support the same process or the same system used in different organisations being configured differently, e.g. different units of measure.	
IC7.3→SC10	Personal
Process participants *create* event data by manually recording activity labels and attributes. In this case, process participants may enter or select the incorrect values by mistake.This will result in sporadic synonymous labels which are syntactically different but have the same semantic meaning.	
IC7.4→SC10	Personal
Process participants *create* event data by manually recording activity labels and attributes. In this case, process participants use their own terminologies, language, abbreviations, etc. based on their experiences or cultural norms. This will cause different syntactic values with same semantic meanings.	

Root Cause	*Related Worlds*
RC3.1→IC6.1→SC10	Social–Material
Different requirements for reporting and different terminologies *shape* the configuration of IT systems in application layer and the design in presentation layer. IT systems are configured and designed based on the requirements and terminologies used by the users, however, this creates inconsistencies when trying to integrate data across different units or organisations.	
RC5.4→IC7.3→SC10	Material–Personal
IT system (application layer) is *modulating* process participants behaviour by allowing them to enter incorrect data or choosing the wrong values. Systems could offer functionality where if users enter incorrect values in a data field a warning is given. This would avoid these sporadic issues as much as possible.	
RC10.2→IC7.4→SC10	Social–Personal
Different languages, cultures used across different organisation units *inculcate* process participants' behaviour in choosing the values for different data fields in the system. For example in some departments they may choose to use an abbreviation for a specific word while in other departments they would write the whole word.	

Homonymous Label: An activity is apparently repeated multiple times within a case (same activity label applied to each occurrence of the activity), but the interpretation of the activity, from a process perspective, differs across the various occurrences. As an example, the act of opening a form to record data and the act of opening the form to review previously entered data are considered the same activity.

Semiotic Content
SC11 This pattern manifests as multiple repetitions of apparently the same activity, i.e. events that have the same activity label occurring in different parts of the process.

Immediate Cause	*Related World*
IC3.2→SC11	Material
IT systems *constraining* the record of event data by capturing events that did not happen in reality. In this case, IT systems re-record all the values on a form even if only one value has been changed. For instance, user opens a form to 'write' an entry, or user opens the same form to 'read' an entry and assigns the same label in the case say of clinician recording vital signs and subsequent reference to the recorded vital signs. This will result in repeated activities which is not the case in reality.	
IC7.3→ SC11	Personal
Process participants *create* event data through manually entering activity labels or selecting activities. In this case, user mistakes (proximity errors) in data entry can create sporadic errors in recorded event data and create incorrect repetition of activities.	
IC7.4→SC11	Personal
Process participants *create* event data through manually entering activity names or selecting activities. In this case a systemic manual entry error may happen because process participants (in different roles and across different departments) use the same terminology (same syntax) to refer to different activity labels.	

Root Cause	*Related Worlds*
RC2.4→IC3.2→SC11	Social–Material
IT systems are designed without consideration of the process flow. The lack of attention to the actual processes and the social interactions in the context of the organisation (e.g. hierarchies in reporting and access to different forms) *shapes* the design of application layer of IT systems as it pertains to logging – the system design does not take into account contextual data about the activities and creates the same activity labels for different types of processes.	
RC5.4→IC7.4→SC11	Material–Personal
System design in application layer *modulates* process participants' behaviour in manual entry of activity labels. In this case, the system allows users to choose the wrong values without creating any warning.	
RC10.2→IC7.3→SC11	Social–Personal
Terminologies and connotations on the social level *inculcates* process participants in manually entering activities and creating the event data. In this case, the same terminology is used by users across different departments to refer to different activities.	

7.2. Using Odigos root cause analysis for targeted data quality improvement

In this section we suggest an approach for dealing with the root causes of data quality issues in event data by constructing a directed network derived from applying the Odigos framework to the 11 log imperfection patterns presented in Section 7. We refer to this directed network, and the related analysis of its vertices, as the Odigos Root Cause (Odigos-RC) reference framework since it can assist in identifying and analysing the possible immediate and root causes of data quality problems, and to subsequently devise a suitable solution i.e. mitigation, control, or cleaning[8].

Accordingly, we formalise a directed network suitable for visualising the relationship between root causes, immediate causes, and data quality problems (imperfection patterns) in the semiotic content. Let RC be the set of root causes, IC be the set of immediate causes, and SC be the set of data quality affecting the semiotic content. We construct a directed network $G = (V, R, f)$ where V is the set of vertices, R is the set of directed arcs, and f is a function that assigns a value to each directed arc. $V = RC \cup IC \cup SC$. $R = \{(v_i, v_j)|v_i \in RC \wedge v_j \in IC\} \cup \{(v_i, v_j|v_i \in IC \wedge v_j \in SC\}$. In this case, the function f represents the frequency of occurrence of the pair of vertices (v_i, v_j) in the observed causal paths. We denote $I(v)$ as the *indegree* of vertex v, and $O(v)$ as the *outdegree* of vertex v.

Here we discuss how the directed network and its different attributes including arc frequencies and in- and outdegrees can be interpreted to provide insights into understanding the root causes of data quality problems.

The approach is predicated on the following assumptions:

(1) Root causes that are involved in multiple causal pathways affect data quality more significantly than root causes that are involved in few causal pathways.
(2) Immediate causes that are related to multiple issues in the semiotic content affect data quality more significantly than immediate causes that are related to few semiotic content issues.

[8] As we explain later, this framework does not present a complete list of data quality problems and their causes but supports an analytical approach for identifying and analysing those problems

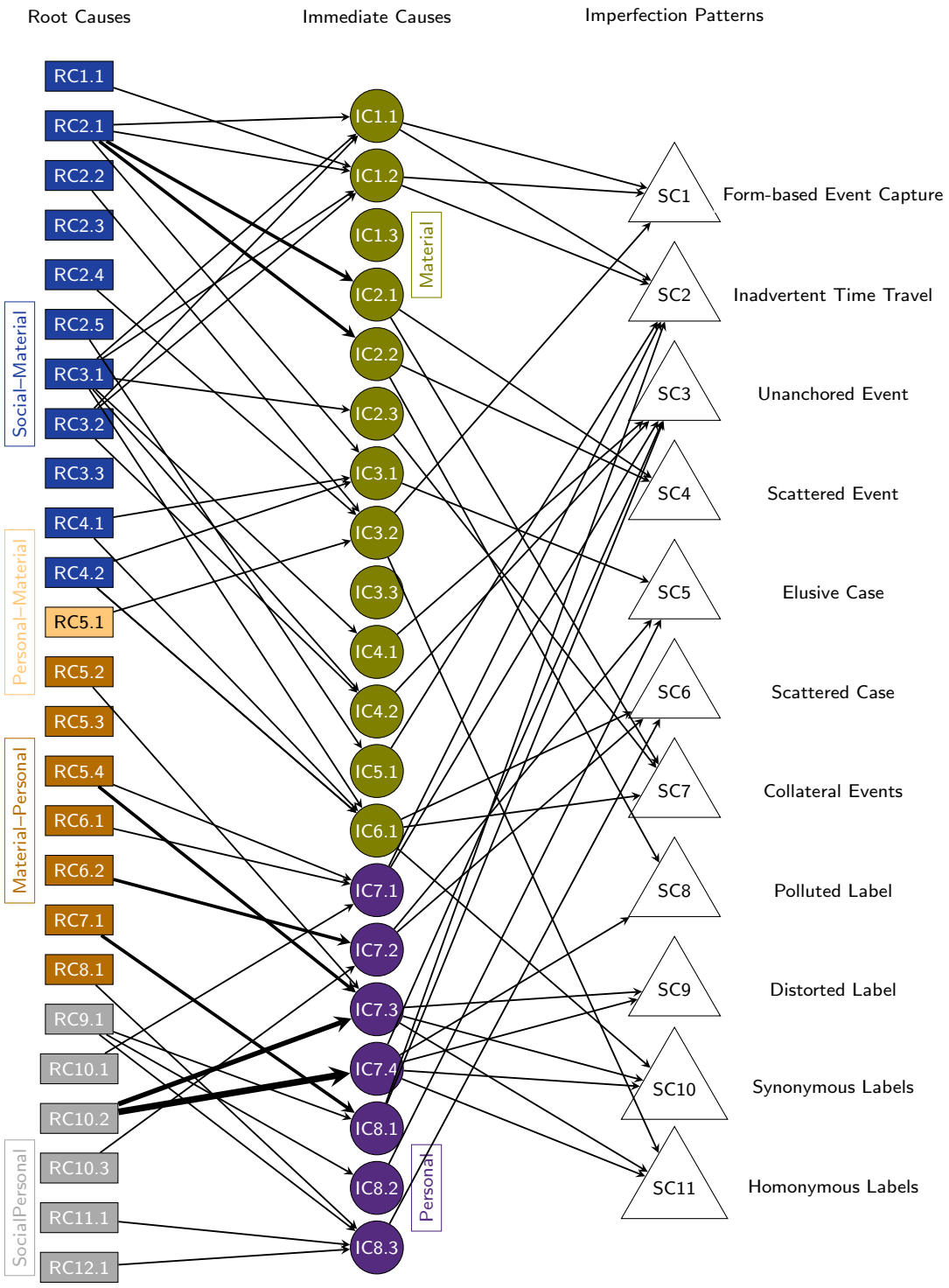

Figure 7.: Directed network representation of causal paths associated with log imperfection patterns. Arc thickness indicates arc frequency (min=1, max=4).

(3) Semiotic content issues related to multiple immediate causes represent data quality issues where it may be more efficient to actually deal with the symptoms and *ex post* clean the data to deal with the problem.

(4) The frequency of any Root cause-Immediate cause pair is a measure of the involvement of that pathway in data quality issues.

If the causes and data quality issues are treated as vertices of a directed network, and the causal pathways are treated as directed arcs, then

- Root cause involvement in causal pathways can be ascertained by calculating the *outdegree* of the associated vertex;
- Immediate cause involvement can be ascertained by calculating the *in-* and *outdegree* of the associated vertex.

Root cause–Immediate cause pair frequency can be represented as an arc weight. (Where the arc weight is greater than 1, the same Root cause–Immediate cause pair is responsible for multiple data quality issues).

Arc frequency represents the number of times the combination of a root cause (RC) and the immediate cause (IC) are repeated in the 11 patterns. The higher the arc frequency, the greater is the importance of addressing the root cause of related data quality issues (rather than simply cleaning the symptoms). However, in each real-world context, the various arcs may be present with different frequencies from those shown in Table 2.

Root Cause	out Degree	in Degree	Immediate Cause	out Degree	in Degree	Pattern
RC1.1	1	3	IC1.1	2	3	SC1
RC2.1	**7**	4	IC1.2	2	**6**	SC2
RC2.2	1	0	IC1.3	0	**4**	SC3
RC2.3	0	2	IC2.1	2	2	SC4
RC2.4	1	2	IC2.2	2	3	SC5
RC2.5	1	1	IC2.3	1	2	SC6
RC3.1	**7**	3	IC3.1	1	3	SC7
RC3.2	3	3	IC3.2	2	2	SC8
RC3.3	0	0	IC3.3	0	2	SC9
RC4.1	2	1	IC4.1	1	3	SC10
RC4.2	3	2	IC4.2	1	3	SC11
RC5.1	1	1	IC5.1	1		
RC5.2	1	**4**	IC6.1	3		
RC5.3	0	3	IC7.1	2		
RC5.4	4	3	IC7.2	2		
RC6.1	1	**7**	IC7.3	3		
RC6.2	2	4	IC7.4	**5**		
RC7.1	2	3	IC8.1	2		
RC8.1	1	1	IC8.2	1		
RC9.1	3	4	IC8.3	1		
RC10.1	1					
RC10.2	**7**					
RC10.3	1					
RC11.1	1					
RC12.1	1					

Table 2.: Degrees of graph vertices

Table 2 gives the *in-* and *outdegrees* of each of the vertices of the directed network. It can be seen that the most frequently occurring root causes are RC2.1 (Logging – traditional audit log instead of being process-aware), RC10.2 (Different process participants may adopt different conventions when recording data attributes), and RC3.1 (Different requirements for recording/reporting across different modules of the system, or different systems design (affecting data type or recording level of date/time attribute)).

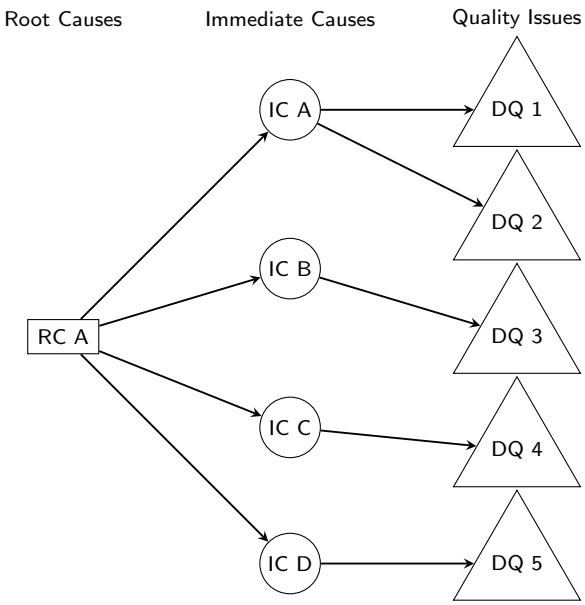

Figure 8.: Single root cause – multiple immediate causes and data quality issues.

The directed graph shown in Figure 7 and arc frequencies listed in Table 2 aid in understanding identified data quality issues and provide a valuable starting point in devising context-sensitive mitigation (prevention), control, and repair actions. The root causes with higher outdegree in Table 2, could help in 1) postulating the existence of associated data quality problems in a real-world context, 2) devising a context-sensitive approach to dealing with those root causes. If a root cause with high outdegree exists in a real-world context, there is a higher chance of having at least one of the associated data quality problems in recorded event data. For example, if, when applying the Odigos-RC framework in a prognostic manner in a particular organisational context, RC2.1 (Logging – traditional audit log instead of being process-aware) exists as a root cause, we would assume that there is high chance of seeing at least some of the associated immediate causes (IC1.1, IC1.2, IC2.1, IC2.2, and IC3.1) and consequently, the log imperfection patterns SC1, SC2, SC4, SC5, SC7 and SC8 in the event log. Conversely, if, when applying the framework in a diagnostic approach, we observe log imperfection patterns SC1, SC2, SC4, SC5, SC7 and SC8 as data quality problems in the event log, we could assume that there is a high chance that RC2.1 exists and is (at least partly) responsible for these data quality problems.

To understand how the existence of a root cause node with a high outdegree could influence a data quality approach, consider the situation illustrated in Figure 8 where multiple data quality problems can be traced back to a single root cause. In such a situation, eliminating the root cause will likely reduce the frequency of the data quality issues, and have an overall positive impact on the quality of the event log. Further, by eliminating the root cause (and associated data quality issues), the effort in pre-processing the log in preparation for process mining will be reduced. This is particularly beneficial where process mining analysis is ongoing and forms a regular part of the organisation's business intelligence activities.

The existence, in a real-world context, of an IC node with a high indegree and outdegree (see Figure 9) implies that the IC is related to multiple root causes and multiple data quality issues. In such a case, dealing with all root-causes (RCs) to solve

the data quality issues (mitigation) is most likely complicated (because there could be multiple causes). Accordingly, controlling the immediate cause could be a more efficient approach (at least in the short term). For example, IC7.3 (related to errors in manual data entries by process participants) has a high indegree (value 7) and a high outdegree (value 3). To deal with data quality problems resulting from IC7.3, the best approach could be to avoid manual data entries and use automation techniques(i.e. controlling IC 7.3) as much as possible rather than dealing with root causes. It should be noted that in a given real-world context, not all the root-causes (RCs in Table 2) for an immediate cause may exist. So, the actual indegree and outdegree values for the immediate cause may be different from Table 2.

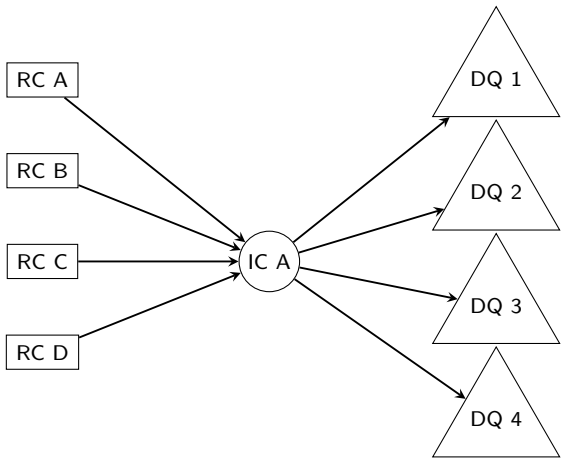

Figure 9.: Multiple root causes and data quality issues – single immediate cause.

Data quality issues with high indegrees (see Figure 10) imply that these quality issues have multiple immediate causes and thus, have a higher chance of occurring in the event log. It also implies that a mitigation approach (resolving all underlying, related RCs) or a controlling approach (dealing with all related ICs) will likely be expensive, time consuming, and require many changes throughout the organisation. In this situation, resolving the data quality problems using data cleaning methods may be the most effective approach. It should be noted that the in- and outdegrees in a real-world context may be different from Table 2 and mitigation, control, and cleaning decisions should be made considering the directed network related to the context.

8. 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. In earlier work (Emamjome et al., 2019), through a survey on process mining case studies, we demonstrated that the current approaches in dealing largely with symptoms of data quality problems have been limiting the impact

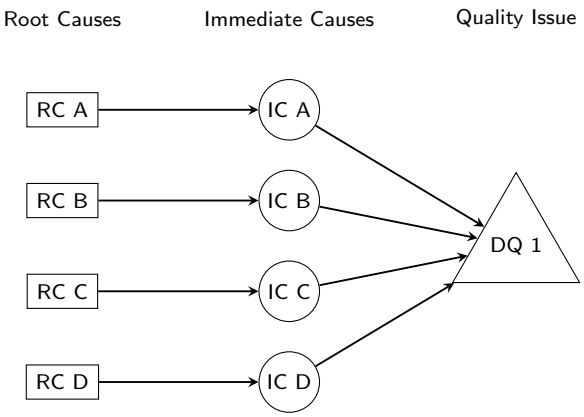

Figure 10.: Multiple root causes and immediate causes – single data quality issue.

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 further development of methodological data pre-processing. By proposing a framework to facilitate identifying root causes of data quality issues in event data, we help analysts 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. We note, as a limitation, that the root causes we have provided are not complete as they were derived from the collective experience of the authors. However, the Odigos-RC framework allows for building, over time, a more and more sophisticated repository of knowledge.

The Odigos-RC reference framework described in this paper is a realisation of Odigos and provides insights into how 11 commonly occurring data quality issues (log imperfection patterns) may be analysed to uncover the root causes that stem from the organisational context. As was noted in (Suriadi et al., 2017), the pattern collection may not be complete, but it is comprehensive. Further, Odigos-RC, as a framework, provides a structured way to accommodate further log imperfection patterns as and when they are described. Odigos-RC provides a means of targeting mitigation and repair activities that are feasible and effective within the organisational context. In some cases, it is most effective to focus data quality improvement actions around resolving root causes. This is indicated where a single root cause is contributory to multiple data quality problems. Conversely, where multiple root causes are contributory to a single data quality issue, it may still be most effective to apply data cleaning techniques to deal with the quality issue symptomatically.

We again point out that the Odigos-RC reference framework presented in this paper was constructed by considering multiple domains, multiple organisational contexts, and multiple event logs. We note, as a final limitation of this work, the lack of an empirical validation of the work through application of the framework in practice. We put this down as future work. It is unlikely that in a specific organisational context, all of the log imperfection patterns, immediate causes, and root causes in Odigos-RC will be present. Therefore, applications of Odigos-RC in practice require identifying the

elements that exist in the context, removing the unrepresented elements, and making mitigation/repair decisions based on the elements that remain (i.e. that are relevant to the context). In a similar vein, Odigos-RC is extensible, thus allowing context specific data quality problems, immediate causes, and root causes to be added to the framework.

Lastly, an opportunity for future work exists in building on preliminary work (Emamjome et al., 2020b) around methodological guidance for applying Odigos in practice.

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, *17*(2), i–xxxii.

Alexander, C. (1977). *A pattern language: towns, buildings, construction.* Oxford university press.

Andrews, R., Emamjome, F., ter Hofstede, A. H., & Reijers, H. A. (2020). An expert lens on data quality in process mining. In *2nd International Conference on Process Mining (ICPM)* (pp. 49–56).

Andrews, R., Suriadi, S., Wynn, M. T., ter Hofstede, A. H., & Rothwell, S. (2018). Improving patient flows at St. Andrew's War Memorial Hospital's emergency department through process mining. In *Business process management cases* (pp. 311–333). Springer.

Andrews, R., van Dun, C. G., Wynn, M. T., Kratsch, W., Röglinger, M., & ter Hofstede, A. H. (2020). Quality-informed semi-automated event log generation for process mining. *Decision Support Systems*, 113265.

Andrews, R., Wynn, M. T., Vallmuur, K., ter Hofstede, A. H., Bosley, E., Elcock, M., & Rashford, S. (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.

Armstrong, P. (1986). Management control strategies and inter-professional competition; the cases of accountancy and personnel management. In D. Knights & H. Willmott (Eds.), *Managing the labour process.* Aldershot Crower.

Ash, J. S., Berg, M., & Coiera, E. (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.

Bose, R. J. C., Mans, R. S., & van der Aalst, W. M. P. (2013). Wanna improve process mining results? In *IEEE symposium on computational intelligence and data mining* (pp. 127–134).

Bose, R. J. C., & van der Aalst, W. M. P. (2010). Trace alignment in process mining: Opportunities for process diagnostics. In *International conference on BPM* (pp. 227–242).

Boudreau, M., & Robey, D. (2005, February). Enacting integrated information technology: A human agency perspective. *Organization Science*, *16*(1), 3–18.

Bozkaya, M., Gabriels, J., & van der Werf, J. M. (2009). Process diagnostics: A method based on process mining. In *International conference on information, process, and knowledge management* (pp. 22–27).

Brown, J. R., & Fehige, Y. (2019). Thought Experiments. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2019/entries/thought-experiment/` (last visited 26/11/2020).

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1997). *Discovering data mining: From concept to implementation.* Prentice Hall PTR New Jersey.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165-1188.

Cheng, H.-J., & Kumar, A. (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 26/11/2020).

D'Adderio, L. (2004). *Inside the virtual product: How organizations create knowledge through software.* Edward Elgar Publishing.

Danermark, B., Ekstrom, M., Jakobsen, L., & Karlsson, J. (2001). *Explaining society: An introduction to critical realism in the social sciences.* Routledge.

Davenport, T. H. (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, F., Andrews, R., & ter Hofstede, A. H. (2019). A case study lens on process mining in practice. In *On the Move to Meaningful Internet Systems: OTM Confederated International Conferences* (pp. 127–145).

Emamjome, F., Andrews, R., ter Hofstede, A., & Reijers, H. (2020b). Signpost-a semiotics-based process mining methodology. In *Proceedings of the 28th european conference on information systems (ecis2020)* (pp. 1–10).

Emamjome, F., Andrews, R., ter Hofstede, A. H., & Reijers, H. A. (2020a). Alohomora: Unlocking data quality causes through event log context. In *European Conference on Information Systems (ECIS).*

Feenberg, A. (2012). *Questioning technology.* Routledge.

Fox, F., Aggarwal, V. R., Whelton, H., & Johnson, O. (2018). A data quality framework for process mining of electronic health record data. In *Ieee international conference on healthcare informatics* (pp. 12–21).

Goes, P. B. (2014). Editor's comments: Big data and IS research. *MIS Quarterly*, *38*(3), iii–viii.

Habermas, J. (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, D. (1998). *Sociological practice: Linking theory and social research.* Sage.

Mans, R. S., van der Aalst, W. M. P., Vanwersch, R. J., & Moleman, A. J. (2013). Process mining in healthcare: Data challenges when answering frequently posed questions. In *Process support and knowledge representation in health care* (pp. 140–153). Springer.

Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in IS research and the implications for replication. *Decision Support Systems*, *115*, A1–A7.

Massa, S., & Testa, S. (2005, July). Data warehouse-in-practice: Exploring the function of expectations in organizational outcomes. *Information and Management*, *42*(5), 709-718.

Mingers, J., & Willcocks, L. (2014). An integrative semiotic framework for information systems: The social, personal and material worlds. *Information and Organization*, *24*(1), 48–70.

Mingers, J., & Willcocks, L. (2017). An integrative semiotic methodology for is research. *Information and Organization*, *27*(1), 17–36.

Mutch, A. (2010). Technology, organization, and structure: A morphogenetic approach. *Organization Science*, *21*(2), 507–520.

Nemati, H. R., & Barko, C. 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.

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

Price, R., & Shanks, G. (2016). A semiotic information quality framework: Development and comparative analysis. In *Enacting research methods in information systems* (pp. 219–250). Springer.

Queensland Audit Office. (2015). *Emergency department performance reporting 3:2014–15.*

Salancik, G. R., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design. *Administrative Science Quarterly*, *23*(2), 224–253.

Strong, D., & Volkoff, O. (2010). Understanding organization-enterprise system fit: A path to theorizing the information technology artifact. *MIS Quarterly*, *34*(4), 731–756.

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

van der Aalst, W. M. P. (2016). *Process mining: Data science in action.* Springer.

van der Aalst, W. M. P., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... others (2011). Process mining manifesto. In *International conference on business process management* (pp. 169–194).

van Eck, M. L., Lu, X., Leemans, S. J., & van der Aalst, W. M. P. (2015). Pm^2: A process mining project methodology. In *International conference on advanced information systems engineering* (pp. 297–313).

Volkoff, O., & Strong, D. (2013). Critical realism and affordances: Theorizing it-associated organizational change processes. *MIS Quarterly*, *37*(3), 819–834.

Volkoff, O., Strong, D. M., & Elmes, M. B. (2007). Technological embeddedness and organizational change. *Organization Science*, *18*(5), 832–848.

Wixom, B. H., & Watson, H. J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, *25*(1), 17-41.

Appendix

Here we provide consolidated definitions of each of the SC, IC, and RC codes.

Code	Affects	Description
SC1	Event data	**Form-based Event Capture** - Groups of events, in the same case,recorded with the same timestamp. Further, these events can be grouped into sets of activities. Such groups of activities occur in multiple cases.
SC2	Event data	**Inadvertent Time Travel** - Infrequent (small number of cases) exhibit incorrect (unexpected) temporal event ordering which arises from an incorrect timestamp attribute value.
SC3	Event log	**Unanchored Event** - The timestamp values of an event log are recorded in a format that is different from that which is expected by the tools used to process the event log.
SC4	Event data	**Scattered Event** - This pattern refers to events in an event log which have attributes that contain further information that can be used to derive new events. In other words, there is information contained within an existing event log that can be exploited to construct additional events, but, the information is hidden in attribute values of several events.
SC5	Event log	**Elusive Case** - No attribute(s) uniquely link events to a case.
SC6	Event log	**Scattered Case** - Event log constructed from multiple source data sets each of which used different (case) identifiers for the same entity. The record sets were amalgamated into a single event log (by the data curator) without properly linking. This results in multiple (partial) cases, i.e. common sets of activities which are subsets of the total set of activities, but no cases with all the activities in the log.
SC7	Event log	**Collateral Events** - Multiple events in a log that refer to a single process step in a case.
SC8	Event data	**Polluted Label** - Event attribute (Activity Label) constructed from mix of boiler plate text and values entered into a form.
SC9	Event data	**Distorted Label** - Event attribute (Activity Label) characterised by existence of two or more values of the attribute that are not an exact match with each other but have strong similarities syntactically and semantically.
SC10	Event data	**Synonymous Labels** - There is a group of values (of certain attributes in an event log) that are syntactically different but semantically similar.
SC11	Event data	**Homonymous Labels** - An activity is apparently repeated multiple times within a case (same activity label applied to each occurrence of the activity), but the interpretation of the activity, from a process perspective, differs across the various occurrences.

Table 3.: Semiotic content

Code	Affected Actor	Affects	Result
IC1.1	IT system - application layer	Logging -Timestamp granularity	Too fine
IC1.2	IT system - application layer	Logging -Timestamp granularity	Too coarse
IC1.3	IT system - application layer	Logging -Timestamp granularity	Mixed
IC2.1	IT system - application layer	Logging - Event granularity	Too fine
IC2.2	IT system - application layer	Logging - Event granularity	Too coarse
IC2.3	IT system - application layer	Logging - Event granularity	Mixed
IC3.1	IT system - application layer	Logging - Event attribute	IT systems do not record both Event ID and Case ID
IC3.2	IT system - application layer	Logging - Event	IT system logging appears to capture event(s) that did not happen in reality
IC3.3	IT system - application layer	Logging - Event	IT systems do not record event(s) that did actually occur in reality
IC4.1	IT system - application layer	Configuration - Date/time	System clocks across multiple systems using different date conventions (e.g. different "day 0" in different systems)
IC4.2	IT system - application layer	Configuration - Date/time	Configuration - Concrete date/time representation (thus forcing some programmatic conversion to a designated format)
IC5.1	IT system - application layer	Automatic data recording	Incorrect attribute value recorded
IC6.1	IT system - data layer	Integration	Event data recorded in multiple, disparate systems – records are not recorded in a single, process-aware system
IC7.1	Process participant	Manual data entry - sporadic	Selecting/typing incorrect attribute value - date/time
IC7.2	Process participant	Manual data entry - sporadic	Selecting/typing incorrect attribute value - case ID
IC7.3	Process participant	Manual data entry - sporadic	Selecting/typing incorrect attribute value - activity label
IC7.4	Process participant	Manual data entry - systemic	Different users may adopt different conventions when manually recording data attributes
IC8.1	Data curator	Extracting event data	Incorrectly extracts values - incorrect date/time mask in ETL
IC8.2	Data curator	Extracting event data	Incorrectly extracts values - attribute values
IC8.3	Data curator	Extracting event data	Merging multiple form fields to create activity label

Table 4.: Immediate causes

Table 5.: Root causes

Code	Pathway	Root cause	Value
RC1.1	Social–Material: System design	Traditional design principles	Forms-based (groups of logically related data items collected on same form
RC2.1	Social–Material: System design	Business requirements	Logging - traditional audit log (instead of being process-aware)
RC2.2	Social–Material: System design	Business requirements	Logging - all fields on form written to log
RC2.3	Social–Material: System design	Business requirements	Logging - only changed values written to log.
RC2.4	Social–Material: System design	Business requirements	Logging does not differentiate between edit and read only mode , implementation does not take into account contextual aspects of the process
RC2.5	Social–Material: System design	Business requirements	System design/implementation lacks suitable data validation to prompt/warn about technical fault
RC3.1	Social–Material: System configuration	Differing requirements across different units of the organisation	Different requirements for recording/reporting across different modules of the system, or different systems design (affecting datatype or recording level of date/time attributes)
RC3.2	Social–Material: System configuration	International companies	Different date conventions or different time zones across multiple systems from which data is to be extracted. System records only local time, not universal time.
RC3.3	Social–Material: System configuration	Process designer in defining the processes	Process designer(s) not recognising link between source event and triggered events
RC4.1	Social–Material: Event log extraction	Multiple disjoint processes (intra and inter-organisation)	Analysis/Question of Interest imposes a “process” where none is formalised
RC4.2	Social–Material: Event log extraction	Siloed organisations	Record linkages between systems not established; or, individual systems are process aware but still have difficulty in linking across systems
RC5.1	Material–Personal: Manual data entry	Presentation layer design	User frequently accesses form (partial completion) either directly or through “Back” button in the case of web-based UI
RC5.2	Material–Personal: Manual data entry	Presentation layer design	System designed so as to allow users to overwrite system generated label/value
RC5.3	Material–Personal: Manual data entry	Presentation layer design	Screen prompts use jargon, language, icons, etc. unfamiliar to the users - users unsure of what constitutes a “correct” data value
RC5.4	Material–Personal: Manual data entry	Presentation layer design	System design/implementation lacks suitable data validation to prompt/warn/prohibit user from entering incorrect value

Continued on next page

Table 5 – continued from previous page

Code	Pathway	Root cause	Value
RC6.1	Material–Personal: Manual data entry	Application layer design	System design/implementation permits/validates date/time format without taking into account temporal constraints of the process
RC6.2	Material–Personal: Manual data entry	Application layer design	System design/implementation does not provide adequate search/validation to identify existing (or non-matching) case identifier
RC7.1	Material–Personal: Event log extraction	Constraints related to data extraction tools.	Data curator's choice (or system forced) in selecting format for ETL, e.g., converting from mm/dd/yyyy format to dd/mm/yyyy.
RC8.1	Material–Personal: Event log extraction	Recording events in level of granularity not suited for analysis	Data curator does not apply a process/case perspective when extracting event log
RC9.1	Social–Personal: Event log extraction	Data curator knowledge in data integration and process mining	Data curator attempts to merge but through lack of technical, domain, or process knowledge, cannot correctly match ALL entities/episodes across multiple record sets
RC10.1	Social–Personal: Manual data entry	Workplace priorities and culture around use of IT	Recording after the fact
RC10.2	Social–Personal: Manual data entry	Different terminologies across different users	Different users (process participants) may adopt different conventions when recording data attributes
RC10.3	Social–Personal: Manual data entry	Work/time pressure or case load	Process participant creates new case rather than searching for existing case
RC11.1	Social–Personal: Event log specification	Poor communication between Analyst and Data curator	Organisational boundaries (lack of communication, lack of transparency (siloed structure)) limits knowledge of what data is available (analyst does not know that linking data exists)
RC12.1	Social–Personal: Event log extraction	Legal and ethical privacy concerns	Some data elements are on purpose hidden (privacy concerns) which hinders reconstructing cases