An Expert Lens on Data Quality in Process Mining

Robert Andrews∗, Fahame Emamjome∗, Arthur H.M. ter Hofstede∗ and Hajo A. Reijers†
∗Queensland University of Technology, Brisbane, Australia
†Utrecht University, Utrecht, The Netherlands

Abstract—The success of a process mining project is highly dependent on the quality of the event log data, the degree to which quality issues are detected, and the way they are resolved. The detection and resolution of data quality issues requires a systematic approach that is aware of the organisational context in which event log data is created. To this end, the Odigos framework has been developed in prior work. The focus of this paper is the validation of this framework through semi-structured interviews with a range of experts in process mining. The experts confirmed the utility of the framework, provided valuable insights into data quality in practical settings, and suggested enhancements to the Odigos framework.

Index Terms—process mining, data quality, Odigos framework, expert validation

I. INTRODUCTION

Event log quality is a critical success factor in any process mining project [1]. Process mining project methodologies that refer to event log quality [2]–[4] make provision for quality issues to be addressed primarily in the pre-processing phase. This has the caveat that further consideration may be required in later phases as event logs are manipulated to address different types of analysis. Curiously, such methodologies pay scant attention to practical elements such as 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 [5]. In many process mining case studies, researchers limit data pre-processing to merely transforming raw event data to a format that can be consumed by process mining tools, and to uncritically report analysis outcomes. We refer to this garbage in - gospel out effect as naive process mining [6]. As is pointed out in [5], identifying the root causes of quality issues in event logs helps researchers to deal effectively with quality issues and, thence, to derive informed insights from their analysis. However, existing approaches to data quality and log cleaning (e.g. [7], [8]) are more focused on treating data quality symptoms (in a given log) than on recognising the root causes of those issues. In [6], the authors propose the notion of informed process mining, which involves a consideration of the context in which a process executes as a means of identifying root causes of event log quality issues. To deal with data quality issues in a systematic way, the Odigos framework [9] guides process mining researchers in diagnostically and prognostically identifying data quality issues in process-related event data. The main contribution of this paper is a qualitative, interview-based, expert evaluation of the Odigos framework, which confirmed the utility of the framework. Additionally, interaction with the experts (i) enhanced our understanding of the problems with data quality issues and awareness around data quality issues, and (ii) resulted in suggestions for refining the framework in the light of their collective experiences. The remainder of this paper is organised as follows. In Section II we discuss unresolved issues relating to event-data quality, which serve to position the Odigos framework. To make this paper self-contained, we briefly describe the main features of the Odigos framework in Section III. In Section IV we describe our approach to evaluating the relevance and usefulness of the Odigos framework through expert interviews. The main results are presented in Section V. In Section VI we discuss some important themes that emerged through analysis of the interviews. In Section VII we discuss possible limitations of our sampling approach as well as insights arising from the evaluation and directions for future work.

II. RELATED WORK

The importance of data quality for data analysis in general, and for process mining in particular, is evident [10]–[12]. Process mining methodologies such as PDM [2], L∗ [3], and PM2 [4] do not provide methodological guidance for the identification of quality issues in process mining studies, and apart from [10], there is little awareness of the impact of data quality on the findings of process mining studies [5]. Event log data quality frameworks, such as those described in [10] and [1] are useful for labeling identified symptoms of data quality. However, neither helps with identifying the causes of those quality issues, nor with recommending possible remedies. Further, the approach in [1] is rooted in hospital information systems, is not generalised, and as a result does not inform researchers in approaching data quality in domains other than healthcare. In [11], the authors, based on their experience in dealing with multiple event logs, abstract a set of commonly occurring event data quality issues as pattern templates which link the manifestation of each data quality issue to likely underlying causes. The authors show that data quality issues can be detected by searching for the relevant pattern’s signature in the event log. They also discuss the impact on a process mining analysis of each pattern and suggest possible remedies for the detected quality issues. The paper provides valuable insights regarding some common root causes of data quality issues. While they may not span the entirety of event log data quality issues, it is a step towards a systematic, generalisable

978-1-7281-9832-3/20/$31.00 ©2020 IEEE
DOI 10.1109/ICPM49681.2020.00018
approach to dealing with data quality issues. Preliminary work towards operationalising this approach was given in [13]. The metrics-based approach to assessing data quality during the preparation and planning stage of process mining projects in [14] highlights how data quality issues can affect the findings of a process mining analysis. The authors argue 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 but do not support this with methodological guidance. The Care Pathways Data Quality Framework (CP-DQF) [15] uses the quality framework described in [10] to support systematic management (identification, recording, mitigation, reporting) of data quality issues affecting process mining research using electronic healthcare records (EHR). CP-DQF provides a comprehensive list of data quality issues in the context of EHR; however, this framework pays limited attention to uncovering root causes of quality issues and is constrained by its exclusive focus on EHR systems. The 3x3 DQA framework described in [16] also deals with requirements of EHR data from a reuse perspective. The framework takes the relationship between task dependence (fitness for use), data dimensionality and data quality assessment as its core feature. The 3x3 DQA framework is, as the name suggests, a matrix with Task Dependence (broken into three data constructs - completeness, correctness and currency) on one axis, and Data Dimensionality (broken into Time, Variables and Patients) on the other axis. Each cell is operationalised by one or more metrics with the determination of what constitutes “sufficient quality” made by the user in the light of the user’s “understanding of the data, the clinical phenomena being examined, and the methods of analysis used” [16, p. 9]. This framework was assessed by expert interview with results of quantitative and qualitative assessments of responses suggesting that the framework was not intuitive, that the operationalised constructs of EHR data quality need to be improved and the cognitive overhead of interpreting the complex guideline logic was a barrier to practical use. The fact that the framework is specific to EHR data could be considered a further limitation.

III. BACKGROUND - ODIGOS FRAMEWORK

The Odigos framework (see Figure 1) suggests a systematic approach to dealing with data quality issues in event logs in both diagnostic and prognostic ways. Based on the principles of Semiotics [17], the Odigos framework provides a systematic understanding of the context of process mining. 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. Figure 1 shows the semiotic content, i.e. actual processes, event data and event logs, in the center of framework. The Odigos framework [9] defines the Social world in terms of both organisational context (situational) and wider social context beyond the organisation (macro) that can influence people and IT systems. The Material world in the context of process mining refers to all IT systems used to support the process. The Odigos framework identifies the different layers (presentation, application, and data) of IT systems as being important when considering data quality issues. The Personal world includes the Process participants who carry out processes, and the Data curator who is responsible for generating event logs from the recorded event data for use by analysts. The central tenet of the Odigos framework is that the semiotic content (processes, event data, event logs) results from direct actions arising from the individual worlds (indicated by the connote, create, and constrain arrows in Figure 1) which are ultimately driven by interactions between the worlds (indicated by the Inculcates, Modulates, and Shapes arrows in Figure 1). Thus, the Odigos framework captures the process mining context. By understanding the context, it is possible to understand/explain the semiotic content, including data quality. In this paper, we refer to the interactions between worlds and semiotics using ‘paths’ (see Table 1 for a complete listing). For example, path 4, the interaction between the Material (IT system) world Modulating the Personal world (Process participants) influencing semiotics (recording event data), could explain why values in a particular field do not bear any relation to the label or database column definition. Process Participants have employed a ‘workaround’ to capture process-critical data with an IT system design that does not properly support the process.

IV. METHODOLOGY

The Odigos framework [9] provides a high level conceptualisation of the process mining context. It aims to help process mining researchers discover the root causes of quality issues in event logs. However, for such a framework to be successfully applied in process mining methodologies, further steps are required to confirm the usability and relevance of the framework. In line with principles given in [18] and [19] a qualitative approach, using semi-structured interviews with process mining experts, was used to explore the usability and relevance of the Odigos framework, and purposive (or judgement) sampling was used to select participants. Interviews
are one of the main data collection methods in qualitative research and purposive sampling is relevant as insights and expert knowledge to properly consider the framework is held by only certain members of the process mining community. The sample included researchers and practitioners deemed to be experts in the field of process mining and data quality.

### A. Expert Profile

We selected a panel of 15 experts of whom 11, coded as E1..E11, agreed to participate. Of the 4 who did not participate, 2 said they were unable to participate due to time constraints and 2 did not reply to the invitation. Selection criteria included: (i) demonstrated experience in practical applications of process mining, e.g. published case studies or employment in a senior role related to process mining; and (ii) demonstrated interest in event log data quality, e.g. publications in this area; or (iii) recommendation from one or more of the selected experts. The experts (a) were distributed across the globe – 3 continents and 6 countries, (b) have a combined research and practice experience of approx. 130 years (average experience = 10.25 years), (c) work across multiple domains (the predominant application area being healthcare followed by education and manufacturing, and (d) have, collectively, 240 process mining publications and 12 (referred) publications relating to event data quality¹.

### B. Interview Process

We used a semi-structured approach in designing the interview protocol [20]. As our aim was to gauge the usefulness and relevance of the Odigos framework through linking to the experts’ experience, our interview approach was based on a protocol that was pretested to ensure questions would be understood and properly interpreted, would yield the appropriate objectives, and would be scoped to encourage open-ended input. The interviewer ensured that all questions in the protocol were covered during the interview. However, digressions to related topics of discussion were permitted in order to increase the richness of the information captured. The interview included an introduction to the Odigos framework through sets of examples. Participants were free to ask questions or comment during this time. Next, the interview continued using three different sets of questions. The first set of questions focused on the specific role of the Social, Personal and Material worlds (inner arcs of the framework) in creating data quality problems. The second set of questions examined the interactions between worlds (outer arcs of the framework) and their role in creating quality issues in an event log. The third set of questions were about the overall relevance and potential implications of the framework for a process mining project.

### C. Interview Analysis

We combined both quantitative and qualitative approaches in analysing the interviews. The interview questions required the experts to rank the causal paths of influence in terms of usefulness, and also to provide examples from their experience. Using the usefulness ranking in combination with the number of examples in each given arc (Table I), we derived, in the first stage of our analysis, an initial quantitative estimation of the usefulness of each arc. Qualitative analysis of the responses provided a richer understanding of the experts’ views on each causal path and the importance of the arcs in explaining data quality issues. The qualitative analysis was supported by Nvivo as a data analysis and evidence management tool and involved coding the interviews inductively to identify emerging themes. The coding scheme was based on the Odigos framework and the structure of the questionnaire (discussed in Section IV-B). Using open and axial coding approaches, we then identified 76 unaggregated codes representing emerging themes within and beyond the Odigos framework’s original elements. The themes that emerged in relation to the causal paths are discussed in more detail in Section V and relate to the Social, Personal, and Material worlds of the framework. The themes discussed in relation to the outer arcs are represented in Table I. The themes which were not within the scope of the Odigos framework are further discussed in Section VI. Emergent Themes.

### V. The Validation Results

The first set of interview questions revolved around the issue as to whether the three worlds of the Odigos framework cover the possible direct causes of data quality problems as observed by the experts. Almost all the experts found the list of presented direct causes to be comprehensive. Further, they were able to elaborate on the Social, Personal, and Material worlds in contributing to data quality issues as experienced by them.

#### A. Social World

The importance of the Social world, and more specifically, the organisational context, was highlighted by all the interview participants. The participants, based on their individual experiences, mentioned the following elements of the social world resulting in data quality problems in event logs:

- Organisation culture (E2, E4, E5). The same system with the same functionalities is used differently in different organisations.
- Social pressures and agreements in the work environment (for example in a hospital emergency department, where the highest priority is delivering treatment, with less attention being paid to recording tasks in the IT systems) (E1, E8)
- Management style (having a process manager role will decrease data quality issues) (E5)
- Organisation structure (focus on specialisation rather than workflow and the processes) (E5). E5 also mentioned that in an organisation with matrix structure management in comparison to a hierarchical structure management, there is a lower chance of having data quality problems in event data. “[...] usually in a matrix structure you have “vertical” department managers and “horizontal” process

¹Based on data from https://dblp.uni-trier.de/
manager. By granting mandate and budget explicitly also to a process owner it could be more likely that processes, workflow and event logs get more and better attention. Thus raising the chance of better data quality.” (E5)²

- Performance criteria and reward systems with a process-focus. “The company did not have target on process execution. They were not organised to reward people based on processes” and “The reward system also can influence on the execution and record of the processes and creation of event logs”. (E5)

- Legal requirements: including performance regimens or government privacy regulations. (E2, E4, E7)

- Cultural differences between countries (E4). For example, the same processes, using the same systems, would be done differently in different countries with different working cultures.

- Consideration of time and temporal nature of business processes in the framework. Processes may change faster than the IT systems which support these processes. “Processes are in a state of flux” and “in process mining processes are considered static”. (E6)

B. Personal World

The Odigos framework recognises two roles as part of the Personal world: Process participants and Data curators. While the experts acknowledged these roles, they mentioned a richer set of relevant roles, including:

- Process designers. They configure the processes in the systems. “If they use the wrong label for the processes and put wrong data, or use inconsistent naming or they forget to log some certain data […] or if they enter a process with a typo and correct it later”. (E2)

- Process analyst. “The way [the] process analyst asks for an event log has [an] influence on what they get and the meaning they generate from it. And how the process analyst looks at data”. (E5)

- Customers. For example, patients or loan applicants if they directly enter data. (E3)

- Data administrator or IT team. It was argued that this role should be separated from the role of data curator. (E2, E4, E6)

- Data curators are not necessarily lone operators, but can form a “chain”. (E6, E7)

- Bots and intelligent systems. “So it is worth to think of IT systems as a new actor, they are not only collecting data. They are also an actor involved, even more important than the participants”. (E6)

 Almost all of the experts agreed with the importance of these roles in terms of root causes of data quality issues. One of our eleven experts (E9), however, suggested that the role of process participants in creating data quality issues can be minimised by better designing IT systems to avoid, for example, human errors and control [flow] variation of the processes.

C. Material World

All of the experts agreed on the important role IT systems play in causing data quality issues in event logs. They men-

---

²Quoted from the expert’s email, in response to our request for feedback on the paper, 21st of August 2020.
tioned, based on their own experiences, the following data quality considerations in relation to IT systems:

- Interoperability between different systems in use, and the problem of linking data from multiple sources. (E1, E6, E7, E8)
- The interoperability between systems in use and the tools used by data curators to derive the data (data curation systems). (E1)
- Information architecture and the design of applications (e.g. system or application errors) (E3, E6). One of the experts (E6) also mentioned that data quality issues could be created as a result of an error in linking applications in the systems, and gave the example of a hospital system that creates automatic (ward) bed requests at 4am when the emergency department interface is connected. If the ED system has a bed request without the request time field entered, the bed request system automatically populates the field with the current time (4am).
- Mixed granularity of events and timestamps recorded. (E1, E11)
- The compatibility of IT systems in recording event data with the requirements of process analysis (E5, E11) (for example the appropriate granularity of time stamps). “The IT systems are only focused on a piece of information but not the entire system and time in which value is created” and “Most systems are not designed to capture event logs”. (E5)
- Timing of system updates by software vendor. (E6)

D. Different Paths of Influence

In the second section of the interview, questions focused on the nine different causal paths in the Odigos framework relevant to explaining the root causes of data quality issues in event logs. Our purpose firstly was to identify whether participants perceive these causal paths as useful in explaining data quality issues (useful referring to if they had seen data quality issues which could be explained using these causal paths) and, secondly, whether they could provide any examples that would help us to further specify these causal paths in our framework. Table I summarises the results of the interviews on these nine causal paths. Each row in the table refers to a single causal path and (i) summarises the experts’ perceptions of the usefulness of the causal path in explaining the root causes of data quality issues, (ii) gives the number of examples of data quality issues that they experienced in terms of this path, and (iii) provides a thematic analysis of the discussion surrounding experts’ experiences related to the path. Table I shows that experts saw the nine causal paths as (predominantly) very useful or somewhat useful in explaining data quality issues. In the following sections we discuss the main themes that emerged and the examples provided in relation to each path.

Path 1: Social-process participant in process execution

Table I shows that eight out of eleven of our experts agreed that path 1 is a very useful causal path in explaining data quality issues. The seven examples given by the interviewed experts highlighted the importance of organizational culture (different ways of doing the same process with the same system (E2), freedom in doing the tasks and not adhering to any formal processes (E5), deviation from formal processes because of customers’ pressure (E11)), temporal social processes (for example change of the processes during night vs day shifts (E6) or use of new drugs by physicians when those drugs are not yet defined in the systems (E3)) and performance criteria (in an emergency department, patients who could be seen earlier were made to wait close to 4 hours (E6)).

Two experts (E3, E10) raised concerns about differentiating between data quality issues and infrequent aberrant process behaviour. We will further explore this in Section VI.

Path 2: IT systems-process participants in process execution

Table I shows that eight of the eleven interviewed experts agreed that this path is very useful, three ranked it as somewhat useful, and one participant commented that this is not useful according to their experience with data quality issues. Three of the interviewees mentioned that IT systems (both applications and interfaces) not supporting the business processes are the main cause for data quality issues in this category (E5, E9, E10). For example in a loan process, the system did not allow multiple offers for the same client and every time a client accepted an offer the other offers needed to be deleted from the system manually, only then the loan offer could be confirmed (E4).

Path 3: Social-process participants in recording event data

Eight out of eleven experts agreed that path 3 is very useful in explaining data quality issues and nine different examples were given by the experts. Two of the common themes between different respondents were registration after the fact and batch recording (E2, E4, E5, E8, E9, E10). Some experts (E1, E4) provided several examples that reflected that process participants had to deal with different social pressures and priorities (e.g. in healthcare, patients and their treatment are prioritised, in a mortgage management context focus is on customer service) rather than with recording tasks in systems. Other examples provided by the experts reflected the role of performance criteria in the way process participants record the data (E1, E5). For example in a manufacturing context, maintenance personnel’s performance is measured according to the number of repairs they complete or attend to. Consequently, these process participants did not pay attention to recording the activities accurately during each case, but only logged these for a case after they finished it (E1). Experts also mentioned organisational norms and training in the use of the IT systems as another factor affecting process participants’ recording of the tasks (E1, E8). One of the main related examples mentioned was medical staff in hospitals and their resistance to using IT systems (E1, E6, E8, E9).

Our experts emphasised the importance of designing systems that support recording the workflow and guide the process participants through processes. They also emphasised having performance criteria which are not inconsistent with workflow management (E5).

Path 4: IT systems-process participants in recording event data

Almost all experts (ten out of eleven) agreed that path 4
is very useful in explaining quality issues in event logs. In relation to this path the experts mentioned that IT systems are not supporting the processes (E1, E4, E5, E8, E11). Different reasons were identified for this lack of support including (i) database design (E4), (ii) terminologies used in systems being different from the terminologies used by process participants (E8), (iii) lack of automation allowing human errors in entering data (E1, E2), as one of the experts mentioned “sometimes they [process participants] have to log as an extra activity, sometimes they log when they save, if the system is focused on workflow then the logs are recorded automatically” (E5), and (iv) not having a user friendly interface. For example, in a hospital using SAP, users had to manually enter data. However, there were many separate forms and data fields, many of which were mandatory. This caused users to enter one word or a few letters in each field to allow progression through the form, perhaps assuming that they will complete it later (E8, E9).

Path 5: Social-data curators in extracting event logs This path was considered very useful by nine out of eleven experts, two of them ranked it as somewhat useful, with three examples being given by them. Themes relating to data curators included (i) lack of knowledge of process mining requirements (E1, E2, E4, E8, E10), (ii) having domain knowledge (E2, E7), and (iii) privacy concerns (E3, E4, E5, E6, E7, E8). Examples of data quality problems were anonymisation of data in such a way that it conflicts with the aims of process mining (E5), for example privacy concerns leading to the changing of timestamps (possible re-identification of patients) (E6) or the aggregation of data without separate case IDs (E4).

Path 6: IT systems-data curators in extracting event logs This path was considered very useful by ten out of eleven experts. They mentioned different factors including data integration problems between different IT systems (E1, E3, E6, E7), different clocks or time settings of different devices (E11), interoperability issues between IT systems and data extraction tools (for example extracting data to the old version of MS-Excel which ended up in many missing rows (E1)), complexity of data extraction and software vendors’ power over this (E8), chain of data curators with consequent data provenance chain (E6, E7). “It can be difficult for data curators to link data from different source systems so they may end up doing probabilistic rather than deterministic linkage” (E6).

Path 7: Social-IT systems in recording event data The influence of the social context on IT systems was considered as very useful by eight out of eleven process mining experts with nine examples given in relation to this path. Participants mentioned several factors including the impact of performance criteria in designing the systems (E2, E5), IT systems designed with goals other than workflow management (E4, E9), siloed organisations where data is compartmentalised and multiple disconnected systems are used by each department (E7, E8), various terms used within the organisation to refer to the same concept (E1), and IT systems designed with i) varying timestamp configurations (future timestamp is assigned as a default value (E11)) and ii) different case ID format rules (in a call centre a dummy case ID is assigned by a call centre agent and has the same format as a customer number (E11)). Examples include using performance measures for patients’ length of stay in emergency departments in hospitals (E6) or defining a KPI around resolution time for IT tickets (E2). In both scenarios, process participants exploited the system configuration to meet the defined KPI without actually achieving those KPIs.

Path 8: Process participants-IT systems in recording event data Process participants influencing the recording of event data in IT systems was considered as useful by seven out of eleven experts, with five examples provided. In relation to this path, the experts mentioned factors such as audit functionality which can be activated by process participants (E7), misusing admin access to the system (E2, E10) which results in deviation from the standard process path, and manual data entries by process participants (E4).

Path 9: Data curators-IT systems in recording event data The influence of data curators on IT systems design was considered useful by six out of eleven experts. A number of them felt that in the definition of path 9, the role of data curator should be changed to data administrator (E2, E4, E6). One of the experts mentioned that data quality issues resulting from path 9 could also be explained by path 7 (E4). One expert stated that it is important for a process analyst to acquire knowledge (either from data curator or data admin) about recent/planned changes and updates to IT systems prior to asking data from data curators (E8).

E. Overall relevance and effectiveness

We asked the experts to rate the overall framework in terms of 1) its effectiveness in providing explanations for data quality problems in process mining, and 2) the relevance of the approach in improving process mining methodologies in the data pre-processing stage. All participants agreed that the framework is very effective or somewhat effective in explaining quality issues and that it is very relevant or somewhat relevant for the pre-processing stage. Expert E11 suggested that the framework is somewhat effective and it needs to be explained with examples for someone “to get into thinking [about] the model”. In relation to the relevance of the framework in the data pre-processing stage of process mining methodologies, one of the experts (E4) said “It is a must [...], there is no other approach in existing process mining studies”. Experts also mentioned that the framework (i) helps to ask the right questions (E2, E8), (ii) provides the landscape of data quality issues (E6), (iii) helps researchers to look into the right place (E6, E5), and (iv) assists with finding the actual problem (E11). “Also you can identify the questions that you can ask from data curator and anticipate problems, gives complete sets of questions” (E8). Four of the process mining experts (E1, E3, E6, E11) mentioned that the Odigos framework can be used to classify data quality issues.

Experts also suggested that the framework would be more useful if it were more detailed (E1, E2, E7). Three of them suggested that the framework should prioritise data quality
issues, i.e., what are the quality issues that affect the process mining findings (E2, E3, E10). All experts agreed with the statement that there is need for a supporting methodology and one expert said that the Odigos framework can be used to improve existing data extraction and pre-processing methodologies (E8). Overall, we found strong support from all of the process mining experts on the effectiveness and relevance of the Odigos framework.

VI. EMERGENT THEMES

Here we discuss themes that were not anticipated by the interviewers but emerged from the interviews. We reflect on how they will be considered in the further development of the Odigos framework.

New roles As we discussed in relation to the personal world (section V-B), experts introduced new roles involved in the creation of event data and the introduction of data quality issues. These new roles include the process designer (E2), process analyst (E5), customer as process participant (E3), data administrator (E2, E4, E6), chain of data curators (E6, E7), (i.e. where there is more than one data curator involved in extracting event logs), and intelligent systems such as bots (E6). The Odigos framework definition of ‘process participant’ is not exclusive to employees, so any role which is involved in recording data and completing tasks using the IT systems, such as a customer or patient, can be considered in this definition. We agree with expert E2 that the process designer role should be considered in relation to data quality issues. Rather than only this addition, we argue that it is then also important to include the role of formal processes and how processes are defined and embedded in IT systems, in the definition of the Material world. As one of our experts (E5) mentioned, the role of process analyst is a user of the Odigos framework. Therefore, we do not need to explicitly consider this role in the definition of the Personal world. We agree with experts E6 and E7 that the data curator could be more than one person. Accordingly, we now extend the definition of data curator to include any person or groups of people who are involved in extraction of data and transferring it to the form of event log(s) for the process analyst. One of our experts (E6) mentioned the emerging role of bots and intelligent systems in the creation of event data and data quality issues in event logs. Intelligent systems and bots could be considered in the Odigos framework in terms of the material world. Advances in technology (material world) may change, but will not diminish, the role and significance of process participants and social structures.

What is data quality? Several experts raised the question of what actually constitutes a data quality issue, and, in particular, if data representing (possibly infrequent) variations in process behavior should be defined as a data quality problem (E2, E3, E6 and E10). “I have [the] more philosophical question whether you need to include the processes. If people are doing the processes in a different way, is that a quality issue?” (E10).

The original motivation behind the Odigos framework was to look beyond symptoms of data quality issues (meaning any anomaly in a data set which hinders the standard analysis processes using process mining tools). We argue that any data quality symptom in an event log is potentially representing a reality. It should be investigated to decide how it can be dealt with. We agree with E6 that “we should get rid of the mindset that data quality is an inconvenience that should be done quickly to get to the analysis phase, exploring data quality is about understanding data”.

Technology advancement Technological advancement could either improve data quality issues or introduce new ones. Expert E7 mentioned that causal paths 1–4 would be less important with technology development and automation. “At the moment [it] is a problem but this is a problem where technology is not developed. For example, in banks you would not have that problem but in healthcare it is always behind in terms of technology” (E7).

Solutions to data quality problems Experts mentioned several solutions to existing data quality problems. Having process-aware information systems and designing systems which guide process participants was one of the main solutions (E2, E4, E5, E9). “The big task is to ask designers and business managers to have the right metrics and to design process-aware systems.” (E5). Expert E9 was of the opinion that in less stressful work environments, a process-aware system with a well designed user interface would improve recording of the tasks with consequent improvement in the quality of event data. However, in contexts with higher work pressure, expert E9 felt that merely improving the IT systems would not solve the data quality problems. More clarity about privacy requirements imposed by government regulations and using better encryption methods was another factor mentioned by expert E7. The importance of communication between process analyst and data curator was added by expert E9 as a solution to deal with data quality issues: “It is important to say what we [process analyst] need otherwise they [data curators] just give us a chunk of data. Maybe some issues can be fixed by them easily if we clarify what we need” (E9).

VII. DISCUSSION

In this paper, we validated the Odigos framework, an approach for identifying causes of data quality issues affecting process mining. We interviewed eleven process mining experts and, through these interviews, gained additional insights into data quality and the usefulness of the Odigos framework. For this study, purposive sampling was chosen with a view of reaching a targeted sample quickly and because, in this instance, sampling for proportionality was not the primary concern. We note that such an approach increases the likelihood of getting the opinions of the target population at the possible expense of overweighting opinions of subgroups in the target population that are more readily accessible. We also note that as the interviews with experts were conducted over a period of several weeks, the conduct of the interviews changed
slightly as we got more familiar with the process and better at explaining concepts underlying the framework to the experts.

Overall, the framework was perceived as useful by the experts who confirmed that such a framework, together with a supporting methodology to guide process mining researchers to go beyond data quality symptoms, fills a gap in the field. Our discussions with the experts unearthed interesting experiential examples, which provided further insight into the causal paths of the framework. They also broadened our perspectives on the three worlds of the framework.

An interesting insight that followed from an interview with one of the experts was that the removal or mitigation of the root cause of a data quality issue is contingent upon the organisational context and thus on an understanding of the causal paths and their ends (the worlds involved).

In some of the examples provided by the experts, we could see that both the Social and Material worlds played a role in the way process participants use the IT systems. To identify what is the best mitigation plan in these scenarios, careful consideration should be paid to the specific context and how plausible changes to these worlds would impact on the way IT systems are used. For example, in a hospital emergency department, the work pressure and the social agreements on prioritising patient welfare is non-negotiable. This may mean that mitigation measures are more likely to be successful (in this organisational context) if changes can be made in the material world, i.e. having IT systems that can accurately and automatically capture activities without compromising process participants’ social world obligations. The feasibility of this solution, however, needs to be assessed against other social world considerations. One can think of implementation cost, risk management, training provision in the use of current IT systems, process redesign, etc. This may then determine the ‘balance point’ of mitigation efforts. That is, emphasis towards one end of the causal path with some effort applied at the other. Thus, the Odigos framework facilitates development of mitigation measures that address root causes of data quality issues tailored to a given organisational context. In the same vein, the recognition that the context at either end of a causal path has changed may be a trigger to (prognostically) re-examine the impact of the change on data quality with early revision of mitigation efforts.

In some of the examples given by the experts, we found multiple causal paths from the Odigos framework that explained the problems. However, we could argue that in any context one of these has the greatest explanatory power. We also note that, it is possible that causal paths may be ‘chained’, which may reflect an ancestor sequence of root causes.

Given the acceptance of the Odigos framework by the experts, future work will focus on development of methodological guidance in the use of the framework for purposes including (i) diagnostic use - identifying root causes of identified data quality issues, (ii) prognostic use - assessing the organisational context to anticipate data quality issues, (iii) remediation - devising data cleaning strategies to rectify data quality issues in such a way that the goals of the process mining analysis are not compromised, and (iv) prevention - using the framework to understand the organisational context and develop prevention strategies that best fit with the imperatives of, and interactions between each of the social, material, and personal worlds.

ACKNOWLEDGEMENT

The authors would like to thank the process mining experts for their valuable feedback and for making time for the interview in their busy schedules.

REFERENCES

[1] R. Mans, W. van der Aalst, R. Vanwersch, and A. Moleman, “Process mining in healthcare: Data challenges when answering frequently posed questions,” in Process Support and Knowledge Representation in Healthcare. Springer, 2013, pp. 140–153.

[2] M. Bozkaya, J. Gabriels, and J. van der Werf, “Process diagnostics: a method based on process mining,” in Int. Conf. on Information, Process, and Knowledge Management. IEEE, 2009, pp. 22–27.

[3] W. van der Aalst, A. Adriansyah, A. De Medeiros, F. Arcieri, T. Baier, T. Blickle et al., “Process mining manifesto,” in Int. Conf. on BPM. Springer, 2011, pp. 169–194.

[4] M. van Eck, X. Lu, S. Leemans, and W. van der Aalst, “PMF: A process mining project methodology,” in CAiSE. Springer, 2015, pp. 297–313.

[5] R. Andrews, M. Wynn, K. Vallmuur, A. ter Hofstede, E. Bosley, M. Elcock, and S. Rashford, “Leveraging data quality to better prepare for process mining: An approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in Queensland,” IJERP, vol. 16, no. 7, p. 1138, 2019.

[6] F. Emamjome, R. Andrews, and A. ter Hofstede, “A case study lens on process mining in practice,” in OTM Confederated Int. Conf.s “On the Move to Meaningful Internet Systems”. Springer, 2019, pp. 127–145.

[7] R. Bose and W. van der Aalst, “Trace alignment in process mining: Opportunities for process diagnostics,” in Int. Conf. on BPM. Springer, 2010, pp. 227–242.

[8] H.-J. Cheng and A. Kumar, “Process mining on noisy logs–can log sanitization help to improve performance?” Decis. Support Syst., vol. 79, pp. 138–149, 2015.

[9] F. Emamjome, R. Andrews, A. ter Hofstede, and H. Reijers, “Allohomora: Unlocking data quality causes through event log context,” in Proc. 28th Eur. Conf. on Inf. Syst. (ECIS). AIS, 2020.

[10] R. Bose, R. Mans, and W. van der Aalst, “Wanna improve process mining results?” in IEEE CIDM. IEEE, 2013, pp. 127–134.

[11] S. Suriadi, R. Andrews, A. ter Hofstede, and M. Wynn, “Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs,” Information Systems, vol. 64, pp. 132–150, 2017.

[12] W. van der Aalst, “Process mining: Discovering and improving spaghetti and lasagna processes,” in IEEE CIDM, 2011. IEEE, 2011, pp. 1–7.

[13] R. Andrews, S. Suriadi, C. Ouyang, and E. Poppe, “Towards event log querying for data quality,” in OTM Confederated Int. Conf.s “On the Move to Meaningful Internet Systems”. Springer, 2018, pp. 116–134.

[14] R. Andrews, C. van Dun, M. Wynn, W. Kratsch, M. Röglinger, and A. ter Hofstede, “Quality-informed semi-automated event log generation for process mining,” Decis. Support Syst., p. 113265, 2020.

[15] F. Fox, V. Aggarwal, H. Whelton, and O. Johnson, “A data quality framework for process mining of electronic health record data,” in IEEE Int. Conf. on Healthcare Informatics (ICHI). IEEE, 2018, pp. 12–21.

[16] N. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng, “A data quality assessment guideline for electronic health record data reuse,” Egesm, vol. 5, no. 1, 2017.

[17] D. Savan, An Introduction to Cs Peirce’s Full System of Semeiotic. Toronto Semiotic Circle, 1988.

[18] M. Marshall, “Sampling for qualitative research,” Family practice, vol. 13, no. 6, pp. 522–526, 1996.

[19] M. Tongco, “Purposeful sampling as a tool for informant selection,” Ethnobotany Research and applications, vol. 5, pp. 147–158, 2007.

[20] R. Longhurst, “Semi-structured interviews and focus groups,” Key methods in geography, vol. 3, no. 2, pp. 143–156, 2003.