Data Governance for Managing Data Quality in Process Mining

Completed Research Paper

Kanika Goel
School of Information Systems
Queensland University of Technology
Brisbane, QLD, Australia
goelk@qut.edu.au

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

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

Abstract

Process mining, a specialised form of data-driven process analytics, is concerned with evidence-based process improvement. Process mining relies on process data, which often suffers from data quality issues that may be hard to detect and rectify. Data governance, recognised as a business capability, was recently introduced to manage data, including its quality, to maximise data's tactical value. Interestingly, no tailored data governance approach for managing process-data quality exists. The paper bridges this gap by introducing a data governance framework, the ImperoPD framework, for process mining with a focus on data quality. We use a capability-based approach and conduct a theoretical review of 75 papers to identify 20 capabilities an organisation should possess to implement process-data governance successfully. The framework is validated for its utility and comprehensiveness by 11 data governance experts. It contributes to an understanding of what is required to implement a data governance program for process mining.

Keywords: Data governance, process mining, data quality, business capability.

Introduction

Data governance refers to processes and practices concerning the formal management of data within an organisation (Mosley et al. 2010). As of 2019, 2.5 quintillion bytes of digital data is produced everyday, which is expected to increase tenfold by 2025 (Dhillon 2019). A priority for organisations is to manage this growing business asset. Businesses recognise that data presents endless and transformative opportunities. However, the volume and complexity of data that organisations are faced with, can be challenging. Furthermore, if this insurmountable quantity of data is not managed properly, it may present considerable risks (Abraham et al. 2019; Dhillon 2019). Data governance has been recognised as a *business capability* (Ladley 2019), which aims to formulate a corporate wide agenda that maximises the value of data and manages the risks related to data (Abraham et al. 2019). Remediating data quality issues has been identified as one of the major drivers of data governance (Ladley 2019). *The cost of bad data is an astonishing 15% to 25% of revenue for most organisations* (Redman 2017).

Process mining is a specialised form of data-driven process analytics where data about the execution of processes is collected and analysed to uncover the real behaviour and performance of business operations (van der Aalst 2016). It is an area of growing significance (Reinkemeyer 2020) with many industries adopting process mining to gain knowledge of underlying business processes.

However, process mining insights are dependent on the quality of input data (Andrews et al. 2020a; Wynn and Sadiq 2019) – the maxim *garbage-in, garbage-out* holds here as well. Inaccurate input data may lead to misleading or erroneous process mining results, which is why data pre-processing is significant for dependable process mining insights (Andrews et al. 2019). While essential, data pre-processing has been recognised as a time-consuming step (Wynn and Sadiq 2019), often requiring 80% of a process analyst's time (Press 2016), which can be costly (de Murillas et al. 2017). Given the tedious and laborious nature of data pre-processing as well as the unique data requirements for process mining (van der Aalst et al. 2011), i.e., data recording, data collection, data format, data cleaning and data analysis, a tailored data management program is necessary that can deal with root causes of data quality problems in process mining. A data governance framework which focuses on process mining requirements can thereby enable organisations to better plan and manage data quality problems optimising their data pre-processing efforts and deriving value from process data.

Despite the significance and need for data governance, it is an under-researched area (Al-Ruithe et al. 2019; Tiwana et al. 2013). Furthermore, our initial review of process mining studies using search keywords 'process mining' AND 'data governance' showed that process mining studies have not explicitly mentioned data governance capabilities to improve process mining practices. This paper aims to build the theoretical foundations to bring data governance capabilities to the field of process mining. We view data governance as an organisational capability (Ladley 2019) and adopt a capability-based approach to answer the research question: *What are the required capabilities to implement data governance for process mining?* We review prior literature to distil 20 capabilities required to govern process-data quality and these constitute our data governance framework, referred to as *ImperoPD*. Next, we validate and refine the ImperoPD framework based on interviews with eleven data governance experts. The final framework enables an understanding of what data governance for managing data quality for process mining entails. It also helps organisations understand the key areas they need to focus on to implement data governance and obtain reliable process mining insights. The data governance framework proposed in this paper as a research contribution falls under type I, the 'theory for analysing' of Gregor (2006). According to Gregor (2006), this type of theory is the starting point to further analyse and understand a new field of research.

The remaining paper is structured as follows. First, the key concepts related to this study are explained. Then the details related to the literature review method are presented followed by the data governance framework for process mining from a capability perspective, ImperoPD. The next section presents the interviews results resulting in validation and refinement of ImperoPD. The paper concludes with contributions, limitations, and future research directions.

Background

Process mining is a specialised form of data-driven process analytics, which analyses data recorded about the execution of processes to identify the behaviour and performance of business operations (van der Aalst 2016). The quality of input data is critical for accurate process mining insights (van der Aalst et al. 2011; Wynn and Sadiq 2019). This is reinforced by a recent process mining survey by Deloitte according to which 51% of respondents mentioned data quality as paramount for a successful process mining initiative (Galic and Wolf 2021). However, managing data quality issues relevant for process mining requires considerable time and efforts from the process analysts, which is costly (Suriadi et al. 2014). Further, data quality issues can also result in misleading insights, which can risk business operations (Wynn and Sadiq 2019). The significance of data quality for process mining is evidenced through a growing stream of research in this area. Data quality issues that can impact process mining have been identified (e.g. (Suriadi et al. 2017)), and techniques to detect (e.g. (Fischer et al. 2020)) and repair (e.g. (Dixit et al. 2018)) data quality issues have been developed. A number of case studies also highlight the efforts involved and the significance of managing data quality issues to obtain dependable process mining insights (e.g. (Andrews et al. 2020c)).

Data governance is the planning, oversight, and control over management of data and data-related resources (Mosley et al. 2010). It aims at implementing a corporate wide agenda and maximising the value of data assets in an organisation (Abraham et al. 2019; Mosley et al. 2010). The DAMA international framework advocates management of the lifecycle of data creation, transformation, and transmission, to ensure that resulting information meets the needs of the data consumers in the organisation (Mosley et al. 2010). Data governance has been recognised as a new business capability (Ladley 2019) to derive value from data. The significance of data governance can be evidenced with growing research in this area. For example, data governance frameworks for big data (Kim and Cho 2018) and supply chain management processes in SMEs (Barrenechea et al. 2019) have been proposed. More recently, Baijens et al. (2020) proposed a governance framework for data analytics, in which the significance of data governance is highlighted. At the same time, Abraham et al. (2019), through a rigorous structured literature review of prior data governance literature, developed a conceptual framework for data governance with six key dimensions. The framework provides the conceptual foundations for data governance and enables approaching data governance in a structured manner.

In this paper we suggest that by considering process mining data requirements in an organisation's data governance framework, we can limit existing data quality problems and improve the outcomes of process mining by preventing them from occurring in the first place. Therefore, a data governance framework can result in proactive actions that prevent data quality problems. To develop the data governance framework, this paper adopts a capability-based approach to identify *what are the capabilities that an organisation needs to possess to implement a successful data governance plan* for process mining. We use the key dimensions comprising data governance proposed in the conceptual framework proposed by Abraham et al. (2019), i.e., organisational scope, data scope, and governance mechanisms, to support the extraction and synthesis of capabilities from the literature. Furthermore, while data governance can have multiple domains such as security, architecture, and more, the focus in this paper is on data quality as that is crucial for obtaining accurate process mining insights.

The data governance framework proposed in this paper provides an understanding of the capabilities an organisation needs to have for successful governance of data quality for process mining. The framework can be adapted to an organisation's needs and enables prioritisation of capabilities required for data governance.

A Survey of Process Mining Literature

The previous section presented the need to develop a data governance framework to manage data quality issues pertinent to process mining. We reviewed the process mining and data quality literature to distil key aspects related to data governance. Our review reveals that no article explicitly mentions data governance practices for process mining. However, key literature on the application of process mining discusses data quality issues (as requirements for data governance in process mining) and mechanisms to address those issues (implicit data governance practices). In this review, we used the aforementioned literature to distil the organisational capabilities for implementing data governance for process mining.

Literature Review Design

A theoretical review was conducted to draw on existing work related to process mining and process-data quality. This review relies on existing conceptual and empirical studies to develop a conceptual framework or model (Paré et al. 2015), which in our case was the data governance framework for process mining. We followed the approach proposed by Webster and Watson (2002) to obtain relevant literature. The steps followed in our literature review were: (i) extract representative process mining and process-data quality literature (Webster and Watson 2002), (ii) determine a selection strategy (Paré et al. 2015), (iii) develop coding guidelines (Paré et al. 2015), and (iv) perform coding and analysis.

The main objective of this study is to identify implicit data governance practices and data governance requirements in the process mining context, which could be translated to data governance capabilities to manage process-data quality. We focused on representative literature and used the dimensions proposed by Abraham et al. (2019) as a lens to extract relevant themes. In other words, the dimensions proposed by Abraham et al. (2019) provided guidance related to what to look for in the literature regarding data gov-

ernance. At the same time we were open to new themes that may be unique to process mining but not captured in the aforementioned dimensions. We chose the database Scopus as the starting point to extract peer-reviewed academic literature. Scopus is an interdisciplinary database and is the largest abstract and citation database of peer-reviewed literature (Aghaei Chadegani et al. 2013). Furthermore, Scopus has been shown to have representative peer-reviewed literature (Bergman 2012; Ghasemi and Amyot 2016). As our search strategy, we looked for the combinations of "data quality" + "process mining", "data pre-processing" + "process mining"and "process mining" + "preprocessing" in the title and abstract of papers. Through this search, we extracted 119 records. Next, we reviewed the articles in detail to include only those articles that were investigating data quality in the context of process-centric data. The result of this stage of filtering was a set of 48 papers. To further reduce the probability of representative publications not being included in the final set of papers, we also conducted a backward and forward search on the filtered set of papers. The final set of papers to be analysed constituted 75 articles in total.

Coding and Analysis

A multi-phased abductive coding approach was applied for data analysis. The abductive coding approach (Timmermans and Tavory 2012) attempts to analyse data combining both inductive and deductive approaches. Since, we are using (Abraham et al. 2019) as the initial analysis framework and we also want to identify the specific data governance capabilities for process mining, the abductive approach was considered suitable for this study. NVivo 12.0 was used as the qualitative data analysis tool to perform coding. The coding proceeded in multiple phases. In Phase 1, any direct or indirect mention of a statement that was considered suitable to govern the quality of process data was captured as a node[1]. Recall, the dimensions of the framework proposed by Abraham et al. (2019) had an influence on the data captured related to data governance, however, we were open to new dimensions. For example, the statement "the onus is usually on a process analyst to identify, assess and appropriately remedy data quality issues so as to avoid inadvertently introducing errors into the data while minimising information loss" (Wynn and Sadiq 2019) was coded as a *process analyst* node. In Phase 2, the coded nodes were grouped together into themes (a higher order node). For example, nodes such as *process analyst* and *business analyst* were grouped together into a theme *roles and responsibilities*. While grouping the nodes a capability perspective was taken into consideration in adherence to the research question. This resulted in a synthesis of 20 capabilities. In Phase 3, we used sense-making to further group the capabilities into business areas. In the end, the literature-based data governance framework consisted of five business areas, each with four capabilities as represented in Figure 1. Throughout the coding process inter-coder reliability was maintained by having two coders involved. Furthermore, the use of NVivo assisted in maintaining a transparent trail of evidence.

ImperoPD: Data Governance Framework for Process Mining

Figure 1 presents the result of our synthesis (coding, analysis and abstraction) of the literature which presents the capabilities required for successful governance of data quality for process mining. We call this framework *ImperoPD*[2]. Our coding and analysis revealed 20 capabilities which need to be read with the prefix "ability to". For example, *ability to* develop strategy-driven data policies, has been identified as a capability. The capabilities were grouped into five business areas: Business Strategy Management, Process Management, Information Technology Management, Organisation and Project Management, and People Management. Each of these areas and corresponding capabilities are described next.

Business Strategy Management

Business strategy management refers to the high level plan of action (Cheng et al. 2017; Mosley et al. 2010) designed to address the data governance needs for process mining in an organisation. Four capabilities are distilled for this business area: develop strategy-driven data policies, identify data standards, define performance measures, and create a prognosis approach to data quality.

Develop Strategy-driven Data Policies: This organisational capability conveys the need to develop

[1]In NVivo, a node is similar to a folder.
[2]"I govern" is one of a number of meanings of "impero" in Latin and PD is short for "Process-Data".

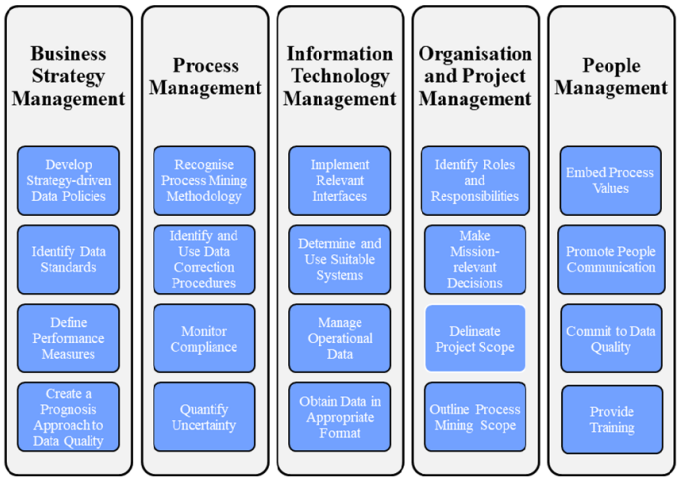

Figure 1. ImperoPD Framework

policies for collection, use, and analysis of data, which align with business strategy. Data policies provide guidelines and rules regarding the creation, acquisition, storage, and permissible use of data (Abraham et al. 2019; Mosley et al. 2010) to maintain data quality. For creation of quality data, policies may enable automated collection of data (Andrews et al. 2020c; Laine et al. 2015), adherence to external rules and regulations (Wang et al. 2018), and development of user friendly system design (Lanzola et al. 2014). Involving stakeholders when designing related systems (Wynn and Sadiq 2019) is encouraged. For acquisition and storage, policies regarding standardised and transparent processes (Laine et al. 2015), documentation of precise semantics for processes and systems (Bose et al. 2013; Laine et al. 2015), recording data fit for purpose (Andrews et al. 2020c), systematic logging of data (Wynn and Sadiq 2019), and having defined quality controls (Laine et al. 2015) are mentioned. For data integrity, policies should inform segregation of duties and authorised access to data (Lanzola et al. 2014).

Identify Data Standards: This capability requires an organisation to identify standards, which it expects the data to comply with. Data standards ensure that the representation and use of data is consistent throughout the organisation (Mosley et al. 2010). Maintaining data standards in accordance with the objectives of the organisation can assist with creating and maintaining data at desired levels of quality. The literature reveals several data quality standards, which need to be considered when using data for process mining, such as, completeness, accuracy, confidentiality, preciseness, and timeliness (Fischer et al. 2020; Verhulst 2016). Furthermore, the literature analysis revealed the significance of understanding data quality requirements (e.g., correct format of timestamps (Fischer et al. 2020) and correct ordering of events (Dixit et al. 2018)), data collection requirements (e.g., duration for which data is analysed (Andrews et al. 2020c) and temporal constraints (Lanzola et al. 2014)), and interface design (e.g., user friendly interfaces (de Murillas et al. 2017)). Finally, defined processes of data preparation to obtain data of an appropriate standard were also mentioned in the literature.

Defining Performance Measures: Performance measures aim at evaluating the quality of data for process mining analyses. This capability will enable organisations to set a benchmark for the quality of data. The literature makes mention of data quality metrics and their formulation (Andrews et al. 2020b; Fischer et al. 2020) as performance measures and advocates the use of established maturity levels for event logs (van der Aalst et al. 2011) to determine the quality of an event log. Metrics offer a formal way of measuring and quantifying data quality standards (Heinrich and Klier 2015; Pipino et al. 2002). For example, Andrews et al. (2020b) propose timestamp precision as a metric which calculates the average, minimum, and maximum granularity of timestamps. Similarly, Fischer et al. (2020) propose the number of timestamps with duplicate values to calculate a uniqueness metric.

Developing a Prognosis Approach to Data Quality: An effective data strategy to ensure high quality data should enable a prognostic approach (Emamjome et al. 2020) in dealing with data quality issues. A

prognostic approach helps anticipate data quality issues in event data based on an understanding of organisational and technological context. This may help resolve root causes of these quality issues (e.g., stakeholder behaviour, the use of different terminology, human errors (Laine et al. 2015)), or adjust the objectives of data analysis. Having a prognostic approach is a strategic capability in the data governance framework and it helps with planning and implementing other capability areas due to its role in preventing and mitigating data quality problems.

Process management

This business area captures capabilities related to the processes for creating and managing quality data for process mining. Our synthesis revealed four capabilities in this area: recognise process mining methodology, identify and use data correction procedures, monitor compliance, and quantify uncertainty.

Recognise Process Mining Methodology: There exists a variety of methodologies for conducting a process mining project (Emamjome et al. 2019). Different methodologies have varied foci, thus, the choice of the best approach for a specific process mining project should be based on the research questions, contextual requirements, and the quality of available data. Maintaining a well-defined, repeatable methodology for a process mining project is a capability, which can improve efficiency in dealing with data quality issues (Saltz et al. 2018). Methodological stages in relation to data quality include planning and justifying data quality issues for a process mining project, modelling high level processes prior to data collection and analysis (Andrews et al. 2018b), understanding underlying data models (Andrews et al. 2020c), identifying related data sources and data collection methods (Andrews et al. 2020c; Perimal-Lewis et al. 2016), defining the unit of analysis (Andrews et al. 2020c), data integration (Fortin et al. 2015; Perimal-Lewis et al. 2016), data anonymisation (Andrews et al. 2020c), quality assessment (Andrews et al. 2018b; Andrews et al. 2019) and cleaning data (Martin et al. 2019; Tavazzi et al. 2020).

Identify and Use Data Correction Procedures: The ability to identify and use data correction procedures is a capability related to detection, quantification, and repair of data quality issues in a log. These procedures can be documented or presented in the form of a model (e.g., flowchart, BPMN model). Guiding the construction of an event log is beneficial for achieving data quality improvement (Jans et al. 2019). The synthesis of process mining studies revealed the following categories of procedures in relation to data quality: specific filtering techniques (Conforti et al. 2016) (e.g., case-level filtering (Suriadi et al. 2014) and filtering erroneous data (Wynn and Sadiq 2019)), clustering techniques (Andrews et al. 2020c), and techniques and methods for: (i) determining adherence to data quality standards (Lanzola et al. 2014), (ii) detection and assessment of data quality issues (Dixit et al. 2018; Fischer et al. 2020; Ramos-Gutiérrez et al. 2021), and (iii) repair of data quality issues (Dunkl 2013; Ekici et al. 2019).

Monitor Compliance: Compliance monitoring requires organisations to define measures and procedures to ensure that the quality of data meets organisational and project goals as well as the identified performance measures. There is a need to monitor how well the data is complying with data requirements so that appropriate actions can be taken to improve the quality of data for future use. Fox et al. (2018) propose a Care Pathway-Data Quality Framework (CP-DQF) that can be used to monitor data quality issues. The authors suggest maintaining an issue register to record and report on compliance with data quality issues.

Quantify Uncertainty: Uncertainty quantification refers to an organisation's capability to identify existing and potential data quality problems and to plan for their resolution. Uncertainty quantification can be used with a process mining project or more generally at the organisational level. The synthesis of the process mining literature revealed the following areas in relation to uncertainty quantification: data provenance with a focus on recording metadata (Laine et al. 2015), organisational and historical data for consistency checking and issue identification (Laine et al. 2015), visualising deviations and different cohorts of interest in data sets as another approach to issue identification and management (Andrews et al. 2020c; Lismont et al. 2016), and identifying issues at the organisational level (such as resistance to use IT systems) which can assist in overcoming data quality problems.

Information Technology Management

IT systems and information architectures play a pivotal role in creating and maintaining high quality data. Many data quality issues in a process mining context result from the design of IT systems either in the presentation, application or database layer (Emamjome et al. 2020; Suriadi et al. 2017). In addition, for the purpose of process mining the format of data recorded by various IT systems should be supported by process mining tools and techniques. Accordingly, a data governance framework for process mining has to encompass the subsequent capability areas within information technology management.

Implement Relevant Interfaces: The interfaces allow users to interact with the application and database layers of an IT system. Review of process mining studies reveals that many data quality problems in event data are created as a result of poor design of interfaces and inconsistencies between the design, tasks and users' requirements (Lanzola et al. 2014). For example, Suriadi et al. (2017) identified data quality issues resulting from form-based design of user interfaces. Given the significance of appropriate interfaces, the literature suggests engaging users in design of interfaces and communicating data quality requirements (Lanzola et al. 2014; Perimal-Lewis et al. 2016), focusing on process-aware interface design (Andrews et al. 2020a), and implementing quality controls for data entry (Lanzola et al. 2014) as organisational capabilities, which could improve data quality of event data for the purpose of process mining.

Determine and Use Suitable Systems: Database systems consist of program code which supports business rules and processes. To ensure that high quality data is recorded for the purpose of process mining, the design and configuration of the system should support the processes and reflect data requirements for process mining. For example, activity labels or granularity of recorded events are concerns which can be addressed through the design of the logical layer and have significant impact on data quality (Suriadi et al. 2017). Also, inconsistencies between the design of the logical layer and the current state of processes can result in data quality problems in event data (Andrews et al. 2020a). "*It may happen that a system is designed and developed for a specific purpose, but after some time it starts to be used also for other ones. A gap between the original design and the actual exploitation may lead to data misuse and erroneous results*" (Lanzola et al. 2014, p.167). Furthermore, how organisational performance criteria are defined and embedded in design of IT systems can change the way data is recorded and hence can affect data quality (Andrews et al. 2018a). Process automation is also recognised as an important aspect in determining the level of data quality of event data (Miclo et al. 2015).

Manage Operational Data: Organisations need to have the capability to gather and maintain quality data for process mining analysis. Our review of process mining studies shows that data integration across different databases is one of the common sources of data quality problems (Andrews et al. 2018a; Andrews et al. 2020c). "*[T]he original sources of data come in great variety, differing in structure depending on the nature of the application or process under study. The standardization of this phase represents a challenge, given that a lot of domain knowledge is usually required in order to carry it out.*" (de Murillas et al. 2017, p.573). To derive event data with an acceptable level of quality, transaction data should be recorded at the right level of granularity across different databases, and proper linkages should be defined for the purpose of data integration (Suriadi et al. 2017). Recording contextual data and metadata about events also have been mentioned by process mining researchers as one of the approaches to improve and manage data quality (Diba et al. 2020; Laine et al. 2015). The use of conceptual data models and data ontology to support database design can also help process mining researchers to achieve a higher level of data quality (Andrews et al. 2020c; Wang et al. 2018).

Obtain Data in Appropriate Format: Organisations need to be able to maintain data in an appropriate format for process mining. Data can be present in open standards accepted by the process mining community, e.g., data can be maintained in the MXML standard (a simple XML specification to maintain audit trails of process-aware information systems (van Dongen and van der Aalst 2005)) or the XES format (an open standard for storing and managing event data (Verbeek et al. 2010)). Data can also be maintained in the form of semantic data models such as UML class diagrams (Suriadi et al. 2014), which can help ensure that the data is captured in the right format. Reference data (Abraham et al. 2019), i.e., data that can be used to classify or categorise other data may also be present. This reference data can assist various process mining stakeholders to make sense of data and confirm that it is interpreted correctly (Andrews et al. 2020c; Wang

Process Mining Role	Responsibilities
Project Sponsor	Funds the process mining project.
Process Owner	Accountable and responsible for the outcomes of the process.
IT Administrator	Clarifies questions about the data and provides a data dictionary.
Data Specialist or Data Curator	Compiles data using ETL tools.
Process Analyst	Cleaning data, analysing data, and answering process-centric questions using data.
Project Manager	Scopes the project and defines realistic milestones.
Domain Expert	Subject matter expert who assists in identifying questions for analysis.
Privacy Officer	Manages the privacy and ethics related to the project.
Process Participant	Consumes the data and reports data quality issues.

Table 1. Structural Roles and Responsibilities

et al. 2018).

Organisation and Project Management

Organisation and Project Management determines the reporting structure, accountabilities, expanse of the scope of data governance, and governance around data sharing. All the capabilities contribute to having a defined organisational and project structure necessary for maintaining data of appropriate quality for process mining analysis. Each of the capabilities are defined next.

Identify Roles & Responsibilities: It is important that an organisation is able to identify dedicated roles and responsibilities for the governance of data quality for process mining. Structural decisions can foster operational excellence of Business Process Management (BPM) initiatives (Hernaus et al. 2016), which include process mining initiatives. The literature communicates three important roles: process owner, process analyst, and process participants (Emamjome et al. 2020; Wynn and Sadiq 2019). A process owner is given decision autonomy and responsibilities regarding the process and is accountable for the outcomes of the process (Willaert et al. 2007) and a process analyst analyses process data. Process participants skilled in problem solving, process improvement, and decision techniques, contribute to creation of data (Kohlbacher and Gruenwald 2011). Since many roles were not uncovered from the reviewed academic literature, we conducted a further search of the non-academic literature, which revealed specific roles for running process mining projects (de Boer 2020; Rozinat 2017). The roles and responsibilities are summarised with their responsibilities in Table 1.

Make Mission-relevant Decisions: An organisation needs to identify which role has the authority to make decisions (Abraham et al. 2019) for process mining projects. Analysis of the literature indicates that the final decision making authority lies with the process mining project owner (Andrews et al. 2019; Suriadi et al. 2014). The project owner defines the objectives and this influences the decisions made by others involved in projects at different points in the life cycle of an event log, e.g., a data curator, a process analyst, and a privacy officer. A data curator makes decisions regarding the kind of queries that need to run to retrieve necessary data, a process analyst makes decisions regarding techniques that can be used to analyse the process, and a privacy officer makes decisions regarding the privacy-preserving technique that can be applied to the data set. In general, the decision making power of all roles needs to be clearly articulated so that data quality is not negatively affected.

Delineate Project Scope: Organisations need to identify the breadth of data governance or the unit of analysis for process mining projects. The scope of projects can be inter-organisational or intra-organisational (Abraham et al. 2019). Intra-organisational refers to the scope of data governance within an organisation. In intra-organisational process mining projects, organisations need to select appropriate business processes (Andrews et al. 2020c; de Murillas et al. 2017), evaluate the sources from which data will be obtained (Vanbrabant et al. 2019), identify mechanisms to overcome differences in recording of data across multiple systems (e.g., varying timestamp granularity) (Suriadi et al. 2014), and decide on how to deal with schema and instance level problems (Vanbrabant et al. 2019). Inter-organisational refers to the scope of data governance across organisations. In addition to prior concerns, the organisation needs to prepare to handle

Question	Sample Data Required
Is the aim of the project to discover process models? (Andrews et al. 2020c)	Event logs of the processes whose models need to be discovered (van der Aalst 2016)
Is the aim of the project to analyse performance of the process models? (Mans et al. 2012)	The key performance indicators (KPIs) of the organisation (Hompes et al. 2016)
Is the aim of the project to check for compliance of the current process? (Mannhardt and Blinde 2017)	A normative process model and KPIs (van der Aalst 2016)
Is the aim of the project to enhance current processes? (Mans et al. 2012)	KPIs and quality criteria that can serve as benchmarks (van der Aalst 2012; van der Aalst 2016)
Is the aim of the project to analyse the performance of cohorts? (Andrews et al. 2020c)	Event log attributes that define different cohorts (Schönig et al. 2016).

Table 2. Sample Questions and Data

challenges associated with blending data from overlapping processes (Andrews et al. 2020c). Furthermore, regardless of scope, the organisation needs to define ethical requirements regarding data sharing (Kurniati et al. 2019).

Outline Process Mining Scope: Organisations need to be able to identify the questions that process mining projects aim to answer. The process mining scope has an influence on the type of data for process mining analysis. Column 1 of Table 2 displays sample questions retrieved from the literature and column 2 lists sample data required for such analysis. This capability also enables organisations to maintain appropriate data for future process mining use (Andrews et al. 2020a). For instance, if an organisation is interested in analysing performance of resources, the systems need to be designed in a way that they record resource information.

People Management

People Management refers to the collaboration among members of an organisation and its culture, which are significant to maintain data quality for process mining. Capabilities related to this area is discussed next.

Embed Process Values: To manage process data of appropriate quality, organisations need to embed process-centric thinking or a BPM culture (Andrews et al. 2020a), which enables a shared understanding of process values. A BPM culture fosters a certain set of values that supports process-centric objectives (vom Brocke and Sinnl 2011). The values include: (i) customer orientation – proactive and responsive attitude towards the needs of recipients of process output, (ii) excellence – the orientation towards continual improvement and innovation to achieve superior process performance, (iii) responsibility – commitment to the objectives of a process and accountability towards the outcomes of a process, and (iv) teamwork – positive attitude towards cross-functional collaboration (Schmiedel et al. 2015; vom Brocke and Mendling 2018). It is through the cultivation of this process-centric mindset that actors take actions which contribute to process mining objectives.

Promote People Communication: Organisations need to promote communication to create awareness about data governance capabilities within the organisation (Abraham et al. 2019; Lomas 2010) as well as improve data quality for process mining (Andrews et al. 2020a; Wynn and Sadiq 2019). Communication among people should enable an organisation to get access to the right data (Andrews et al. 2020a) and use correct techniques to manage data quality (Andrews et al. 2019). Communication enables dissemination of protocols about the secure, confidential, and legitimate use of data for process mining analysis. Timely communication among members also encourages sharing of knowledge and experience as well as building an understanding of the key data quality issues and solutions (Mosley et al. 2010). Two-way communication between the process analysts and business experts assists in uncovering business insights (De Weerdt et al. 2013), while continual interaction with domain and data experts allows identification of insights related to the quality of data (Andrews et al. 2019).

Commit to Data Quality: To obtain data of high quality, a commitment to data quality needs to be embedded in the culture of the organisation. Poor quality of input data will result in inaccurate process mining insights (Suriadi et al. 2017; Wynn and Sadiq 2019). A shared understanding of the need to maintain data

quality will enable people in the organisation to commit to actions required to maintain data quality. According to Mosley et al. (2010), promoting data quality commitment encourages the necessary buy-in of stakeholders in the program, which in turn increases the chances of success of such a program (in this case a process mining program).

Provide Training: Organisations need to provide training to equip stakeholders with the necessary knowledge and skills required for effective implementation of the data governance program (Abraham et al. 2019). According to Andrews et al. (2020a), training should enable people to use the related IT systems and understand new processes. Further, training should allow people to have a shared understanding of data quality standards and metrics (Wynn and Sadiq 2019). Training is also expected to contribute to the smooth execution of the data governance program (Mosley et al. 2010).

Expert Validation

ImperoPD has been drawn from a comprehensive review of the literature. To validate its utility and comprehensiveness, a qualitative approach, using semi-structured interviews, was deployed. Interviews are an important data collection method in qualitative research (Myers 1997). According to Mabry (2008, p.318) semi-structured interviews allow for "*probative follow-up questions and exploration of topics unanticipated by the interviewer, facilitate development of subtle understanding of what happens in the case and why*". Purposive sampling was used to select the participants for the interview (Marshall 1996).

Expert Profile

We selected 17 participants, of whom 11 (coded as P1 to P11) agreed to participate. Linkedin was the portal that was used to connect with participants. Selection criteria included: (a) some prior experience with analytics and (b) at least three years of experience with data governance. Furthermore, we also considered a variety of data governance roles when approving our participants. Additionally, we reached out to participants working in companies across diverse areas such as banking, retail, consultancy, and education. The final participants included four data governance managers, two data governance consultants, two data governance analysts, two data governance technical leads, and one national data and analytics lead who is also a representative of DAMA. The participants offered consultancy services or worked for companies such as Woolworths, Suncorp, Univerity of Queensland, Price Waterhouse Coopers, Deloitte, Bolton Clarke, FutureFund, ANZ bank, Australian Energy Market, and KPMG. All participants had undertaken at least one data governance initiative at an organisation and had experience with data governance tools. While all participants had experience with analytics, five participants had experience with process mining as well. Those who were familiar with process mining had experience with tools such as DISCO and Celonis.

Interview Design and Analysis

Semi-structured interviews were conducted to gain the participants' opinion on the utility and comprehensiveness of the framework. An interview protocol was designed that was pretested to ensure that the questions are easily understood, cover the objectives of the study, and allow for open-ended input. Each interview lasted between 45 and 60 minutes. The interview started with generic questions around the participants' role, their opinion on data governance, advantages of data governance, and areas downplayed in data governance. Next, process mining and the data requirements for process mining were explained to them. Following this, their opinion on the need for a data governance framework for process mining was obtained. Then, ImperoPD was presented followed by questions related to this framework. The questions covered, among others, their overall impression of the framework, if they would use the framework in practice, if anything is missing in the framework, if anything is not relevant to data governance and hence the framework, if they would group capabilities differently, and possible renaming of capabilities which have industry relevance. The interview concluded with any remaining comments regarding the framework and the next steps they expect to see regarding further development of the framework. A qualitative analysis of the responses of the interviews was conducted to understand the utility and comprehensiveness of the framework as well as to understand future research directions.

Findings

The first part of the interview revolved around the basic concepts related to data governance, the need for data governance, and challenges in the field of data governance. Analysis of the first part of the interview is presented under the heading *About Data Governance*. Next, the responses are analysed to convey the *Utility of the Framework* and the *Comprehensiveness of the Framework*. Following this, suggestions for improvements are discussed and our *Refined ImperoPD Framework* is presented.

About Data Governance

It was interesting to note that the definition of data governance cited by the majority of participants was "*data being fit for use*" (P1,P2,P4–P11). Data governance is considered essential for "*building trust in data*" (P1–P11). Furthermore, data governance allows having "*appropriate control over the use of data*" (P1,P2,P4–P11). Eight out of 11 participants indicated "*data quality as the main driver*" (P1,P2,P4-P6,P8–P11) for data governance and one said that "*60% of a data analyst's time is spent in finding what data to use*" (P7). The participants also acknowledged the need for data governance because of "*increasing regulations industry needs to abide by*" (P1–P3,P5–P8,P9,P11).

Additionally, several interesting areas that are perceived to be downplayed in the area of data governance were revealed. Seven of the 11 participants indicated *lack of consensus or understanding around data governance* (P1,P2,P4,P7–P10) as a challenge, which is also downplayed. In fact one participant mentioned "*even with vendors there is no single understanding of what data governance means*" (P4). Additionally, participants (P1,P4,P5) appreciated the presence of the data management body of knowledge (DMBOK), however, they found it very broad and not clear. Furthermore, definition of key performance indicators (P1–P5,P7,P9–P11), people management (P3), collecting sufficient metadata (P1,P2,P4–P6,P9,P11), and understanding data governance holistically (P1,P2,P4,P7–P10) were other areas seen to be downplayed in data governance. We hence inferred that guidance on foundations of data governance and what it entails remains a challenge and needs further elaboration for practical implementation of data governance.

The findings reinforce the motivation of this study, which aims to identify capabilities an organisation needs to have to manage quality data for process mining analysis. This is also in line with data quality being recognised as the main driver for data governance by 82% of the participants.

Utility of Framework

One of the main objectives of the interviews was to understand the utility of the framework. Seven participants communicated the need for a data governance framework that can manage the quality of process data (P1,P2,P4,P5,P7,P9,P11). The seven participants included all five participants who had experience with process mining. The remaining four participants indicated the need to customise an existing data governance framework to address the data requirements for process mining. In either case, the responses demonstrate the need to have a data governance framework addressing the data quality challenges faced by process mining. Additionally, all eleven participants considered the framework useful, as it provides an understanding of the different components required for process mining. Seven participants said that the framework is "*practical for industry*" (P1,P2,P4,P5,P7,P8,P11). Eight participants (P1–P5,P9–P11) appreciated the presence of business areas, as it provides an overview of the key areas a business needs to focus on. One participant said "*it provides a visual overview of the efforts involved in data governance..... which is* [otherwise] *difficult to sell to top management*" (P5). One participant also commented that the framework prevents a "*one-size-fits-all approach*" (P4), which is prevalent in industry but not true for data governance.

Comprehensiveness of Framework

All eleven participants found the framework *comprehensive and detailed*. One participant (P3) appreciated the inclusion of people management, which according to them is often ignored while implementing data governance. "*The framework is comprehensive. I like the inclusion of people management as it is crucial but ignored*" (P3). Three participants (P1,P2,P4) valued the presence of process-related requirements (under process management), which they found essential for managing process data. Three participants

(P5,P7,P9) found that the framework provided a *step-by-step approach* and a detailed overview of the key elements required for data governance for process mining.

Refined ImperoPD Framework

In addition to comments related to utility and comprehensiveness of the framework, we sought feedback for its potential improvement. The proposed improvements resulted in a refined data governance framework, presented in Figure 2. First, it was suggested by three participants (P5,P6,P8) to reconsider the name of the capability 'quantify uncertainty' as it covers details more than just quantification. Uncertainties are also communicated to relevant stakeholders and are planned for resolution. Given the wider scope of the capability we decided to change the name to *manage uncertainty*. In the IT management business area, it was suggested by seven participants (P1,P2,P4–P6,P9,P11) to make metadata management explicit, as it involves significant efforts and is crucial for appropriate analysis of data. "*While you mention metadata, it would be nice to see it in the framework as its management involves considerable effort*" (P4). We cover metadata under the capability *manage operational data*, however, to make its management explicit we decided to rename the capability to *manage operational data and metadata*. In the organisation and project management business area, one participant (P2) suggested renaming *make mission-relevant decisions* as it did not convey the true meaning of the capability. Considering the comment and in line with the description of the capability, we renamed this capability to *assign decision-making authority*. The changed capability names are displayed in yellow in Figure 2.

In addition to feedback on capabilities, we received constructive feedback on the entire model. First, one participant (P9) indicated the need for explicit recognition of change management for each business area as it is crucial for implementing data governance in a successful manner. "*All the capabilities presented in the framework are relevant, however, they require change and project leaders to be able to manage this change*" (P9). Data governance indeed brings change to the organisation, which does require appropriate change management strategies (Panian 2010). We view change management as a business activity supporting the implementation of data governance. This is why we added change management as an overarching concept in the framework to explicitly recognise it as an activity to happen in conjunction with various capabilities related to data governance. Second, two participants (P4,P10) suggested rearranging business areas such that business strategy management and organisation, and project management are presented consecutively. According to the participants, these are two areas the organisation will look at first, which influence the other business areas in turn. Therefore, we rearranged the business areas in line with the comment. Furthermore, one participant (P7) suggested more explicit recognition of the keyword 'data' in the framework to reinforce the significance of data in its implementation. Many capabilities already utilise the keyword 'data', however there was one obvious area where we could be more explicit: we changed the name of the business area *process management* to *process-data management*. This brings the fact we are actually dealing with process data more to the foreground.

Overall, the ImperoPD framework was considered detailed and useful as it provided a holistic overview of capabilities required to manage the quality of data for process mining. Five participants (P1,P2,P4,P5,P7) indicated that the framework will allow organisations to prioritise capabilities depending on available resources and organisational maturity. Finally, three participants (P3,P7,P11) brought forth the need for review of the framework at regular intervals to ensure that data governance remains an ongoing process.

Conclusion

Process mining is a specialised form of data-driven process analytics that suffers in practice from a wide range of data quality issues (Suriadi et al. 2017; Wynn and Sadiq 2019). Detecting and rectifying these data quality issues can be a costly and time-consuming effort. Data governance has been recognised as a business capability that aims to formulate a corporate wide agenda to manage data, including the quality of data. It is an area of growing research attention with data governance frameworks being developed for areas such as supply chain management and electronic health records. Surprisingly, there exists no data governance framework for process mining. This paper overcomes this gap and presents a data governance framework for managing data quality for process mining. A theoretical literature review of 70 papers is conducted and an

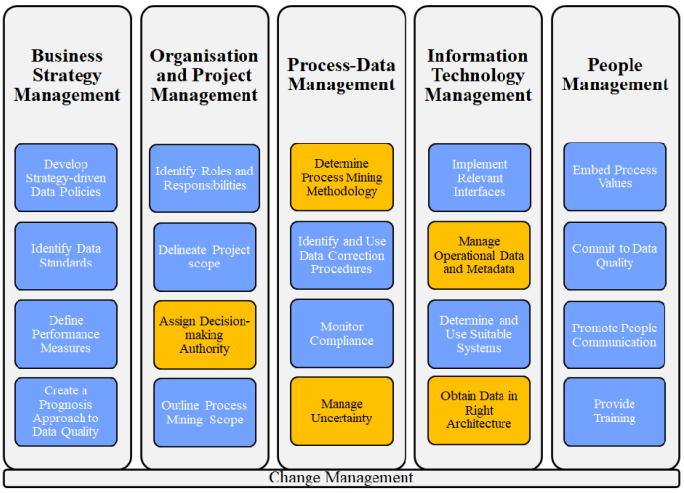

Figure 2. Refined ImperoPD Framework.

inductive coding approach is used to synthesise key capabilities for governance of data quality for process mining. In total, 20 capabilities across five business areas are distilled, synthesised, and presented. The framework was validated by 11 data governance experts who confirmed its comprehensiveness and utility, and whose input led to refinements.

To the best of our knowledge, ImperoPD is the first data governance framework in the field of process mining. From a practical perspective, the framework enables organisations to understand *what* is required for implementation of data governance and allows them to approach the implementation of data governance in a systematic manner (recall that lack of an understanding of what data governance entails was identified as a challenge by the experts). The framework enables organisations to conduct a capability analysis and therefore build new capabilities or upgrade the existing capabilities to maintain high quality process data. From a theoretical perspective, the work presented falls in the category 'theory of analyzing' as proposed by Gregor (2006), because the framework entails an understanding of *what* data governance for managing quality of data for process mining comprises.

We do not argue that our framework is complete but believe that it is comprehensive due to its basis in the literature and feedback from knowledgeable experts. Further, we acknowledge that our framework focuses on data quality only and that there are other aspects related to data governance. However, data quality has been identified as a critical issue in obtaining reliable process mining insights (Wynn and Sadiq 2019), and hence we, as well as the experts interviewed, considered it to be a good starting point to build a data governance framework for process mining. In future, we plan to develop guidelines to operationalise the framework and create a methodology for its implementation. Additionally, the framework can be expanded to include other areas of data governance such as security and storage for which grounded theory approach may be used.

References

Abraham, R., Schneider, J., and vom Brocke, J. 2019. "Data governance: A conceptual framework, structured review, and research agenda," *International Journal of Information Management* (49), pp. 424–438.

Aghaei Chadegani, A., Salehi, H., Yunus, M., Farhadi, H., Fooladi, M., Farhadi, M., and Ale Ebrahim, N. 2013. "A comparison between two main academic literature collections: Web of Science and Scopus databases," *Asian Social Science* (9:5), pp. 18–26.

Andrews, R., Emamjome, F., ter Hofstede, A. H., and Reijers, H. A. 2020a. "An Expert Lens on Data Quality in Process Mining," in: *Proceedings of the Int. Conf. on Process Mining,* Padua, Italy, pp. 49–56.

Andrews, R., Suriadi, S., Wynn, M., ter Hofstede, A. H., and Rothwell, S. 2018a. "Improving patient flows at St. Andrew's War Memorial Hospital's emergency department through process mining," in: *Business Process Management Cases,* Springer, pp. 311–333.

Andrews, R., van Dun, C. G., Wynn, M. T., Kratsch, W., Röglinger, M., and ter Hofstede, A. H. 2020b. "Quality-informed semi-automated event log generation for process mining," *Decision Support Systems* (132), p. 113265.

Andrews, R., Wynn, M. T., Vallmuur, K., ter Hofstede, A. H., and Bosley, E. 2020c. "A comparative process mining analysis of road trauma patient pathways," *Int J Environ Res Public Health* (17:10), p. 3426.

Andrews, R., Wynn, M. T., Vallmuur, K., ter Hofstede, A. H., Bosley, E., Elcock, M., and Rashford, S. 2018b. "Pre-hospital retrieval and transport of road trauma patients in Queensland: A process mining analysis," in: *Business Process Management Workshops,* Sydney, Australia, pp. 1–12.

Andrews, R., Wynn, M. T., Vallmuur, K., ter Hofstede, A. H., Bosley, E., Elcock, M., and 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," *Int J Environ Res Public Health* (16:7), p. 1138.

Baijens, J., Helms, R. W., and Velstra, T. 2020. "Towards a Framework for Data Analytics Governance Mechanisms," in: *Proceedings of the European Conference on Information Systems,* Online, p. 81.

Barrenechea, O., Mendieta, A., Armas, J., and Madrid, J. M. 2019. "Data Governance Reference Model to streamline the supply chain process in SMEs," in: *Proceedings of the Int. Conf. on Electronics, Electrical Engineering and Computing,* IEEE, pp. 1–4.

Bergman, E. M. L. 2012. "Finding citations to social work literature: The relative benefits of using Web of Science, Scopus, or Google Scholar," *The Journal of Academic Librarianship* (38:6), pp. 370–379.

Bose, R. J. C., Mans, R. S., and van der Aalst, W. M. 2013. "Wanna Improve Process Mining Results?," in: *Proceedings of the IEEE Symp on Computational Intelligence and Data Mining,* Singapore, pp. 127–134.

Cheng, G., Li, Y., Gao, Z., and Liu, X. 2017. "Cloud data governance maturity model," in: *Proceedings of the IEEE Int. Conf. on Software Engineering and Service Science,* Beijing, China, pp. 517–520.

Conforti, R., La Rosa, M., and ter Hofstede, A. H. 2016. "Filtering out infrequent behavior from business process event logs," *IEEE Transactions on Knowledge and Data Engineering* (29:2), pp. 300–314.

de Boer, L. 2020. "4 Process Mining Project Roles You Need…Plus One You Don't," (available at https://www.signavio.com/post/process-mining-project-roles/; accessed Feb. 20, 2021).

de Murillas, E. G. L., Hoogendoorn, G., and Reijers, H. A. 2017. "Redo log process mining in real life: data challenges & opportunities," in: *Int. Conf. on Business Process Management,* Barcelona, Spain, pp. 573–587.

De Weerdt, J., Schupp, A., Vanderloock, A., and Baesens, B. 2013. "Process Mining for the multi-faceted analysis of business processes—A case study in a financial services organization," *Computers in Industry* (64:1), pp. 57–67.

Dhillon, G. 2019. "The Importance Of A Data Governance Framework," (available at https://www.forbes.com / sites / forbestechcouncil / 2019 / 05 / 22 / the - importance - of - a - data - governance - framework / #551202f3ee85; accessed Jan. 20, 2021).

Diba, K., Batoulis, K., Weidlich, M., and Weske, M. 2020. "Extraction, correlation, and abstraction of event data for process mining," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* (10:3), p. 1346.

Dixit, P. M., Suriadi, S., Andrews, R., Wynn, M. T., ter Hofstede, A. H., Buijs, J. C., and van der Aalst, W. M. 2018. "Detection and interactive repair of event ordering imperfection in process logs," in: *Proceedings of the Int. Conf. on Advanced Information Systems Engineering,* Tallinn, Estonia, pp. 274–290.

Dunkl, R. 2013. "Data improvement to enable process mining on integrated non-log data sources," in: *Proceedings of the Int. Conf. on Computer Aided Systems Theory,* Spain, pp. 491–498.

Ekici, B., Tarhan, A., and Ozsoy, A. 2019. "Data Cleaning for Process Mining with Smart Contract," in: *Proceedings of the Int. Conf. on Computer Science and Engineering,* Samsun, Turkey, pp. 1–6.

Emamjome, F., Andrews, R., and ter Hofstede, A. H. 2019. "A case study lens on process mining in practice," in: *Proceedings of the OTM Conference,* Rhodes, Greece, pp. 127–145.

Emamjome, F., Andrews, R., ter Hofstede, A. H., and Reijers, H. A. 2020. "Alohomora: Unlocking Data Quality Causes Through Event Log Context." In: *Proceedings of the European Conference on Information Systems,* Online, pp. 611–620.

Fischer, D. A., Goel, K., Andrews, R., van Dun, C., Wynn, M. T., and Röglinger, M. 2020. "Enhancing Event Log Quality: Detecting and Quantifying Timestamp Imperfections," in: *Proceedings of the Int. Conf. on Business Process Management,* Online, pp. 309–326.

Fortin, É., Gonzales, M., Fontela, P. S., Platt, R. W., Buckeridge, D. L., and Quach, C. 2015. "Improving quality of data extractions for the computation of patient-days and admissions," *American journal of infection control* (43:2), pp. 174–176.

Fox, F., Aggarwal, V. R., Whelton, H., and Johnson, O. 2018. "A data quality framework for process mining of electronic health record data," in: *Proceedings of the IEEE Int. Conf. on Healthcare Informatics,* New York, USA, pp. 12–21.

Galic, G. and Wolf, M. 2021. *Global Process Mining Survey 2021*. Deloitte, pp. 1–36.

Ghasemi, M. and Amyot, D. 2016. "Process mining in healthcare: a systematised literature review," *International Journal of Electronic Healthcare* (9:1), pp. 60–88.

Gregor, S. 2006. "The nature of theory in information systems," *MIS quarterly* (30:3), pp. 611–642.

Heinrich, B. and Klier, M. 2015. "Metric-based data quality assessment—Developing and evaluating a probability-based currency metric," *Decision Support Systems* (72), pp. 82–96.

Hernaus, T., La Rosa, M., Mendling, J., and Reijers, H. A. 2016. "How to go from strategy to results? Institutionalising BPM governance within organisations," *Bus Process Manag J* (22:1), pp. 173–195.

Hompes, B. F., Buijs, J. C., and van der Aalst, W. M. 2016. "A generic framework for context-aware process performance analysis," in: *Proceedings of the OTM Conference,* Rhodes, Greece, pp. 300–317.

Jans, M., Soffer, P., and Jouck, T. 2019. "Building a valuable event log for process mining: an experimental exploration of a guided process," *Enterp. Inf. Syst.* (13:5), pp. 601–630.

Kim, H. Y. and Cho, J.-S. 2018. "Data governance framework for big data implementation with NPS Case Analysis in Korea," *Journal of Business and Retail Management Research* (12:3), pp. 36–46.

Kohlbacher, M. and Gruenwald, S. 2011. "Process orientation: conceptualization and measurement," *Bus Process Manag J* (17:2), pp. 267–283.

Kurniati, A. P., Rojas, E., Hogg, D., Hall, G., and Johnson, O. A. 2019. "The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database," *Health Informatics Journal* (25:4), pp. 1878–1893.

Ladley, J. 2019. *Data governance: How to design, deploy, and sustain an effective data governance program,* Academic Press.

Laine, S., Soikkeli, J., Ruohonen, T., and Nieminen, M. 2015. "Timestamp accuracy in healthcare business process improvement," in: *Proceedings of the Int. Conf. on Information Quality,* Cambridge, USA.

Lanzola, G., Parimbelli, E., Micieli, G., Cavallini, A., and Quaglini, S. 2014. "Data quality and completeness in a web stroke registry as the basis for data and process mining," *J Healthc. Eng.* (5).

Lismont, J., Janssens, A.-S., Odnoletkova, I., vanden Broucke, S., Caron, F., and Vanthienen, J. 2016. "A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways," *Comput. Biol. Medicine* (77), pp. 125–134.

Lomas, E. 2010. "Information governance: information security and access within a UK context," *Records Management Journal* (20:2), pp. 182–198.

Mabry, L. 2008. *Case study in social research,* Sage London, pp. 214–227.

Mannhardt, F. and Blinde, D. 2017. "Analyzing the Trajectories of Patients with Sepsis using Process Mining," in: *Proceedings of the RADAR+ EMISA@ CAiSE,* Essen, Germany, pp. 72–80.

Mans, R. S., van der Aalst, W. M., Vanwersch, R. J., and Moleman, A. J. 2012. "Process mining in healthcare: Data challenges when answering frequently posed questions," in: *Proceedings of the Process Support and Knowledge Representation in Health Care,* Tallinn, Estonia, pp. 140–153.

Marshall, M. N. 1996. "Sampling for qualitative research," *Family practice* (13:6), pp. 522–526.

Martin, N., Martinez-Millana, A., Valdivieso, B., and Fernández-Llatas, C. 2019. "Interactive data cleaning for process mining: a case study of an outpatient clinic's appointment system," in: *Business Process Management Workshops,* Vienna, Austria, pp. 532–544.

Miclo, R., Fontanili, F., Marquès, G., Bomert, P., and Lauras, M. 2015. "RTLS-based Process Mining: Towards an automatic process diagnosis in healthcare," in: *Proceedings of the IEEE Int. Conf. on Automation Science and Engineering,* Gothenburg, Sweden, pp. 1397–1402.

Mosley, M., Brackett, M. H., Earley, S., and Henderson, D. 2010. *DAMA guide to the data management body of knowledge,* Technics Publications.

Myers, M. D. 1997. "Critical ethnography in information systems," in: *Proceedings of the Int. Conf. on Information Systems and Qualitative Research,* Pennsylvania, USA, pp. 276–300.
Panian, Z. 2010. "Some practical experiences in data governance," *World Academy of Science, Engineering and Technology* (62:1), pp. 939–946.
Paré, G., Trudel, M.-C., Jaana, M., and Kitsiou, S. 2015. "Synthesizing information systems knowledge: A typology of literature reviews," *Information & Management* (52:2), pp. 183–199.
Perimal-Lewis, L., Teubner, D., Hakendorf, P., and Horwood, C. 2016. "Application of process mining to assess the data quality of routinely collected time-based performance data sourced from electronic health records by validating process conformance," *Health Informatics Journal* (22:4), pp. 1017–1029.
Pipino, L., Lee, Y., and Wang, R. 2002. "Data quality assessment," *Communications of the ACM* (45:4), pp. 211–218.
Press, G. 2016. "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," (available at https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=358ed8696f63; accessed Dec. 10, 2020).
Ramos-Gutiérrez, B., Varela-Vaca, Á. J., Ortega, F. J., López, M. T. G., and Wynn, M. T. 2021. "A NLP-Oriented Methodology to Enhance Event Log Quality," in: *Enterprise, Business-Process and Information Systems Modeling - 22nd International Conference, BPMDS 2021, and 26th International Conference, EMMSAD,* vol. 421. Melbourne, Australia, pp. 19–35.
Redman, T. C. 2017. "Seizing Opportunity in Data Quality," (available at https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/; accessed Nov. 20, 2020).
Reinkemeyer, L. 2020. *Process Mining in Action: Principles, Use Cases and Outlook,* Springer Nature.
Rozinat, A. 2017. "Skills and Roles Needed For Your Process Mining Project," Fluxicon. (Available at https://fluxicon.com/blog/2017/02/skills-and-roles-needed-for-your-process-mining-project/; accessed Dec. 15, 2020).
Al-Ruithe, M., Benkhelifa, E., and Hameed, K. 2019. "A systematic literature review of data governance and cloud data governance," *Personal and Ubiquitous Computing* (23:5-6), pp. 839–859.
Saltz, J., Hotz, N., Wild, D., and Stirling, K. 2018. "Exploring Project Management Methodologies Used Within Data Science Teams," in: *Proceedings of the Americas Conference on Information Systems,* LA, USA.
Schmiedel, T., van den Bergh, J., Willems, J., and Deschoolmeester, D. 2015. "Culture in business process management: how cultural values determine BPM success," in: *Handbook on BPM 2, Strategic Alignment, Governance, People and Culture, 2nd Ed,* Springer, pp. 649–663.
Schönig, S., Cabanillas, C., Jablonski, S., and Mendling, J. 2016. "A framework for efficiently mining the organisational perspective of business processes," *Decision Support Systems* (89), pp. 87–97.
Suriadi, S., Andrews, R., ter Hofstede, A. H., and Wynn, M. T. 2017. "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs," *Inf. Syst.* (64), pp. 132–150.
Suriadi, S., Mans, R. S., Wynn, M. T., Partington, A., and Karnon, J. 2014. "Measuring patient flow variations: A cross-organisational process mining approach," in: *Proceedings of the Asia Pacific Business Process Management,* Brisbane, Australia, pp. 43–58.
Tavazzi, E., Gerard, C. L., Michielin, O., Wicky, A., Gatta, R., and Cuendet, M. A. 2020. "A Process Mining Approach to Statistical Analysis: Application to a Real-World Advanced Melanoma Dataset," in: *Process Mining Workshops - ICPM,* vol. 406. Padua, Italy, pp. 291–304.
Timmermans, S. and Tavory, I. 2012. "Theory construction in qualitative research: From grounded theory to abductive analysis," *Sociological theory* (30:3), pp. 167–186.
Tiwana, A., Konsynski, B., and Venkatraman, N. 2013. "Information technology and organizational governance: The IT governance cube," *Journal of Management Information Systems* (30:3), pp. 7–12.
van der Aalst, W. M. et al. 2011. "Process Mining Manifesto," in: *Business Process Management Workshops,* France, pp. 169–194.
van der Aalst, W. M. 2012. "Process mining: making knowledge discovery process centric," *ACM SIGKDD Explorations Newsletter* (13:2), pp. 45–49.
van der Aalst, W. M. 2016. *Process Mining - Data Science in Action, Second Edition,* Springer, pp. 3–23.
van Dongen, B. F. and van der Aalst, W. M. 2005. "A Meta Model for Process Mining Data," in: *Proceedings of the EMOI-INTEROP,* Portugal, p. 30.

Vanbrabant, L., Martin, N., Ramaekers, K., and Braekers, K. 2019. "Quality of input data in emergency department simulations: Framework and assessment techniques," *Simulation Modelling Practice and Theory* (91), pp. 83–101.

Verbeek, H., Buijs, J. C., van Dongen, B. F., and van der Aalst, W. M. 2010. "XES, XESame, and ProM 6," in: *CAiSE Forum,* Hammamet, Tunisia, pp. 60–75.

Verhulst, R. 2016. "Evaluating quality of event data within event logs: an extensible framework," MA thesis. Eindhoven University of Technology.

vom Brocke, J. and Mendling, J. 2018. "Frameworks for business process management: a taxonomy for business process management cases," in: *Business Process Management cases,* Springer, pp. 1–17.

vom Brocke, J. and Sinnl, T. 2011. "Culture in business process management: a literature review," *Bus Process Manag J* (17:2), pp. 357–378.

Wang, Y., Hulstijn, J., and Tan, Y.-h. 2018. "Regulatory supervision with computational audit in international supply chains," in: *Proceedings of the Int. Conf. on Digital Government Research,* pp. 1–10.

Webster, J. and Watson, R. T. 2002. "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly* (26:2), pp. 13–23.

Willaert, P., van den Bergh, J., Willems, J., and Deschoolmeester, D. 2007. "The process-oriented organisation: a holistic view developing a framework for business process orientation maturity," in: *Proceedings of the Int. Conf. on Business Process Management,* Brisbane, Australia, pp. 1–15.

Wynn, M. T. and Sadiq, S. 2019. "Responsible process mining-A data quality perspective," in: *Proceedings of the Int. Conf. on Business Process Management,* Vienna, Austria, pp. 10–15.