








On the Origin of Questions in Process Mining Projects

Francesca Zerbato¹ , Jelmer J. Koorn² , Iris Beerepoot² ,
Barbara Weber¹ , and Hajo A. Reijers² 

¹ University of St. Gallen, St. Gallen, Switzerland
{francesca.zerbato,barbara.weber}@unisg.ch
² Utrecht University, Utrecht, The Netherlands
{j.j.koorn,i.m.beerepoot,h.a.reijers}@uu.nl

Abstract. In line with the growing popularity of process mining, several methodologies have been proposed to guide the conduct of process mining projects. Such methodologies reason that process mining projects start with a concrete question. However, in practice we observe projects with a different starting point, often aimed at exploring the data. Existing methodologies provide limited aid in such situations, and as a result, we wonder: how are questions developed *within* process mining projects? In this paper, we present the results of an interview study that sheds light on question development in process mining. We provide insights from expert interviewees, resulting in six recommendations that enhance existing methodologies. In doing so, we present concrete examples of how process mining analyses can support question formulation and refinement.

Keywords: Process mining · Question development · Interview study · Process mining methodology

1 Introduction

Process mining brings together a variety of methods and techniques for the analysis of process execution data recorded in event logs [1]. Enterprises conduct process mining analyses as part of projects aiming at improving, standardizing, and automating their business processes. Several methodologies have been proposed to guide the planning and execution of process mining projects [8]. These include, among others, the L* lifecycle model [1], the Process Mining Project Methodology (PM²) [6] and the question-driven methodology [14].

Most existing methodologies recommend starting process mining projects by defining “concrete research questions” [6], which, in turn, can be used to guide the extraction and analysis of event logs to find answers [1]. Some methodologies also recommend iteratively refining questions through exploratory analyses [6], suggesting that process mining analyses and their findings can themselves contribute to the development of questions. However, a good research question is not

always available at the start of a project [6]. Sometimes, projects are not triggered by concrete questions but are driven by the availability of data [2]. Moreover, analysts may need to familiarize themselves with the data before being able to formulate concrete questions that they can answer with process mining [16]. As a result, starting a process mining project with a concrete question is not as straightforward as it may sound.

While it is clear that questions play a crucial role in process mining projects [8], research has provided little insight into how questions are developed. Descriptions of case studies focus on answering a specific question, providing a limited picture of question formulation and refinement. In this paper, we aim at closing this gap by looking into the development of process mining questions through the eyes of experts in the field. In detail, we focus on the following research question: *how are questions developed within process mining projects?* We are particularly interested in finding out whether analysts typically start a project with a question, and if not, how they formulate such a question.

To investigate this research question, we followed an empirical approach and engaged in an interview study with experts to learn how they develop questions in their work practices. This empirical approach is fitting as experts can provide insights into how they deal with this issue in practice. Such insights are usually difficult to obtain as it would require significant effort and resources to, for example, organize direct observations. In addition, the literature describing process mining case studies, such as those surveyed by [8], suffers from a reporting bias, meaning it often only reports on the most relevant questions and the related findings. Thus, they provide limited insights into question development.

With this study, we contribute to an improved understanding of question development in two ways. First, we describe how questions are developed within a process mining project as reported by our interviewees. Here, we describe how specific process mining analyses can support question formulation and refinement. Second, we draw on the results of the study and propose a set of recommendations that enhance process mining methodologies, by demonstrating how questions are developed throughout process mining projects. Our findings contribute to existing research by providing concrete ways of supporting the development of questions in process mining projects. We propose a set of recommendations to enhance existing methodologies with question development steps for practitioners who are in charge of overseeing process mining projects.

This paper is structured as follows. In Sect. 2, we introduce background concepts. Section 3 describes our research method that led to the findings presented in Sect. 4. Then, in Sect. 5, we present the recommendations. We close with Sect. 6, where we discuss the limitations of this work and future research plans.

2 Background

In this section, we look at questions in the context of process mining methodologies and discuss the terminology used in this paper.

To support the use of process mining in research and industry, various methodologies have been proposed to guide the execution of process mining

projects. These methodologies include, amongst others, the Process Diagnostics Method [5], the L* lifecycle model [1], and PM² [6]. As summarized in [8], process mining methodologies generally adopt the following structure: (1) definition of questions and goals, (2) data collection, (3) data pre-processing, (4) mining & analysis of results, (5) stakeholder evaluation, and (6) implementation. In this paper, we are mainly concerned with the first step: defining questions and goals, in which objectives and research questions are specified by process analysts in collaboration with organizational stakeholders and domain experts.

The first step has been described differently throughout the literature, since the term *question* has been associated with different levels of abstraction and specificity. For example, in PM² the authors define a “research question” as “a question related to the selected process that can be answered using event data” [6], which can be “abstract” or “concrete” based on the setting. In [13], the authors define *types* of frequently-posed questions (FPQs) in healthcare, differentiating between “generic” and “specific” ones. Generic questions concern general process mining problems, e.g., “What happened?”, while specific ones address specific healthcare needs, e.g., “Do we comply with internal guidelines?”.

Looking at the start of a project, there are a number of aspects that are described: (i) goals/objectives, (ii) questions, and (iii) scope. The methodology by Erdogan and Tarhan [10] addresses all three elements. They are defined as follows: “The *scope* of a project indicates what the process is, when it starts and where it ends, and for which processes and which patients. The *goals* of a project may be related to improving KPIs (e.g., time, cost, risk, and quality). A set of concrete performance-driven *questions* are used to determine the way to assess or achieve these goals” [10, p.5]. Here, examples of questions are “How does the process look like?” or “Where are the bottlenecks in the process?” [10, p.11] Other methodologies focus more on questions, such as the question-driven methodology for analyzing emergency room processes [14], in which process mining projects are prescribed to start from a list of FPQs designed by domain experts. An example of FPQ is “What is the process for treating patients with different diagnoses?” [14, p.4]. Yet others emphasize a subset of elements questions and goals [1,6], or omit all three elements entirely [5]. From the literature, it is clear that questions play a central role at the start of a process mining project, and they largely influence how the project develops. However, none of the considered methodologies describes how a question should be formulated and refined.

In this paper, we adopt a number of definitions of questions. First, we relate to the definition of question given in PM² [6] reported above. In addition, we borrow the concept of “exploratory” questions from literature [4]. With *exploratory* questions, we refer to questions that do not necessarily correspond to a specified goal but focus on understanding the data, discovering patterns, and generating hypotheses. An example is “What happens in this process?”. Exploratory questions are opposite to *confirmatory* questions, which aim at testing a specific hypothesis, such as “Is the invoice process delayed on weekends?”. We consider this distinction relevant as it allows us to understand better how the degree to which a question is exploratory influences how it is formulated and refined.

3 Research Method

To understand how process mining questions are developed, we interviewed experts who have participated in process mining projects. In this section, we describe our research method. First, we describe the data collection. Then, we elaborate on the data analysis based on qualitative coding.

3.1 Data Collection

In this section, we cover the study *design*, where we describe the set-up of the study, the *setting*, where we elaborate on the execution of the study, and the *participants*, where we report on the participants selection and demographics.

Study Design. The interview study presented in this paper is part of a broader study in which we collected data using three methods: (1) a questionnaire, (2) think-aloud, and (3) interviews. The questionnaire was designed to capture the demographics of participants on three primary matters: area of occupation, level of experience, and project experience. It consisted of 18 closed questions:

- Six questions captured basic demographics, including the sector and position in which the participant was employed at that time.
- Seven questions focused on the experience of the participant in terms of: process mining, business intelligence, and data science/engineering.
- Five questions focused on the practical experience with using process mining tools and conducting process mining projects and event log analyses.

Participants were then invited for a virtual session which consisted of two parts: think-aloud and interview. In the first part, they were asked to engage in a realistic process mining task using think-aloud [9]. The task concerned an analysis of the road traffic fine management event log [7] guided by a high-level question asking to investigate circumstances and reasons for not paying a fine.

In this paper, we focus on the information collected in the interviews that were conducted in the second part of the session. Here, participants were interviewed using a semi-structured interview guide. The interview guide consisted of four parts: (1) activities and artifacts, (2) goals, (3) strategies, and (4) challenges. The first part, *activities and artifacts* focused on the steps that participants perform and the information they gather when engaging in process mining analyses. In the second part, *goals*, participants were invited to provide details on their analysis objectives and the amount of exploratory work they typically engage in. This flowed into the third part, *strategies*, where participants were asked about specific plans of actions they follow to achieve their goals. Finally, in the *challenges* part, participants could reflect on the obstacles they run into during the analysis and what kind of support might aid in overcoming them. In each of the four parts, participants were asked to reflect on the interview questions in two contexts: the recently performed process mining task and the broader context of their work practices and experiences. This constant comparison allowed us to better understand how experts work in process mining projects.

Setting. The data was collected between May 1st and July 28th, 2021. Participants were invited for virtual one-on-one sessions with the first author to ensure that the task and the interview were conducted in the same way for all participants. In this session, the participant was granted access to a remote desktop environment with the materials, i.e., data, protocol, and tools. The think-aloud part of the session took roughly 40 min and was recorded through screen capture and voice recording. After this, the participant was asked to report their answers to the guiding question in a post-task questionnaire. Then, the interview was conducted. This lasted roughly 30 min per participant, resulting in a total of 1046 min of audio recording. The audio records were transcribed verbatim for coding purposes by the first author. We note that the example statements from the participants reported in Sect. 4 have been edited to exclude pauses, fillers, and repetitive words. Participants were informed that they could ask questions during the session. Also, they were encouraged but not required to finish the process mining task. In addition, the successful or unsuccessful completion of the process mining task was not relevant for the interview; the task served as a basis to discuss how the participants performed a process mining analysis, but the task itself has little intrinsic value in the context of this study.

Participants. Participants were approached via email through the professional network of the authors, encouraging the recipients to forward the email to anyone else interested. For this paper, we target process mining experts. In detail, we consider the following inclusion criteria. Participants must (1) have analyzed at least two real-life event logs¹ in the two years prior to the study, (2) perceive themselves as knowledgeable with at least one of the process mining tools available for the process mining task and (3) have participated in at least two process mining projects having the goal to analyze process data for a customer. Such criteria allowed us to exclude participants without practical experience in customer projects. The final sample selected for this study consists of 33 participants.

The potential biases we identified that could play a role in the study sample were captured in the demographics questionnaire. Of the 33 selected participants, half ($n = 16$) work as an academic and the other half ($n = 17$) as a practitioner. Overall, participants have an average of 5.6 years of experience in process mining and have experience in data science ($n = 32$) and business intelligence ($n = 30$). Finally, participants hold diverse roles, such as process analyst, process mining consultant, product manager, senior researcher, and Ph.D. candidate.

3.2 Data Analysis

For the analysis of the interview data, we followed a qualitative coding approach [15]. The coding was performed by a team of three of the authors. First, all the members of the coding team individually studied the interview data to get an understanding of the content. Separately, they developed ideas for a possible

¹ In contrast to synthetic logs, real-life event logs are logs obtained from the execution of real-life processes, such as those provided by the IEEE Taskforce in Process Mining: <https://www.tf-pm.org/resources/xes-standard/about-xes/event-logs>.

coding structure. Then, a meeting was held to pitch the different coding structures and merge the ideas. The main coding structure that emerged revolved around steps that process analysts follow in formulating, answering, or refining questions while conducting an analysis in the context of a process mining project. Two authors were tasked with coding the different steps; one focused on the individual steps and the other focused on the relationships between them. The third coder verified the codes of the first two and disagreements were discussed.

As we aimed to study how process mining questions are developed but did not have an initial hypothesis or framework to start from, we used an *inductive approach*. From the transcripts, we used *open coding* [15] to arrive at an initial set of codes. The codes were mainly descriptive of the different steps observed or parts of them. We then used *thematic analysis* to organize our codes [15], as it helped us to identify clusters. Specifically, we clustered the codes based on whether they described a possible start point or endpoint for a process mining analysis. On the highest level, we identified two starting points: “Question” and “No Question”, and three endpoints “Not a process mining question”, “Question answered”, and “New question generated”. This formed the basis for our analysis. In the following step, we started from the codes representing the starting points and endpoints and iteratively searched for dependencies between them and the remaining codes. This led us to discover two high-level themes: analysis and analysis strategies. The *analysis* theme captured different kinds of process mining analyses that the interviewees performed in different settings. Examples are “exploratory analysis” and “pre-defined analysis”. The second theme includes *analysis strategies*, i.e., common approaches not specific to process mining that helped interviewees progress in a process mining project. For example, “evaluate hypothesis” and “explore beyond the question”.

More details on the participants and the data analysis can be found online on <https://doi.org/10.5281/zenodo.6984229>.

4 Results

In this section, we present the question development process emerging from the analysis of the interview data. First, we provide an overview of the whole process, which is depicted in Fig. 1. Then, we go into the details of the question formulation and refinement phases respectively in Sect. 4.1 and Sect. 4.2.

From our analysis, we learned that a question developed in the context of a process mining project can undergo three main phases: question formulation, refinement, and answering. Question formulation concerns posing a question about the process under analysis. Question refinement involves transforming an existing question into another one that can be more specific or easier to answer. Question answering deals with finding an answer to a given question.

Our analysis revealed that such phases originate from two different starting points, *No Question* and *Question*, depicted as orange-filled circles in Fig. 1. In the first case, process analysts do not have a question at hand and need to formulate one. Usually, they start by directly looking at the event log and gathering data-driven insights that can lead to questions, for example, with the help

of “Exploratory analysis”. In the second case, analysts start with a previously formulated question and plan their analysis based on it (“Plan analysis based on question”). Then, they engage in one or more iterations of “Process mining analysis”, to either refine or answer the question. Usually, analysts transition from refining a question to answering it based on the findings of their analyses.

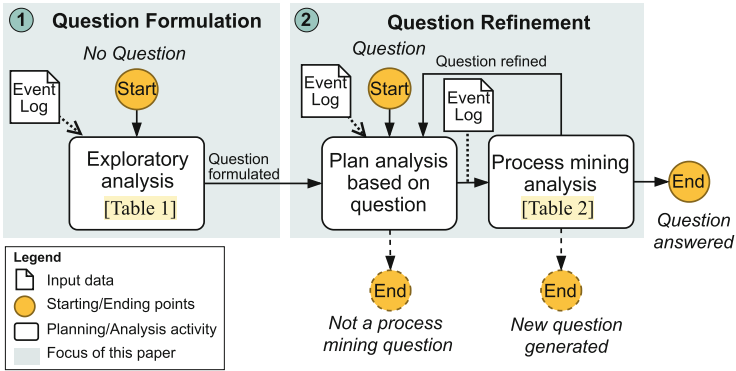


Fig. 1. Overview of the question development process showing different phases of developing a question in the context of a process mining project.

Similar to starting points, we identified three main endpoints, shown as orange-filled circles in Fig. 1: *Question answered*, *Not a process mining question*, and *New question generated*. The first one, *Question answered*, captures the expected end of a process mining analysis. The other two endpoints, depicted with dashed borders in Fig. 1, show possible “exits” from the process of developing one question. The endpoint *Not a process mining question* covers the situation in which a question is not suited to be answered with process mining analyses and, thus, is typically discarded or answered with other analyses. Instead, endpoint *New question generated* shows the generation of new questions, which can occur as a consequence of gaining process knowledge during the analysis.

In the remainder, we focus on the ① question formulation and ② refinement phases. We discuss what analysis activities can support them and how. For each phase, we report on input, factors, and typical steps using example statements from our interviewees. Input are the data at the disposal of the analysts. In this paper, we assume that analysts have an event log available for both phases. Factors are possible influences on question development, which can be the cause for a specific step (e.g., a low level of process thinking maturity) or can affect the choice of one step over another (e.g., the availability of domain knowledge). Steps are different analysis activities.

4.1 Question Formulation: From Event Logs to Questions

Based on our analysis, we define question formulation as the phase of question development that begins without question and concerns deriving and posing a question about the process under analysis.

Analysts start their analysis without question, for example, when business stakeholders are new to process mining and are curious to know what process mining technology can achieve. This is often prompted by the (broad) availability of event data, typical of data-driven projects [2], e.g., *“I have experienced a customer who has 40 GB of data: ‘See what you can find’”* (p11). Indeed, although process mining methodologies prescribe starting with concrete questions (cf. Sect. 2), formulating questions at the start of a project can be difficult [6]. This is partly due to the required participation of stakeholders. One interviewee explained that: *“it is very often hard to identify the correct question. So, sometimes the correct question is just given by process owners, stakeholders, etc., but other times we are just interested in finding out patterns in the event log”* (p36).

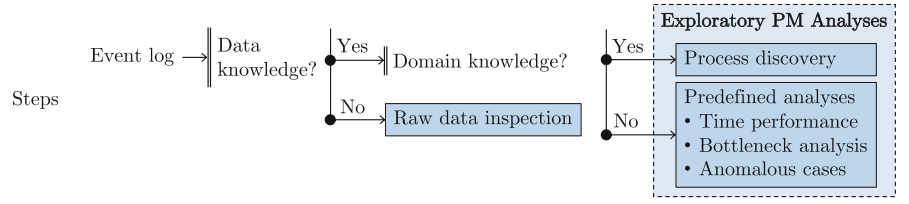
In this setting, process analysts usually start by analyzing an event log to derive data-driven insights or hypotheses that can be discussed with the stakeholders to formulate questions that are aligned with their expectations (cf. Table 1).

Factors. The absence of a question seems to be caused by low levels of process thinking and of process mining maturity, as p11 narrated *“it [the lack of questions] ‘happens more often if the customer is immature in process work and they do not invest any time in the process mining task’* (p11). It may occur in *“projects where people still have limited knowledge of what is possible”* (p39) or *“don’t have any clue of what the process is”* (p40). In such cases, stakeholders do not pose questions but are curious to know *“What is the process that I have?”* (p40). Moreover, at this stage, process analysts may have not yet gathered knowledge about the data and may have little domain knowledge and process knowledge to formulate process mining questions by themselves, especially if they are external to the organization. Thus, they follow different steps based on whether they know the data or have process and domain knowledge, e.g., through external resources such as documentation or from the stakeholders.

Steps. Interviewees reported following two main analysis steps when dealing with the absence of questions, namely (i) raw data inspection and (ii) exploratory analyses. Both steps are helpful in generating data-driven insights and hypotheses that can inspire new questions and assessing what kinds of analyses can be performed on the given event logs. Still, analysts seem to combine such steps based on the factors mentioned above, i.e., their knowledge of the data, the domain, and the process.

Some interviewees reported that, when lacking data and domain knowledge, it is easier to start from the “raw” event log to learn about the structure of the data, typical attribute values, and the underlying data models, if available. This raw data inspection helps analysts gather knowledge about the data and

Table 1. Question formulation. Starting with *No Question* and using “Raw data inspection” and “Exploratory PM Analyses” to derive insights and hypotheses.

Analyses Supporting Question Formulation	
Inputs	Event log
Factors	Process thinking maturity, process mining maturity, data knowledge, domain knowledge, process knowledge
Steps	 <pre> graph LR A[Event log] --> B{Data knowledge?} B -- Yes --> C{Domain knowledge?} B -- No --> D[Raw data inspection] C -- Yes --> E[Exploratory PM Analyses] C -- No --> D subgraph E [Exploratory PM Analyses] F[Process discovery] G[Predefined analyses • Time performance • Bottleneck analysis • Anomalous cases] end </pre>

estimate what analyses can be done on it. Interviewees also reported validating the data and its quality to ensure that it *“is really usable”*, and they can avoid *“working with information that’s completely useless”* (p15). Both data structure and quality can be a starting point for finding analysis questions related to data-driven issues or (new) data extractions. For example, p34 narrated that *“usually I will take time to analyze data quality. [...] the data quality step would help me to see a lot of problems and maybe guide me to the solution.”* (p34).

Moreover, analysts can conduct exploratory process mining (PM) analyses to understand the process and the context in which it is enacted and find insights from the data that lead to hypotheses that can *“inspire the stakeholders about what they could have as a question”* (p12). In process mining, a big part of exploratory analyses is covered by process discovery. In this setting, analysts often exploit the visual artifacts generated by process discovery algorithms as a basis for discussing with stakeholders and developing data-driven hypotheses and questions. One interviewee described this process as *“My way of doing process mining is to explore the dataset as I did now [process mining task] but with more time for reflections. I will take two or three hours in my office alone and try to make sense of the dataset. Next, I will get out some questions [...] and have interactive sessions with the data owner to understand things”* (p34).

Next to process discovery, analysts can resort to predefined analyses, i.e., ready-to-use or “standard” analyses aimed to gather information about process descriptives, which seem particularly helpful in case of limited domain knowledge. Such analyses include user-specified steps that analysts implement based on their experience or are provided by process mining vendors as *“sets of standard hypotheses and analyses behind”* that *“make your life easier so that you don’t start with an empty piece of paper. So, we look at the standard analyses, and that’s always something we bring to the first workshop”* (p33). Interviewees provided examples of predefined analyses, indicating time performance and bottleneck analysis as the most common ones, followed by anomalous, non-compliant cases, and control flow. In particular, process performance and anomalous cases seem to be *“the most common perspective that one can look at while doing process*

mining without having any additional information about the context” (p36) as opposed to, for example, analyzing resource behavior which *“does not give a lot of interesting insights in an exploratory setting. Because typically [...] they [the resources] are anonymized”* (p18). If process knowledge is available, the standard KPIs defined within the organization are also included in predefined analyses.

4.2 Question Refinement: Refining Questions with Process Mining

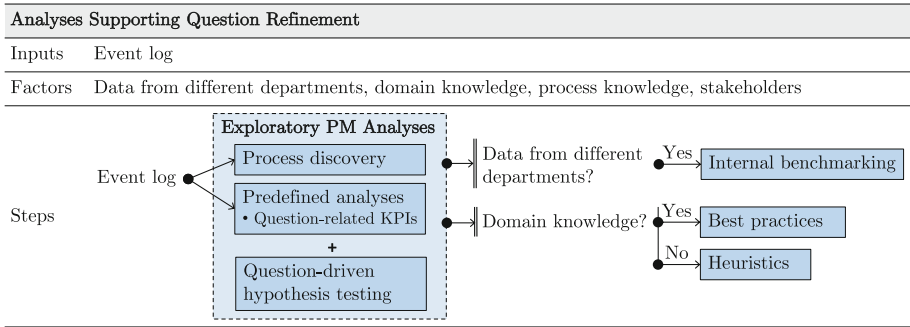
Based on our analysis, we define question refinement as the situation in which analysts start from a question, either posed by stakeholders or resulting from a previous question formulation phase, that needs to be refined. When provided with a question, analysts typically plan their analysis based on the kind of question being asked, as explained by one interviewee: *“the kind of analysis you do very much depends on the main question that is asked”* (p16).

In an ideal setting, analysts are provided with process-related questions, i.e., questions that presume an underlying notion of process control flow and can be suitably answered with process mining techniques. Process-related questions can be exploratory or confirmatory (cf. Sect. 2), which determines if questions need to be refined and, if so, with what analysis steps. Indeed, interviewees reported that *“it’s part of the strategy to adapt the analysis based on the question”* (p9).

However, our findings revealed that analysts can also find themselves dealing with data-related questions, i.e., questions that do not assume an underlying notion of process control flow. This happens, for example, when *“the problem is not directly a process problem but has a more statistical nature”* (p34). In this scenario, analysts can choose to refine the question or perform analyses other than process mining, as exemplified by the *Not a process mining question* endpoint in Fig. 1. Below, we discuss the different steps that analysts can follow to refine process-related questions, which are summarized in Table 2 together with related inputs and factors. We then provide examples of data-related questions.

Refining Process-Related Questions. Interviewees reported several examples of exploratory process-related questions, spanning from general ones such as *“What does this process look like?”* (p16), to questions aimed at investigating a particular process aspect, e.g., performance *“Can we check which machines have bottlenecks?”* (p26) or *“Where do cases spend a lot of time?”* (p19). Usually, exploratory questions require analysts to iteratively refine them, as opposed to confirmatory questions that can be answered more directly. To refine questions, analysts can narrow the analysis focus by *“building hypotheses related to the question that confirm or reject certain possible causes”* (p39) or by identifying patterns in the data that are linked to the question. Such approaches typically lead to “partial answers” that can be confronted with the stakeholders and potentially used to refine the question. One interviewee described the refinement of questions as follows: *“So, start with the first question. So, let’s get some data, get some partial answers, then refine the questions and do it again and maybe two or three or four times and then you converge to something that is robust, that makes sense and that can be used in a much more general way”* (p25).

Table 2. Question refinement. Starting with a *Question* to refine using different kinds of “Exploratory PM Analyses” to narrow the analysis space around the question.



Factors. The availability of stakeholders is a determinant factor for question refinement since business stakeholders often bring domain and process knowledge that analysts can combine with their prior experience to interpret the data and the question. Usually, process knowledge and domain knowledge are exchanged during interactive sessions, where analysts “*understand a lot of the business process with business people*” and learn how to avoid “*silly questions or putting down silly hypotheses*” (p25). Stakeholders also provide crucial feedback on the results and refinement steps. The availability of data from different departments can instead enable the use of benchmarks to narrow the analysis space, which we can also see as a form of refinement.

Steps. Not surprisingly, interviewees reported engaging in exploratory analyses to refine questions, as described in Sect. 4.1. However, given that analysts have both the event log and a question at their disposal, they can combine data-driven analyses, such as process discovery and predefined analyses with question-driven hypotheses to test on the data. Hypotheses are made in different ways, for example by “*finding positive and negative examples related to the questions*” (p39) or by following “*this CRISP thing, right? [...] picking out hypothesis right from the question and then creating something that I can reject or validate this hypothesis*” (p12). Compared to the analyses carried out in the absence of a question, which focus on “*finding interesting things in the data*” (p12), these exploratory analyses are driven by the need to identify parts of the event log that are relevant to the question. For example, predefined analyses can focus on KPIs associated with the question or can help test a given hypothesis.

Our interviewees reported different steps they use to narrow the analysis space around the question. Some interviewees mentioned following a data-driven approach, exploiting the data from different organizational departments for internal benchmarking. Internal benchmarks help narrow the analysis space by allowing analysts to identify critical steps on which to focus as one interviewee explained: “*you most of the time, already know from a benchmark where the critical process steps are, and you can deep dive into those few steps and see if it’s really an issue*” (p23). Other analysts rely on their own domain knowledge and

narrow the analysis space with the help of **best practices** and “good cases”. Best practices are particularly useful to refine questions around improvement opportunities, as they hint towards improvements, e.g., “*We look at the positive cases to see whether there’s a gap in the way our clients work. So, if there’s some best practices missing. So, if we notice that the clients we’re working with didn’t do a step that a lot of phone companies do, we say, ‘look, this would be a good idea for you.’ So, you can look for improvements.*” (p31). While best practices require domain knowledge, analysts can rely on general principles or heuristics to focus on specific parts of the event log. For example, some experts reported focusing on finding cases related to the question within “*the mainstream behavior [of the process] because it’s more supported and makes it easier to reject or validate a hypothesis*” (p41). The mainstream behavior can be separated with the help of heuristics such as the Pareto principle that allows focusing “*where you have more flesh on the bone. I use this 80/20 Pareto principle at the beginning because if you start looking at all the variants, you get lost*” (p26).

Refining Data-related Questions. Data-related questions are about event data but do not explicitly relate to the process control flow. One interviewee explained that data-related questions arise when stakeholders “*don’t have that ‘normal’ process idea, because they don’t get all the information of the status of all the activities in the process*” (p40).

From our interviews, we identified several examples of data-related questions asking, for instance, “*why there are data quality issues*” (p41) or “*what is the percentage of cases that do that*” (p19) or “*how many different activities*” (p37) a log contains. While the first exemplifies a data quality question, the other two require looking into specific data attributes or measurable KPIs. Such questions are often addressed with non-process mining analyses since “*Excel or SQL queries are just much more efficient than trying to do it with process mining*” (p19).

Still, stakeholders ask questions such as “*Could you predict if they are going to pay or not?*” (p7) that analysts can refine and transform into process-related questions. One interviewee described the iterative refinement of data-related questions into process-related ones as follows: “*So, the first thing is that they [the stakeholders] don’t know what ‘process’ means [...] So, all the questions are data-related. So, the first thing you need to do is drive them to the process-related questions. Of course, the data is also interesting, but the process-related analysis is what you can do. And then after that, when you show them some results, like process models, they start to understand what process-related analysis is. And then the questions start to shift. So, it’s never a one, two, or not even three iterations. The first one [iteration] I am sure that is going to be questions to be answered with predictive analysis or data mining, or machine learning*” (p7). In the interviews we found evidence that process discovery results, such as process models are used to refine data-related questions into process-related questions, but we couldn’t derive the detailed steps making up this scenario.

Overall, our findings provide evidence that process mining analyses are used to support question both formulation and refinement. One interviewee well-summarized the relevance of using process mining for question formulation saying that “the nice idea of process mining is that it allows us to detect new research questions” (p40), for example, based on the insights and hypotheses gathered through exploration. Another one remarked how question refinement is intrinsic to iterative analyses: “you start from a hypothesis, get some data, look at the data, go back, refine the hypothesis, get some additional data... So, you repeat the analyses for an entire year” (p25). However, in both settings, close interactions between process analysts and domain experts seem crucial because “without domain knowledge, you won’t achieve much or nothing at all” (p39), especially if organizations are less mature in process mining and thinking.

5 Discussion

In this section, we incorporate our findings into existing process mining methodologies and propose six recommendations (R1–R6) based on our findings with the aim to enhance the current body of knowledge from the perspective of question development. Figure 2 shows, in the dashed boxes connected by arrows, the phases that are typically prescribed by existing process mining methodologies (adapted from [8]). In the blue boxes, we illustrate how the results of Sect. 4 fit into existing methodologies. In addition, we depict as blue arrows with circled numbers recommendations R1–R4, which relate to newly added flows. Recommendations R5–R6 are more generic and, thus, not depicted.

R1 Use process mining to formulate questions. Our interviewees remarked that process mining can provide substantial value in *formulating* questions.

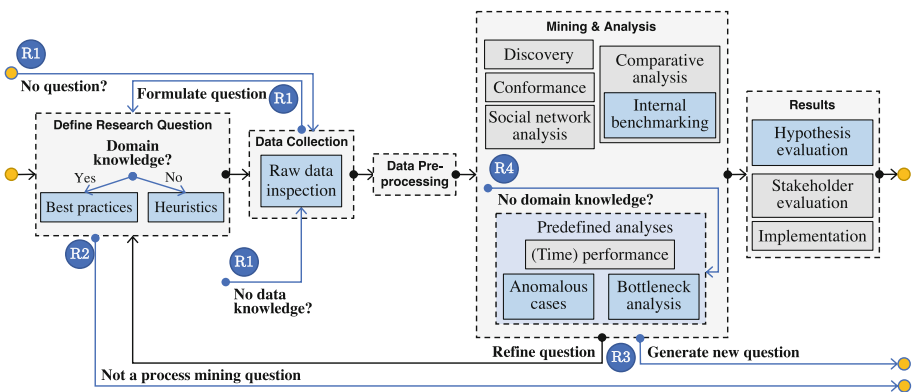


Fig. 2. Incorporation of our findings (blue boxes and arrows) into existing process mining methodologies. Start and end events are depicted as orange-filled circles. (Color figure online)

They reported that questions often emerge from discussions between analysts and stakeholders, who exchange knowledge with the help of process mining tools and artifacts. Process mining tools can be used by process analysts in an exploratory manner to inspect the raw data and “get a feeling” of what interesting directions to investigate are. Based on this, analysts can identify the most fruitful questions that can be formulated. Indeed, the experts in our study indicated that formulating questions without the data at hand is hard. On the other hand, process discovery tools and visual artifacts help spark process thinking among stakeholders. This eases the exchange of the domain and process knowledge required to formulate questions.

- R2 Evaluate if process mining techniques are appropriate to answer the formulated research question.** Our interviewees remarked that we should not always assume that a provided question is a process mining question. This is a good time to reflect on what to do next. As guidance, one might check whether the question at hand fits one of the following frequently asked questions in process mining [13]: (1) What happened? I.e., discovering the process model, (2) Why did it happen? I.e., finding the root causes for a particular situation, (3) What will happen? I.e., predicting process executions in specific circumstances, (4) What is the best that can happen? I.e., finding improvements. When the question does not match these templates, other types of analysis might be more suitable to answer it. For example, the experts indicated that some questions can be solved efficiently with the help of traditional data querying and manipulation languages.
- R3 Document the refinement of questions and the generation of new questions.** During the *Mining & analysis* phase, existing questions are often refined based on new insights. In such a case, it is important to revisit the earlier steps of question formulation and data collection, as also outlined in the PM² methodology [6]. The experts in our study also indicate that entirely new questions can be raised during this project phase. It is good practice to document the development of questions both to inform the stakeholders as well as for proper (academic) reporting. Indeed, keeping track of the questions and the analyses performed to answer them can help streamline the analysis process, identify cause-effect relationships among different questions and ease the answering of questions in future analyses.
- R4 Use predefined analyses to get started.** The use of predefined analyses can help generate questions and spark discussion on what to analyze more in-depth. Most process mining tools have predefined analyses built-in. Usually, these predefined analyses concern time performance, bottleneck analysis, and anomaly detection. Predefined analyses can be a good way to explore potential pain points or improvement opportunities and could be linked to “standard” process mining use cases [3] to help organizations with less process mining maturity get started with projects.
- R5 Value the collaboration between process analysts and stakeholders.** Process mining methodologies describe the value of the collaboration between process analysts and stakeholders in later project phases [6]. Our

results show that such collaboration brings value also in the *early* phase of question formulation. Experts do indicate that setting up a collaboration can be difficult, especially when projects are conducted in organizations with low process mining maturity. However, they advise working interactively and incorporating the stakeholders' knowledge to avoid "getting lost" and "putting down trivial questions". Later on, predefined analyses and, to some extent, agile discovery can also be done interactively. Experts described this phase as an agile collaboration, developing the work in a similar fashion as Scrum sprints.

R6 Align the question and the analysis. From our findings, we observe the importance of *aligning the question with the appropriate analyses*. Questions can include exploratory or confirmatory aspects, which may influence what analyses are conducted. This, in turn, can explain why process mining projects take a different course based on the application domain. *Exploratory* questions are usually formulated in contexts where prior knowledge of the process is scarce, and the main objective is to discover the process. We found evidence that in some domains such as healthcare [13] exploratory questions already bring much value in promoting process thinking since healthcare information systems are often not process-aware [11]. Instead, *confirmatory* questions intend to verify specific hypotheses, as we emphasized in Fig. 2 by adding a blue box in the *results*. Usually, formulating confirmatory questions requires deep process understanding, which is not always available. Our results reveal that such questions are common in domains such as auditing, where questions put a strong focus on the detection of non-compliance. Although we cannot claim that the course of a process mining project depends only on the questions and their nature, we believe that aligning questions with possible analyses and their outcomes could help organizations assess what they can and cannot achieve with process mining technology.

With these recommendations, we aim at enhancing existing methodologies with tangible examples that show how process mining analyses can support question development. We emphasize that process mining brings value not only for answering (concrete) questions but also for question formulation and refinement.

6 Conclusion

In this paper, we have looked into the development of questions within process mining projects. Drawing on 33 interviews with process mining experts, we have gained insights into how specific process mining analyses can support question formulation and refinement. Then, based on the interview findings, we have proposed six recommendations that enhance existing methodologies with concrete steps supporting question development within process mining projects.

Limitations. Our findings emerged from retrospective interviews and, therefore, are subject to validity threats typical of interview studies, such as reactivity, respondent bias, and researcher bias [12]. We mitigated these risks by:

(1) using a well-developed and pilot-tested interview guide, (2) coding the data with multiple authors, and (3) guaranteeing the anonymity of the participants. Moreover, we note that the set of recommendations presented in this paper may not be complete. Thus, we cannot exclude that additional ones emerge when asking different groups of experts. To mitigate this risk, we considered a sample size of 33 process mining experts with diverse backgrounds and we elaborated on themes that repeatedly emerged across the interviews.

Future Work. In the future, we will refine and extend the list of recommendations considering (i) factors that affect question development in specific settings and application domains and (ii) insights from literature on question development in the broader context of data analysis. We will also conduct a user evaluation to assess the generalizability and practical relevance of our findings.

References

1. van der Aalst, W., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011. LNBP, vol. 99, pp. 169–194. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_19
2. Van der Aalst, W.M.: Process mining: discovering and improving Spaghetti and Lasagna processes. In: IEEE Symposium Computational Intelligence Data Mining (CIDM), pp. 1–7. IEEE (2011)
3. Ailenei, I., Rozinat, A., Eckert, A., van der Aalst, W.M.P.: Definition and validation of process mining use cases. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011. LNBP, vol. 99, pp. 75–86. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_7
4. Behrens, J.T.: Principles and procedures of exploratory data analysis. *Psychol. Methods* **2**(2), 131 (1997)
5. Bozkaya, M., Gabriels, J., Van der Werf, J.M.: Process diagnostics: a method based on process mining. In: International Conference on Information, Process, and Knowledge Management, pp. 22–27. IEEE (2009)
6. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.P.: PM²: a process mining project methodology. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) CAiSE 2015. LNCS, vol. 9097, pp. 297–313. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19069-3_19
7. Eindhoven University of Technology: Road traffic fine management process (2015). Data retrieved from 4TU ResearchData
8. Emamjome, F., Andrews, R., ter Hofstede, A.H.M.: A case study lens on process mining in practice. In: Panetto, H., Debruyne, C., Hepp, M., Lewis, D., Ardagna, C.A., Meersman, R. (eds.) OTM 2019. LNCS, vol. 11877, pp. 127–145. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33246-4_8
9. Ericsson, K.A., Simon, H.A.: Protocol Analysis: Verbal Reports as data. MIT Press, Cambridge (1984)
10. Gurgen Erdogan, T., Tarhan, A.: A goal-driven evaluation method based on process mining for healthcare processes. *Appl. Sci.* **8**(6), 894 (2018)
11. Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., et al.: Process mining for healthcare: characteristics and challenges. *J. Biomed. Inform.* **127**, 103994 (2022)
12. Padgett, D.K.: *Qualitative Methods in Social Work Research*, vol. 36. Sage, Thousand Oaks (2016)

13. Rojas, E., Munoz-Gama, J., Sepúlveda, M., Capurro, D.: Process mining in health-care: a literature review. *J. Biomed. Inform.* **61**, 224–236 (2016)
14. Rojas, E., Sepúlveda, M., Munoz-Gama, J., Capurro, D., Traver, V., Fernandez-Llatas, C.: Question-driven methodology for analyzing emergency room processes using process mining. *Appl. Sci.* **7**(3), 302 (2017)
15. Saldaña, J.: *The Coding Manual for Qualitative Researchers*. Sage, Thousand Oaks (2021)
16. Zerbato, F., Soffer, P., Weber, B.: Initial insights into exploratory process mining practices. In: Polyvyanyy, A., Wynn, M.T., Van Looy, A., Reichert, M. (eds.) *BPM 2021. LNBP*, vol. 427, pp. 145–161. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85440-9_9