UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI DEL TERRITORIO E DELLE COSTRUZIONI

RELAZIONE PER IL CONSEGUIMENTO DELLA LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

Data management and data handling in the context of process mining – cases of application to the audit processes of the European Court of Auditor

SINTESI

RELATORI

Prof. Ing. Davide Aloini
Dipartimento di ingegneria dell'energia dei sistemi del territorio e delle costruzioni

Ing. Pierluigi Zerbino
Dipartimento di ingegneria dell'energia dei sistemi del territorio e delle costruzioni

Ing. Emanuele Fossati
European Court of Auditors

IL CANDIDATO

Danilo De Pascalis
depascdanilo@gmail.com

Sessione di Laurea Magistrale del 16/06/2021

Data management and data handling in the context of process mining – cases of application to the audit processes of the European Court of Auditor

Danilo De Pascalis

Sommario

L'auditing è sottoposto a forti regolamentazioni, questo incide sull'utilizzo che i revisori possono fare delle fonti su cui si basano per i propri giudizi personali; spesso i revisori non possono utilizzare tutte le fonti a loro disposizione perché altrimenti il risultato di audit potrebbe essere contestabile. Questo problema è ancora più evidente in campo digitale, il modo in cui i revisori possono utilizzare i dati è limitato. Dati di scarsa qualità influenzano tale problematica, di conseguenza anche tecniche come il Process Mining (PM) sono private del loro potenziale. L'obiettivo di questo elaborato è quindi capire come problemi di data quality (DQI) impattino sul lavoro dei revisori e l'uso che questi ultimi fanno delle fonte digitali. Utilizzando tecniche di PM, tramite affiancamento e contatto diretto, sono stati quindi seguiti alcuni casi studio alla European Court of Auditors (ECA).

I risultati ottenuti hanno evidenziato ulteriori problemi nelle procedure di audit che dovranno essere riviste per essere coerenti ed in linea con le nuove tecnologie.

Abstract

The auditing is heavily regulated, this affects the use auditors can make of the sources they rely on for their personal judgments; often auditors cannot use all the sources at their disposal because otherwise, the audit result could be questionable. This problem is even more pronounced in the digital field; the way auditors can use data is limited. Poor quality data affects this issue, as a result, techniques such as Process Mining (PM) are also deprived of their potential. The goal of this paper is therefore to understand how data quality issues (DQI) impact the work of reviewers and their use of digital sources. Using PM techniques, through shadowing and direct contact, some case studies have been followed at the European Court of Auditors (ECA).

The results obtained have highlighted further problems in the audit procedures that will have to be revised to be consistent and in line with new technologies.

1 Introduction

With the implementation of digital, audit techniques also need to evolve and improve. The audit field is difficult to implement due to its highly regulated nature, so the use of digital techniques and tools must be well structured.

Some techniques that are digitally based, such as the use of Artificial Intelligence (AI) or Robotic Process Automation (RPA), are currently in their infancy in the audit[1] field, Process Mining (PM) is still considered an emerging technique, its use is still uncommon especially in the audit field. To be able to use PM techniques it is necessary to transform data into event logs, this is often in contrast with the need in the audit field to have potentially non-controversial evidence.

In this context, there is the problem of the quality of data to be used as input to PM techniques.

The purpose of this master's thesis is to understand and manage the data quality problem in PM and how this affects the work of the auditors. The problem statement of this research is:

- How can data quality issues in Process Mining affect the audit process and related evidence?

Understanding where DQIs are most present and addressing them would improve the use of PM in auditing: developing PM in greater depth would allow auditors to further reduce the time spent analyzing processes; using the results of PM analysis as evidence would reduce the need to identify "alternative sources" of evidence.

The analysis work was carried out at the European Court of Auditors in Luxembourg. It has been structured in three parts: the first concerns the study of the procedures used in ECA. The second one is the study of the existing literature to identify the data quality problems, in the field of PM, currently known. The main source from which this paper draws in this regard is the work done by Van Scheepstal in his paper “Data quality within Process Mining in the auditing context”. The third one, starting from implementation cases observed at the ECA, aims to identify the data quality issues in the real scenarios observed, with the objective of understanding how they are handled and solved by the ECALab team. These issues will then be evaluated from the auditor's perspective to better understand how the auditor makes use of the information and analysis results.

[1] An audit is an independent examination, it provides third-party assurance to various stakeholders that the subject matter is free from material misstatement (Wikipedia)

2 Background

Auditors typically examine processes through non-automated methods such as process documentation, interviews with people, and inspecting samples. This is time-consuming and does not guarantee that the actual problems will be detected.

PM is designed to discover, monitor, and improve real processes by extracting knowledge from event logs available in information systems. The starting point for PM is an event log. Each event in such a log refers to an activity (i.e., a well-defined step in some process) and it is related to a particular case (i.e., a process instance). The events in a case are represented in the form of a trace, i.e., a sequence of unique events (Van der Aalst, Data Science in Action, 2016).

Using PM the auditors can focus on compliance questions, like segregation of duties and process deviations. The advantage of using PM is that the analysis can be much faster. Furthermore, the auditors can analyze the full process, not just samples, and, therefore, achieve a higher assurance. They can focus on the deviations and better identify the true risks for the organization (Fluxicon, Process Mining Book, 2020).

Often the requirements with which data is stored in IT systems are different from the requirements data must have in order for it to be used in PM techniques. In addition, data is often subject to quality issues. Data quality refers to the level of information the data has, both qualitatively and quantitatively. Data is considered to be of high quality if it correctly represents the object to which it relates, and generally whether the information they contain is suitable for its intended uses.

Currently, the present literature divides DQIs into two categories: problems regarding process characteristics and deficiencies due to the event log. In the first case, we have problems arising from deviations in the underlying business processes and information systems, these problems are often the most impactful and difficult to manage because they involve multiple factors and often cannot be solved because they are inherent in the information systems. The second category concerns the quality of the event log, these issues are easier to manage but also time-consuming for the analyst. In table 1 the issues studied in the literature are reported and briefly described.

	Type of issue	Explanation
Process characteristics	Voluminous data	A company can produce immense amounts of data
	Case heterogeneity	A process may have a high number of different scenarios that are difficult to handle
	Granularity	The level of detail of data is different from the level of detail necessary for the analysis
	Concept drift	Business processes change in the meantime that the analyzes are carried out
	Object centric data	IT System does not relate to a certain process to a certain object
Deficiencies due to the event log	Missing data	In an event, log misses mandatory information which is needed for process mining analysis
	Incorrect data	The information available in the event log is logged incorrectly
	Imprecise data	The entries in the event logs are too common or rough
	Irrelevant data	The data in the event log is irrelevant to the applicable analysis

Table 1 - Type of issues known from the literature

However, it is precisely the quality of event data that has been identified as a major problem in the practical application of PM.

In "Process Mining Manifesto" different criteria are used to define the quality of the recorded events, a classification of 5 quality levels is also given in table 1: event logs that do not match reality or have too much-missed data do not allow the PM analysis. On the contrary, reliable and complete event logs allow for optimal results. The first level is often linked to manual data entry, while the last level is linked to recorded automatically and systematically. It is clear that the result of PM analysis is strongly influenced by the input data. Therefore, the ability to identify and resolve DQIs in order to use PM techniques becomes critical even outside the audit field.

Level	Characterization
★★★★★	Highest level: the event log is of excellent quality (i.e., trustworthy and complete) and events are well-defined. Events are recorded in an automatic, systematic, reliable, and safe manner. Privacy and security considerations are addressed adequately. Moreover, the events recorded (and all of their attributes) have clear semantics. This implies the existence of one or more ontologies. Events and their attributes point to this ontology. *Example:* semantically annotated logs of BPM systems.
★★★★	Events are recorded automatically and in a systematic and reliable manner, i.e., logs are trustworthy and complete. Unlike the systems operating at level ★★★, notions such as process instance (case) and activity are supported in an explicit manner. *Example:* the events logs of traditional BPM/workflow systems.
★★★	Events are recorded automatically, but no systematic approach is followed to record events. However, unlike logs at level ★★, there is some level of guarantee that the events recorded match reality (i.e., the event log is trustworthy but not necessarily complete). Consider, for example, the events recorded by an ERP system. Although events need to be extracted from a variety of tables, the information can be assumed to be correct (e.g., it is safe to assume that a payment recorded by the ERP actually exists and vice versa). *Examples:* tables in ERP systems, events logs of CRM systems, transaction logs of messaging systems, event logs of high-tech systems, etc.
★★	Events are recorded automatically, i.e., as a by-product of some information system. Coverage varies, i.e., no systematic approach is followed to decide which events are recorded. Moreover, it is possible to bypass the information system. Hence, events may be missing or not recorded properly. *Examples:* event logs of document and product management systems, error logs of embedded systems, worksheets of service engineers, etc.
★	Lowest level: event logs are of poor quality. Recorded events may not correspond to reality and events may be missing. Event logs for which events are recorded by hand typically have such characteristics. *Examples:* trails left in paper documents routed through the organization ("yellow notes"), paper-based medical records, etc.

Table 2 - Maturity levels for event logs (Van der Aalst, W., Adriansyah, A., De Medeiros, A., Arcieri, F., Baier, T., Blickle, T., 2011)

3 Methodology

The methodology is based on a hybrid approach:

- use of the top-down approach: the main audit procedures present in ECA was studied, in particular, the "Performance Audit Manual" (PAM) and the "Financial and Compliance Audit Manual" (FCAM) in order to understand how audit is carried out within ECA. The present literature on DQIs associated with PM was studied.

- use of the bottom-up approach: shadowing ECALab analysts in the data processing phases, in the process analysis using PM techniques, and in direct contact with the auditors in order to understand how the results of data analysis are then used during the audit process.

In detail the bottom-up approach:

- Data were processed:
    - extraction and inspection phase: the objective of this activity is to understand if the data extracted/received is suitable for the PM analysis or needs to be supplemented.
    - Cleaning and transformation of data into event logs were performed. In both phases, the DQIs present were identified and compared with those studied in the literature. Data quality issues arising from data processing were solved. How the ECALab team manages DQIs has been documented.
- Process Mining analysis was then performed in order to get a complete view of the process and understand if further data processing was necessary. The goal is to understand whether and how DQIs affect a comprehensive PM analysis.
- The results of the analysis were presented to the auditors. It was studied whether and how the auditors used the processed data. The auditors' use of the information from the analysis was studied.

Generally, an ECA analysis has a duration of 12 months, not being therefore possible to analyze an end-to-end process, different projects have been used in order to have a complete project proxy.

4 Data Analysis

Through the shadowing, it was possible to learn about the analysts' *modus operandi* and how auditor requests are handled and resolved by the ECALab team. This process was found to be unstructured, so it was formalized through a BPMN (fig 1).

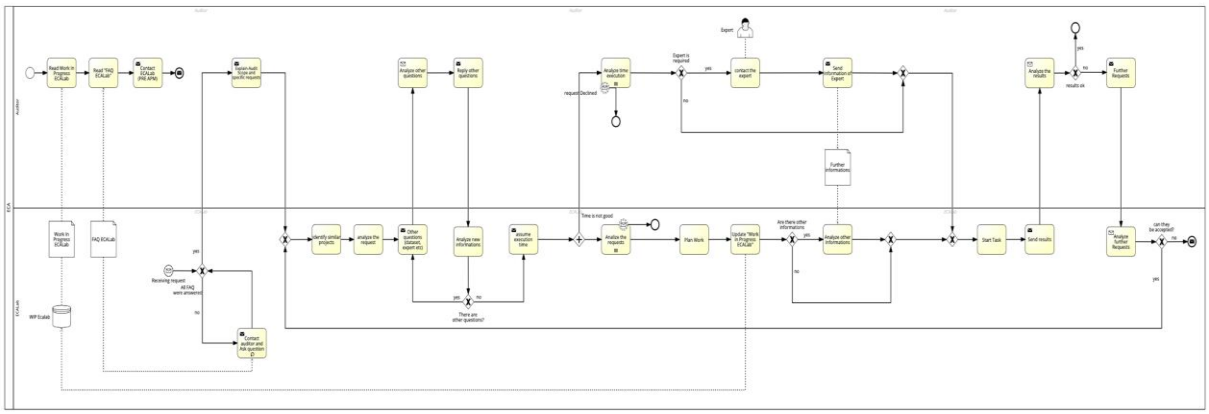


Figure 1 - BPMN ECALab request

4.1 Extraction and inspection phase

The first phase is database extraction/reception. Once the auditor's request is received, the ECALab team inspects the data for an initial assessment based on the audit objectives. It is then determined if the data is sufficient or needs to be supplemented with other sources. In this phase the DQIs encountered are often related to voluminous, incorrect, and imprecise data, but if in the first case this problem is easily identifiable in a qualitative way through an assessment "minimum data needed/tot data received", as regards the incorrect and imprecise data these are identified only during the processing of the dataset. This involves a long work of preprocessing for the analyst: as the analyst increases the knowledge of the given dataset he is able to identify more easily problems of this type, however, this activity is time-consuming. Regarding the missing data instead, this problem is generally limited to a few data in the dataset and is not solved directly, but through the use of statistics, in fact, a large amount of data allows an appropriate analysis through the use of estimates that must be validated by the statisticians of the ECA.

4.2 Transformation phase

Once the data were received and inspected, using Anaconda software, the data were cleaned so that the reviewer could use the analyses appropriately to better understand the context and any limitations of the data received. Such data processing, however, makes the received data contestable in the audit field. Each processing on the data involves a small approximation on it, these approximations add up for each processing, which makes the data contestable in the audit field.

Once the data had been inspected and cleaned, the transformation of the data into "synthetic" event logs was then performed. The data received were in fact not ready for analysis of PM, it was, therefore, necessary a strong pre-processing to make the dataset

usable. Communication with the auditee was not easy as the tables and data types required are far from the way the auditee's information system stores data.

The problems encountered in this phase were more difficult to solve, in this case, a high volume of data, incorrect or missing data create problems. The transformation was done through Python software, the transformation code is also made available to the auditee.

During this phase there is poor communication between the analyst and reviewer, generally, the reviewer has little affinity with the data science domain, which causes misunderstandings between the two levels based on what is possible with the dataset and what the reviewer would like. Continuous alignments are then required, but these often do not lead to the reviewer's desired outcome.

The main problem faced was a problem of granularity, the level of granularity was therefore low and/or variable, this did not allow an analysis too deep as the time required to perform detailed analysis with this level of granularity would be extremely high.

Other data received were instead subject to problems of Object centric data, this problem has been addressed using integrations of tables through Excel or Python, the results of these solutions are often not optimal from the point of view of PM because even if the data are not subject to changes, the process displayed in the software PM is fragmented or not very linear, thus losing useful information from an analysis PM.

4.3 Analysis phase

Through the Disco and Prom6.9 tools, a process analysis was then conducted using PM techniques. This analysis provided insight into whether the transformation of the data into synthetic event logs was adequate and/or whether further processing was needed. The transformation activity is the most complicated activity and requires good experience on the part of the analyst as well as excellent knowledge of the auditee's domain and business processes. Communication between auditor and analyst would need to be constant at this stage to integrate the auditor's in-depth knowledge of the context with the analyst's analytical capability. Activities and the "happy path" were then identified, as well as deviations and critical points such as bottlenecks or compliance issues.

In-depth analysis using PM techniques also revealed an additional problem, the problem of concurrency. This problem stems from the PM algorithms that are still unable to properly identify parallel or exclusive XOR choices, meaning that some tasks are placed in parallel or skipped because the software is unable to recognize the right sequence. This obviously

involves some problems at the moment of the analysis. This problem is easily solved through proper knowledge of the process, but sometimes it is not easy to detect.

4.4 Analysis results

The results of the analyses were then discussed with the reviewers who could either request further analysis and investigation or use the results obtained. The ECALab analyst must then provide the reviewer with qualitative analysis on the approximations made to the data to arrive at the given result. This step, like troubleshooting, is not structured, nor is it possible to provide a quantitative result; it is all based on the analyst's experience and audit objectives, which makes the process highly subjective. Generally, auditors use the results of the analysis not as evidence but as confirmation or as a path to audit inquiries. While it is true that such information cannot be used as evidence, it is equally true that it allows for a more precise focus. This limitation in the use of data is due to the various approximations that the data must undergo, the various processing makes the data questionable from the audit point of view.

5 Conclusions

The study of audit procedures revealed a lack of integration of digital processes into the audit process. ECA procedures have not yet integrated the possible benefits of digital analyses, the auditors that to date rely on ECALab for such analyses are extremely small in number, a process integration would help the ECA system to improve this critical aspect. The first step in this regard has been taken with the formalization of the ECALab requests procedure, this will then need to be implemented and improved as the flow of requests increases and expands.

During the data preparation phase, it was evident that the data received by the ECAlab team is often from legacy systems, such systems are unlikely to store data in a format that can be used in PM. Such datasets are problematic to clean and transform, they often need to be integrated with other data. Because of the poor aptitude of such datasets in PM, it is often necessary to use "synthetic" event logs. However, this issue is external to the ECA and the auditors have no power over it. Auditees often use information systems that are not up to date, this leads to data storage problems that can only be resolved through long hours of pre-processing by analysts, and the result may not fit the audit objective. Therefore, auditors often prefer to use other sources for their evaluations.

The main issues addressed were related to data transformation. In order to create "synthetic" event logs, the data underwent heavy preprocessing by analysts. Data subject to processing often undergoes approximations that, although very small, add up at each stage.

The analysis phase showed that DQI affected datasets can be used by the analyst, who with a more or less long process is able to guarantee results. Results from low-quality input are not very useful and in fact, may be confusing. The main objective of the PM in the audit field is to show the processes as they really happen and to identify the real deviations present within the audited processes. Data that does not allow for these findings in a clear manner is detrimental to the purpose of the audit. But, in order that the analysis of PM, with a dataset affected by DQI, is useful it is necessary a deep knowledge of the domain. An experienced analyst with a good knowledge of the domain is able to face and evaluate DQIs without affecting the overall analysis, and to have a qualitative idea of the approximations needed to continue with the analysis, DQI problems could mislead an inexperienced analyst, making the analysis useless.

From the auditor's perspective, information from PM analyzes is usable: auditors can request further insights from the auditee using their own professional skepticism as a reason. However, reviewers cannot use the data as evidence because of the treatments and resulting approximations the data must undergo to be usable in PM.

The research has therefore highlighted how the DQIs do not allow the use of data in the audit field as audit evidence, every approximation that the data undergoes increases the degree of contestability of the source risking invalidating the audit result and the final judgment of the auditor. Until auditee information systems can provide data of high enough quality to warrant its use in PM software, PM techniques will likely remain in the background of auditors' work. The research question was then answered.

Anyway, this condition does not downplay the role that the use of data, even if affected by DQI, has in the audit field and as support for the work of the auditor. Because while it is true that analyses cannot be used as evidence, it is also true that a PM analysis makes it easy to identify real process weaknesses and deviations that would otherwise be time-consuming and perhaps never discovered by the auditor. In fact, the auditor can use the results of the PM analysis as the basis for his or her professional judgment, without explicitly saying that such information came from a PM analysis.

It is evident how the lack of flexibility that the audit sector allows when it comes to audit evidence. In this way, auditors use digital tools as confirmation rather than evidence, limiting

the potential for analysis. The regulations in the audit field have not yet been updated and adapted to the digital evolution, this if on the one hand does not allow the use of digital techniques, on the other hand, does not encourage the auditors in the use of these techniques thus postponing the updating and digitization of audit procedures. The auditors, not being "protected" in the use of digital, cannot use the data as evidence, they use the analysis mainly as confirmation. On the other hand, it's also fair to point out that reviewers are not very digitally inclined: transformation codes can indeed be validated and evaluated, but most reviewers don't have the appropriate skills to be able to read or write a code. This is another reason why processed data is rarely used as evidence.

More accurate code assessment guidelines would indeed allow auditors to use data, however, this would shift the focus to the code, so auditors need to improve their digital knowledge. It would therefore be advisable also to use compilation standards. However, it is fair to point out that such practices should be addressed to the auditee rather than auditors themselves, an international compilation standard would greatly help the auditors' work.

In any case, for an “official” audit with audit reports, the findings are first sent to the auditee, so if process mining results are used as audit evidence, the correctness of the information must be verified by the auditee.

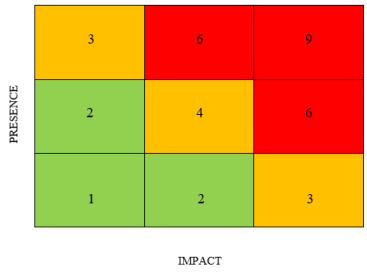
There is a fundamental problem with this, stemming from the failure to measure data reliability. The quality of the data, before and after treatment, cannot be measured. While the data received from the auditee may be taken for "granted" from a reliability perspective, this is not true after the same data has been processed. Having the ability to measure the reliability of the data after it has undergone processing would in fact allow the data to be used as audit evidence without the risk of challenge. The audit sector would benefit from having a measure of reliability: analysts, auditors, and even auditees would be able to compare themselves on objective data instead of bureaucratic talk. A measure of trustworthiness would allow analysts' processes to be integrated as true support for audit work and use the information from the analyses as audit evidence.

6 Future work

Regarding the data reliability issue, an interesting future development would be the ability to create a risk management framework to calculate the reliability of the data. This would allow that tool to be used to evaluate data, clarifying whether that data is contestable or not in the audit field. The goal is to have a usable and globally recognized tool as a standard for

assessing the reliability of a data/dataset, then understand whether the data is usable, usable with reservations, or not usable as audit evidence. Currently, such a framework cannot be developed accurately due to the lack of some information necessary for its construction.

The framework provided below is only an example of usage but has never been tested in a real context.



Through a 3x3 risk matrix (tab. 3), it would be possible to define whether or not the data has undergone treatments and how these have impacted the reliability of the data in the audit field.

Reliability index = impact x presence

An example of the Evaluation Criteria could be:

- Impact:
    - Low: The data has not been approximated
    - Medium: The data undergoes approximations that limit its use
    - High: the data has been subjected to approximations that affect its use in the audit field
- Presence:
    - Low: less than 35% of data is subject to the problem
    - Medium: less than 70% of data is subject to the problem
    - High: more than 70% of data is subject to the problem

At this point, the reliability calculation of the data would be needed.

This figure cannot be derived due to the lack of information regarding:

- Correlation between data quality issues

- Aggregation of data quality problems in a dataset: we are not currently able to give an effective weight to the single problem identified

- Impact and approximation are calculated based on the professional judgment of the analyst, this is subjective

- Data reliability and subsequent approximations: each step adds uncertainty that is relatively easy to resolve and quantify, but this causes overall uncertainty. The problem of successive approximations becomes greater with each step.

- Additional restrictions include ISO standards and legislative restrictions.