Multimodal Provenance-based Analysis of Collaboration in Business Processes

Maria Luiza Falci, Andréa Magalhães, Aline Paes, Vanessa Braganholo, Daniel de Oliveira

Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brazil
marialuizafalci@id.uff.br
{andrea, alinepaes, vanessa, danielcmo}@ic.uff.br

Abstract. Modeling business processes as a set of activities to accomplish goals naturally makes them be executed several times. Usually, such executions produce a large portion of provenance data in different formats such as text, audio, and video. Such a multiple-type nature gives origin to multimodal provenance data. Analyzing multimodal provenance data in an integrated form may be complex and error-prone when manually performed as it requires extracting information from free-text, audio, and video files. However, such an analysis may generate valuable insights into the business process. The present article presents MINERVA (***M**ultimodal bus**IN**Ess p**R**o**V**enance **A**nalysis*). This approach focuses on identifying improvements that can be implemented in business processes, as well as in collaboration analysis using multimodal provenance data. MINERVA was evaluated through a feasibility study that used data from a consulting company.

Categories and Subject Descriptors: H.2.1 [**Database Management**]: Logical Design; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords: Multimodal Data Provenance, Provenance Data, Business Process Data.

1. INTRODUCTION

With business globalization over the last decade, organizations recognized the importance of formally modeling their processes [ABPMP Brazil 2011; Reijers 2021]. Some several languages and notations can be used to model a business process. The most known is BPMN[1] (Business Process Model And Notation), an OMG[2] (Object Management Group) standard. As globalization demands, organizations need to be always improving so they can stay competitive. Thus, one of their needs resides in process improvement. Improving processes requires a detailed analysis of their design and execution. For that, historical data are necessary. This type of data is called *Provenance* [Freire et al. 2008].

Provenance data describe the origin of a specific piece of data, as well as the processes that transformed it from its original form into its final state. Provenance data make it possible that, when finding a data fragment with unexpected behavior, one can search the origin of data and the transformations performed until the current moment. Provenance data present its own semantics (following a W3C standard named PROV[3] [Groth and Moreau 2013]) and need to be searchable, which is different from traditional log files produced by business process engines. In the business process analysis context, provenance data enable identifying which collaborator contributes to a specific activity or process,

[1] https://www.omg.org/spec/BPMN/2.0/About-BPMN/-LastAccess:30/08/2021
[2] https://www.omg.org/about/index.htm-LastAccess:30/08/2021
[3] https://www.w3.org/TR/prov-overview/-LastAccess:30/08/2021

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This study was also financed in part by CNPq and FAPERJ.

for example. Furthermore, provenance data inform data sources used in processes, which activities are attributed to which collaborators, *etc.* Efficient collaboration is the key in process improvement [Mendibil et al. 2002]. Discovering how people collaborate on the execution of a process may open several opportunities for process improvements, including reallocating people among tasks, adding new people for specific tasks, *etc.*

However, this kind of analysis is not trivial, since provenance data in business processes are naturally multimodal, *i.e.*, they can be represented in different formats (*e.g.,* event logs, texts, audio, *etc.*). All of this information should be considered during the collaboration analysis and process improvement. Multimodal provenance analysis offers the possibility to obtain knowledge through the integration of a wide range of heterogeneous data from diverse data sources to enrich collaboration analysis.

Despite the great potential of multimodal provenance data for understanding users' collaboration in business processes and supporting process improvement, research in this direction remains scarce. The majority of the existing approaches found in the literature use only event logs as input data. These approaches analyze collaboration in business processes, but they are not designed to handle heterogeneous data. For example, [Van Der Aalst et al. 2005] discover social networks based only on event logs, and they are capable of capturing only the names of employees who are directly allocated to each activity. [Ferreira and Alves 2011] also discover social networks from event logs, and they affirm that these social networks can get too complex and challenging to analyze and comprehend. To deal with this problem, they propose to group users from the social network in communities, enabling the analysis and visualization of the social network in different abstraction levels. [Zhao and Zhao 2014] present a survey of approaches that use event logs to extract the organizational structure, social network, roles in the process, and resource allocation in business processes. All of the approaches mentioned earlier do not consider information within e-mails, documents, messages from apps, audio records, *etc.* Thus, organizations that use these approaches may miss essential optimization opportunities.

Capturing and analyzing multimodal provenance data in the business process context is an open yet important problem. Even though there are provenance representation standards such as the W3C PROV [Groth and Moreau 2013], which can be extended and applied to different contexts, there are no standards for provenance capturing and processing. Also, most existing approaches rely on capturing and querying provenance in a structured form, and thus do not deal with multimodal provenance. Thereby, in order to use multimodal provenance data to analyze collaboration and identify possible improvements in business processes, this article proposes an approach called MINERVA (***M**ultimodal bus**IN**Ess p**R**o**V**enance **A**nalysis*). MINERVA aims at extracting provenance data from a plethora of formats, including event logs, free-form text, audio, and video. MINERVA is based on Natural Language Processing (NLP) techniques and graph databases to perform the analysis. Evaluation using actual business process data from the company "dheka Consultoria"[4] shows the benefits and the potential of the proposed approach.

This article is an extension of a conference paper published in the Proceedings of the 2020 Brazilian Symposium on Databases (SBBD) [Falci et al. 2020]. In this extended version we enriched the Experimental Evaluation Section with a new analysis. We have also added a Background section that discusses business processes and multimodal provenance, and improved the related work section. The remainder of this article is structured as follows. Section 2 presents background information. Section 3 discusses related work. Section 4 details the proposed approach. Section 5 presents the experimental evaluation and discusses the results. Finally, Section 6 concludes this article.

[4] https://www.dheka.com.br/-LastAccess:30/08/2021

2. BACKGROUND: BUSINESS PROCESSES AND MULTIMODAL PROVENANCE

Although many people think that the purpose of a company is to make profit, the heart of a company is on the development of high-quality products and services to attract new and keep current customers. High-quality products and services can only be provided if there is a well-defined Business Process with activities executed by humans or machines. Such activities can be represented through a Business Process Model using Business Process Model and Notation[5]. Business Process Management (BPM) can guide organizations to manage their business processes aiming to acquire better results [ABPMP Brazil 2011].

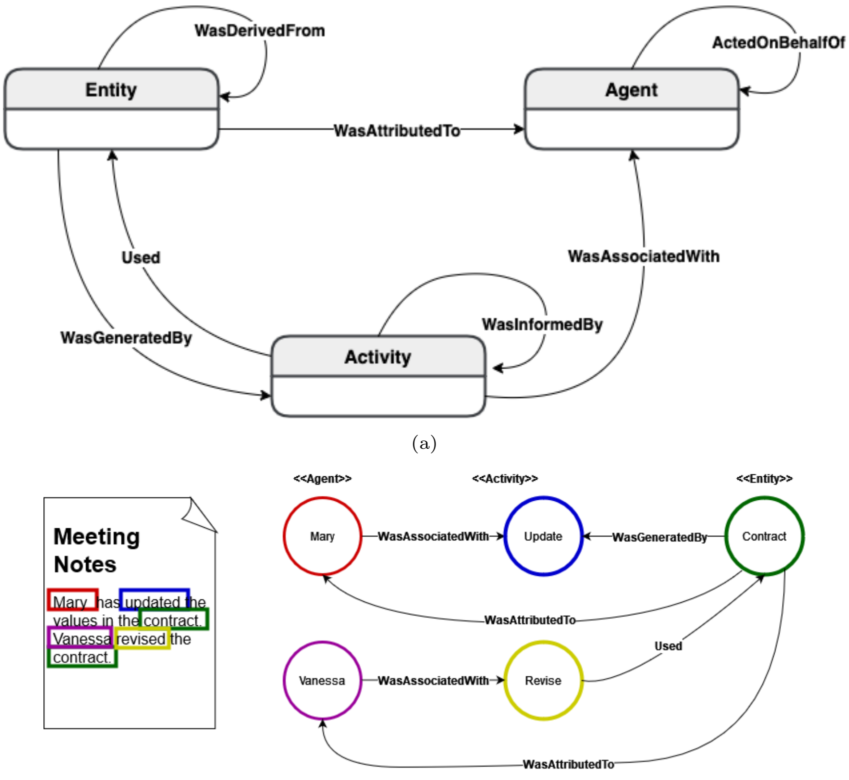
Business Processes can be executed as many times as needed, and each execution is named a "process execution instance". Each instance may have distinct metadata, *e.g.* different start and end times, agents responsible for their execution, artifacts produced and consumed during its execution, *etc.* Business Process Management Systems (BPMS) can support these executions, and they commonly store data about each instance in event logs. Event logs are traditionally plain text files where each line contains data about a specific activity execution. These logs have to be analyzed to extract knowledge about the process to discover how they can be improved. However, it is worth noticing that these files are not searchable, *i.e.*, one has to write a parser to extract useful information from files before performing queries over data.

Despite the utility of event logs, many valuable data of a business process execution instance is found in other types of files. For example, the content of an e-mail may provide helpful information about an instance, such as produced artifacts, the data derivation path, which person is responsible for a specific activity, *etc.* This information is also found in phone calls, text in instant message Apps (*e.g.*, Whatsapp, Telegram), videos, and audio recordings. For this reason, it is fundamental to capture this historical information from these heterogeneous files. In this article, we call this type of data "Multimodal Provenance Data". Multimodal provenance can also be classified as "implicit provenance" [Neves et al. 2017] since it involves data that was not clearly captured and structured to serve as provenance data. For example, log files are designed to contain historical information about the process, while free-text files (*e.g.*, an e-mail) are not. However, an e-mail may contain important provenance data to analyze. Once a company aims at improving its business processes, multimodal provenance data have to be captured and stored, and later can be analyzed in order to find possible improvements.

Although it is far from trivial to capture multimodal provenance data (since it involves extracting elements from free text, audio, and video files), we can benefit from the existing solutions for storing and querying provenance data. There is a W3C provenance standard, named PROV [Groth and Moreau 2013], that can be used to model this type of provenance data. PROV allows for users to represent provenance data in terms of entities, agents, activities, and multiple types of relationships as presented in Figure 1a. An entity is a physical, digital, or conceptual representation of a thing. In this article's context, an entity may be a file, a document, a video, *etc.* Activities are actions that occur in a specific period and act upon entities. For example, data transformations, modifications in a document or a contract are defined as activities. Finally, an agent is an actor or software agent responsible for performing activities, own an artifact, *etc.* Although PROV is an agnostic model (not tied to a specific domain of knowledge) it can represent the data derivation path and people that can influence it. In the context of multimodal provenance, the provenance graph can be generated based on the content of free text files. In Figure 3b, annotations of a specific meeting mention that "Mary has updated the contract" and that "Vanessa has revised the contract". Both sentences state that there was a transformation in one artifact (*i.e.*, contract), but this kind of information may not be represented in event logs.

Although in the literature several different approaches assume that only one individual is responsible

[5] `https://www.omg.org/spec/BPMN/2.0/About-BPMN/-LastAccess:30/08/2021`



Mary has updated the values in the contract. Vanessa revised the contract.

(a)

(b)

Fig. 1: PROV Data Model (a) and an example of Multimodal Provenance (b).

for performing a process activity [Schönig et al. 2018], recently the BPM research area has started to pay attention to social aspects [Ariouat et al. 2017] of the process execution. Even though an activity may be associated with only one individual in the event logs, in real life, it can be influenced by many individuals, who can collaborate towards the same goal. In this context, the multimodal business process provenance data becomes strategic since it reflects the reality of the individuals who have actually collaborated to execute an activity and contributed to the artifact construction.

3. RELATED WORK

This section discusses existing approaches related to the social aspect of the business process execution data. To the best of our knowledge, none of the existing approaches use multimodal provenance data. Thus, in this section, we focus on approaches that use event logs to perform collaboration analysis.

Zhao and Zhao [2014] do not propose a new approach in their work. Instead, the authors provide an overview of process mining literature with a focus on the organizational perspective. Although the survey papers focus on organizational mining, not all of them are directly related to the social aspect. Additionally, the approaches presented in the papers they survey use only event logs as input data.

Van Der Aalst et al. [2005] aim at discovering social networks from event logs. In the discovery process, the authors use the field from the event log that represents which person was associated with each activity. Besides, the authors also consider the order of activities (they are ordered according to the execution timestamp). Thus, they can understand the hand-over of work from one user to the next. According to the authors, if a hand-over *e.g.,* from user "a" to user "b" occurs more frequently than other (*e.g.,* from user "a" to user "c", when "b" and "c" play identical roles on the organization), it may be inferred that "a" and "b" have a stronger interaction than "a" and "c". The authors present a case study with real data from Dutch National Public Works Department that employs 1000 civilians.

The case study used the particular process of handling invoices. Although this work represented a step forward, it just considers data represented in log files, which may reduce opportunities for optimizations in the business process.

Ferreira and Alves [2011] focus on the organizational perspective of process mining. However, instead of discovering social networks from event logs, which can lead to very complex models when applied to real data, the authors present an approach to discover communities on event logs. To accomplish this goal, they apply hierarchical clustering techniques to users' data and create communities that allow for the analysis and visualization of the social network created in different abstraction levels. They evaluate their approach with a case study with real data from a hospital.

Different from the aforementioned related work, Schönig et al. [2018] do not use only event logs as input. In their approach, each event must contain an explicit reference to the enacted task and the "operating resource" (the person who is responsible for a task), but also the organizational background knowledge (which should explicit the roles, capabilities, and membership from each employee). Afterwards, these data are mined to compose teams to work on collaborative process activities.

Schönig et al. [2018] state that prior research in this area assumes that only one person is related to each process activity, and does not consider that activities can be collaborative. The authors propose a two-phase approach that extracts (from the event log) data about people working in a collaborative activity, and then finds out their characteristics (*e.g.,* skills and roles). After that, a post-processing phase is executed to obtain the most informative team compositions. The authors evaluate their proposal with real-life event logs.

However, none of the approaches mentioned earlier consider using data extracted from heterogeneous formats neither structure such data using provenance representations as W3C PROV. To bridge this gap, we propose MINERVA, which is detailed in the next section.

4. PROPOSED APPROACH: MINERVA

Business process data can be analyzed aiming at different targets. The most common target is to understand the process itself, trying to find possible improvements that can be performed, or other analyses (*e.g.*, analyses of collaborations and process improvements). To illustrate this analysis necessity, one example would be to investigate the average quantity of employees related to processes, or which individuals have contributed to more than one process. In this scenario, a process manager could try to answer the questions "Which individuals have influenced more than one process instance?", or "Which process had more than one employee related to it?". The answer to these questions could help the process manager to comprehend details about the process instances. Then he could make decisions in case the reality is not according to his expectations. Traditionally, this type of analysis uses as input only event logs. However, business process execution produces and consumes different types of data in multiple granularities, *e.g.*, e-mails, the audio of a meeting, data extracted from platforms that support activity executions, *etc.* All provenance data from the process is stored in these multiple heterogeneous files and should be considered in this analysis.

To enable this type of analysis, provenance data must be extracted from these heterogeneous sources and structured in a searchable form. In this article, we propose MINERVA, which aims at solving this problem. The proposed architecture of MINERVA is presented in Figure 2, and it is composed of three layers: (i) External Data Sources, (ii) Processing Layer, and (iii) Data Layer.

The *External Data Sources* layer contains all data products produced during the execution of a business process, such as event logs, documents, audios, e-mails, *etc.* Implicit provenance data must be extracted from them. The problem that arises is that provenance data are represented in different formats. MINERVA processes texts from e-mails, texts from instant message applications (*e.g.*, WhatsApp, Telegram, Signal), meeting videos and audio, and free-text comments that can be written in process execution support platforms (*e.g.*, Pipedrive, Trello). These comments commonly contain fundamental provenance data, which can explain the collaboration course during the process execution.

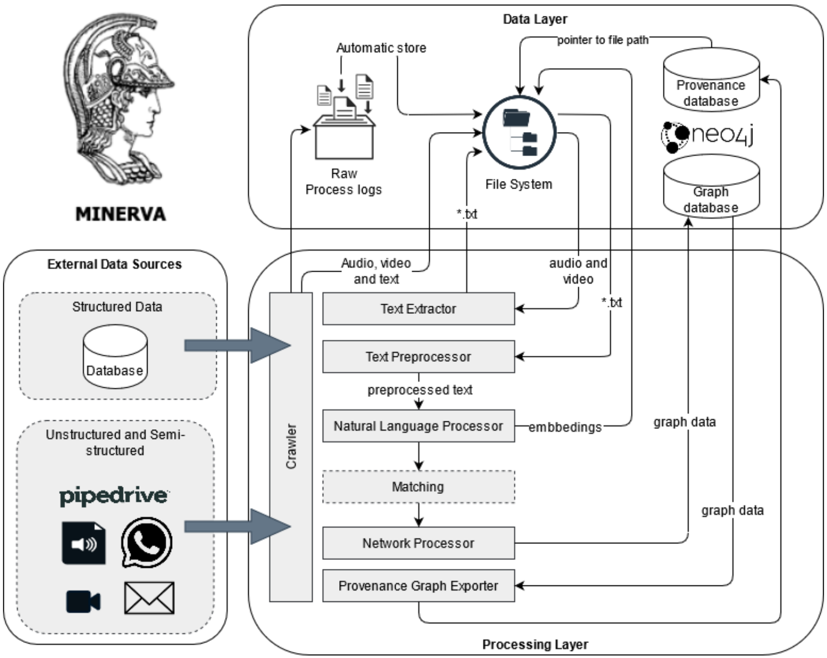


Fig. 2: The proposed architecture of MINERVA.

In the *Processing Layer*, all content imported from external data sources is captured by the *Crawler* component and stored on the *Data Layer*. The *Crawler* is the agent that receives a list containing different endpoints of resources to access. We assume that the Crawler can process data from different data sources. As soon as the *Crawler* accesses these endpoints, it downloads all content to the file system. Since such content may contain audio and video, a text extraction must be performed. The *Text Extractor* converts the audio files and the audio extracted from video content into text. These data are also stored in the file system and referenced on the provenance database (just a pointer to the path where the file is stored). The storage (File System) resembles a Data Lake [Miller 2018], because the files are stored in their raw format, before the transformations that will be performed in the next step. Once the content of all non-structured files is in text-form, the Text Processor component cleans special symbols and specific characters, breaks the text into tokens, and removes stop-words. Stop words are those without intrinsic semantic information. Next, Once the content of all non-structured files are in text-form, the *Natural Language Processor* identifies elements as substantives and verbs on the text, which will be used to identify the actions performed during the execution of the business process and the collaboration network. The *Matching* component is optional, and it can be used to match terms from texts extracted from audio and video files. This matching component is required since MINERVA needs to integrate heterogeneous text files, written independently, and thus each having their vocabulary. Equivalent terms in multiple texts are not necessarily identical, but they have a level of similarity. Euzenat and Shvaiko [2013] discuss that syntactic and semantic dimensions can be considered when defining the similarity level of two terms. Mestre and Pires [2014] and Ribeiro et al. [2016] propose approaches to solve problems related to matching.

Finally, after selecting the words that are nouns and verbs, the *Network Processing* component creates the collaboration graph. The proper nouns collected by the *Natural Language Processing* component are used to create the network's nodes, and two types of edges are created: "isRelatedTo" from entity nodes to process instance nodes; and "belongsTo", from process instance nodes to company nodes. It is worth mentioning that the network structure contains the same agents of the provenance graph. However, it has relationships (e.g., isRelatedTo) that are not part of the PROV recommendation. An entity extraction module, such as the one proposed by Han et al. [2019], can be used in this component. With this approach, entity names are extracted and related to the process instance from where the comments texts were extracted. Additionally, the process instance is connected to the company that owns this process. A collaboration network is created, and it is possible to observe the employees who have collaborated on each process instance.

As an example, in the sentence “Mary called John”, the *Natural Language Processing* component creates two nodes [Mary, John], and edges between “Mary”, “John” and the process instance are created, to indicate that these two entities are related to the process instance from where the comment was extracted. The complex network created represents a social network with the actors from processes executions. Proper nouns from this social network may not be mentioned on the event logs. Even though the activity was associated with a specific individual in practice, in theory (or in the event logs registry), this person may not be formally related to the activities that she executed in practice. In general, process specifications (and their logs) contain only roles instead of employee names, for example.

Besides generating the collaboration graph, the *Network Processing* component also analyzes the role of each network node (that represents people or organizations), which has details as node centrality, incoming links, and outgoing links [Perez 2020] [Boccaletti et al. 2006]. With this analysis, it is possible to understand better the influence that each person had on the execution of a specific process or activity [Cross et al. 2004]. Finally, the *Provenance Graph Export* component generates a *JSON* file that contains the graph generated according to the model mentioned earlier. In the end, the *Data Layer* contains provenance data (currently in a relational database in PostgreSQL DBMS), the file system that stores the raw data files and embeddings, and the graph database that stores the collaboration network.

4.1 Implementation

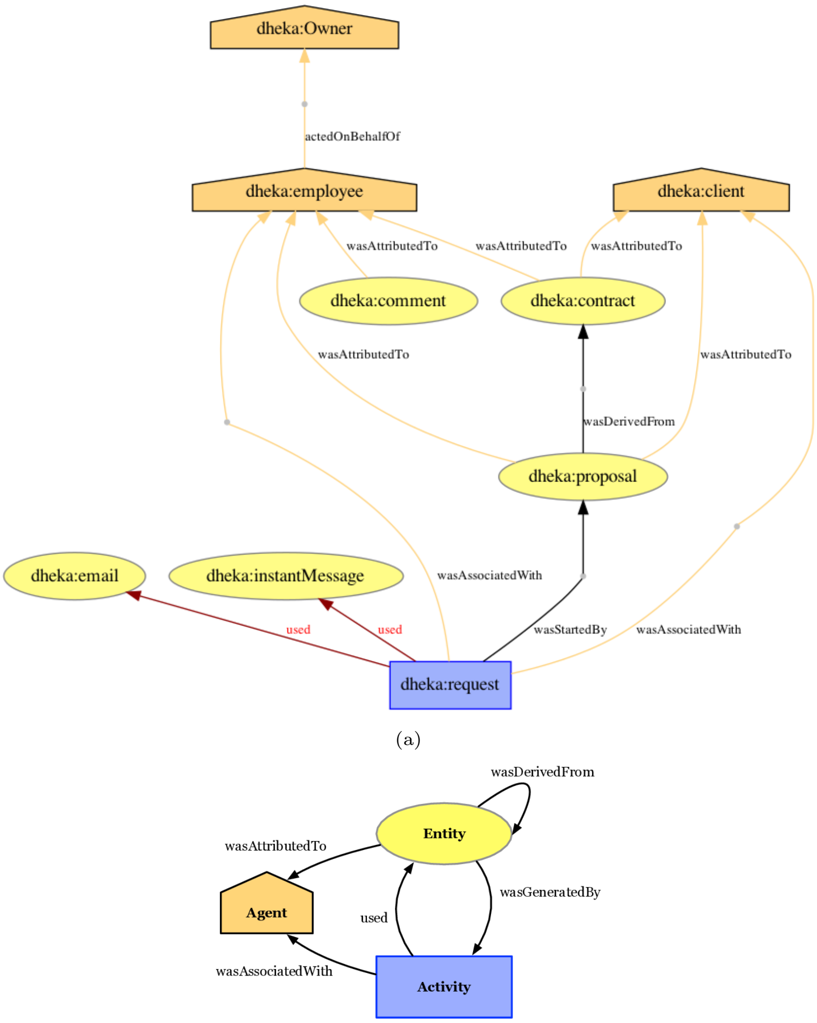
We have implemented a proof of concept of MINERVA in Python and Cypher. In its current version, MINERVA uses the Google Cloud Speech API[6] as the *Text Extractor* component. This API can convert audio and video to text, which can later be processed by the *Natural Language Processor*. We currently support texts in Brazilian Portuguese only. In our current implementation, we have not implemented the *Matching* component of the architecture. The reason is that in our proof of concept the text is usually well structured, and so for now there is no need for such a component.

The *Natural Language Processor* component can be implemented in multiple ways, but in the current version, it uses Spacy [Honnibal and Montani 2017], which is a natural language processing tool. We use Spacy to normalize the text by using lemmatizing, removing stopwords, and marking POS (Part of Speech) categories of words, *etc.* The texts are also saved in a vectorial form (*i.e.,* embeddings), so it could be processed by other machine learning algorithms in the future, if needed.

The *Provenance Graph Export* component is implemented using Cypher in Neo4J. It generates the collaboration graph and exports it to JSON. This provenance graph is validated using existing services such as ProvToolBox[7]. This service automatically checks if the structure of the provenance graph is compatible with PROV, which is a W3C recommendation. To store the graph, we currently use Neo4J. Although there are other forms to query provenance graphs (*e.g.,* using Prolog [de Oliveira et al. 2017]), we have chosen Neo4J for its simplicity.

[6] https://www.labnol.org/code/20280-google-cloud-speech-api/ - Last Access: 30/08/2021

[7] `https://lucmoreau.github.io/ProvToolbox/-LastAccess:30/08/2021`



(a)

(b)

Fig. 3: The proposed Provenance Model (a) and PROV model (b).

5. EXPERIMENTAL EVALUATION

Case Study. To evaluate MINERVA, we have chosen a case study with data from a real company ("dheka Consultoria"). The data is related to the selling process, which is spread across event logs, comments in Pipedrive, and e-mails. In the case study (CRM platform), each potential sale (that can be further confirmed or not) is considered an instance of a sales process. Each instance has its provenance data about the sales process activities execution, such as employees associated with each activity, data about the client that is related to that sale, and free-form comments.

The sales process' provenance data model is presented in Figure 3a. As shown in the Figure, the data derivation path starts with a (*request*) for proposal. This request is associated with *e-mails* and (*instantMessages*). Based on a (*request*), a business (*proposal*) is created, which has relationship with a (*client*) and one or more (*employees*). This proposal can evolve to a (*contract*) (after being revised and signed). The employee is also associated with (*comments*), that one can write about the process in free-form text. Usually, the data analysis about the sales process activities uses only event logs. However, as the proposed approach focuses on data analysis that can bring additional information, the comments in the free-form text (in Pipedrive or e-mails) are also analyzed.

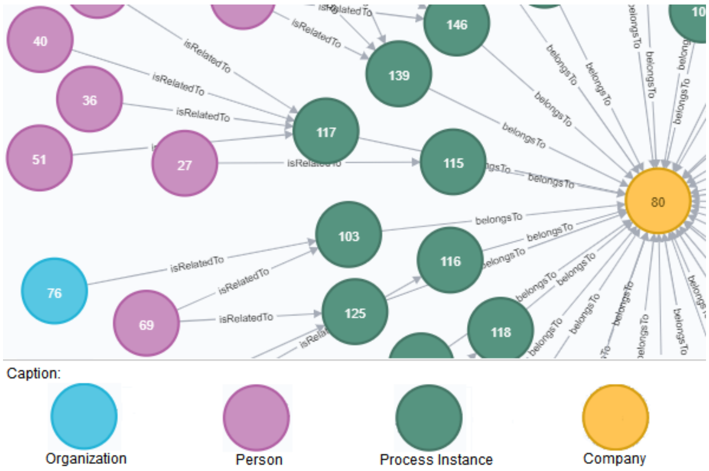


Fig. 4: Fragment of the Collaboration Network.

The comments are downloaded directly from Pipedrive in CSV (Comma Separated Values) format. Each comment may present more than one sentence. Comments can also have abbreviations and acronyms, that were previously added to a dictionary, which replaces the abbreviated words with their non-abbreviated form. After this text transformation, each comment is used as input to Spacy [Honnibal and Montani 2017] in Portuguese (the comments' native language), and Spacy extracts nouns and verbs from each sentence. Neo4j is used to create the complex network. The network's nodes, edges, and queries about the complex network are created using the Cypher query language. Individuals' names and organizations' names are used to create the nodes. The edges created are "isRelatedTo", when an entity node is related to a process instance node, and "belongsTo" when a process instance node belongs to a company node.

After creating the graph database, the graph shows all the people and organizations that influence the process execution instances, as presented in Figure 4. When the complete collaboration network is built (just a fragment is presented in Figure 4), it is possible to update the business process model to consider people that collaborated with the process execution and who were not officially allocated to the process before the execution of MINERVA. Officially, each process instance is related to one person only. As presented in Figure 4, in some cases (as on process instance 117), we can identify that four people are related to the process.

Data Analysis. The goal of the data analysis is to find possible improvements based on collaboration data extracted from multimodal provenance. The images of the complex networks presented in this section are anonymized to avoid the exposition of real data from employees, clients, and organizations. The analysis of the business process social network presented in this section does not have as the main goal to find out new activities or to "rediscover" the business process. However, it is possible to use the data and analysis to accomplish this goal. We have interviewed a company manager to ask for questions that should be answered using the generated collaboration graph. The answers to these questions potentially improve the process, or at least can be used to update it with current implicit collaborations. The selected questions are:

(1) Q1 - *Which individuals have influenced more than one instance?* - By answering this question, it is possible to investigate the people who have more influence on the sales processes. Initially, according to the activities execution data, only one person is associated with each process instance (and all the process activities). When analyzing the comments in free-form text, new proper nouns from people related to each process instance were found. However, as many names can be related to a process instance, and one person can be related to more than one process instance, it may be interesting to investigate who is related to more than one instance. This way, we investigate

nodes with type "Person" who have more than one outgoing link, so they have a higher degree in the network, and have a stronger influence than nodes with just one outgoing link.

(2) Q2 - *Which projects had more than one related entity (people and organizations)?* - By answering this question, it is possible to investigate which processes instances are related to more people and organizations. This question is important to help investigate instances that required more effort to be executed, maybe because they are complex and more collaboration is needed, or because more people asked for assistance during the execution course of the process (*e.g.* to answer questions). In case an instance presents an unexpected behavior, it may be interesting to ask further questions and investigate if something can be improved in future instances.

Performed Queries and Results. The social network graph built based on the case study has 101 nodes and 108 edges (where blue nodes correspond to "organizations", purple nodes represent "people" (or collaborators), green nodes represent "process instances", and yellow nodes represent "companies"). Since it is not a small graph, if one analyzes it manually, it will be a tedious and error-prone analysis. This way, it is necessary to perform queries and get results using a Database Management System, in this case, Neo4J (a graph-based database system). Several queries were developed (in the Cypher query language) to answer the aforementioned questions. In the following topics, we present these queries and discuss the results.

(1) Q1 - *Which people have influenced more than one instance?* The submitted query (which is shown below) in Cypher searches for nodes with the type *Person* who had more than one relation *isRelatedTo* with instances.

```
MATCH p=(u:Person)-[r:isRelatedTo]->(:Project)-->()
WITH u, count(r) AS count
WHERE count > 1
MATCH x=(u)-[r:isRelatedTo]->(v:Project)-->()
RETURN x;
```

By analyzing the produced graph, presented in Figure 5, it is possible to observe that from the 51 nodes of the "Person" type, only eight of them have collaborated in more than one instance. This result provides insights about which people have more influence on the instances, and further questions can be investigated afterward, *e.g.,* "Why among 51 people, only eight people have collaborated to more than one instance?", "Are the roles played by these people the same in the different instances?", "How have these people collaborated in each instance? Was this influence similar in both instances?". It is worth noticing that there is a bipartite subgraph composed of nodes 70, 44, 45, 146, and 139. Since 70, 44, and 45 are individuals and 146 and 139 are process instances, we can conclude that these three individuals always work as a group in those. In future projects, these three individuals could be allocated together.

(2) Q2 - *Which projects had more than one related entity?* The query in Cypher is presented following:

```
MATCH p=(u:Person)-[r:isRelatedTo]->(t:Project)-->()
WITH t, count(r) AS count
WHERE count > 1
MATCH x=(u)-[r:isRelatedTo]->(t:Project)-->()
RETURN x;
```

Figure 6 shows the query result. Is it possible to observe that from 40 instances of projects, only 14 present more than one entity related to it ("Person" and/or "Organization"). This result shows that most of the projects are small-scale ones, *i.e.*, they require a reduced number of individuals

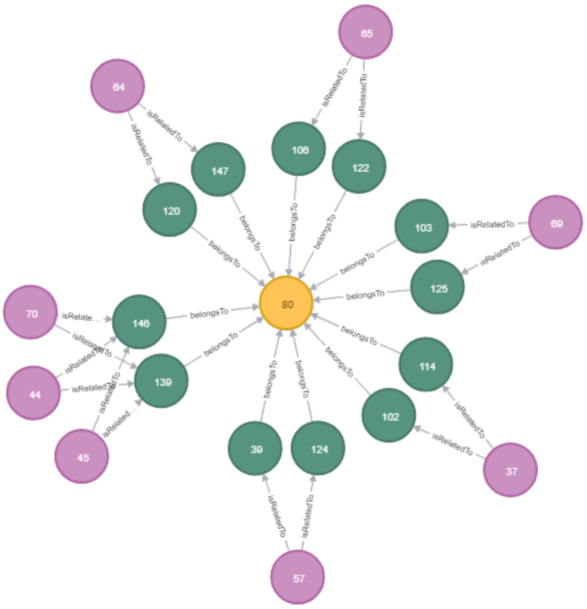


Fig. 5: Individuals who have contributed to more than one process instance.

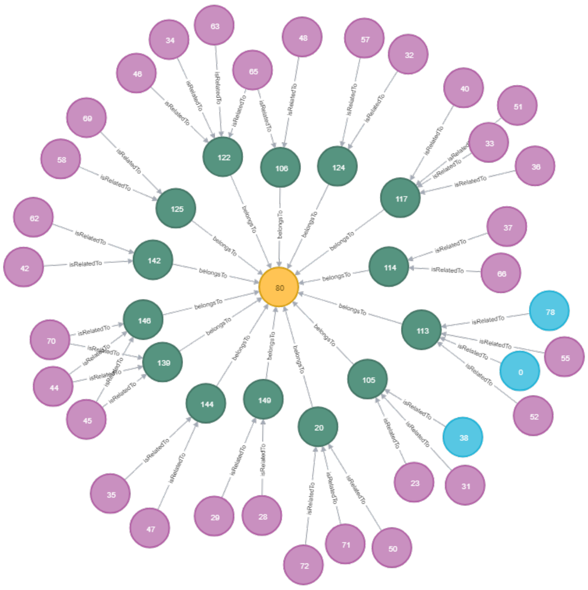


Fig. 6: Instances who had more than one related entity.

to run them. In addition, a few projects require more than two individuals or organizations (*e.g.*, 113 and 105), which possibly means that these projects require more attention from the company. Also, based on this resulting graph, further questions can be asked, *e.g.*, “Why do these instances have more than one related entity, while the majority has just one?”, “Why do some instances are related to organizations and others are not?”, “Is the complexity of an instance related to a higher amount of people related to it?”.

6. FINAL REMARKS

The approach proposed in this article, named MINERVA, focuses on analyzing provenance data from business process executions. Differently from the current mainstream, it considers a particular type of provenance data: multimodal provenance. The provenance of a business process can be found in heterogeneous forms: event logs, texts, audio, video, *etc.* Capturing and querying heterogeneous multimodal provenance data is an open, yet significant problem. To the best of the authors' knowledge, the approach presented in this article is the first attempt to process and integrate multimodal provenance data.

MINERVA was evaluated using data from the business processes of a real company named "dheka consultoria". In this evaluation, we created and analyzed a collaboration network generated based on the textual comments extracted from the Pipedrive system. This social network enables complementary analysis about the collaboration network considering important knowledge implicit in the business model. In the current version of MINERVA, the complex network created does not support relationships (or edges) between two entities. In future work, we plan to support this type of relation. As an example, in the sentence "Mary called John", besides the relations already created in the present version, a relationship between "Mary" and "John" will be created, and the edge weight will be influenced by the verb "call".

In the Experimental Evaluation section, after Q1 and Q2 are answered, we have mentioned questions that can be submitted in order to deepen the investigations performed. These questions were not answered in the present article. However, in future work, they can be answered using additional data and analysis, in order to aggregate more knowledge about the processes. Another advantage of the proposed approach is that the provenance model is W3C PROV compatible, *i.e.*, it can be exported from MINERVA and imported in any approach that follows the W3C provenance standard. The provenance model enables the origin of data to be investigated, as well as to track all the data transformation processes (*e.g.*, which contract was generated by each proposal and who wrote the contract). Additionally, the effective use of a graph database allows for the data to be analyzed from a complex network point of view. In future work, new case studies will be considered, especially the ones with audio and video.

REFERENCES

ABPMP Brazil. Bpm cbok v3. 0: Guia para o gerenciamento de processos de negócio-corpo comum de conhecimento, 3ª edição . ABPMP Brazil, 2011.

Ariouat, H., Hanachi, C., Andonoff, E., and Benaben, F. A conceptual framework for social business process management. *Procedia Computer Science* vol. 112, pp. 703–712, 2017.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. Complex networks: Structure and dynamics. *Physics reports* 424 (4-5): 175–308, 2006.

Cross, R. L., Cross, R. L., and Parker, A. *The hidden power of social networks: Understanding how work really gets done in organizations.* Harvard Business Press, 2004.

de Oliveira, W. M., Ocaña, K. A. C. S., de Oliveira, D., and Braganholo, V. Querying provenance along with external domain data using prolog. *J. Inf. Data Manag.* 8 (1): 3–18, 2017.

Euzenat, J. and Shvaiko, P. *Ontology Matching, Second Edition.* Springer, 2013.

Falci, M. L., Magalhães, A., Braganholo, V., Paes, A., and de Oliveira, D. Análise de colaboração em processos de negócio por meio de sgbds de grafos e dados de proveniência multimodais. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados.* SBC, Porto Alegre, RS, Brasil, pp. 169–174, 2020.

Ferreira, D. R. and Alves, C. Discovering user communities in large event logs. In *International Conference on Business Process Management.* Springer, pp. 123–134, 2011.

Freire, J., Koop, D., Santos, E., and Silva, C. T. Provenance for computational tasks: A survey. *Comput. Sci. Eng.* 10 (3): 11–21, 2008.

Groth, P. and Moreau, L. W3C PROV - An Overview of the PROV Family of Documents, 2013. W3C Working Group Note. Available at `https://www.w3.org/TR/prov-overview/`.

Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., and Sun, M. Opennre: An open and extensible toolkit for neural relation extraction. *arXiv preprint arXiv:1909.13078*, 2019.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.

Mendibil, K., Little, D., and Macbryde, J. Managing processes through teamwork. *Business Process Management Journal* vol. 8, pp. 338–350, 10, 2002.

Mestre, D. G. and Pires, C. E. Efficient entity matching over multiple data sources with mapreduce. *Journal of Information and Data Management* 5 (1): 40–40, 2014.

Miller, R. J. Open data integration. *Proc. VLDB Endow.* 11 (12): 2130–2139, 2018.

Neves, V. C., Oliveira, D. D., Ocaña, K. A. C. S., Braganholo, V., and Murta, L. Managing provenance of implicit data flows in scientific experiments. *ACM Trans. Internet Technol.* 17 (4), Aug., 2017.

Perez, B. C. *Analyzing and Modeling Complex Networks: Patterns, Paths and Probabilities.* Ph.D. thesis, University of Duisburg-Essen, Germany, 2020.

Reijers, H. A. Business process management: The evolution of a discipline. *Comput. Ind.* vol. 126, pp. 103404, 2021.

Ribeiro, L. A., Schneider, N. C., de Souza Inácio, A., Wagner, H. M., and von Wangenheim, A. Bridging database applications and declarative similarity matching. *Journal of Information and Data Management* 7 (3): 217–217, 2016.

Schönig, S., Cabanillas, C., Di Ciccio, C., Jablonski, S., and Mendling, J. Mining team compositions for collaborative work in business processes. *Software & Systems Modeling* 17 (2): 675–693, 2018.

Van Der Aalst, W. M., Reijers, H. A., and Song, M. Discovering social networks from event logs. *Computer Supported Cooperative Work (CSCW)* 14 (6): 549–593, 2005.

Zhao, W. and Zhao, X. Process mining from the organizational perspective. In *Foundations of intelligent systems.* Springer, pp. 701–708, 2014.