A case study on designing business processes based on collaborative and mining approaches

João Carlos de A.R. Gonçalves, Flávia Maria Santoro, Fernanda Araujo Baião

NP2Tec – Núcleo de Pesquisa e Prática em Tecnologia
Departamento de Informática Aplicada – DIA Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

{João.goncalves,flavia.santoro, fernanda.baiao}@uniriotec.br}

Abstract. *Companies invest a significant amount of time and resources to discover and represent how they work into business processes models. However, traditional process mapping has been done in an ad hoc manner and tends to be resource-intensive and time-consuming due to the informal and ambiguous collection of process information. The Story Mining Method aims to address those problems by the union of free-form narratives about processes and the usage of Text Mining and Natural Language techniques for text translation into process models. This paper presents a case study of the method, detailing its implementation as well as major issues found on a practical scenario within an organization.*

Resumo*. As organizações têm investido uma quantidade significativa de tempo e recursos para descobrir como trabalham e representar em modelos de processos de negócios. No entanto, a modelagem tradicional de processos é feita de forma ad hoc e tende a ser uma tarefa cara e demorada, devido à coleta informal e ambígua de informações. O método proposto neste artigo visa resolver estes problemas através de uma abordagem baseada em narrativas livres sobre processos e do uso de mineração de texto e técnicas de linguagem natural para a tradução de texto em modelos de processos. Este artigo apresenta um estudo de caso da aplicação deste método detalhando sua execução.*

1. Introduction

Process improvement initiatives start with the challenge of articulating existing (as-is) business processes. Process models are the basis for analyzing existing business processes in an organization and further getting them better. Furthermore, they also play an important role in bridging the business domain to the Information Technology (IT) domain, representing a fundamental tool for IT architecture planning and Service-Oriented Architecture implementation [Woodley and Gagnon, 2005].

Traditionally, companies invest a significant amount of time and resources to discover and represent how they work into business processes models. Very often, however, the outcome is not the one expected: process models present inaccuracies, and by the time companies complete this task, the processes have quite likely evolved, thus making the recently-obtained process models obsolete. It is therefore very important not to spend too much time to discover, analyze and represent the current state.

These problems are typically due to lack of perceived value, insufficient resources, faulty methodology, or inadequate tooling [Verner, 2004]. According to Wang et al.

(2009), challenges in business process discovery include the complexity of the enterprise and the interactions among its units, the inaccuracy and incompleteness of available business information, and the rate of changes of the enterprise business. Those authors affirm that traditional process mapping has been done in an ad hoc manner and tends to be resource-intensive and time-consuming due to the informal and ambiguous collection of process information [Alvarez, 2002].

Indeed, one could expect that as-is process are well-known in an enterprise. In practice, however, process knowledge is tacit – it exists in the minds of those individuals who actually participate in the process – and local - each participant has a local view of the process [Verner, 2004]. Therefore the author states that process discovery deals with transforming the organizational understanding of current business processes from tacit to explicit. The first step for process discovery is to identify individual process activities; the second step addresses the shape of the process, that is, its control flow (entry and exit points, sequential flow of activities, decision points, forks and joins). This information is essentially visual.

Other approaches for process discovery are based on applying formal methods and theories, such as linear programming, cost optimization, computational experiments, and probability theory, to generate process models. However, the use of these technologies does not suffice to reach a comprehensive business process representation. Xu et al. (2007) assert that without full communication with business participants and a methodological instruction for the integration of existing IT techniques, the result will lead to distortion. Thus, we argue that an adequate approach for process design needs to combine collaborative people interaction with computational techniques [Gonçalves et al., 2009]. This paper presents the first results of the application of such an approach, and a case study.

The paper is organized as follows: Section 2 describes the Story Mining method; and the supporting tools for it; Section 3 presents the case study performed at an organization and its results; Section 4 compares business processes discovery proposals and Section 5 concludes the paper.

2. The Story Mining Method

Our previous work [Gonçalves et al., 2009] proposed a method for business process elicitation based on Group Storytelling and using Text Mining and Natural Language Processing (NLP) Techniques for analyst support. In that work, business people involved in a process execution describe their way of acting through stories, in a collaborative way using the GroupStoryTelling tool [Leal et al., 2004]. Story Mining is composed of three phases that start from concrete facts told by participants, continuing towards achieving abstractions and classifications of these facts, and ending with a process model. Its second phase comprises a sequence of task mining tasks, with the generation of a proto-model of the process. Automatic generation of proto-models is an essential part of the method, as it gives a better understanding to the modeler of the “process knowledge” contained at the story repository, thus making process elicitation and modeling easier and trustworthy. In the third phase, the analyst refines and validates the model with the story tellers, creating the final model. Figure 1 depicts the Story Mining method.

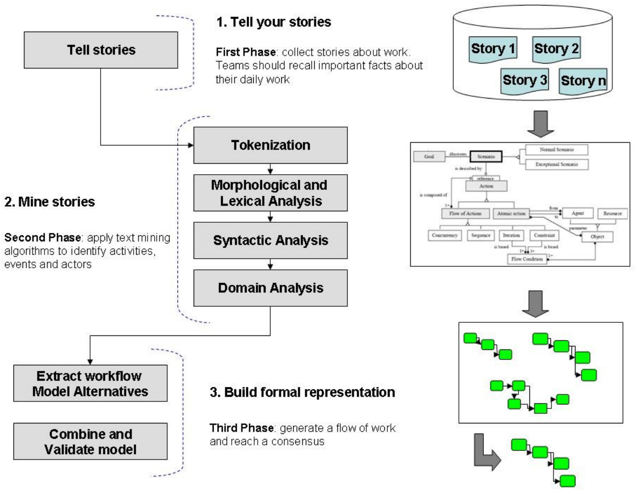

Tell stories
1. Tell your stories
First Phase: collect stories about work. Teams should recall important facts about their daily work
Story 1
Story 2
Story 3
Story n
Tokenization
Morphological and Lexical Analysis
Syntactic Analysis
Domain Analysis
2. Mine stories
Second Phase: apply text mining algorithms to identify activities, events and actors
Goal
illustrates
Scenario
Normal Scenario
Exceptional Scenario
is described by
reference
Action
is composed of
1+
Flow of Actions
Atomic action
from
to
Agent
Resource
parameter
Concurrency
Sequence
Iteration
Constraint
Object
is based
is based
1+
1+
1+
Flow Condition
Extract workflow Model Alternatives
Combine and Validate model
3. Build formal representation
Third Phase: generate a flow of work and reach a consensus

Figure 1. The Story Mining Method [Gonçalves et al., 2009]

The Story Mining method combines the free expression of knowledge of Group Storytelling approach together with the automatic extraction of process elements.

In [Gonçalves et al., 2009] we pointed to a wide array of techniques to be selected and applied at each different phase of the method. In this section we describe each phase in detail, and specify the techniques and algorithms used in the practical evaluation and case study presented in this paper.

"Tell your stories" Phase. At this initial phase, groups of tellers are selected based on their professional background and their involvement with specific tasks, processes and organizational structures within an organization.

Facilitators must be chosen, based on their relationship with the tellers and their skill on solving conflicts and problems that may appear during the collaborative storytelling process. Finally, the modelers must be defined, due to their expertise with Business Process notations and modeling experience.

"Mine stories" Phase. After the stories are told and loaded in the repository, the automatic extraction part begins. There are several steps required for conducting knowledge extraction (i.e., process models) from the texts of stories, in a process that is called Text Mining (TM) [Feldman and Sanger, 2007]. In fact, the TM process comprises several techniques from the NLP area, and its sequence of steps may be designed as a workflow. While designing our approach, we explored the TM algorithms implementations within the Biguá library [Oliveira, 2008] and the NLP functions within the NLTk framework [Bird and Loper, 2004].

Due to the great variety of techniques, algorithms and programs that may be applied for conducting each TM/NLP step, a scientific workflow management system (SWfMS) was needed. A SWfMS enabled the execution and comparative analysis of several instances of our workflow, in which we varied the techniques, programs and/or argument values set.

In our work, we adopted the VisTrails SWfMS [Callahan et al., 2006], due to its advanced functionalities for result visualization and comparison, as well as its native extensibility to invoke external legacy functions from the Bigua and NLTK repositories. Figure 2 illustrates our designed workflow on top of Vistrails interface.

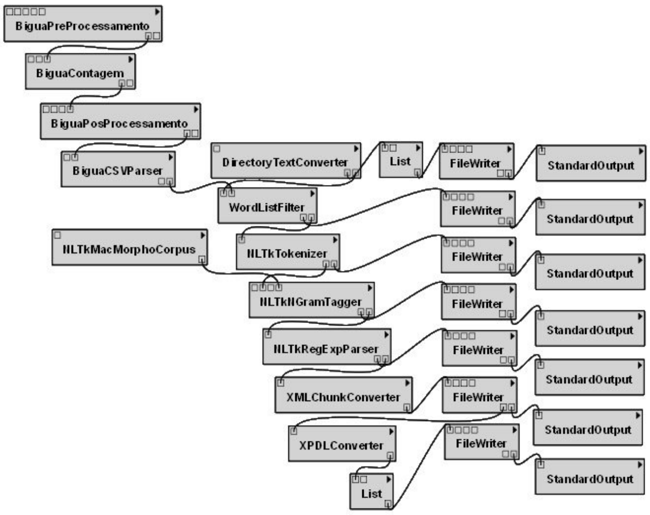


Figure 2. The Story Mining workflow

We shall cover each part of the TM workflow in the following sub-sections:

Tokenization

The aim here is to select which parts of the story's text are relevant for the process and extract them. First we use an external source of knowledge specific about the process theme, ranging from a simple list of words relevant to complex structures like ontologies.

At the present implementation, a simple list of relevant words acquired from documents related to the story was used, applying Text Mining techniques on them, in order to extract words based on their TF/IDF (Term Frequency/Inverse Document Frequency) value.

For the case study, we used all words extracted, regardless of frequency values. A stemming algorithm was applied to them, in order to improve its usage for the filtering of input text sentences.

The applied algorithms were proposed by the Bigua TM function library [Oliveira, 2008], such as the RSLP Stemmer [Orengo and Huyck, 2001 algorithm we have adopted for stemming Portuguese words extracted from the domain documents.Afterwards, the relevant story excerpts are processed and a list of its words and sentences is generated.

Morphological and Lexical Analysis

At this phase, the list of sentences previously extracted will be classified based on its morphological and lexical characteristics. A Trigram Tagger algorithm was used for this purpose, and a language specific tagged corpus of documents containing general texts, such as newspapers, magazines (The MAC-MORPHO Corpus [Aluisio et al., 2004]).

The tagger will perform an initial structuring of the extracted text, to be used in the next phases of the mining process.

Syntactic Analysis

The previous list of words and sentences is now tagged and classified according to their grammatical role. Using a NLP technique called Shallow Parsing [Osborne, 2000], specific patterns of tagged words are searched, and the words are further classified into

Noun Phrases, Verb Phrases, Sentences, and so on. A regular expression grammar was developed to aid this tagging task.

For process activity identification, this task focuses on sentences containing a Verb Phrase (VP) and its additional elements, like Noun Phrases as subjects and other constituents.

Domain Analysis

The output from the previous phase may be used for extracting a process model. For this goal, we need to establish correspondences between textual elements and process elements. In order to accomplish this, we have used the CREWS scenario metamodel [Achour, 1998], due to the fact that a scenario is defined as "a behavior limited to a possible set of interactions with a purpose, occurring between different actors. An analogy with a process model seems evident.

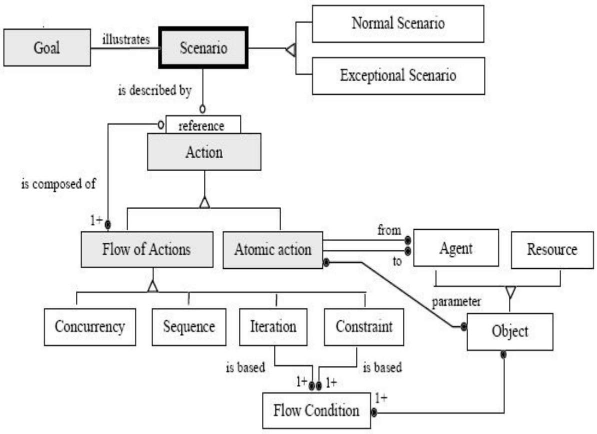


Figure 3. CREWS scenario metamodel [Achour, 1998] adapted for process model

We have conducted a practical application of the proposed method, focusing on Activities (Described as Actions and Flows of Actions in Figure 3) and Actors (described as Objects at Figure 3). Based on this analogy, the process elements are extracted from the structured text and two outputs are produced.

The first output produced is a structured text file containing the log of the extraction process. It shows each activity mined in a structured way, being the "raw data" for the refined proto-model generation.The second output is a BPMN proto-model [BPMN, 2008], generated from the first text file, using the XPDL schema. It aims at depicting the extracted knowledge for the modeler to visualize and use it properly.

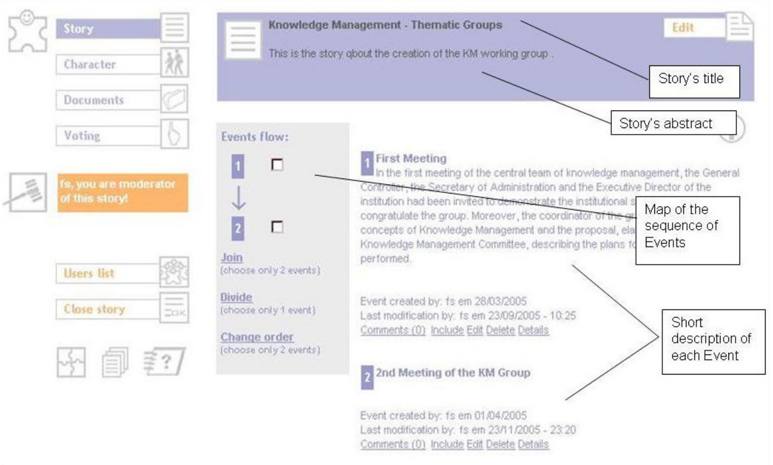
"Build formal representation" Phase. After the proto-model is generated in XPDL, the modeler can verify which process elements were discovered and can look at the proto-model as a "snapshot" of the specific process knowledge present at the story repository.

The proto-model can also be shown to the participants in order to present them the different process elements and workflow alternatives that were extracted from the story repository. The objective here is to reach a consensus about how the final process model should look like. The final process model can then be composed by the modeler, utilizing the proto-model and the tellers' remarks and commentaries.

3. Case Study: Undergraduate and Graduate Course Enrollment at DIA/UNIRIO

In this section a case study is described, performed at the Department of Applied Informatics of the Federal University of the State of Rio de Janeiro. The tellers were selected among the university faculty, staff and students of both undergraduate and graduate levels.

We used a version of the TellStory [Leal et al., 2004] collaborative tool, with some modifications specifically implemented to support the mining process, while allowing free expression of knowledge through Group Storytelling.


Knowledge Management - Thematic Groups

This is the story qbout the creation of the KM working group .

fs, you are moderator of this story!

1 First Meeting
In the first meeting of the central team of knowledge management, the General Controller, the Secretary of Administration and the Executive Director of the institution had been invited to demonstrate the institutional s congratulate the group. Moreover, the coordinator of the g concepts of Knowledge Management and the proposal, ela Knowledge Management Committee, describing the plans fo performed.

Event created by: fs em 28/03/2005
Last modification by: fs em 23/09/2005 - 10:25

2 2nd Meeting of the KM Group

Event created by: fs em 01/04/2005
Last modification by: fs em 23/11/2005 - 23:20

Figure 4. The TellStory tool for group storytelling

The selected group used the collaborative tool for about one month, telling their different views about the enrollment process through narratives, including commentaries on other tellers' story parts, uploaded relevant documents about the process, as well as reading the entire collaborative story text, looking for insights for its own parts.

After the storytelling phase, the main source of input text was selected, the story's events. Afterwards, the relevant documents were also downloaded from TellStory and used as the main input for the Text Mining workflow algorithms.

The final process model discovered through the StoryMining method was compared to a process model that was built using traditional interview and manual process design techniques. Table 1 summarizes some quantitative statistics of both models. PM1 was the Course Enrollment sections of the manually-created model, while PM2 was the one resulted from Story Mining. Figures 5 and 6 illustrate an excerpt of both process models. The Portuguese language was used for this case study, so the elements were translated to English.

The activities present at the figures below may be different, but they depict the same part of the process. While the manually-created model has a focus on general and broad activities, the other seems to portray detailed elements that can be used later for the design of models in greater detail.

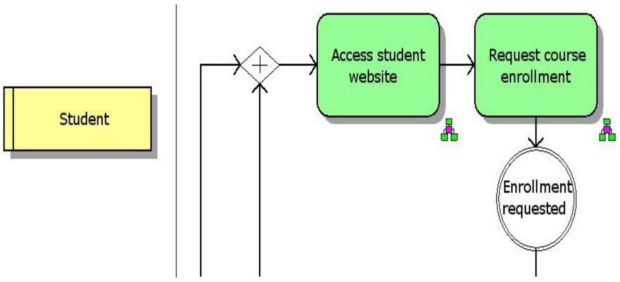


Figure 5. An excerpt of the manually-created process model

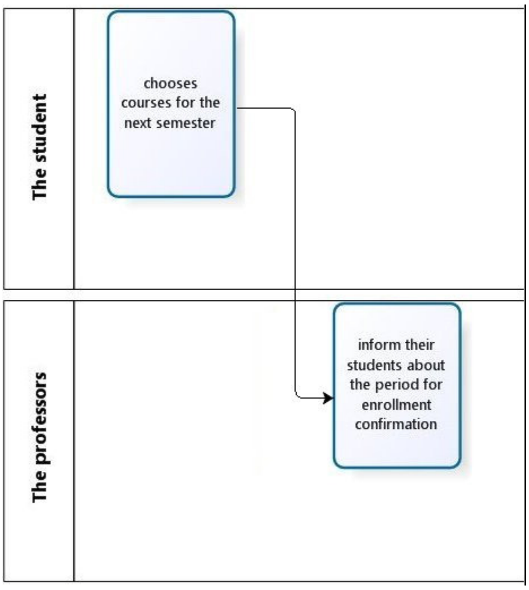

The student
chooses courses for the next semester
The professors
inform their students about the period for enrollment confirmation

Figure 6. The corresponding model excerpt discovered through Story Mining

The Story Mining method was able to automatically discover 8 out of the 21 process activities of the manually-created process model (for example, "Request course subscription"). We observed that the major part of the remaining 13 activities which were not discovered by Story Mining represented process activities that were not perceived by the users (for example, tasks involving interaction with information systems – such as "Inform availability of Course Enrollment Report" and administrative tasks – such as "File Enrollment Report"). Table I illustrates the results of the mining process.

TABLE I. Result comparison

Statistics	PM1	PM2
Total # of activities	21	51
# of coincident activities	8	21
# of non-coincident activities	13	30

PM2 also contained 30 activities discovered by StoryMining which were not detailed in the original process PM1. A further analysis on these activities showed that they represented real experiences lived by participants during previous process executions that were not explicit or documented (for example, "The special student may choose to apply for isolated courses" which represents situations where non-students (defined as "special students") request to enroll for an independent course.

An unexpected result during the case study was the richness of previously unknown activities of the process itself (ex: the "special students" part of the process) as well as elements, extracted as activities by the method, but later identified as business rules (Ex: "The period for course enrollment is defined each year")

In the next section, we present other research work on process discovery and compare with our approach.

4. Related Work

Process discovery has been addressed through human and automatic approaches. Human approaches are typically interviews or workshops conducted in order to collect relevant information from all the roles involved in the process execution. The problems with those techniques are that it is hard to capture the necessary detail about the process and it takes too long to hold the entire group. Even if the necessary detail is reached, it is hard to register it since there are a series of relationships among process elements, making it difficult to get them all together. Moreover, they are highly dependent on the abilities of the analysts who carry out the sessions [Hickey and Davis, 2004; Zowghi and Coulin, 2005].

Automatic approaches consist of computer-supported knowledge discovery techniques from information systems, called process mining [Aalst and Weijters, 2005]. Process mining has been focusing on discovery, i.e., deriving information about the original process model, the organizational context, and execution properties from enactment logs. Hence, process mining techniques work well on structured processes with little exceptional behavior and strong causal dependencies between the steps in the process [Aalst and Gunther, 2007].

However, the majority of real-life processes are not executed within rigid and inflexible behavior. "The most popular solutions for supporting processes do not enforce any defined behavior at all, but merely offer functionality like sharing data and passing messages between users and resources. Examples for these systems are ERP (Enterprise Resource Planning) and CSCW (Computer-Supported Cooperative Work) systems, custom-built solutions, or plain E-Mail" [Aalst and Gunther, 2007]. Process mining techniques for less structured environments need to come over with a high level view on the process, abstracting from details. As far as our proposal deals with unstructured information collected from participants of the process, we expect overcome at some level the problems discussed by [Aalst and Gunther, 2007].

Xu et al. (2007) propose a method that tries to address three aspects: (1) divide-and-conquer the complex enterprise business system to simplify the discovery problem; (2) harmonize the macroscopic and microscopic information provided by different business participants; (3) reduce the burden for incrementally updating and maintaining the discovery results due to the dynamic business evolution and provide a flexible verification approach to make these results trustworthy. The method is divided into three layers: the Component layer, the Operation Integration layer and the Operation layer, as shown in Figure 7.

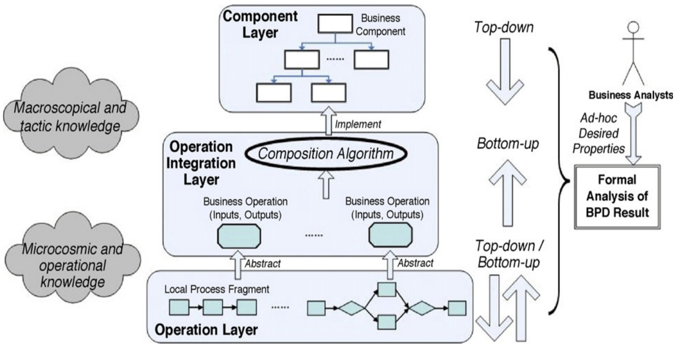

Component Layer
Business Component
Macroscopical and tactic knowledge
Implement
Operation Integration Layer
Composition Algorithm
Business Operation (Inputs, Outputs)
Business Operation (Inputs, Outputs)
Microcosmic and operational knowledge
Abstract
Abstract
Local Process Fragment
Operation Layer
Top-down
Bottom-up
Top-down / Bottom-up
Business Analysts
Ad-hoc Desired Properties
Formal Analysis of BPD Result

Figure 7. A three-layered method for business processes discovery [Xu et al., 2007]

The method proposed by [Xu et al., 2007] is concerned to a larger problem of identifying the components of processes in different levels of detail. Our proposal might fit as a support for the operation layer described by the authors.

Wang et al. (2009) propose a methodology called Policy-Driven Process Mapping (PDPM) for extracting process models from business policy documents. The application of PDPM is restricted to organizations with well-defined business policies, while our proposal may be applied in any context. Besides, PDPM cannot be fully automated because identifying and correcting syntactical and semantic errors and checking process completeness require human intervention and business expertise and the identification of process elements requires significant domain knowledge from the business analysts. The authors are currently investigating algorithms that can assist process analysts with policy analysis tasks using text mining techniques, which go in the direction of our proposal.

Recent studies [Ghose et al., 2007, Ingvaldsen et al., 2005; Ingvaldsen, 2006; Sinha et al., 2008] point to another approach for automatic process discovery. Instead of system logs, they focus on plain text process descriptions, relying on the application of Natural Language Processing and Text Mining techniques [Feldman and Sanger, 2007] for information extraction. This enables automatic process elicitation from documents such as interview reports, which commonly occur in organizations practice. We also intend to extend our proposal to include document and intranet mining in order to improve the models obtained.

5. Conclusions

This paper presents the practical implementation and experience with a process design method combining free-form narratives and automatic extraction of knowledge. The results presented show that Story Mining is useful, especially for the initial information gathering phase of process modeling, where little information, if any at all, is available about the process to be modeled.

The method was able to find a great amount of new activities of the process that were not present in the traditionally-created model, which represented tacit knowledge that was not recognized as a business formal rule or documentation. The approach of collaboratively telling stories and sharing experiences about a known situation in which all participants have already been involved in for several times probably helped in gathering more details about the process that was further extracted by StoryMining. The method can arguably broaden the range of its possible applications, making it able to improve already existing process models at an organization.

Further research will focus on the improvement of the mining process, with the application of advanced NLP techniques as Anaphora Resolvers, for example. Moreover, alternative forms of formal representation, different from the XPDL file, will be explored in order to improve the usage of the information extracted by the modeler and analyst.

References

AALST, W. M. P., GUNTHER, W., "Finding Structure in Unstructured Processes: The Case for Process Mining," acsd, pp.3-12, Seventh International Conference on Application of Concurrency to System Design (ACSD 2007), 2007.

AALST, W.M.P., WEIJTERS, A., “Process Mining” Process-Aware Information Systems: Bridging People e Software through Process Technology, Wiley & Sons, 2005.

ACHOUR, C. B., Guiding Scenario Authoring. 8th European-Japanese Conference on Information Modelling and Knowledge Bases, Finland, 1998.

ALUISIO, S., PINHEIRO, G.M., MANFRIM, A.M.P, OLIVEIRA, L. H. M., GENOVES JR, L. C., TAGNIN, S. E. O., “The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools” In: LREC 2004. Proceedings of LREC, 2004, Lisboa, Portugal, p. 1779-1782.

ALVAREZ, R.. Discourse Analysis of Requirements and Knowledge Elicitation Interviews. In: Proceedings of 35th Hawaii International Conference on System Sciences, 255, 2002.

BIRD, S., LOPER, E. “NLTk: The Natural Language Toolkit”; Proc.the ACL demonstration session, 2004.

BPMN – Business Process Modeling Notation, available from: http://www.bpmn.org/Documents. Last accessed on June 2008.

CALLAHAN, S. P., FREIRE, J., SANTOS, E, SCHEIDEGGER, C. E., SILVA, C. T., and VO, H.T., VisTrails: visualization meets data management” In: Proceedings of the 2006 ACM SIGMOD, New York, NY, 2006, 745-747

FELDMAN, R., SANGER, J., The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambrigde University Press, 2007.

GHOSE, A., KOLIADIS, G., CHUENG, A. “Process Discovery from Model and Text Artefacts”, in IEEE Congress on Services, 2007.

GONÇALVES, J. C. A. R., SANTORO, F. M., BAIÃO, F. A. “Business Process Mining from Group Stories”; Proc. 13th International Conference on Computer-Supported Cooperative Work in Design., Santiago, Chile, 2009.

HICKEY, A., DAVIS, A., “A unified model of Requirements Elicitation”. Journal of Management Information Systems 20 (4), 65-84, 2004.

INGVALDSEN, J.E., GULLA, J.A., SU, X., et al. “ A Text Mining Approach to Integrating Business Process Models and Governing Documents”, in Proceedings of the Workshop on Interorganizational Systems and Interoperability of Enterprise Software and Applications, 2005.

INGVALDSEN, J.E. “Information Retrieval for Organizational Process Insight”, in Proceedings of the CAISE’06 Doctoral Consortium, 2006.

LEAL, R. P., BORGES, M.R.B., SANTORO, F. M., Applying Group Storytelling in Knowledge Management. IX International Workshop on Groupware (CRIWG), Lecture Notes in Computer Science 3198, 34-41, 2004.

ORENGO, V. M., HUYCK, C. R.. “A Stemming Algorithm for the Portuguese Language” In 8th International Symposium on String Processing and Information Retrieval (SPIRE). Laguna de San Raphael, Chile, 2001.

OLIVEIRA, D., MiningFlow: Adding Semantics to Text Mining Workflows (in Portuguese), Master Dissertation, COPPE/UFRJ, Brazil, 2008.

OSBORNE, M. “Shallow parsing as part-of-speech tagging,” Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7, 2000.

SINHA, A., PARADKAR, A., KUMANAN, P., et al. “An Analysis Engine for Dependable Elicitation of Natural Language Use Case Description and its Application to Industrial Use Cases”, IBM Research Report RC24712, 2008.

VERNER, L., “The Challenge of Process Discovery”, BPTrends May, 2004.

XU, K., LIU, L., WU, C., “A three-layered method for business processes discovery and its application in manufacturing industry”. Computers in Industry 58 (2007) 265–278.

WANG, H.J., ZHAO, J.L., ZHANG, L.J., “Policy-Driven Process Mapping (PDPM): Discovering process models from business policies”. Decision Support Systems 48 (2009) 267–281.

WOODLEY, T., GAGNON, S., BPM and SOA: Synergies and Challenges, In Proceedings of 6th International Conference on Web Information Systems Engineering, LNCS, New York, NY, USA, 2005.

ZOWGHI, D., COULIN, C., “Requirements Elicitation: A Survey of Techniques, Approaches, and Tools”, in Engineering and Managing Software Requirements, edited by A. Aurum, C. Wohlin, Springer: USA, 2005.