UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Humberto Luiz de Oliveira Dalpra

PROV-Process: Proveniência de Dados Aplicada a Processos de Desenvolvimento de Software

Juiz de Fora
2016

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Humberto Luiz de Oliveira Dalpra

PROV-Process: Proveniência de Dados Aplicada a Processos de Desenvolvimento de Software

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientadora: Regina Maria Maciel Braga Villela
Coorientador: Victor Ströele de Andrade Menezes

Juiz de Fora
2016

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Dalpra, Humberto Luiz de Oliveira.

PROV-Process: Proveniência de Dados Aplicada a Processos de Desenvolvimento de Software / Humberto Luiz de Oliveira Dalpra. -- 2016.

158 f.

Orientadora: Regina Maria Maciel Braga Villela

Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2016.

1. Engenharia de Software. 2. Proveniência de Dados. 3. Processos de Software. I. Braga, Regina Maria Maciel, orient. II. PROV-Process: Proveniência de Dados Aplicada a Processos de Desenvolvimento de Software.

Humberto Luiz de Oliveira Dalpra

PROV-Process: Proveniência de Dados Aplicada a Processos de Desenvolvimento de Software

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Aprovada em 23 de agosto de 2016.

BANCA EXAMINADORA

_____

Prof.ª Regina Maria Maciel Braga Villela, D. Sc. - Orientador
Universidade Federal de Juiz de Fora

_____

Prof. Victor Ströele de Andrade Menezes, D. Sc. - Coorientador
Universidade Federal de Juiz de Fora

_____

Prof.ª Fernanda Cláudia Alves Campos, D. Sc.
Universidade Federal de Juiz de Fora

_____

Prof. Antônio Tadeu Azevedo Gomes, D. Sc.
Laboratório Nacional de Computação Científica

Juiz de Fora

2016

Aos meus pais e a minha esposa.

AGRADECIMENTOS

A Deus, por me conduzir pelos caminhos certos mesmo quando as adversidades superavam as perspectivas de um futuro melhor. Por me abençoar, me proteger, me ensinar e cuidar, zelosamente, de minha vida.

A minha esposa Gabriella, pela paciência, companheirismo, dedicação, incentivo e amor incondicional de todos os dias, mesmo nos momentos mais difíceis.

Aos meus tios Sérgio e Mônica, e ao meu primo (irmão mais novo) Victor, pelo incentivo, acolhimento, ensinamentos e amor com que sempre me trataram.

Aos meus pais, Sandra e Geraldo, por sempre acreditarem em meus sonhos.

Aos meus orientadores, Prof.ª Regina Braga e Prof. Victor Ströele, pela dedicação, paciência, incentivo e auxílio no decorrer de todo o curso de Mestrado.

Ao professor Marco Antônio pelas oportunidades e suporte dados a mim, desde a graduação, tanto profissionalmente quanto academicamente.

Aos membros da Banca Examinadora pelo trabalho de avaliação.

Aos amigos que fiz durante esta jornada, Tassio, Marcos e Welington, os quais foram fundamentais neste período. Pelo companheirismo, apoio, risadas e boas histórias que passamos juntos. #foreveralone

Aos alunos do NEnC, pela contribuição e auxílio sempre que requisitados. Em especial ao Leandro Simões, Weiner Oliveira, Philipe Marques, Guilherme Martins, Camila Paiva e Claudio Lelis.

Aos professores e colegas do Mestrado, por todos os ensinamentos.

À Universidade Federal de Juiz de Fora, pelo apoio financeiro através da bolsa de Mestrado.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

RESUMO

O processo de desenvolvimento de software pode ser definido como um conjunto de atividades, métodos, práticas e transformações utilizadas para desenvolver e manter o software e seus produtos associados. A descrição simplificada deste processo é denominada modelo de processo, no qual definem-se as atividades para o desenvolvimento do software, as especificações dos produtos de cada atividade e a indicação dos papéis das pessoas envolvidas. A execução destes processos gera dados importantes sobre o mesmo. A análise devida do histórico destes dados pode resultar na descoberta de informações importantes, as quais podem contribuir para o entendimento de todo o processo e, consequentemente, colaborar para a melhoria deste. A palavra proveniência refere-se a origem, fonte, procedência de um determinado objeto. Em termos computacionais, proveniência é um registro histórico da derivação dos dados que pode auxiliar no entendimento do dado e/ou registro atual. Este trabalho apresenta a proposta de uma arquitetura que, através do uso de modelos de proveniência de dados, aliado a um modelo ontológico e técnicas de mineração de dados, visa identificar melhorias nos processos de desenvolvimento de software e apresentá-las ao gerente de projetos por meio de uma ferramenta. Esta ferramenta, através da importação dos dados de execução de processos, alimenta um banco de dados relacional, modelado conforme a especificação de um modelo de proveniência de dados. Estes dados são carregados em um modelo de Ontologia e em um arquivo de mineração de dados. Assim, os dados são submetidos a uma máquina de inferência, no modelo ontológico, e também a análise de um algoritmo que integra regras de classificação e associação, na mineração de dados. O resultado desta análise apresenta indícios de pontos de melhorias no processo de desenvolvimento de software. A arquitetura proposta baseia-se em trabalhos relacionados, os quais foram selecionados a partir da execução de uma revisão sistemática.

Palavras-chave: Proveniência de dados, Ontologia, Mineração de Dados

ABSTRACT

The software development process can be defined as a set of activities, methods, practices and transformations used to develop and maintain the software and its related products. A simplified description of this process is called process model, which defines the activities for the development of software, product specifications of each activity and the indication of the roles of the people involved. The implementation of these processes generates important data on it. The proper analysis of the history of this data may result in the discovery of important information, which can contribute to the understanding of the process and therefore contribute to its improvement. The word provenance refers to the origin or source a particular object. In computer terms, provenance is a historical record of the derivation of data that can assist in the understanding of the data and / or the current record. This dissertation presents a proposal for an architecture that, through the use of data source models, combined with an ontological model and data mining techniques, aims to identify improvements in software development processes and present them to the project manager. This tool, by importing the process execution data, feeds a relational database, modeled based on a provenance model. These data are loaded into an ontology model and into a data mining file. Upon this loading, the data are processed by an inference machine, considering the ontological model, and also by an algorithm that integrates classification and association rules in data mining. The result of this analysis can presents points to improvements in the software development process. The proposed architecture is based on related work, which selected from the execution of a systematic review.

Keywords: *Provenance data, Ontology, Data Mining*

LISTA DE FIGURAS

Figura 1: Diagrama para modelar a proveniência prospectiva e a proveniência retrospectiva (LIM *et al.*, 2010) 22
Figura 2: Representação dos nós do OPM (Moreau *et al.*,2011). 24
Figura 3: Nós e dependências causais [adaptado de Moreau *et al.* (2011)]. 24
Figura 4: Representação dos nós do Modelo PROV. 26
Figura 5: Entidades e relacionamentos do PROV. 26
Figura 6: A organização do PROV (GROTH e MOREAU, 2013). 27
Figura 7: Modelo Definido (ou Declarado) (COSTA, 2015). 35
Figura 8: Modelo Inferido (COSTA, 2015). 36
Figura 9: Fluxograma PROV-Process. 54
Figura 10: Modelo gráfico do processo de software. 55
Figura 11: Fluxograma de processos. 57
Figura 12: Arquitetura PROV-Process. 58
Figura 13: DTR PROV-Process. 60
Figura 14: Classes e propriedades do PROV-O (BELHAJJAME *et al*. 2012). 61
Figura 15: Property chain wasAssociatedWith. 62
Figura 16: Ontologia PROV-Process. 63
Figura 17: Atividades que influenciaram a geração de outras atividades. 65
Figura 18: Agentes que influenciaram uma atividade. 66
Figura 19: Atividades e agentes manipulados pelo agente DotNet. 66
Figura 20: Atividade onde um artefato (entidade) foi consumido. 67
Figura 21: Arquivo data.arrf. 68
Figura 22: Lista de instâncias. 70
Figura 23: Lista de atividades, agentes e entidades de uma instância. 71
Figura 24: Detalhe atividade com inferências. 72
Figura 25: Visualização Ontologia. 73
Figura 26: Exibição dos dados de mineração. 74
Figura 27: Pesquisa ACM 124
Figura 28: Pesquisa ISI 125
Figura 29: Pesquisa Scopus 125
Figura 30: Pesquisa IEEE 126

Figura 31: Artigos retornados pela *string* de busca 129
Figura 32: Artigos selecionados após a eliminação de repetições e análise do título 131
Figura 33: Artigos selecionados após leitura do resumo dos artigos selecionados 132

LISTA DE TABELAS

Tabela 1: Especificação de documentos do PROV [Adaptado de (GROTH e MOREAU, 2013)] 28
Tabela 2: Comparativo OPM X PROV [Adaptado de BIVAR et al. (2013)]. 29
Tabela 3: Provenance from Log Files: a BigData Problem. 46
Tabela 4: Prime: A Methodology for Developing Provenance-Aware Applications. 47
Tabela 5: Provenance of Software Development Processes. 48
Tabela 6: Comparativo de características das propostas. 49
Tabela 7. Arquivo padrão – Parte 1 64
Tabela 8. Arquivo padrão – Parte 2 64
Tabela 9: Objetivo da avaliação 78
Tabela 10: Tipos de entrevistas [WOHLIN et al., 2012 apud SILVA, 2015] 80
Tabela 11: Resultados da avaliação 85
Tabela 12: Resultados da avaliação quantitativa do formulário I 94
Tabela 13: Resultados da avaliação quantitativa do formulário II 101
Tabela 14: Definição de palavras chave 121
Tabela 15: Total de resultados obtidos mediante a aplicação de cada filtro 128

LISTA DE ABREVIATURAS

API	*Application Programming Interface*
CBA	Classificação Baseada em Associações
CSV	*Comma-Separated Values*
GQM	*Goal/Question/Metric*
HTML	*Hypertext Markup Language*
IEEE	*Institute of Electrical and Electronics Engineers*
OPM	*Open Provenance Model*
OWL	*Web Ontology Language*
PICO	*Population, Intervention, Comparison, Outcome Measure*
RDF	*Resource Description Framework*
SGBD	Sistema de Gerenciamento de Banco de Dados
W3C	*World Wide Web Consortium*
XML	*eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO 16

1.1 MOTIVAÇÃO 17

1.2 PROBLEMA 18

1.3 HIPÓTESE 18

1.4 OBJETIVOS 19

1.5 ESTRUTURA DO TRABALHO 19

2 PRESSUPOSTOS TEÓRICOS 21

2.1 PROVENIÊNCIA DE DADOS 21

2.1.1 OPM – Open Provenance Model 23

2.1.2 PROV 25

2.1.3 Comparativo entre os Modelos OPM e PROV 29

2.2 PROCESSO DE SOFTWARE 30

2.2.1 Modelo de Processo 31

2.2.2 Execução de Processo 32

2.3 ONTOLOGIAS 33

2.3.1 Inferência 34

2.4 MINERAÇÃO DE DADOS 36

2.4.1 Regras de Associação 37

2.4.2 Regras de Classificação 39

2.4.3 Algoritmo CBA 40

2.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO 41

3 TRABALHOS RELACIONADOS 43

3.1 CONSIDERAÇÕES FINAIS DO CAPÍTULO 51

4 PROV-PROCESS 52

4.1 INTRODUÇÃO 52

4.2 ARQUITETURA 56

4.2.1 Base de Dados 58
4.2.2 Ontologia 59
4.2.3 Módulo Importador dos Dados de Execução 63
4.2.4 Módulo Gerador da Ontologia 64
4.2.5 Módulo para Mineração de Dados 67
4.2.6 Visualização dos Resultados 69
4.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO 75
5 AVALIAÇÃO 76
5.1 ESTUDO DE CASO 76
5.2 PLANEJAMENTO DO ESTUDO DE CASO 77
5.2.1 Objetivos 78
5.3 SELEÇÃO DOS INDIVÍDUOS 79
5.4 COLETA DE DADOS 80
5.5 MÉTRICAS 81
5.6 ANÁLISE DOS RESULTADOS 82
5.6.1 Análise do PROV-Process 82
5.6.2 Análise do processo das empresas 87
5.7 AMEAÇAS A VALIDADE 103
5.8 CONSIDERAÇÕES FINAIS DO CAPÍTULO 104
6 CONSIDERAÇÕES FINAIS 106
6.1 VISÃO GERAL 106
6.2 CONTRIBUIÇÕES 107
6.3 LIMITAÇÕES 108
6.4 TRABALHOS FUTUROS 109
REFERÊNCIAS 111
APÊNDICE I – REVISÃO SISTEMÁTICA 120
APÊNDICE II – FORMULÁRIO DE CARACTERIZAÇÃO DO PARTICIPANTE 144

APÊNDICE III – FORMULÁRIO I - AVALIAÇÃO EMPRESA I 146

APÊNDICE IV – FORMULÁRIO II - AVALIAÇÃO EMPRESA II 149

APÊNDICE V – FORMULÁRIO III - AVALIAÇÃO PROV-PROCESS 152

APÊNDICE VI – DETALHAMENTO DA BASE DE DADOS 156

1 INTRODUÇÃO

A gestão de dados está crescendo em complexidade, considerando seu uso em aplicações em grande escala, distribuídas e com dados heterogêneos que necessitam integração. Metadados, que descrevem os dados utilizados e gerados por estas aplicações, são essenciais para remover a ambiguidade dos dados e viabilizar a sua reutilização (SIMMHAN *et al.*, 2005a).

Proveniência de Dados é definida por Buneman (2001) como a descrição da origem de uma parte de um dado e o processo pelo qual este passou. A captura da proveniência pode ser feita considerando tanto os dados quanto os processos que deram origem a estes dados.

Um processo pode ser definido como uma abordagem sistemática para criar um produto ou realizar alguma tarefa (OSTERWEIL, 1987). Partindo deste princípio, a busca pela melhoria da qualidade de seus produtos tem motivado muitas organizações a investirem na definição e melhoria de seus processos. Apesar disto, o aumento do volume de dados gerados e utilizados por processos impacta diretamente na qualidade dos mesmos. Uma das formas para se verificar a qualidade de dados existentes é através da utilização de técnicas e modelos de proveniência de dados.

De acordo com Simmhan (2007), a proveniência de processos envolve a descrição das tarefas que fazem parte de um processo. Em determinados processos, para que se tenha informação dos dados geradores de um produto, é importante que registre-se cada dado consumido pelo processo. Na proveniência de um processo, é necessário capturar a descrição da execução de todas as tarefas efetivadas para que se tenha a informação relativa ao sucesso ou insucesso destas tarefas durante a execução. O uso de proveniência auxilia também na execução dos processos (MILES *et al.*, 2011), permitindo ao usuário refinar as regras de filtragem aplicadas para coleta dos dados (GHOSHAL *et al.*, 2013).

As duas principais características da proveniência de um conjunto de dados são o dado ancestral, a partir do qual os mesmos evoluíram, e o processo de transformação destes dados ancestrais, os quais resultaram nos dados atuais (SIMMHAN *et al.*, 2005b). Assim, por meio da proveniência de dados é possível certificar a qualidade dos dados com base nestes dados ancestrais e derivações. Neste sentido, o rastreamento de fontes de erros permite a reprodução automatizada de derivações para atualizar um conjunto de dados. Proveniência também é essencial para o domínio de negócios, podendo ser utilizada no detalhamento da fonte de dados

em um *Data Warehouse*, acompanhamento da criação de propriedade intelectual, entre outros (SIMMHAN *et al.*, 2005b).

Alguns domínios exigem que a proveniência seja armazenada em vários níveis de granularidade, o que acarreta a necessidade de adotar uma abordagem flexível pelo sistema de proveniência. A utilização de conjuntos de dados abstratos, que se referem a dados independentes de granularidade ou formato, fornecem essa flexibilidade. Isso faz com que a coleta de proveniência independa da qualidade ou de uma representação do conjunto de dados. O custo de coleta e representação da proveniência pode ser inverso à sua granularidade e isso vai desempenhar um papel importante em relação a granularidade (FOSTER *et al.*, 2003) (ZHAO *et al.*, 2004).

Para obtenção dos benefícios das informações de proveniência, é necessário a captura destas, bem como a modelagem e armazenamento de modo integrado para posterior consulta (MARINHO, 2011). Com base nesta necessidade, a principal contribuição deste trabalho consiste em propor “uma camada de proveniência que permita o armazenamento e a captura dos dados de execução dos processos. Estes dados serão utilizados na identificação de registros que possam apoiar as tarefas de modelagem, instanciação e melhoria de processos, por meio do uso da ontolgia e mineração de dados”. Nesta camada, tanto para o armazenamento quanto para a captura dos dados de proveniência será utilizado o modelo PROV (GROTH, 2013). O uso de ontologia e mineração de dados deve-se ao fato da distinção das informações buscadas por meio do uso de ambas, as quais tem por objetivo informações específicas acerca do processo de desenvolvimento de software. Embora exista a especificidade citada, as informações obtidas por meio da ontologia e mineração de dados, são complementares.

1.1 MOTIVAÇÃO

Barreto (2011) cita que, devido à realização de processos de software similares em uma mesma organização, existe a possibilidade de utilização da experiência adquirida durante a execução destes processos, visando definir melhores ações e políticas a serem adotadas em execuções futuras. Partindo desta afirmação, o desenvolvimento da presente dissertação tem como principal motivação a definição, criação e validação de uma abordagem que permita a análise dos dados utilizados em execuções passadas de processos de software (dados de proveniência de processos

de software), de forma a prover mecanismos que facilitem a tomada de decisão de gerentes de processos, auxiliando-os na identificação das melhores ações e políticas a serem adotadas para as próximas execuções de instâncias do processo.

Para a captura de dados de proveniência podem ser utilizados modelos de proveniência citados na literatura, tais como OPM (MOREAU *et al*., 2011) e PROV (GROTH e MOREAU, 2013). Acredita-se que a adaptação destes modelos para processos de software possa ser realizada para garantir o armazenamento/captura e posterior análise destes dados de execução dos processos. Para esta fase de análise dos dados, duas formas possíveis podem ser citadas, sendo elas o uso de Ontologias (e dos mecanismos de inferência oferecidos por elas) e de mineração de dados. O uso destas abordagens pode possibilitar a descoberta de informações sobre o processo executado, o que auxilia o gerente do processo na identificação por exemplo, de um menor tempo de execução ou uma maior eficiência dos resultados.

1.2 PROBLEMA

Partindo da motivação apresentada anteriormente, no contexto da análise dos dados de proveniência de processos de software, surge o problema tratado neste trabalho: a necessidade de obter e analisar informações implícitas a respeito de execuções anteriores de processos de software, de forma a permitir uma melhor análise do processo de software que é realizado por uma organização e facilitar a tomada de decisão de gerentes de processos, no sentido de melhorar a mesma em execuções posteriores.

1.3 HIPÓTESE

Com base no probelma mencionado anteriormente, tem-se como hipótese deste trabalho: *O armazenamento e posterior análise dos dados de proveniência de processos de software, utilizando um modelo de proveniência de dados, ontologias e técnicas de mineração de dados, é capaz de oferecer informações adicionais sobre o processo analisado, facilitando a tomada de decisão por parte do gerente de projeto sobre ações de melhoria para a execução das próximas instâncias do mesmo*.

1.4 OBJETIVOS

Este trabalho tem como objetivo principal definir, implementar e avaliar uma abordagem composta por uma camada de proveniência, que permite o armazenamento dos dados de proveniência de processos, e a posterior análise dos mesmos através do uso de Ontologias e mineração de dados.

A fim de alcançar o objetivo proposto, este foi subdividido nos seguintes objetivos específicos:

1. Analisar as propostas existentes na literatura, que tratam sobre a utilização de proveniência de dados no contexto de processos de desenvolvimento de software;
2. Definição de modelos para a captura/armazenamento dos dados de proveniência de processos de software;
3. Definição das técnicas de análise dos dados de proveniência capturados sobre o processo de software em questão;
4. Desenvolvimento de ferramental para apoiar os dois objetivos anteriores;
5. Avaliação da abordagem proposta para verificação se esta atende ao objetivo principal citado anteriormente.

1.5 ESTRUTURA DO TRABALHO

O presente trabalho está organizado em 6 capítulos. No capítulo 2 são descritas as principais tecnologias e conceitos envolvidos na proposta, sendo apresentados dois dos principais modelos de proveniência, PROV (GROTH, 2013) e OPM (MOREAU, 2008), bem como as principais regras de mineração e a técnica a ser utilizada neste trabalho. Garijo e Gil (2014) afirmam que ambos os modelos de proveniência, descritos no capítulo 2, já foram utilizados com sucesso na captura de proveniência em diferentes sistemas e suas definições básicas são de domínio independente e extensível para acomodar fins específicos de aplicação. O capítulo 3 apresenta trabalhos descritos na literatura, similares à proposta deste. O capítulo 4 apresenta, em detalhes, a proposta deste trabalho, apresentando as principais funcionalidades, a aplicabilidade e a arquitetura com os principais componentes. Já o capítulo 5 apresenta a avaliação da proposta do

trabalho. Por fim, no capítulo 6 são apresentadas as considerações finais, explicitando o objetivo e as limitações da proposta.

2 PRESSUPOSTOS TEÓRICOS

Neste capítulo são descritas as principais tecnologias e conceitos utilizados na elaboração desta proposta, incluindo proveniência de dados, processo de software, Ontologias e mineração de dados.

2.1 PROVENIÊNCIA DE DADOS

Proveniência de dados é um registro da história da derivação dos dados, que possibilita a reprodutibilidade, interpretação dos resultados e diagnóstico de problemas (LIM *et al.*, 2010).

Existem vários contextos onde se pode descrever proveniência de dados. Buneman *et al.* (2001) define proveniência de dados, no contexto de sistemas de banco de dados, como a descrição da origem dos dados e do processo pelo qual este chegou à base de dados. Lanter (1991) refere-se à linhagem de produtos derivados em sistemas de informação geográfica (SIG) como a informação que descreve os materiais e transformações aplicadas para determinar os dados. Greenwood *et al.* (2003) expande a definição de proveniência de Lanter (1991) ao considerá-la como a gravação de metadados de processos de fluxos de trabalho de experimentos. Proveniência pode também ser associada não apenas com produtos de dados, mas com os processos que permitem a criação destes dados.

Lim *et al.* (2011), afirma que a proveniência pode ser capturada de forma prospectiva e retrospectiva. A forma prospectiva captura a especificação abstrata do workflow como uma receita para derivação de dados futuros. Já a proveniência retrospectiva captura a execução do workflow e as informações de derivação dos dados fornecem informações importantes para a análise de resultados.

A Figura 1 apresenta um modelo capturado por meio de um diagrama de entidades e relacionamentos, aplicado a fluxos de trabalho científicos. A parte esquerda corresponde ao modelo de proveniência prospectiva e a parte direita corresponde ao modelo de proveniência retrospectiva (LIM *et al.*, 2010).

O modelo da Figura 1 é uma evolução do modelo OPM proposto por Moreau *et al.* (2011), o qual prevê apenas a captura da proveniência retrospectiva. Alves (2013) utiliza a

modelagem de proveniência prospectiva para coletar informações da especificação de um *workflow*.

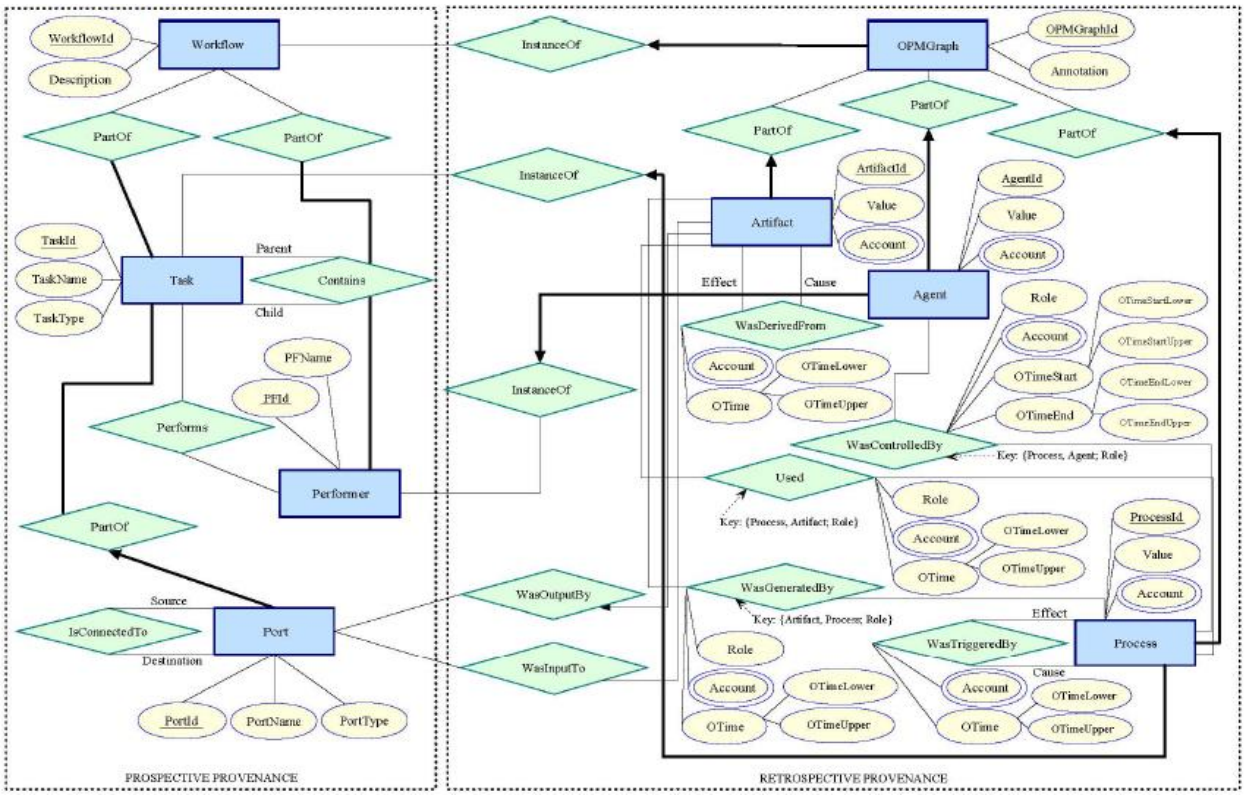


Figura 1: Diagrama para modelar a proveniência prospectiva e a proveniência retrospectiva (LIM *et al.*, 2010)

Em Moreau *et al.* (2011) é definida uma Ontologia OWL para capturar os conceitos do OPM e as inferências válidas neste modelo. Essa Ontologia é denominada *Open Provenance Model Ontology* (OPMO), e utiliza o OPMV (*Open Provenance Model Vocabulary*), um vocabulário leve, para descrever os conceitos fundamentais do modelo OPM. Este vocabulário não permite que inferências sejam realizadas, logo, permitir a completa expressividade dos conceitos OPM e inferências, são o objetivo da Ontologia OPMO. Porém, algumas regras e inferências não foram especificadas neste modelo, haja vista que tratam da junção da proveniência prospectiva e retrospectiva. Especificar essas regras de completude, considerando inclusive a proveniência retrospectiva juntamente com a prospectiva, para derivação de conhecimento implícito, são propostas no trabalho apresentado por Alves (2013).

A utilização da proveniência de dados conta com vários modelos propostos na literatura, sendo os dois principais o OPM (MOREAU, 2008) e o PROV (GROTH, 2013). Nas subseções a seguir (2.1.1 e 2.2.2) estes modelos são detalhados.

2.1.1 OPM – Open Provenance Model

Durante o '*Second Provenance Challenge*', em meio as discussões entre as equipes, compostas por treze integrantes, identificou-se que havia uma concordância sobre uma representação padrão para o núcleo (core) dos dados de proveniência. Como resultado, na sequência, em Agosto de 2007, em Salt Lake City, um modelo de dados foi elaborado pelos autores e liberado como 'Open Provenance Model' (OPM v1.00) (MOREAU *et al.*, 2008).

No dia 19 de junho de 2008, cerca de vinte participantes assistiram ao primeiro trabalho utilizando o OPM, realizado após a *International Provenance and Annotation Workshop* (IPAW'08), para discutir a especificação OPM. A ata da oficina e recomendações foram publicadas, e levaram à atual versão (v1.1) do modelo (MOREAU *et al.*, 2008).

De acordo com o OPM, gráficos, artefatos, processos e agentes são identificados por identificadores únicos e as arestas de dependência causais são identificados por suas fontes, destinos e papéis (para aqueles que têm papéis) (LIM *et al.*, 2010).

O OPM é um modelo projetado para atender aos seguintes requisitos (MOREAU *et al.*, 2011):

- Permitir que informações de procedência possam ser trocadas entre os sistemas por meio de uma camada de compatibilidade, com base em um modelo de proveniência compartilhado.
- Permitir que os desenvolvedores criem ferramentas que operem em tais ações de modelos de proveniência.
- Definir o modelo de uma maneira precisa à tecnologia a ser utilizada.
- Apoiar uma representação digital de proveniência para qualquer "coisa", produzidas, ou não, por sistemas de computador.
- Definir um conjunto de regras que identificam as inferências válidas e possam gerar gráficos de proveniência.

O OPM baseia-se em três tipos de nós, os quais podem ser visualizados na Figura 2, sendo detalhados abaixo:

Figura 2: Representação dos nós do OPM **(**Moreau *et al.*,2011**)**.

- Artefato: Parte imutável de estado, que pode ter um corpo físico em um objeto físico ou uma representação digital em um sistema de computador.
- Processo: Ação ou série de ações executadas ou causadas por artefatos, e que resultam em novos artefatos.
- Agente: Entidade Contextual agindo como um catalisador de um processo, permitindo, facilitando, controlando ou afetando sua execução.

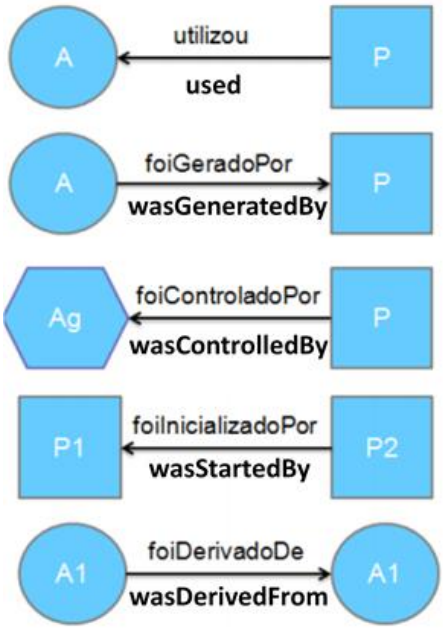


Figura 3: Nós e dependências causais [adaptado de Moreau *et al.* (2011)].

Conforme Moreau *et al.* (2011), para capturar as dependências entre os artefatos, processo e agentes, utiliza-se uma aresta, que representa a dependência causal entre a sua fonte, denotando o efeito e o seu destino, que denota a causa. O OPM possui 5 relações de dependência causal, sendo respectivamente *Used* (Utilizou), *wasGeneratedBy* (FoiGeradoPor), *wasControlledBy*

(FoiControladoPor), *wasStartedBy* (FoiInicializadoPor) e *wasDerivedFrom* (FoiDerivadoDe). A Figura 3 apresenta a representação dos nós e dependências causais (MOREAU *et al.*, 2011).

O modelo OPM é mais simples, orientado a captura de proveniência de processos, por isso considerado um modelo apenas de proveniência retrospectiva. Um modelo mais amplo, denominado PROV, permite a captura tanto de proveniência centrada em processo quanto centrada em entidade ou centrada em agente (ALVES, 2013).

2.1.2 PROV

O PROV (BELHAJJAME *et al*. 2012) é um modelo de proveniência especificado pelo W3C (W3C, 2015), que objetiva expressar a proveniência de dados através da descrição das entidades, atividades e agentes envolvidos em produzir, entregar ou enunciar um determinado objeto.

O PROV, assim como o OPM, utiliza um grafo para representar as informações de proveniência. O OPM foi projetado para ser aplicável a dados científicos ou coisas imateriais, tais como decisões. Por conta de ser uma especificação do W3C, tornou-se mais focado na web, ainda que os conceitos fundamentais sejam bastante gerais.

Assim, da necessidade de maior especificidade e menor nível de abstração surge o PROV, que traz a relação de derivação, que liga diretamente a entidade derivada ao seu antecessor (BELHAJJAME *et al*. 2013). Este novo modelo, mais específico, traz novas nomenclaturas a dois vértices e uma nova representação a um vértice, além de apresentar dez tipos de arestas distintas. O vértice representado por um círculo, no PROV é denominado Entidade, no OPM o mesmo é chamado Artefato. Já o vértice representado por um retângulo é denominado Atividade, no OPM o mesmo é chamado de Processo (BIVAR *et al.* 2013). O vértice representado por um hexágono, indicando um agente, no modelo OPM, é representado por um pentágono no modelo PROV (BELHAJJAME *et al*. 2012). Na Figura 4 é possível visualizar a representação dos termos no modelo.

Figura 4: Representação dos nós do Modelo PROV.

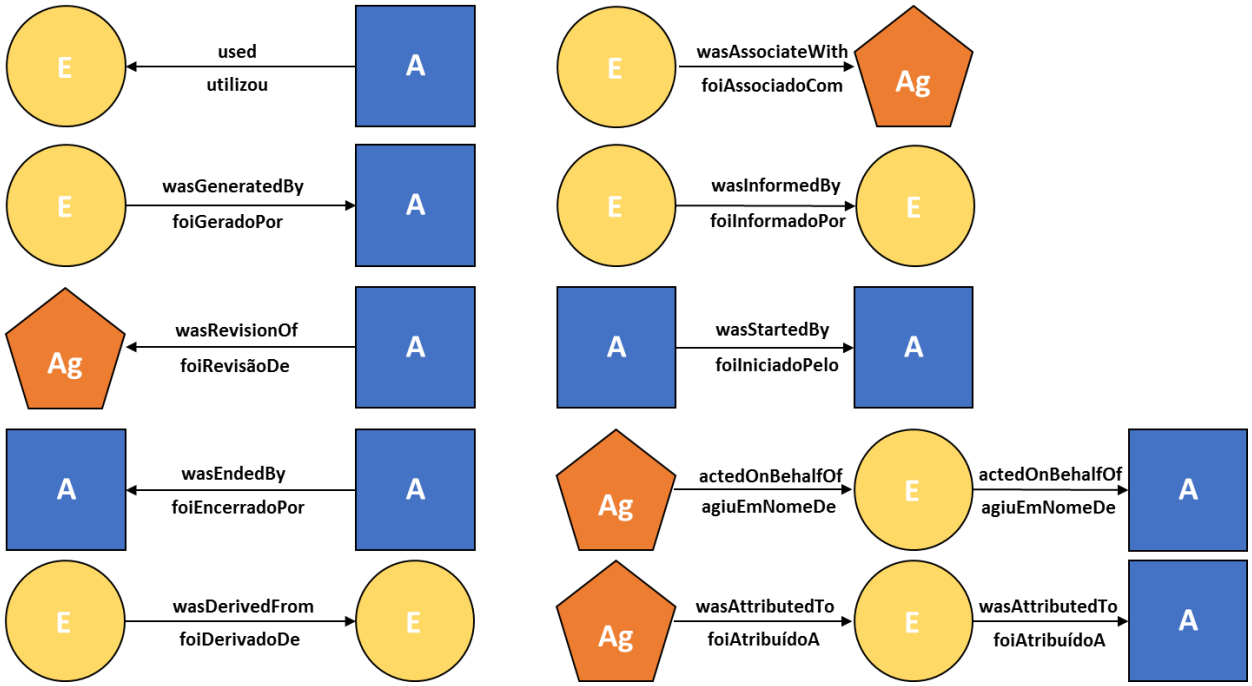


Figura 5: Entidades e relacionamentos do PROV.

As dependências causais no PROV, da mesma forma que no OPM, são representadas por arestas entre seus nós, as quais são direcionadas do efeito para a causa. Na Figura 5 é possível observar uma representação dos nós e dependências causais.

De acordo com Moreau e Missier (2013), as dependências causais do PROV são respectivamente:

- *used*: Relaciona atividades, afirmando que uma atividade usou outra atividade.
- *wasGeneratedBy*: Relaciona entidades a atividades e indica que uma entidade foi gerada por uma atividade.
- *wasAssociatedWith*: Relaciona atividades e agentes, indicando que uma determinada atividade foi associada a um determinado agente.

- *wasAttributedTo*: Relaciona entidades e agentes e indica que uma entidade foi atribuída a um determinado agente.
- *actedOnBehalfOf*: Relaciona agentes indicando que um agente tem autoridade ou responsabilidade por outro agente.
- *wasRevisionOf*: Relaciona entidades, registrando que uma entidade foi derivada de outra com o caráter corretivo, por exemplo, de correção de erro.
- *wasDerivedFrom*: Da mesma forma que a anterior, essa dependência causal relaciona entidades, no sentido de que uma entidade foi originada da outra. Esta derivação tem o caráter evolutivo, e não corretivo como a anterior.
- *wasInformedBy*: Relaciona atividades implicando que uma atividade informada foi gerada pela atividade que a informou, porém essa atividade é desconhecida ou não é de interesse.
- *wasStartedBy*: Relaciona atividades e entidades indicando que uma atividade iniciou uma entidade. A diferença entre esta dependência causal para *wasGeneratedBy* é que a *wasGeneratedBy* cria a entidade, o que implica em dizer que ela não era existente antes dessa relação ocorrer enquanto a *wasStartedBy* inicia uma atividade já existente previamente.
- *wasEndedBy*: Da mesma forma que a aresta anterior, relaciona atividades a entidades, registrando que uma atividade finalizou uma entidade.

A Figura 6 mostra a organização do PROV e como os documentos dependem uns dos outros. Na sequência, a Tabela 1 detalha a utilização dos documentos de acordo com o público específico.

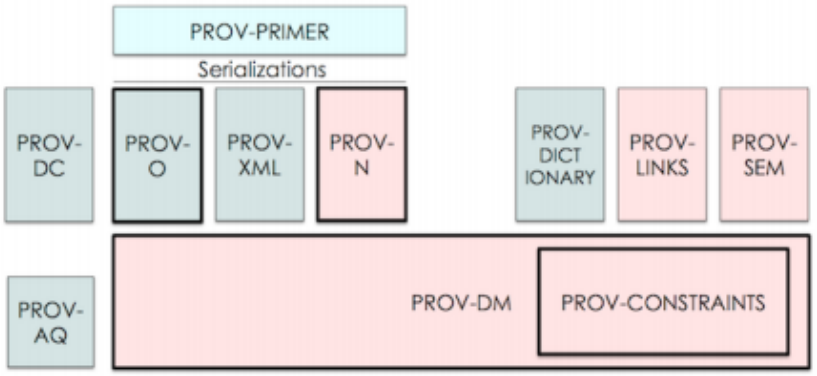


Figura 6: A organização do PROV (GROTH e MOREAU, 2013).

Tabela 1: Especificação de documentos do PROV [Adaptado de (GROTH e MOREAU, 2013)]

Parte	Público	Documento
1	Usuários	PROV-PRIMER é o ponto de entrada para PROV, oferecendo uma introdução para o modelo de dados de proveniência. Este é o lugar por onde começar, e para muitos pode ser o único documento necessário.
2	Desenvolvedores	PROV-O define uma Ontologia leve OWL2 para o modelo de proveniência de dados. Este destina-se ao *Linked Data* e Web Semântica.
3	Desenvolvedores	PROV-XML define um esquema XML para o modelo de dados de proveniência. Este é destinado a desenvolvedores que precisam de um XML nativo para serialização do modelo de dados PROV.
4	Avançado	PROV-DM define um modelo de dados conceitual de proveniência, incluindo diagramas UML. PROV-O, PROV-XML E PROV-N são serializações de modelo conceitual.
5	Avançado	PROV-N define uma notação legível para o modelo de proveniência. Isto é usado para fornecer exemplos dentro do modelo conceitual, bem como utilizado na definição de PROV-CONSTRAINTS.
6	Avançado	PROV-CONSTRAINTS define um conjunto de restrições sobre o modelo PROV, que especifica a noção de proveniência válida. São direcionados principalmente aos implementadores de validadores.
7	Desenvolvedores	PROV-AQ define como usar mecanismos baseados na Web para localizar e recuperar informações de proveniência.
8	Desenvolvedores	PROV-DC define um mapeamento entre Dublin Core e PROV-O.
9	Desenvolvedores	PROV-DICTIONARY define construções para expressar o dicionário de proveniência de estruturas de dados de estilo.
10	Avançado	PROV-SEM define uma especificação declarativa em termos de lógica de primeira ordem do modelo de dados PROV.
11	Avançado	PROV-LINKS define extensões para PROV para permitir a ligação de informação de proveniência através de faixas de descrições de proveniência.

2.1.3 Comparativo entre os Modelos OPM e PROV

Os modelos OPM e PROV possuem certas semelhanças, sendo possível, segundo BIVAR *et al.* (2013), fazer um mapeamento entre determinados conceitos chave de ambos os modelos. Na Tabela 2 são apresentadas as características, semelhanças e diferenças entre os respectivos modelos.

Tabela 2: Comparativo OPM X PROV [Adaptado de BIVAR et al. (2013)].

	OPM	**PROV**
Vértices	• Artefato • Processo • Agente	• Entidade • Atividade • Agente
Relacionamentos	5	10
	• *used* • *wasGeneratedBy* • *wasControlledBy* • *wasTriggeredBy* • *wasDerivedFrom*	• *used* • *wasGeneratedBy* • *wasAssociatedWith* • *wasAttributedTo* • *actedOnBehalfOf* • *wasRevisionOf* • *wasDerivedFrom* • *wasInformedBy* • *wasStartedBy* • *wasEndedBy*
	• *wasGeneratedBy:* liga processos a agentes	• *wasAssociatedWith:* liga atividades a agentes
	• *wasTriggeredBy: indica que um processo foi disparado por outro.*	• *wasInformedBy:* sua função é mostrar que uma atividade particular transmite alguma coisa para outra atividade.
		• *wasStartedBy:* equivalente ao *wasTriggeredBy*, porém pode ocorrer não somente entre duas atividades, mas também entre uma atividade e uma entidade, afirmando que a entidade é iniciada pela atividade.
	• Não mapeados	• *wasEndedBy:* finalização do processo

		• *actedOnBehalfOf:* relações de delegação e associação entre agentes, entidades e atividades. • *wasAttributedTo:* idem ao anterior.
Modelos	• Mais simples e mais orientado a captura de proveniência de processos; • Possui em sua documentação, regras de completude e inferências bem definidas.	• Mais amplo e permite a captura tanto de proveniência centrada em processo quanto centrada em entidade ou centrada em agente; • Ainda não estão definidas regras de completude e inferência como na OPMO.

Conforme pode-se observar na Tabela 2, o modelo PROV mostra-se mais detalhado quanto ao armazenamento dos dados de proveniência, dando ênfase nas responsabilidades dos agentes nos itens de proveniência, por meio de relações específicas para os mesmos, as quais não possuem equivalência no modelo OPM. Assim, por mostrar-se um modelo mais abrangente, o modelo PROV foi escolhido para ser utilizado no PROV-Process.

2.2 PROCESSO DE SOFTWARE

Processo de software pode ser definido como “um conjunto de atividades, métodos, práticas e transformações que as pessoas usam para desenvolver e manter software e seus produtos associados”, segundo a definição apresentada por Paulk (2009). Porém, existem outras diversas definições para processo de software, as quais são descritas a seguir, com o objetivo de auxiliar no entendimento deste conceito.

Fuggetta (2000) define processo de software como “um conjunto coerente de políticas, estruturas organizacionais, tecnologias, procedimentos e artefatos necessários para conceber, desenvolver, implantar e manter um produto de software”. Para Acuna *et al*. (2000), um processo de software pode ser definido como um conjunto de atividades necessárias para produzir um sistema de software, o qual é executado por um grupo de pessoas, organizadas de acordo com uma determinada estrutura organizacional, e que utiliza o apoio de ferramentas”. Já Sommerville

(2004) define processo de software como um conjunto de atividades e resultados associados que resultam em um produto de software.

Reis (2003) descreve o ciclo de vida de processos de software em 6 fases: provisão de tecnologia, análise de requisitos, projeto do processo, instanciação do processo, simulação do processo, execução do modelo de processo e avaliação do processo. O foco desta dissertação é na fase de execução e análise de processos.

Existem diversas ferramentas para execução dos processos. Estas ferramentas permitem a execução de processos de negócio como um todo, tendo em vista a similaridade dos ciclos de vida, e tendo como domínio o desenvolvimento de software. Como exemplos de ferramentas para apoiar a execução de processos, pode-se citar: Bonita Open Solution (BONITASOFT, 2014), Bizagi (BIZAGI, 2015), FLUIG (FLUIG, 2015), Intalio BPMS (INTALIO, 2014), eClarus Business Process Modeler for SOA Architects (ECLARUS, 2015). Estas ferramentas também apoiam a reexecução e auditoria de processos.

Desta forma, a captura dos dados de proveniência pode ser feita a partir do uso destas ferramentas e os dados capturados podem ser utilizados para a melhoria da reexecução e auditoria destes processos. No entanto, para que isso possa ser feito, é importante especificar o modelo de processo utilizado, além de detalhar os passos envolvidos na execução de processos.

2.2.1 Modelo de Processo

O modelo de processo é uma descrição simplificada do processo onde são definidas as atividades para o desenvolvimento do software, as especificações dos produtos de cada atividade e a indicação dos papéis das pessoas envolvidas (Sommerville, 2004).

Existem diversas notações para a modelagem de processos, onde pode-se citar como exemplo o BPMN (Business Process Model and Notation) (OMG, 2011) e o SPEM (Software & Systems Process Engineering Metamodel) (OMG, 2008).

O BPMN (*Business Process Modelling Notation*) (OMG, 2011) é uma notação padrão para modelagem de processos de negócios, com enfoque na análise de domínio e projeto de sistemas de alto nível. A notação herda elementos de uma série de notações anteriores, combinando à modelagem de processos de negócio, tais como a linguagem de definição de processo XML (XPDL) (WFMC, 2002) e o Diagramas de Atividades da UML (OMG, 2005).

Modelos de processos BPMN são compostos, principalmente por: (i) atividades, denotando eventos de negócios, (ii) itens de trabalho (realizados por seres humanos ou por aplicativos de software) e (iii) itens de controle, que gerenciam o fluxo de controle entre as atividades.

SPEM (*Software & Systems Process Engineering Metamodel*) (OMG, 2008) é um padrão proposto e mantido pelo Object Management Group (OMG). É um metamodelo baseado em MOF (Meta Object Facility) usado para especificar os processos de software. Além disso, define um perfil UML, a fim de fornecer um mecanismo para modelar processos com a linguagem UML.

2.2.2 Execução de Processo

Conforme já dito, a fase de execução de processos de software é realizada com o auxílio de ferramentas para guiar e assistir a realização do processo, onde o modelo do processo instanciado é a base. A execução de processos deve ser capaz de automatizar o processo, suportar o trabalho cooperativo, o monitoramento e o registro do histórico do processo (REIS, 2003).

A automatização do processo refere-se ao apoio à coordenação de atividades e ativação automática destas, as quais podem ser executadas sem intervenção humana. O suporte ao trabalho cooperativo indica apoio à cooperação de pessoas envolvidas em um projeto de software durante o ciclo de desenvolvimento. O suporte ao monitoramento provê diferentes visões do estado da execução do processo, possibilitando a obtenção de informação sobre o andamento das atividades, por parte do gerente de projetos. Já o registro do histórico do processo envolve a coleta de dados da evolução do processo, viabilizando a melhora do processo, caso seja necessário (REIS, 2003).

A construção do mecanismo de execução de processos deve ser embasada na semântica da linguagem que foi utilizada para a modelagem do processo, haja vista que a execução de um processo pode ser realizada a partir do momento em que se obtém um modelo de processo executável (ou instanciado), ou seja, que seja detalhado a ponto de permitir a sua execução por uma máquina (REIS ,2003).

A execução de um processo gera dados que registram as ações realizadas durante o mesmo. A captura e análise destes dados auxilia na identificação do que foi realizado durante a execução de um processo, possibilitando a detecção de problemas ou falhas que, por ventura,

possam ter ocorrido após a finalização deste. Estes registros podem ser utilizados pelos gerentes de projetos como forma de analisar as causas dos problemas, viabilizando a adoção de medidas que evitem a reincidência destes problemas em uma nova instância do processo de software modelado.

2.3 ONTOLOGIAS

Em Ciência da Computação, uma Ontologia define uma especificação formal e explícita de uma conceitualização compartilhada (GUARINO, 1998). Ela permite capturar o entendimento comum de objetos e seus relacionamentos em um determinado domínio. Ontologias fornecem um modelo formal e manipulável deste domínio (GRUBER, 1995). Sua utilização reúne benefícios como reuso, compartilhamento de conhecimento, portabilidade, manutenção e confiabilidade, partindo-se do princípio que elas representam uma conceituação compartilhada.

Com o uso de Ontologias, consegue-se estabelecer uma compreensão comum sobre objetos e os relacionamentos existentes entre eles em um determinado domínio, através de um modelo formal e manipulável. Além disso, a especificação formal do significado dos termos envolvidos na Ontologia possibilita a criação de novos termos, através da combinação dos já existentes (MATOS, 2008). Ontologias são uma das abordagens que podem ser utilizadas para a representação do conhecimento. Além disso, elas permitem que o conhecimento sobre um domínio possa ser lido e processado por um sistema computacional. Adicionalmente, há a possibilidade de utilização de máquinas de inferências (*reasoners*), que oferecem algoritmos através dos quais se consegue derivar novas informações e relacionamentos que anteriormente estavam implícitos na Ontologia inicial (FILHO *et al.,* 2012).

O uso de Ontologias na área de software popularizou-se com a Web Semântica (SOUZA e ALVARENGA, 2004). Assim, Ontologias passaram a ser usadas como uma taxonomia e um conjunto de regras de inferência, através das quais se torna possível capturar o conhecimento que não está explícito na taxonomia (BERNERS-LEE *et al.*, 2001). De acordo com o W3C, a linguagem padrão para definição de Ontologias é a OWL (*Web Ontology Language*) (OWL, 2004). Esta linguagem foi projetada para ser utilizada por aplicações que necessitam processar o conteúdo da informação ao invés de somente apresentar informações para os humanos. Esta

linguagem tem como objetivo suprir as necessidades de uma linguagem de Ontologia para a Web, além de resolver algumas limitações das linguagens anteriores, como XML e RDF.

Utilizando OWL consegue-se descrever classes, descrições de classes, propriedades e suas instâncias. Além disso, como esta linguagem é baseada em lógica descritiva, é possível a utilização de mecanismos de inferência, os quais permitem explicitar conhecimentos que estão implícitos em uma base de conhecimento. Assim, Ontologias descritas utilizando OWL não devem ser consideradas apenas sob o ponto de vista de sua sintaxe, mas também de sua semântica.

2.3.1 Inferência

Uma das características mais relevantes em uma Ontologia é a possibilidade de inferir novo conhecimento através do uso de máquinas de inferência (*reasoner*). Neste contexto, uma máquina de inferência consegue 'inferir', por exemplo, uma hierarquia de acordo com o que foi definido na Ontologia. Assim, a máquina de inferência pode ser utilizada para testar se uma determinada classe é uma subclasse de outra classe declarada na Ontologia. Outra funcionalidade que é oferecida pelas máquinas de inferência é a verificação de consistência da Ontologia. Com base na descrição (restrições) de uma classe, a máquina de inferência pode verificar se é ou não é possível para uma classe possuir quaisquer instâncias. Uma classe é considerada inconsistente quando a mesma não pode ter nenhuma instância. Este tipo de verificação é possível graças ao processamento das restrições relacionadas às classes que foram declaradas na Ontologia (HORRIDGE *et al.*, 2011).

Como exemplo de utilização do mecanismo de inferências, tem-se as figuras 7 e 8 abaixo, as quais apresentam uma pequena parte da Ontologia de Alinhamento de Sequências (*SequenceAlignmentOntology*), utilizada para a implementação da abordagem PL-Science (COSTA *et al., 2015*), outro trabalho do grupo de pesquisa no qual esta dissertação está relacionada. Na Figura 7, tem-se a hierarquia de classes conforme as mesmas foram declaradas (ou seja, o modelo definido). Nota-se nesta figura que as classes *ClustalW2* e *Clustal_Omega* são 'subclasses' da classe *bioinformatics_algoritms* e ocupam uma mesma posição hierarquica. Esta mesma situação ocorre com as classes *PHRAP*_View e *Consed*, e também com as classes *CAP3* e *PHRAP*. Após processar o algoritmo de inferência sobre a mesma Ontologia, que continha as

classes anteriormente citadas, consegue-se visualizar o que é exibido na Figura 8, ou seja, de acordo com as restrições existentes entre as classes, consegue-se visualizar que a classe *ClustalW2* passa a ser uma ‘subclasse’ de *Clustal_Omega* (Figura 8(a)). Já na Figura 8(b), tem-se que as classes *PHRAP*_View e *Consed* são semelhantes. O mesmo ocorre com as classes exibidas na Figura 8(c).

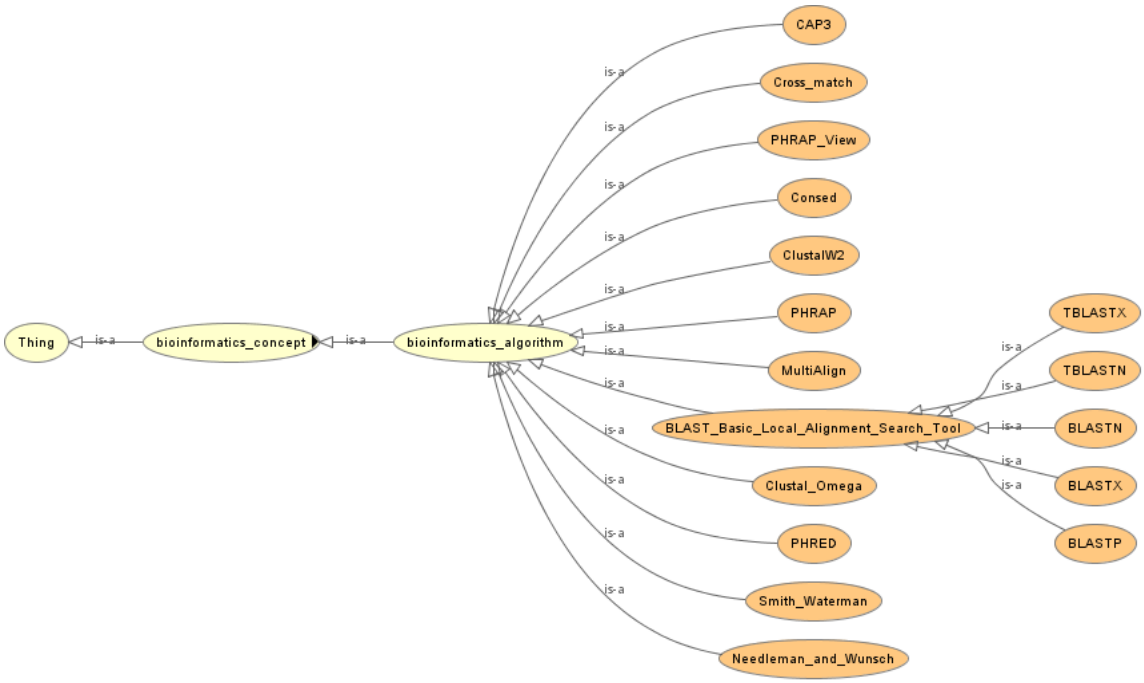


Figura 7: Modelo Definido (ou Declarado) (COSTA, 2015).

Como um exemplo das vantagens do uso de mecanismos de inferência (Figura 8(a)), no caso do domínio em questão, podemos citar a facilidade de obter a informação de que a classe *ClustalW2*, é uma 'subclasse' ou uma 'especificação' da classe *Clustal_Omega*. Isto significa que as instâncias relacionadas à classe *ClustalW2*, na Ontologia, possuem as características das instâncias relacionadas à classe *Clustal_Omega*.

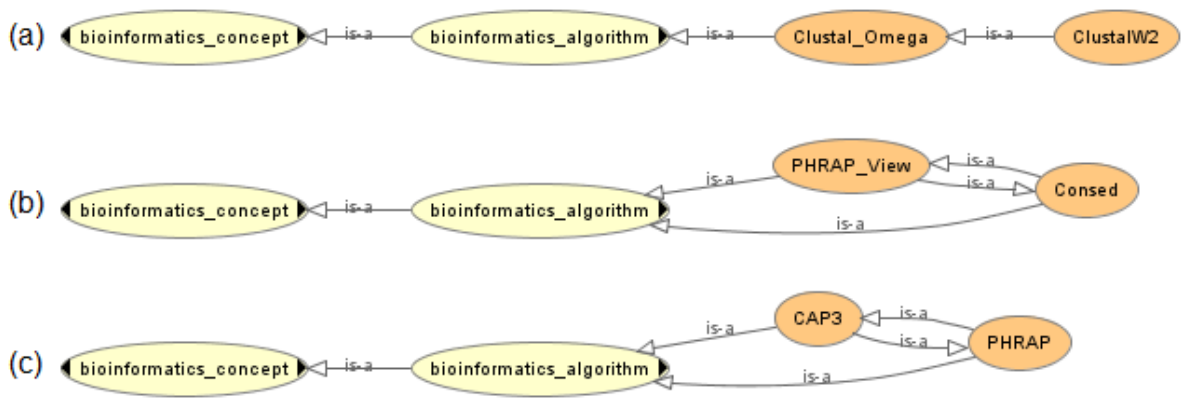


Figura 8: Modelo Inferido (COSTA, 2015).

O uso das máquinas de inferência, executadas a partir de um modelo ontológico alimentado com os dados de execução de processo, resulta na inferência de novo conhecimento acerca dos processos de desenvolvimento de software. Este novo conhecimento apresenta, aos gerentes de projetos, relações implícitas entre indivíduos (componentes, versões, etc.) dos processos de desenvolvimento de software. Nestas relações é possível identificar possíveis falhas no processo, cabendo aos gerentes de projetos, de posse das informações citadas, avaliar e definir a necessidade de adoção de medidas para melhoria do processo de desenvolvimento de software.

2.4 MINERAÇÃO DE DADOS

A mineração de dados (FAYYAD *et al.*, 1996) ou *data mining* é o processo através do qual busca-se extrair novo conhecimento de um conjunto de dados. Consiste na análise de grandes volumes de dados sob diferentes perspectivas, visando a descoberta de informações úteis, as quais, embora disponíveis, normalmente não são visíveis ou dificilmente são encontradas (VASCONCELOS e CARVALHO, 2004). Fayyad (1996) define a mineração de dados como um passo no processo de descoberta de conhecimento, o qual consiste na enumeração de padrões (ou modelos) sobre os dados.

A mineração de dados possui diversas funcionalidades que podem ser aplicadas a diferentes tipos de dados no intuito de extrair diferentes tipos de informações (CABENA *et al.*, 1997) (HAN *et al.*, 2011). Pode-se citar como exemplos de funcionalidades, agrupamento, classificação, regras de associação, dentre outros.

A classificação é o processo de encontrar um modelo (ou função) que descreve e distingue classes de dados. A finalidade dessa técnica é que o modelo seja capaz de prever a classe de objetos cujo rótulo da classe é desconhecido. O modelo é derivado com base na análise de um conjunto de dados de treinamento, ou seja, objetos de dados cuja classe é conhecida.

O modelo utilizado para a classificação e predição de novos objetos pode ser representado de diversas formas, como: regras de classificação se-então, árvores de decisão, redes neurais, máquinas de vetores suporte, dentre outros.

Regras de associação, funcionalidade que será adotada neste trabalho, buscam detectar conjuntos de itens frequentes que ocorrem de forma conjunta na base de dados e formam regras a partir destes conjuntos. Já as *regras de classificação* buscam solucionar o problema de classificação utilizando as regras de associação (ALVES, 2007).

Como os dados de execução de processo, utilizados na elaboração deste trabalho, podem ser separados em diferentes classes, pretende-se utilizar o algoritmo CBA (*Classification Based on Associations* - Classificação Baseada em Associações), o qual trata a classificação por regras de classificação.

Nas subseções a seguir, são detalhadas as regras de associação (subseção 2.4.1) e classificação (2.4.2), as quais compõem o algoritmo CBA. O algoritmo também é descrito na subseção sequente (2.4.3).

2.4.1 Regras de Associação

O principal objetivo das regras de associação é encontrar elementos que indiquem a presença de outros elementos em uma mesma transação. A premissa básica dessa funcionalidade da Mineração de Dados é encontrar padrões frequentes em um conjunto de dados (VASCONCELOS e CARVALHO, 2004).

Os *Padrões Frequentes*, como o nome sugere, são os padrões que ocorrem com frequência no conjunto de dados. Existem muitos tipos de padrões frequentes, incluindo *itemsets*, subsequências e subestruturas. Um *itemset* frequente se refere a um conjunto de itens que frequentemente aparecem juntos em um conjunto de dados transacional. Por exemplo, no conjunto de dados dos consumidores de uma padaria, provavelmente, será encontrado o padrão de compra do item leite e do item pão.

A *subsequência frequente* representa uma sequência de ações que se repetem com frequência no conjunto de dados. Como exemplo, no conjunto de dados de uma loja de informática, uma sequência de ações frequentes dos clientes poderia ser, em primeiro lugar, comprar um computador, seguido por uma câmera digital, e então um cartão de memória.

A *subestrutura* pode se referir a diferentes formas estruturais, tais como grafos e árvores, que podem ser combinadas com *itemsets* ou *subsequências*. Se uma subestrutura ocorre com frequência, é chamado de padrão (frequente) estruturado. Mineração de padrões frequentes leva à descoberta de associações e correlações interessantes que estão escondidas dentro de grandes bancos de dados.

Assim, as regras de associação representam padrões existentes em transações armazenadas, onde o termo transação refere-se aos itens consultados em uma determinada operação de busca. As *regras de associação* visam detectar os itens que ocorrem de forma conjunta na base de dados, definir a frequência de suas subsequências e definir regras a partir destes, as quais podem ser aplicadas, por exemplo, nas áreas de suporte a decisão, análise de dados de vendas e descoberta de tendências (ALVES, 2007).

Assim, por meio do uso de estratégias de mineração de dados, aplicadas a uma base de dados que registra itens adquiridos por clientes, pode-se gerar regras de associação tais como: {bermuda, camiseta}→{chinelo}, a qual indica, com determinado grau de certeza, que o cliente que compra bermuda e camiseta também compra chinelo. O grau de certeza mencionado é definido, em geral, pelo fator de suporte e o fator de confiança (VASCONCELOS e CARVALHO, 2004).

O fator de suporte de uma regra $X \Rightarrow Y$ é definido pela porcentagem de transações que incluem todos os itens do conjunto $X \cup Y$, onde cada item frequente é um conjunto de k itens. O percentual resultante representa a fração das transações que satisfazem tanto o antecedente (subconjunto X de Y tal que X tem $k - 1$ itens) quanto o consequente (item Y - X) da regra. O suporte de uma regra indica sua relevância (FILHO, 2010).

O fator de confiança de uma regra $X \Rightarrow Y$ é definido pela porcentagem de transações que incluem os itens X e Y em relação a todas que incluem os itens de X. Esta regra representa o grau de satisfatibilidade do consequente, em relação às transações que incluem o antecedente. A confiança indica a validade da regra (FILHO, 2010).

Em síntese, a confiança indica a probabilidade do segundo evento ocorrer dada a ocorrência do primeiro elemento. O suporte indica a frequência que os itens da regra aparecem juntos, considerando toda a base de dados. Normalmente, as regras de associação são descartadas por não serem consideradas interessantes, se não satisfazerem suporte e confiança mínimos.

É importante ressaltar que as técnicas de mineração de dados podem produzir centenas e até milhares de padrões ou regras, porém a grande maioria não é interessante. Geralmente, apenas uma pequena parcela das regras e padrões encontrados tem um valor interessante para um determinado usuário. Pode-se afirmar que para que um padrão seja interessante ele precisa ser de fácil compreensão para os humanos; ser válido quando aplicado a novos dados ou dados de testes, ou seja, precisa ter certo grau de confiança em novos dados; precisa ser potencialmente útil; e, por último, precisa ser novo, ou seja, padrões que já são conhecidos não são interessantes.

2.4.2 Regras de Classificação

Os sistemas de classificação englobam a descoberta de relações ou dependências entre variáveis de classe e outras variáveis, sendo estas relações utilizadas na classificação. As relações são armazenadas como modelos de classificação em forma de regras (BEEFERMAN, 1999), árvore de decisões (QUINLAN, 1993) ou formulações matemáticas.

Por ser reconhecido como um problema importante na descoberta de conhecimento implícito em base de dados, atraiu grande atenção dos pesquisadores da área de mineração de dados, tais como LIU *et. al*. (1998), MERETAKIS *et. al* (2000), XIAOXIN, JIAWEI (2003) e vem sendo utilizado na resolução de diversos problemas (XINDONG *et. al.*, 2014), (DELEN, 2014), (VIEIRA *et. al.*, 2012)

Os métodos para a construção de classificadores baseiam-se em heurísticas e algoritmos gulosos, onde pode-se citar como exemplo a árvore de decisão (QUINLAN, 1993) e técnicas estatísticas (LIM *et. al*, 2000). Por meio de um conjunto de regras, as quais codificam relações entre as variáveis de classe e outras variáveis, é possível provisionar novos casos.

As regras de associação vislumbram a obtenção de regras de associação de alta qualidade, que por sua vez serão utilizadas na construção de classificadores (DONG et. al, 1999). A metodologia é conhecida como classificação associativa e possui vantagens tais como lidar naturalmente com valores perdidos e "*outliers*" (ponto com comportamento diferente dos

demais), capturar todo o domínio das relações entre os itens na base de dados. Os classificadores desenvolvidos com base nesta metodologia são robustos e estudos de desempenho demonstram que tais classificadores são altamente precisos e eficientes (ALVES, 2007).

2.4.3 Algoritmo CBA

Para descrever o algoritmo CBA é necessário citar duas técnicas importantes de mineração, sendo respectivamente a mineração de regras de classificação e a mineração de regras de associação. A mineração de regras de classificação tem como objetivo a descoberta de um conjunto de regras, no banco de dados, para formar um classificador preciso, como por exemplo Quinlan (1993) e Breiman *et. al*(1984). Já a mineração de regras de associação encontra, no banco de dados, todas as regras que possuem suporte e/ou confiança mínimos, Agrawal e Srikant (1994). Na mineração de regras de associação, o objetivo não é pré-determinado, diferentemente da mineração de regras de classificação, onde existe apenas um alvo pré-determinado, a classe (LIU *et. al*, 1998).

Liu *et. al.* (1998) propõem a integração destas técnicas, com o objetivo de desenvolver um classificador mais preciso e que resolva problemas nos sistemas de classificação existentes. A integração mencionada foca na mineração de um subconjunto especial de regras de associação, denominado regras de associação de classes (CARs). O algoritmo proposto é chamado CBA (Classificação Baseada em Associações), o qual é composto por duas partes, um gerador de regra (CBA-RG), que baseia-se no algoritmo Apriori (AGRAWAL e SRIKANT 1994) para encontrar regras de associação, e um construtor de classificador (CBA-CB). Os resultados experimentais mostram que o classificador construído desta forma é, em geral, mais preciso do que os produzidos pelo sistema de classificação *state-of-the-art* C4.5 (QUINLAN, 1993).

As regras de classificação são definidas mediante a análise dos dados, as quais são apresentadas como: *Activity=Implementacao 1 2 Agent=VB6 4 1 Type_Activity=ErronoSistema 2 1 generated TipoDESD_name_Erro_no_Sistema_CNT=1 conf:(0.85), (27).*

Nesta regra, os números 1 e 2, constantes entre “Implementacao” e “Agent”, indicam que o valor do atributo Activity está na linha 1 coluna 2, ou seja, o valor "Implementacao" consta nesta posição. O mesmo vale para o atributo Agent=VB6 que está na linha 4 e coluna 1. O valor que é exibido depois de “conf”, refere-se à confiança dessa regra e o último valor entre parênteses trata-se da quantidade de transações que possuem todos esses itens juntos. Sendo assim, na regra

apresentada como exemplo, verifica-se que existem 27 transações na base de dados que possuem os itens *Activity*, *Agent*, *Type_Activity* e *TipoDESD_name_Erro_no_Sistema_CNT*=1.

Confiança é a probabilidade condicional de que as combinações de atributos resultem no parâmetro a ser analisado. A probabilidade dos atributos da regra aparecerem juntos é denominado suporte. No exemplo utilizado, a análise está sendo efetuada em relação a geração de desdobramentos de erro no sistema. A regra apresentada informa que em 85% das vezes em que uma atividade de implementação é executada por um agente VB6, esta refere-se a um desdobramento do tipo erro no sistema.

O cálculo do suporte é dado por SUP = FREQ(A, B, C)/N, onde FREQ(A, B, C) refere-se a quantidade de transações do conjunto de dados nas quais três atributos aparecem juntos e N é o número total de transações do conjunto de dados. Já o percentual de confiança apresentado é resultado do seguinte cálculo: CONF = FREQ(A, B, C)/FREQ(A, B).

Diferentemente da Ontologia que, no contexto deste trabalho permite a realização de inferências para descobertas de relações implícitas entre indivíduos (versões, módulos, integrantes de uma equipe, etc.), a opção pela mineração de dados deu-se pela possibilidade de analisar o percentual de vezes em que a combinação de atributos (módulos, membros de equipe, componentes, etc.) relacionam-se ao parâmetro de desdobramento de erro no sistema. Essa informação viabiliza ao gerente de projetos a indicação de padrões relativos ao tratamento de desdobramentos de erros no sistema, o que não seria possível apenas com o uso da Ontologia.

2.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou as principais tecnologias utilizadas na proposta apresentada nesta dissertação. Foram apresentados os conceitos e exemplos de modelos e execução de processo. Na sequência foram descritos os dois modelos de proveniência existentes, sendo apresentado na subseção seguinte um quadro comparativo entre os mesmos. Além disso, apresentou-se o conceito de Ontologia e a possibilidade de inferir novo conhecimento com base na elaboração de um modelo ontológico. Por fim, foi apresentado o conceito de mineração de dados, bem como uma breve descrição sobre as regras de classificação e associação, utilizadas na formação do algoritmo CBA, o qual será utilizado para minerar os dados obtidos junto as empresas parceiras,

como intuito de identificar informações relevantes para melhoria dos processos de desenvolvimento de software.

3 TRABALHOS RELACIONADOS

Para embasar o trabalho desenvolvido nesta dissertação, foi realizada uma revisão sistemática da literatura, com o objetivo de avaliar as evidências disponíveis sobre o tema proveniência de dados e execução de processos, e seguir uma sequência de passos metodológicos baseados em um protocolo bem definido. Este protocolo apresenta o tema a ser investigado de uma forma específica, pré-definida e com indagações estruturadas, bem como as instruções para a seleção, análise e resumo de trabalhos relevantes (STEINMACHER *et al*., 2013). Um detalhamento maior da revisão realizada é apresentado no Apêndice I.

Para a realização do estudo, foi executado o processo para apoio à condução de estudos baseado em revisão sistemática definido por MONTONI (2007). Este processo baseou-se em duas atividades: (1) Desenvolvimento do protocolo: definição de um protocolo de pesquisa que foi utilizado para conduzir o estudo; (2) Condução da pesquisa: com base no protocolo, o estudo foi conduzido e os resultados obtidos foram validados. Este estudo também pode ser considerado como um "mapeamento sistemático da literatura" (BUDGEN *et al*., 2008; PETERSEN *et al*., 2008) ou uma "*quasi* revisão sistemática" (TRAVASSOS *et al.*, 2008).

Podemos definir como objetivo do estudo realizado: a*través do* ***objeto de estudo*** *proveniência de dados e execução de processos, a* ***intenção/propósito*** *é identificar técnicas, abordagens, métodos, metodologias e ferramentas que tenham como* ***efeito*** *uma melhoria de processos através do uso de proveniência de dados,* ***do ponto de vista de*** *gerentes de projeto, no* ***contexto*** *de processos de desenvolvimento de software.*

Partindo-se deste objetivo, foram elaboradas as questões de pesquisa apresentadas a seguir, no contexto de processos de desenvolvimento de software:

Questão 1: Como a Proveniência de Dados pode ser utilizada na melhoria de processos, independente do domínio de aplicação?

Questão 2: Como as vantagens obtidas no uso da proveniência auxiliam a melhoria de processos?

Questões 3:

- Como a confiabilidade dos dados de proveniência pode ser avaliada?
- Como o volume de dados de Proveniência em Processos pode ser controlado?

Para a especificação do método de busca utilizou-se a abordagem PICO (*Population, Intervention, Comparison, Outcome Measure*) para organizar e estruturar a busca a ser realizada. Com base nesta abordagem, definiu-se:

- **População:** processos;
- **Intervenção:** proveniência de dados;
- **Comparação:** não se aplica;
- **Resultados:** técnicas, abordagens, métodos, metodologias e ferramentas que venham a auxiliar na melhoria de processos através do uso de proveniência de dados, apontando suas vantagens e apresentando formas de controle para avaliar a confiabilidade dos dados de proveniência.

Com a definição dos métodos de busca, foi elaborado o procedimento de seleção e critérios, com o objetivo de definir filtros a serem aplicados para selecionar somente as publicações relevantes ao estudo. O procedimento foi realizado em seis passos:

1. **Busca e Catalogação**: nesta etapa foram realizadas as buscas utilizando uma *string* e as principais bases de busca existentes. Os resultados retornados foram catalogados para análise posterior.

2. **Eliminação de artigos repetidos na mesma base (1° filtro):** para gerenciamento dos artigos retornados mediante a execução da *string* de busca, foi utilizada a ferramenta JabRef (JABREF, 2014), que foi escolhida por atender a critérios como: ser uma ferramenta de código aberto, funcionar em todos os sistemas operacionais, possibilitar importação de arquivos com extensões ‘bibtex’ e ‘txt’ e possuir exportação de bases nos formatos HTML, latex e csv. Para execução deste primeiro filtro foi realizada a importação dos resultados de cada base de forma separada, a fim de identificar repetições.

3. **Eliminação de artigos repetidos entre bases (2° filtro):** para diminuir a possibilidade de artigos repetidos, foi realizada a importação dos resultados de todas as bases de forma unificada, com o intuito de eliminar os artigos repetidos entre as bases.

4. **Seleção de artigos com base nos títulos (3° Filtro):** os trabalhos retornados após a aplicação do 3º filtro foram verificados com base no título, onde os que não se enquadravam no contexto da pesquisa foram eliminados. Como forma de evitar que a inclusão ou exclusão de um determinado artigo com base apenas no título fosse realizada apenas sobre a visão do autor deste trabalho, participaram deste 3º filtro 3 pesquisadores

(dois estudantes de mestrado e um de doutorado). Após avaliação dos títulos dos artigos retornados, dois dos pesquisadores posicionaram-se em relação à inclusão ou exclusão dos mesmos nas referências a serem analisadas. Em caso de empate, o terceiro pesquisador efetivava a avaliação do título e compartilhava sua opinião, e assim se definia a inclusão ou exclusão do artigo da seleção realizada. O fato de haver um número ímpar de pesquisadores envolvidos foi propositalmente determinado para evitar a situação de empate na avaliação realizada.

5. **Seleção de artigos com base na leitura dos resumos (4º Filtro):** como a busca realizada através das *strings* de busca criadas são restritas ao aspecto sintático, isto não garante que todas as publicações selecionadas no passo anterior são úteis para o propósito deste estudo. Partindo deste princípio, os resumos dos artigos selecionados através da análise do título foram lidos e analisados por dois dos três pesquisadores participantes da seleção anterior. Assim como na seleção anterior, o terceiro pesquisador só foi acionado para definição dos casos de empate, onde os pesquisadores participantes tenham opinado de forma divergente sobre a inserção ou deleção de determinado artigo. Para a aplicação deste filtro, foram utilizados os seguintes critérios de exclusão (CE):

- **CE1 –** Publicações que não tratavam de proveniência de dados relacionada a processos;
- **CE2 –** Publicações voltadas especificamente para *workflows* científicos;
- **CE3 –** Publicações não disponíveis para *download*, em sua forma completa, nas bibliotecas digitais, nem através de nenhum outro meio sem custos para o pesquisador;

De acordo com os critérios acima, as publicações obtidas nas máquinas de busca foram selecionadas nesta etapa apenas se NÃO se enquadrassem nestes critérios.

6. **Seleção das publicações relevantes com base na leitura completa do artigo (5º Filtro)**: Neste quinto e último filtro, as publicações selecionadas no passo anterior foram lidas por completo e analisadas usando os mesmos critérios de exclusão do passo 2, haja vista que a análise apenas dos resumos não garante que a publicação será útil para o estudo apresentado neste artigo.

Para nortear a execução da *string* de busca, a fim de verificar a qualidade dos artigos retornados por esta, foram definidos três artigos base, chamados de artigos de controle, os quais

respondem a algumas questões de pesquisa propostas na revisão sistemática e/ou relacionam-se ao que está sendo proposto neste trabalho. Estes são listados a seguir:

- GHOSHAL, D., PLALE, B. Provenance from log files: a BigData problem, Proceedings of the Joint EDBT/ICDT 2013 Workshops, Genoa, Italy, 2013.
- MILES, S., GROTH, P., MUNROE, S., MOREAU, L. PrIMe: A methodology for developing provenance-aware applications, ACM Transactions on Software Engineering and Methodology (TOSEM), v.20 n.3, p.1-42, August 2011.
- WENDEL, H., KUNDE, M., SCHREIBER, A. Provenance of software development processes. In: Deborah McGuinness, James Michaelis, and Luc Moreau, editors, Provenance and Annotation of Data and Processes, v. 6378 of Lecture Notes in Computer Science, p.59-63. Springer Berlin / Heidelberg, 2010.

A seguir serão apresentadas, de forma sintetizada, informações referentes aos artigos de controle. Nas tabelas 3, 4 e 5, são exibidos o título, autores e o ano de publicação do artigo, acompanhados de um resumo do trabalho e as respostas às questões de pesquisa, definidas na revisão sistemática realizada, caso possuam. No Anexo I encontra-se disponível parte da revisão sistemática realizada, a qual traz os dados de todas as publicações encontradas, bem como todos os passos seguidos para realização desta, desde a definição do protocolo de pesquisa até a execução do mesmo.

Tabela 3: *Provenance from Log Files: a BigData Problem.*

Dados da publicação	
Título	**Provenance from Log Files: a BigData Problem**
Autor(es)	**GHOSHAL, D., PLALE, B.**
Ano de publicação	**2013**
Resumo	
Os autores exploram a opção de obter proveniência dos arquivos de log existentes, uma abordagem que reduz a tarefa de instrumentação substancialmente, mas levanta questões sobre refinar enormes quantidades de informação para o que pode ou não ser proveniência. Os autores estudam a troca de facilidades de captura, integralidade de proveniência e mostram que em algumas circunstâncias, a captura através de registros pode resultar em proveniência de alta qualidade.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	

Na identificação de anomalias e erros.
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?
• **Não atrapalha a execução dos processos mediante a sua aplicação** • **Permite ao usuário refinar regras de filtragem** • **A captura de proveniência através de registros pode resultar em proveniência de alta qualidade**
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**
• **Captando nuances de ambiente de execução que poderiam influenciar os resultados de grande escala de análise de dados.** • **Utilizando um sistema baseado em regras, o qual facilita as tarefas de extração de proveniência, pois permite uma flexibilidade na maneira de selecionar informações relevantes e também dá o controle na gestão da granularidade de informações de proveniência.**

Tabela 4: *Prime: A Methodology for Developing Provenance-Aware Applications.*

Dados da publicação	
Título	**PrIMe: A Methodology for Developing Provenance-Aware Applications**
Autor(es)	**MILES, S., GROTH, P., MUNROE, S., MOREAU, L.**
Ano de publicação	**2011**
Resumo	
Os autores propõem uma técnica de engenharia de software, denominada Prime, para adaptar projetos de aplicações para que possam interagir com uma camada mediadora de proveniência, tornando-os, assim, voltados à proveniência. Os autores especificam as etapas envolvidas na aplicação de Prime, analisam a sua eficácia e ilustram seu uso com dois estudos de caso, bioinformática e medicina.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Prime é aplicado a um determinado tipo de caso de uso, questões de proveniência e tecnologias que ajudam a satisfazer os casos de uso, caso o projeto tenha uma forma particular.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Utilização de sistema separado de sua função primária, o qual trata processos e dados na aplicação. Respondendo a cada tipo de questão, proveniência pode ser	

vista como um caso de uso, ao invés de uma mudança no projeto para o benefício dos desenvolvedores.
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**
• **Possibilitando a compreensão de como os dados foram derivados até seu estado atual.** • **Não mencionado no artigo.**

Tabela 5: *Provenance of Software Development Processes.*

Dados da publicação	
Título	**Provenance of Software Development Processes**
Autor(es)	**WENDEL, H., KUNDE, M., SCHREIBER, A.**
Ano de publicação	**2010**
Resumo	
Os autores propõem uma solução para problemas relacionados a falhas nos processos de desenvolvimento de software, com base em tecnologias de proveniência.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado no artigo.** • **Não mencionado no artigo.**	

Como pôde-se observar, os artigos de controle respondem a algumas questões de pesquisa propostas na revisão sistemática e/ou relacionam-se ao que será proposto neste trabalho. A fim de relacionar as informações obtidas junto aos artigos de controle com a proposta deste trabalho, foi criada a Tabela 6, a qual traz uma comparação entre as principais características das respectivas publicações com o que é proposto neste trabalho.

Tabela 6: Comparativo de características das propostas.

Artigos	Proposta do artigo	Proposta deste trabalho
Provenance from Log Files: a BigData Problem	Obter proveniência dos arquivos de log existentes, trazendo questões sobre refinamento de grandes quantidades de informação para o que pode ou não ser proveniência.	A arquitetura proposta traz uma camada para armazenamento dos dados de proveniência. Nesta camada é necessário detectar os dados relevantes à proveniência, para que somente estes sejam armazenados, evitando o registro de informações não relevantes que culminam em um grande volume de dados desnecessários. Os dados armazenados serão utilizados na detecção de melhorias do processo de desenvolvimento de software.
PrIMe: A Methodology for Developing Provenance-Aware Applications	Uma técnica de engenharia de software, denominada Prime, para adaptar projetos de aplicações para que possam interagir com uma camada mediadora de proveniência, tornando-os, assim, voltados à proveniência.	Assim como proposto neste artigo de controle, a camada proposta nesta dissertação também visa mediar dados de proveniência, objetivando auxiliar na melhoria de processos. Essa melhoria é obtida mediante a apresentação ao gerente de projetos, de informações estratégicas do processo, utilizando inferências, obtidas com o uso de Ontologia, e análise dos dados por meio da mineração de dados.
Provenance of Software Development Processes	Solução para problemas relacionados a falhas nos processos de desenvolvimento de software, com base em tecnologias de proveniência.	Na proposta desta dissertação, a busca pela melhoria de processos através do uso da proveniência, enfoca, assim como o artigo de controle, os processos de desenvolvimento de software. O objetivo é que, através da proveniência, seja possível identificar falhas e propor soluções para melhoria dos processos de desenvolvimento de software.

Por meio da execução da revisão sistemática, apresentada nesta sessão, constata-se a viabilidade de desenvolvimento da proposta deste trabalho, haja vista que não foram encontrados na literatura, trabalhos que aliem tecnologias distintas para obtenção de conhecimentos estratégicos sobre o processo de desenvolvimento de software. Dentre as tecnologias mencionadas, englobam-se, além da proveniência de dados, o uso de mineração de dados e um modelo ontológico.

Gunther *et. al.* (2006), utiliza técnicas de mineração de dados em logs que armazenam registros referentes a alterações em processos, com o objetivo de obter informações relativas ao período e a causa que levam a necessidade de mudanças no processo. A proposta deste trabalho é a análise dos dados do processo de proveniência usando mecanismos de ontologias e de inferência, visando a melhoraria do processo de desenvolvimento de software, com base em métricas definidas anteriormente por Gunther *et. al.* (2006) e Gunther *et. al. (*2008*).*

O uso da inferência da ontologia para monitorar as atividades de negócios já foi investigado por Pedrinaci *et. al.* (2008). Os mesmos propõem uma ferramenta, denominada Sentinel, a qual é baseada em tecnologias semânticas, incluindo ontologia para métricas e ferramentas para computação e análise dessas métricas. Este trabalho propõe ainda o uso de ontologias, considerando que, através do uso dos mecanismos de inferência oferecidos pela mesma, pode-se encontrar informações implícitas nos dados de proveniência de processos de desenvolvimento software, como, por exemplo, as relações implícitas entre usuários e artefatos manipulados no processo.

O próximo capítulo apresenta a abordagem desenvolvida e a ferramenta de apoio. Nesta será possível verificar que o uso das máquinas de inferência da Ontologia, aliada a técnicas de mineração de dados, aplicadas a dados de execução dos processos obtidos por uma camada de proveniência, possibilita a apresentação, ao gerente de projetos, de informações sobre a execução dos processos de desenvolvimento de software, as quais auxiliam na tomada de decisão para melhoria dos mesmos. Estas informações advêm de conhecimento estratégico obtido por meio da Ontologia, e também da identificação de padrões dos dados, através do uso de técnicas de mineração de dados.

3.1 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou uma revisão sistemática da literatura, na qual foi possível identificar evidências disponíveis sobre o tema proveniência de dados, por meio dos trabalhos publicados nesta área. Através das questões de pesquisa, foi possível a identificação de publicações que abordam a utilização da proveniência de dados e de mineração de dados na melhoria de processos, suas vantagens e a forma de avaliar a confiabilidade destes dados.

4 PROV-PROCESS

4.1 INTRODUÇÃO

Neste capítulo é apresentada a arquitetura PROV-Process, que tem como meta prover ao gerente de projetos sugestões de melhorias para o processo de desenvolvimento de software. A PROV-Process é composta por uma camada de proveniência, estendida do modelo PROV (GROTH, 2013), uma Ontologia, estendida do PROV-O (LEBO, 2013), e técnicas de mineração de dados.

O objetivo é prover, através do uso de uma camada de proveniência, a identificação de possibilidades de alteração no processo, trazendo melhorias para este como um todo. Abaixo estão listadas as principais contribuições da proposta:

- Para armazenamento destes dados, foi desenvolvido um modelo, estendido do modelo PROV (GROTH, 2013), que é específico para armazenamento de dados de execução de processos de desenvolvimento de software;
- A Ontologia PROV-O (LEBO, 2013) foi também estendida, gerando uma nova Ontologia com novas classes, propriedades e regras específicas para tratar dados de proveniência de processos de software. Essa Ontologia, denominada PROV-Process é alimentada com os dados armazenados no banco de dados do PROV-Process;

Na sequência, estão dispostos os passos para utilização da proposta deste trabalho:

- Através dos registros gerados na execução do processo, são capturados os dados de proveniência, utilizando o modelo PROV;
- Aos dados obtidos aplicam-se tecnologias de web semântica, possibilitando deduções sobre os registros coletados e obtendo indicações de melhorias para os processos em uso, através do uso de Ontologias e máquinas de inferência;
- Em paralelo ao uso de tecnologias de web semântica, utiliza-se mineração de dados, a fim de identificar outras informações relevantes não explícitas nos dados e que possam auxiliar na melhoria do processo;
- Foi utilizada uma API (*Application Programming Interface*) para execução da etapa de mineração de dados diretamente pelo sistema PROV-Process, a qual utiliza o algoritmo CBA (LIU *et. al*, 1998);

- Foi desenvolvida também uma aplicação web para utilização das tecnologias mencionadas, com o objetivo de possibilitar a utilização destas informações pelo gerente de projetos.

A utilização da proveniência de dados conta com vários modelos propostos na literatura (MISSIER *et. al*, 2013, MOREAU *et. al*, 2008, MOREAU e MISSIER, 2013). Alves (2013), com base em Lim *et al.* (2010), desenvolveu um modelo que engloba a proveniência prospectiva e retrospectiva como uma extensão do OPM (MOREAU, 2008). Nosso trabalho adota uma abordagem semelhante. No entanto utilizamos o modelo PROV (GROTH, 2013).

A Figura 9 apresenta o fluxograma de funcionamento da PROV-Process. De acordo com este fluxograma, os dados referentes à execução de processos de desenvolvimento de software devem ser exportados para um arquivo em um padrão previamente estabelecido, conforme representado pelas duas imagens exibidas no início do fluxograma. O padrão para este arquivo foi especificado de acordo com as informações do processo, necessárias ao modelo de proveniência. Os dados serão importados para o banco de dados, modelado de acordo com as especificações do PROV (GROTH, 2013).

Duas abordagens são utilizadas para derivação de informações estratégicas a partir dos dados armazenados: Ontologia e mineração de dados. A Ontologia, estendida do PROV-O (LEBO, 2013), está representada no fluxograma por um arquivo OWL. Na Ontologia, podem ser executadas inferências para obtenção de informações, bem como indicações de melhorias para o processo vigente. Em paralelo ao uso da Ontologia, podem ser utilizadas técnicas de mineração de dados, enriquecendo a gama de sugestões de melhoria. Por fim, as sugestões de melhoria do processo de desenvolvimento de software serão apresentadas ao gerente de projetos, o qual, mediante a análise das mesmas, pode optar pela implantação ou não da sugestão, a qual visa melhorar o processo como um todo.

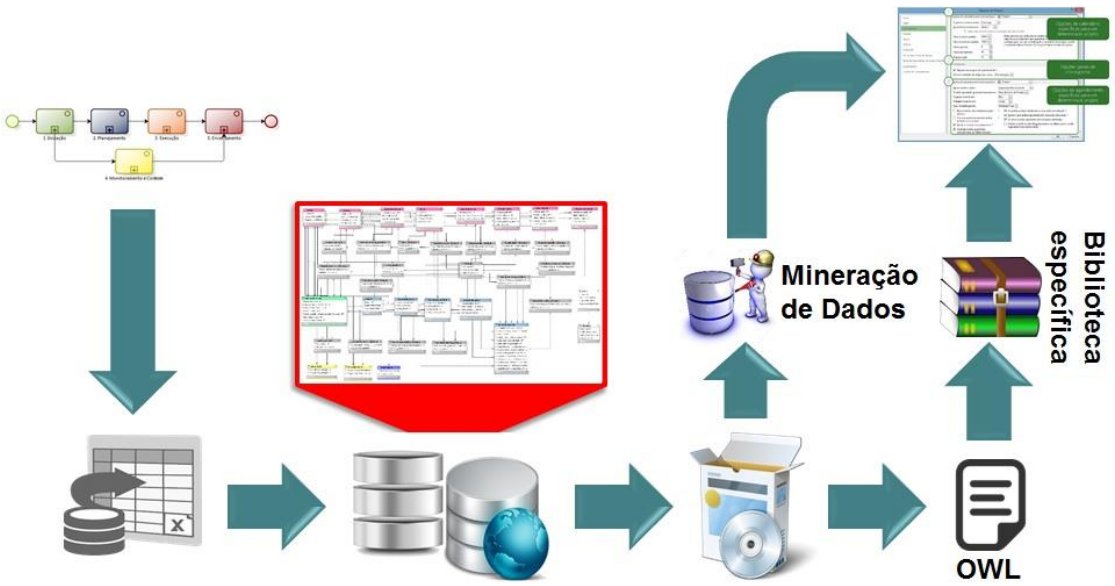


Figura 9: Fluxograma PROV-Process.

Com o intuito de exemplificar o uso da arquitetura, é apresentada, na Figura 10, a modelagem de um processo de software de um projeto baseado no modelo de desenvolvimento de software ágil, proposto por Magdaleno (2013). Este projeto tem uma duração de seis meses a um ano. As tarefas deste processo são exibidas pelos retângulos em azul e estas são atribuídas aos papéis definidos na lateral esquerda do processo, tais como Gestor da Demanda, Gerente de Projetos, Desenvolvedor, entre outros. Assim, todas as tarefas incluídas neste fluxo possuem um nome e um papel, que especifica a função que um ator deverá possuir para ser capaz de realizar a tarefa mencionada. Embora não possa ser visualizado nesta figura, durante a modelagem do processo também foram definidos os artefatos necessários para a execução da tarefa (artefatos de entrada), podendo estes serem mandatórios ou opcionais, e os artefatos gerados na execução.

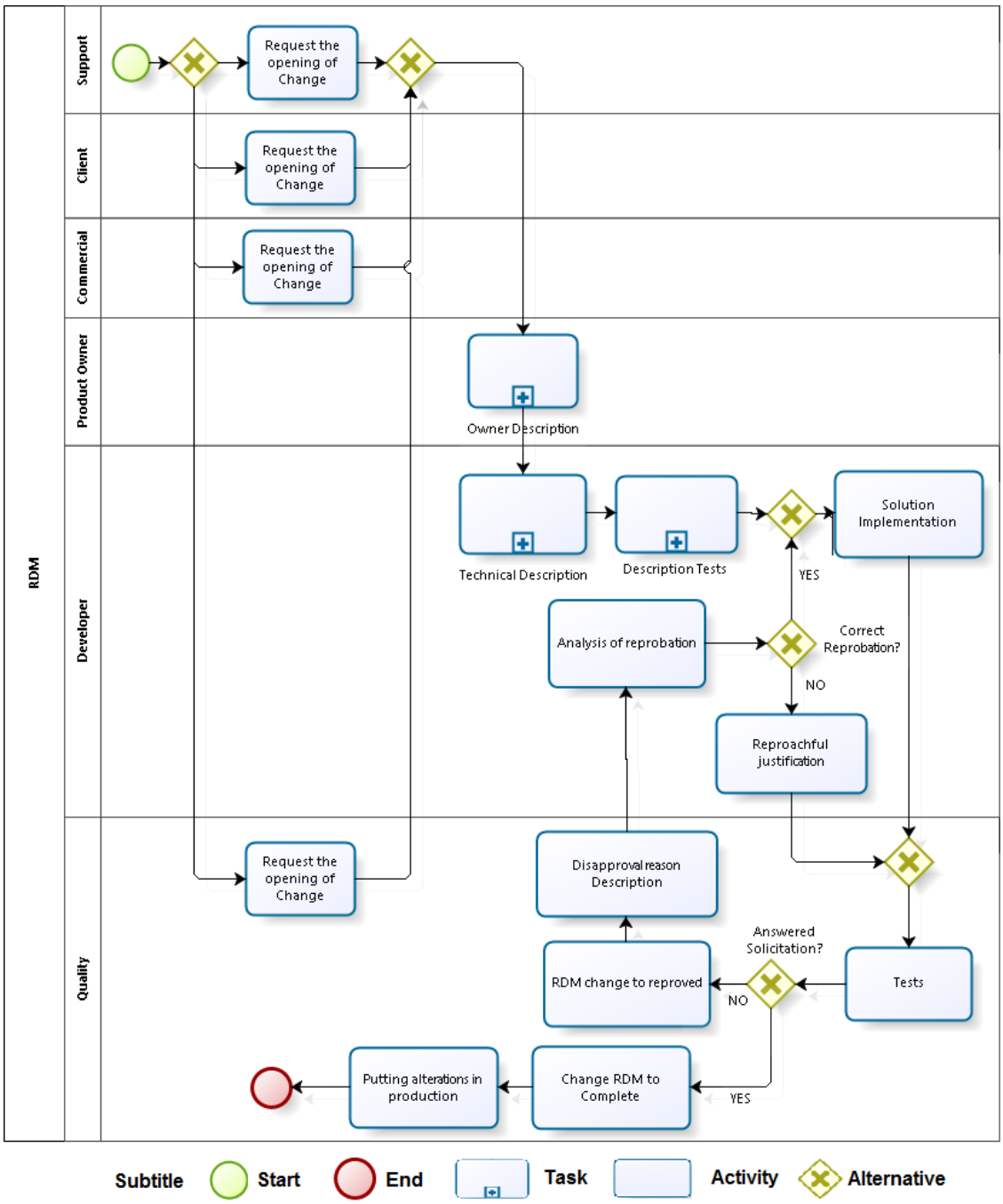


Figura 10: Modelo gráfico do processo de software.

Ainda com base no modelo de processo de software apresentado na Figura 10, pode-se citar os seguintes exemplos de uso da PROV-Process:

1. Analisar o que foi realizado durante a execução de uma determinada tarefa em outras instâncias de um mesmo processo, visando identificar se existe a reincidência desta

tarefa ou tarefas similares a esta. Por exemplo, ao analisarmos a Tarefa 1, “Iniciar atendimento da Demanda”, a proveniência aplicar-se-á ao histórico de todas as tarefas deste tipo que foram realizadas, verificando o que foi feito e as correções efetuadas. Desta forma, seria possível a verificação das correções já aplicadas e o impacto das alterações em outros fluxos, a fim de identificar a origem do erro para tratamento devido.

2. Além da aplicação de técnicas de proveniência em partes/tarefas específicas do processo, estas também podem ser utilizadas no processo como um todo. A proveniência de dados pode ser utilizada, por exemplo, para analisar o histórico de ações realizadas durante a execução do processo, como forma de encontrar os problemas decorridos após a finalização deste. Através de técnicas de proveniência, a análise destes registros de execução pode ser utilizada pelo gerente de projetos como uma forma de analisar as causas dos problemas. Assim, é capaz de evitar que os mesmos problemas voltem a ocorrer em uma nova instância do processo de software modelado.

Na seção seguinte, a arquitetura PROV-Process é apresentada em detalhes.

4.2 ARQUITETURA

Nesta seção é apresentada a arquitetura para o armazenamento e análise dos dados de execução dos processos. Conforme já dito, nesta camada serão capturados e armazenados dados de proveniência utilizando o modelo PROV (GROTH, 2013).

Na Figura 11 é apresentado um fluxograma de processos, que utiliza a camada de proveniência proposta. O fluxo inicia-se na modelagem do processo, que gera um processo definido. Um exemplo de processo definido é apresentado na Figura 10. Após a modelagem do processo, é feita a instanciação do mesmo, resultando no processo instanciado. O processo resultante é executado e os dados de execução são armazenados em um repositório de dados, modelado de acordo com as especificações do PROV (GROTH, 2013). Os dados obtidos a partir da execução do processo geram a proveniência retrospectiva, já os dados obtidos a partir da modelagem do processo compõem a proveniência prospectiva, sendo estes posteriormente tratados e analisados pela camada de proveniência. Assim, a camada deverá, além de capturar os dados de proveniência, prover mecanismos para identificação destes dados que possam auxiliar nas tarefas, sendo sugeridas melhorias para o processo como um todo.

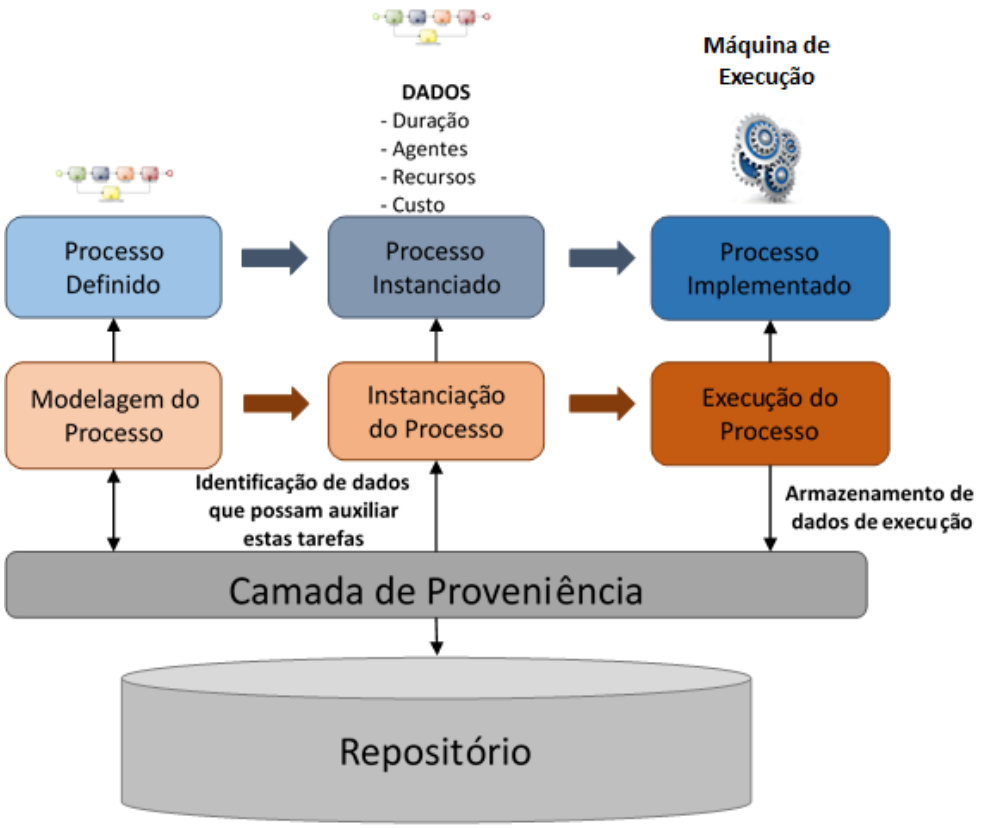


Figura 11: Fluxograma de processos.

Com base no fluxograma da Figura 11 e de acordo com os objetivos da proposta, foi elaborada a arquitetura da abordagem PROV-Process. A Figura 12 detalha a arquitetura, a qual é composta por 3 módulos e um repositório. O módulo importador é o responsável por importar os dados dos processos para o banco de dados. Os dados do banco de dados alimentam a Ontologia e o módulo de mineração de dados. A Ontologia gera uma nova Ontologia já com inferências, se mecanismos de inferência forem utilizados sobre os dados importados. O módulo de mineração de dados, através do uso de regras de associação, também gera novo conhecimento, com base nos dados captados junto ao banco de dados. As informações obtidas, tanto pelas inferências da Ontologia quanto pela técnica de mineração de dados, são apresentadas ao gerente de projetos, para que o mesmo possa analisá-las e tomar decisões, com base nos registros dispostos pelo sistema. Estas informações geradas pela Ontologia e mineração de dados não estão explícitas nos dados de execução do processo e por isso são estratégicas para o gerente de projetos.

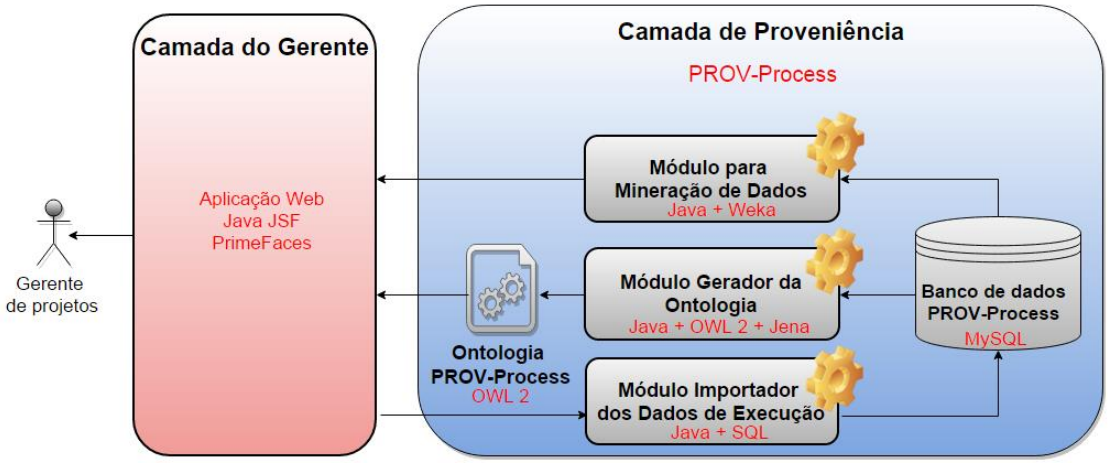


Figura 12: Arquitetura PROV-Process.

A subseção 4.2.1 detalha o modelo utilizado para o armazenamento dos dados de proveniência. A subseção 4.2.2 especifica a Ontologia utilizada como base, bem como a Ontologia estendida para a abordagem. A seção 4.2.3 detalha o módulo de importação desenvolvido para importar os dados para o banco de dados relacional do PROV-Process. A subseção 4.2.4 detalha as inferências realizadas na Ontologia. A 4.2.5 descreve o módulo de mineração de dados, também utilizado para obtenção de registros para sugestão de melhorias dos processos de desenvolvimento de software. Por fim, a subseção 4.2.6 apresenta a aplicação web desenvolvida para uso das tecnologias utilizadas.

4.2.1 Base de Dados

Esta seção apresenta, em detalhes, o repositório de dados da abordagem PROV-Process. O diagrama de tabelas relacionais (DTR) deste repositório é exibido na Figura 13. Este foi criado de acordo com as especificações do PROV-DM (MOREAU *et al.*, 2013). PROV-DM é o modelo conceitual de dados que constitui a base W3C de especificações de proveniência (PROV). Distingue estruturas centrais, formando um núcleo de informações de proveniência, a partir do qual é possível estender estruturas para uso mais específico de proveniência. O modelo PROV-DM está organizado em seis componentes, sendo respectivamente: (1) entidades, atividades e o momento em que são criados, utilizados ou terminam; (2) derivações entre entidades; (3) agentes responsáveis por entidades que foram geradas e atividades que ocorreram; (4) *bundles*, um

mecanismo de apoio à proveniência de proveniência; (5) propriedade de vincular entidades que se referem à mesma coisa; e, (6) coleções que formam uma estrutura lógica para os seus membros (MOREAU *et al.*, 2013).

O diagrama de tabelas relacionais, apresentado na Figura 13, destaca as tabelas relacionadas aos componentes descritos acima, onde as mesmas apresentam colorações distintas, representando os componentes citados. A cor rosa, representa o componente 1 (Entidades e atividades), a cor verde, o componente 2 (Derivações). A cor azul denota o componente 3 (Agentes, Responsabilidade e Influência), já a cor amarela representa o componente 5 (Alternativo) e, por fim, a cor roxa representa o componente 6 (Coleções). O componente 4 (*Bundles*), encontra-se representado no diagrama como atributo da tabela Entity, haja vista que este trata-se de um tipo de entidade. No Apêndice VI é apresentado o detalhamento de cada tabela, bem como seus atributos. Com base no modelo detalhado, os dados de proveniência são armazenados nas tabelas propostas e podem então serem utilizados pela Ontologia PROV-OEXT, descrita a seguir.

4.2.2 Ontologia

Conforme já dito, neste trabalho utilizou-se como base a Ontologia denominada PROV-O (LEBO *et al.*, 2013), a qual foi criada com base no modelo de banco de dados PROV-DM. Esta Ontologia utiliza a codificação OWL2 (*Web Ontology Language*) com base no modelo de dados PROV-DM.

Esta especificação da Ontologia fornece a base para implementar aplicações de proveniência em diferentes domínios, podendo ser utilizada para o intercâmbio e a integração de informações de proveniência gerada em sistemas e contextos diferentes. As classes e propriedades da Ontologia PROV-O são definidas de forma que possam ser usadas diretamente para representar informações de proveniência, sendo possível especializá-las para modelagem de aplicações específicas, detalhando proveniência em diversos domínios. Sendo assim, a Ontologia PROV-O possibilita seu uso em aplicações específicas, além de servir como um modelo de referência para a criação de Ontologias de proveniência de domínio específico, facilitando a modelagem de proveniência interoperável (LEBO *et al.*, 2013).

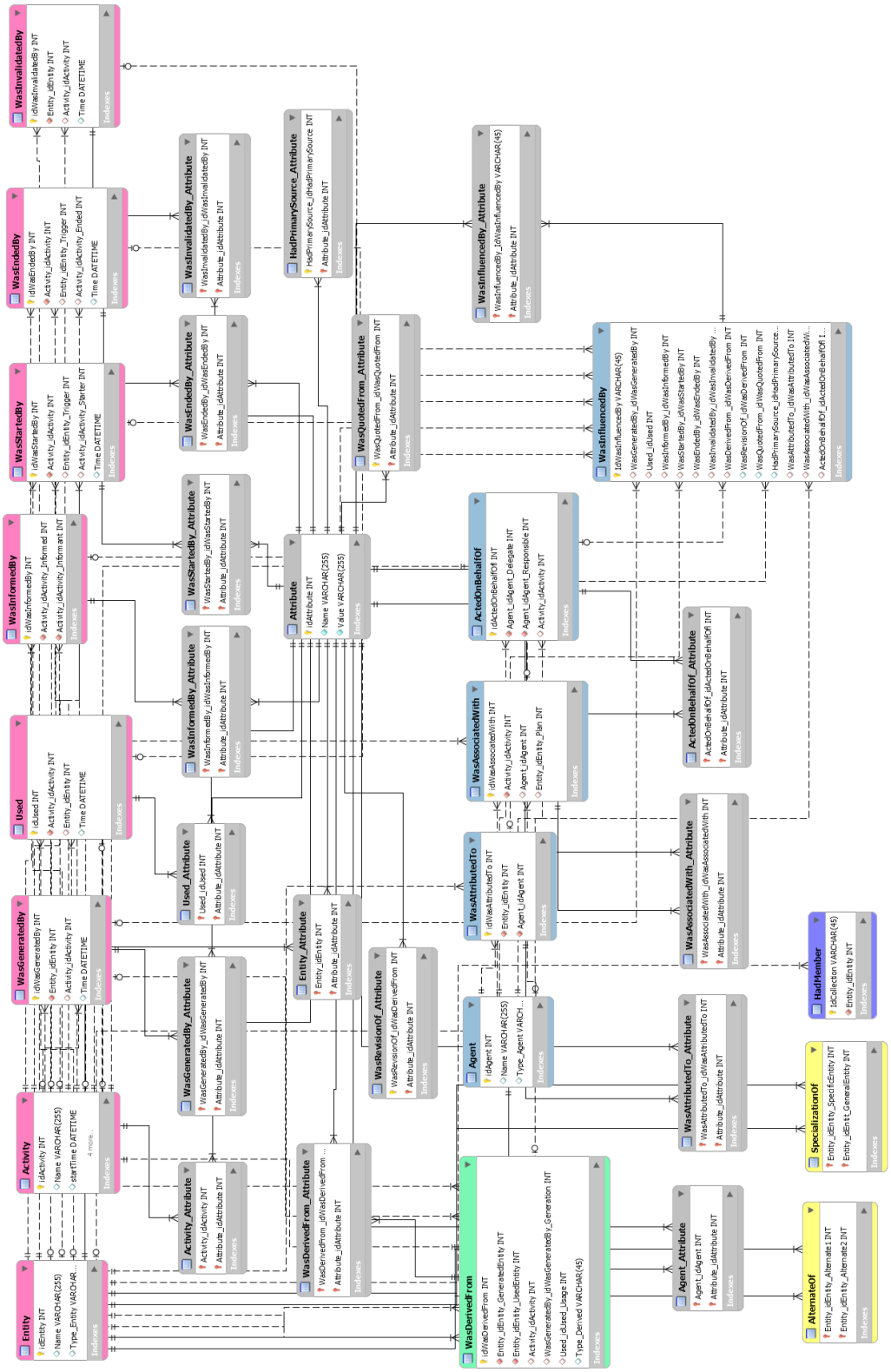


Figura 13: DTR PROV-Process.

Neste contexto, conforme já mencionado, a Ontologia da abordagem PROV-Process foi desenvolvida a partir da Ontologia PROV-O (BELHAJJAME *et al.*, 2013), que por sua vez foi definida com base no modelo de dados PROV-DM. Os vértices do PROV (Entidade, Atividade e Agente) foram definidos como classes, sendo utilizadas *object properties* para representação das relações. As principais classes e propriedades do PROV-O são apresentadas na Figura 14.

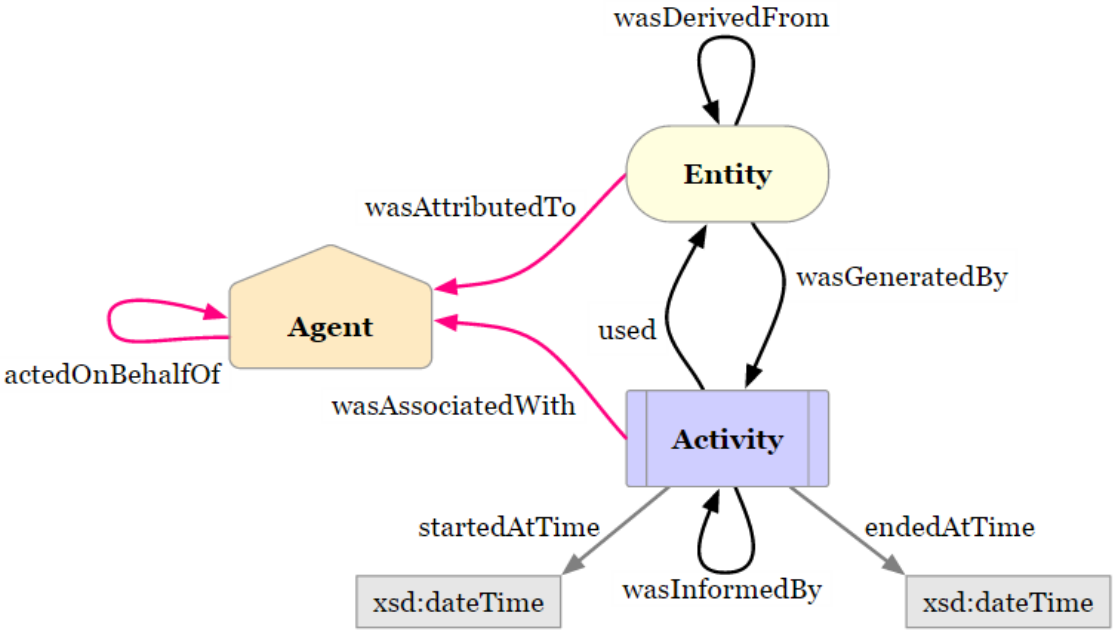


Figura 14: Classes e propriedades do PROV-O (BELHAJJAME *et al*. 2012).

Assim, classes e propriedades do PROV-O podem ser usadas diretamente para representar informações de proveniência ou pode-se especializá-las para a modelagem de aplicações específicas. Com base nisso, foram criadas novas propriedades no PROV-O, originando a Ontologia PROV-Process, a fim de adaptá-la ao domínio de processos de software, possibilitando a inferência de novas informações para melhorar os processos de desenvolvimento de software. Além disso, um grupo de regras que utilizam *property chains*, foi adicionado ao PROV-O, na *data property* 'wasAssociatedWith':

1. **used o wasAttributedTo**

2. **wasStartedBy o wasAttributedTo**
3. **wasEndedBy o wasAttributedTo**

Estas regras estabelecem que, conforme apresentado na Figura 15, se uma atividade usou, foi iniciada por ou foi encerrada por uma entidade e uma entidade foi atribuída a um agente, pode-se inferir que uma atividade está associada a um agente.

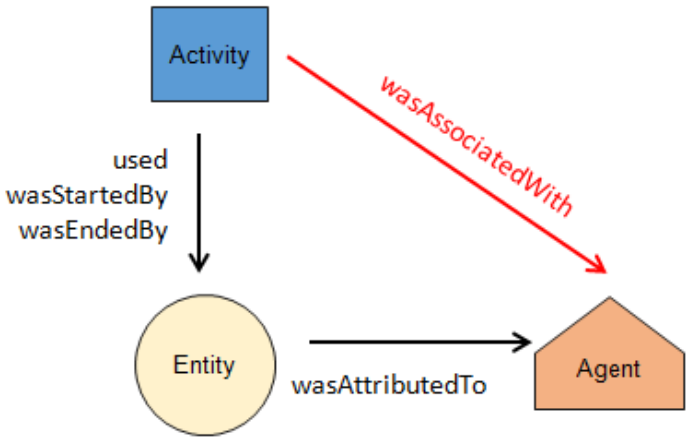


Figura 15: Property chain wasAssociatedWith.

Na PROV-Process, uma *data property* denominada *processInstanceId*, correspondente ao identificador de uma instância / execução a partir do processo principal, também foi inserida. Deve-se notar que os atributos do banco de dados do PROV-Process, devem ser exportados para a Ontologia PROV-Process como novas *data properties*, com seu respectivo valor.

Os dados de execução dos processos também serão exportados para Ontologia PROV-Process. Sendo assim, foram inseridas as *data properties priority*, *processinstanceId*, *typeActivity*, *typeEntity* e *typeAgent*. A *data property priority* refere-se a especificação quanto ao grau de priorização de uma instância, podendo ser classificada como alta, média ou baixa. A *processinstanceId* é o número de identificação da instância. Já a *typeActivity* trata-se da especificação dos tipos de atividades de uma instância, podendo ser, por exemplo, uma solicitação de novo recurso, um erro, uma alteração legal, entre outras. A especificação da parte do sistema onde estão sendo realizadas modificações e/ou implementações é atribuída a *typeEntity*, a qual pode ser preenchida como módulo, componente ou versão. Por fim, a

typeAgent indica o responsável por uma atividade, podendo ser preenchido com “Pessoa”, “Organização” ou “Software”.

Além destas *data properties* todas as inversas do PROV-O (PROV-O, 2016) foram inseridas na Ontologia da abordagem PROV-Process. A Figura 16 exibe parte da Ontologia citada nesta seção, com o objetivo de mostrar o uso e a forma de exibição da mesma. A atividade “Abertura_2937” foi incluída na Ontologia como um indivíduo da classe *Activity*, sendo associada as *data properties*, citadas anteriormente, e a outros indivíduos, utilizando as *object properties* da Ontologia PROV-O.

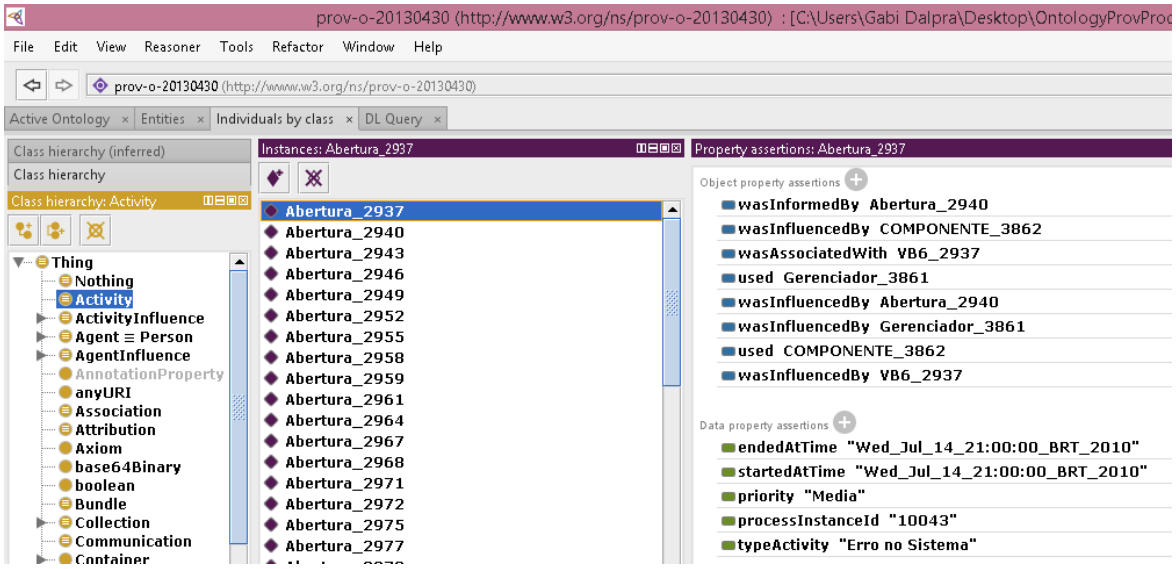


Figura 16: Ontologia PROV-Process.

4.2.3 Módulo Importador dos Dados de Execução

Com o objetivo de viabilizar a utilização da arquitetura por diversas empresas, foi elaborado um arquivo padrão para importação de dados pelo PROV-Process. Para desenvolvimento deste arquivo, foram utilizados como base os dados gerados por duas empresas brasileiras de desenvolvimento de software. Os registros importantes para o uso da camada de proveniência, bem como para o processo de desenvolvimento de software das mesmas foram definidos de comum acordo entre elas e a equipe de especificação do PROV_Process, resultando na padronização do arquivo a ser importado, o qual deve estar no formato ‘.csv’. Assim, é

importante ressaltar que para utilização da PROV-Process, é necessário exportar os dados para o arquivo padrão, o qual pode ser visualizado nas tabelas 7 e 8.

É necessária a exportação dos dados gerados pela empresa para o formato utilizado pela PROV-Process. Como alternativa, é possível a captura de dados diretamente de ferramentas de execução de processos, como o JBPM (RED HAT, 2015). No entanto, este suporte ainda é parcial na arquitetura PROV-Process.

Tabela 7. Arquivo padrão – Parte 1

NUMINSTANCIA	ATIVIDADE	TIPO_ATIVIDADE	PRIORIDADE	INICIO	FIM
3025	Abertura	Erro	Alta	01/01/2014	10/01/2014
3028	Abertura	Novo Recurso	Baixa	05/04/2015	06/04/2015

Tabela 8. Arquivo padrão – Parte 2

RESPONSAVEL	TIPO_RESPONSAVEL	MODULO	COMPONENTE	VERSÃO	DESDOBRAMENTO
Emanuel	Pessoa	Fiscal			
Geovani	Pessoa			15.7	3025

4.2.4 Módulo Gerador da Ontologia

Conforme já dito, os dados armazenados na base relacional são carregados como indivíduos da Ontologia PROV-Process por este módulo. O módulo carrega os dados na Ontologia gerando duas novas Ontologias, uma sem inferência e a outra com inferência[1].

A fim de avaliar a aplicabilidade da Ontologia PROV-Process para processos de software, será apresentado no capítulo 5 um estudo de caso conduzido em duas empresas brasileiras de desenvolvimento de software, as quais serão tratadas como empresa I e II. Para ilustrar a execução do módulo de carregamento de Ontologia, alguns dados deste estudo de caso

[1] Essas duas ontologias são necessárias para comparação do que foi inferido, haja vista que a biblioteca jena (JENA, 2016) não permite a consulta das respectivas inferências.

serão aqui utilizados. O detalhamento do estudo de caso, no entanto, será apresentado no capítulo 5.

Assim, um modelo de fluxo de processos, apresentado na Figura 10, foi criado com base nas informações do processo de desenvolvimento obtidas junto a uma das empresas citadas, detalhando as tarefas e atividades relacionadas.

Após a importação dos dados de execução de processos para o banco de dados, os dados foram introduzidos como indivíduos na Ontologia PROV-Process. A partir deste ponto, utilizando inferências, a derivação de informações estratégicas foi possível. Como exemplo de informações inferidas a partir de dados de proveniência retrospectiva desse processo, podem ser destacados quatro tipos:

1) Atividades que influenciaram a geração de outras atividades, ou seja, conforme destacado na cor vermelha na Figura 17, a abertura de requisição de mudança (id = 1) influenciou a abertura de requisição de mudança (id = 4).

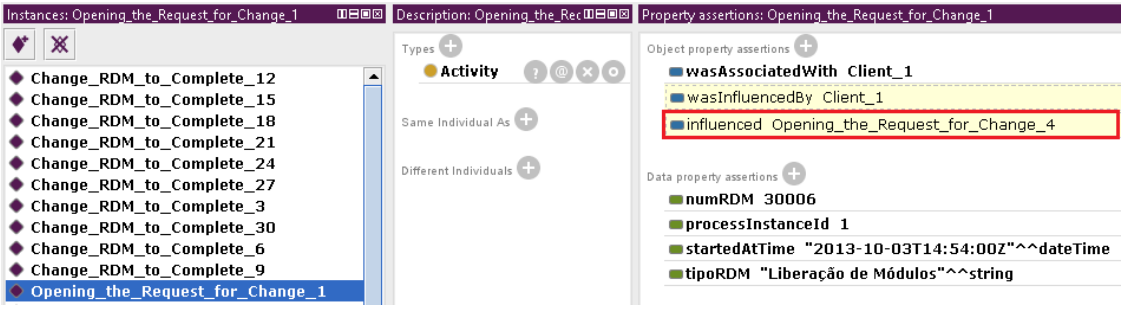


Figura 17: Atividades que influenciaram a geração de outras atividades.

2) Agentes que poderiam ser associados com a atividade de implementação da solução, considerando-se que eles já tratam os artefatos envolvidos nesta atividade em qualquer outra execução do processo. A Figura 18 mostra que a atividade de implementação da solução (ID = 11) foi influenciada pelo agente DotNet (ID = 5), dado que este agente trata artefatos comuns para esta atividade em outros exemplos deste processo.

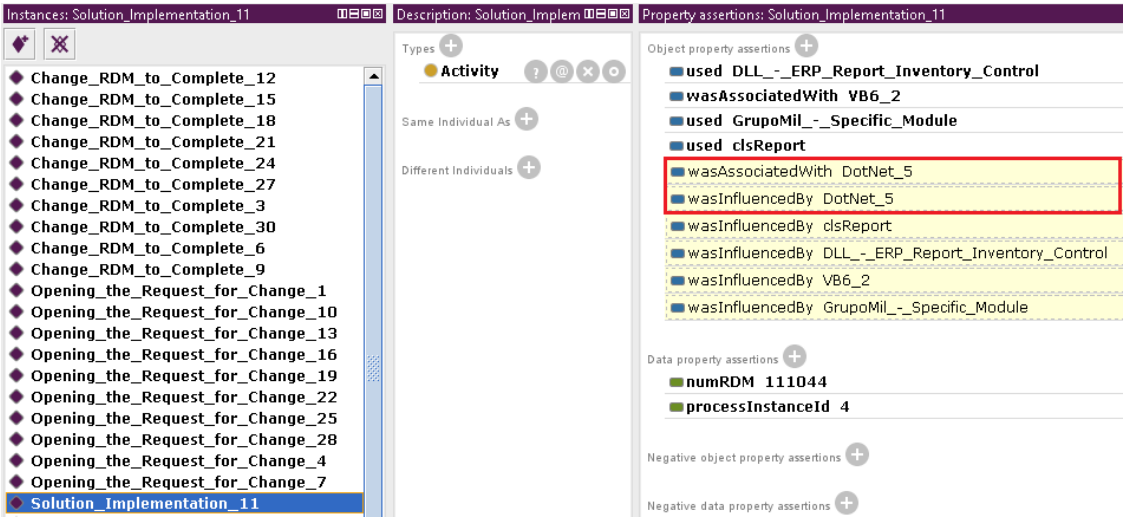


Figura 18: Agentes que influenciaram uma atividade.

3) Uma lista com todas as atividades em que um agente esteve envolvido, bem como os artefatos (entidades) manipulados pelo mesmo, pode ser visualizada na Figura 19. Embora este tipo de informação possa ser obtida por meio de consultas ao banco de dados, utilizando-se a máquina de inferência, esta informação pode ser obtida mais facilmente, com uma consulta SPARQL simples.

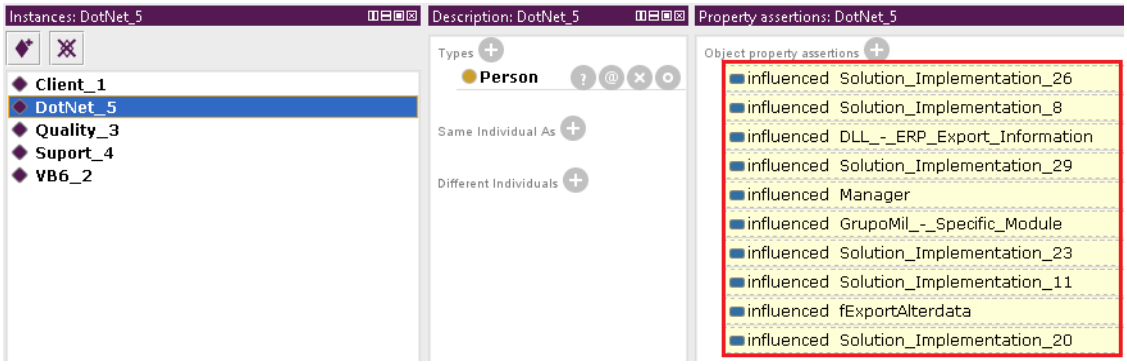


Figura 19: Atividades e agentes manipulados pelo agente DotNet.

4) Uma lista de todas as atividades onde um artefato (entidade) foi consumido, pode ser vista na Figura 20. Embora este tipo de informação também possa ser obtido por meio de consultas ao banco de dados, utilizando a máquina de inferência, obtém-se essas informações de forma mais fácil, com uma consulta SPARQL.

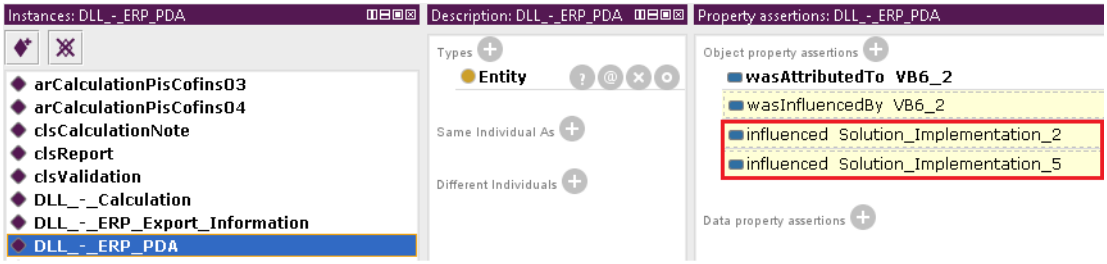


Figura 20: Atividade onde um artefato (entidade) foi consumido.

Informações inferidas a partir da utilização da Ontologia proposta pela abordagem PROV-Process podem auxiliar o gerente de projetos na otimização do processo. As informações inferidas podem sugerir, por exemplo, os agentes e artefatos mais adequados para a solução do tipo de problema relatado.

Paralelo ao uso de Ontologias, também são utilizadas técnicas de mineração de dados para obtenção de informações que possam contribuir para a melhoria dos processos de software. As técnicas utilizadas são detalhadas na subseção seguinte.

4.2.5 Módulo para Mineração de Dados

A partir dos dados armazenados no banco de dados, é possível também utilizar técnicas de mineração de dados para extrair conhecimento estratégico para o gerente de projetos.

Existem ferramentas que auxiliam no processo de mineração de dados, tais como IBM Enterprise Miner (SAS, 2016), Oracle Data Mining (ORACLE, 2016) e Weka (HOLMES *et. al*, 1994). Neste trabalho optou-se pela utilização do Weka por tratar-se de um sistema de código aberto, que contém um conjunto de algoritmos de aprendizado de máquina, e disponibilizar API (HALL *et. al*, 2009) para utilização destes algoritmos, além de estar em constante evolução. Estes algoritmos podem ser aplicados diretamente a um conjunto de dados ou acionados a partir de seu próprio código Java. Para utilização destes algoritmos foi implementado, no sistema PROV-Process, uma API do Weka (HALL *et. al*, 2009). Cabe ainda ressaltar que o Weka contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização (HOLMES *et. al*, 1994).

Para que seja possível a mineração dos dados de proveniência do processo, estes devem estar formatados de acordo com o padrão[2] especificado pela técnica de mineração de dados a ser utilizada no Weka (HALL *et. al*, 2009). Para isso, um arquivo com a extensão '.arrf' é gerado pelo sistema PROV-Process, conforme exibido na Figura 21.

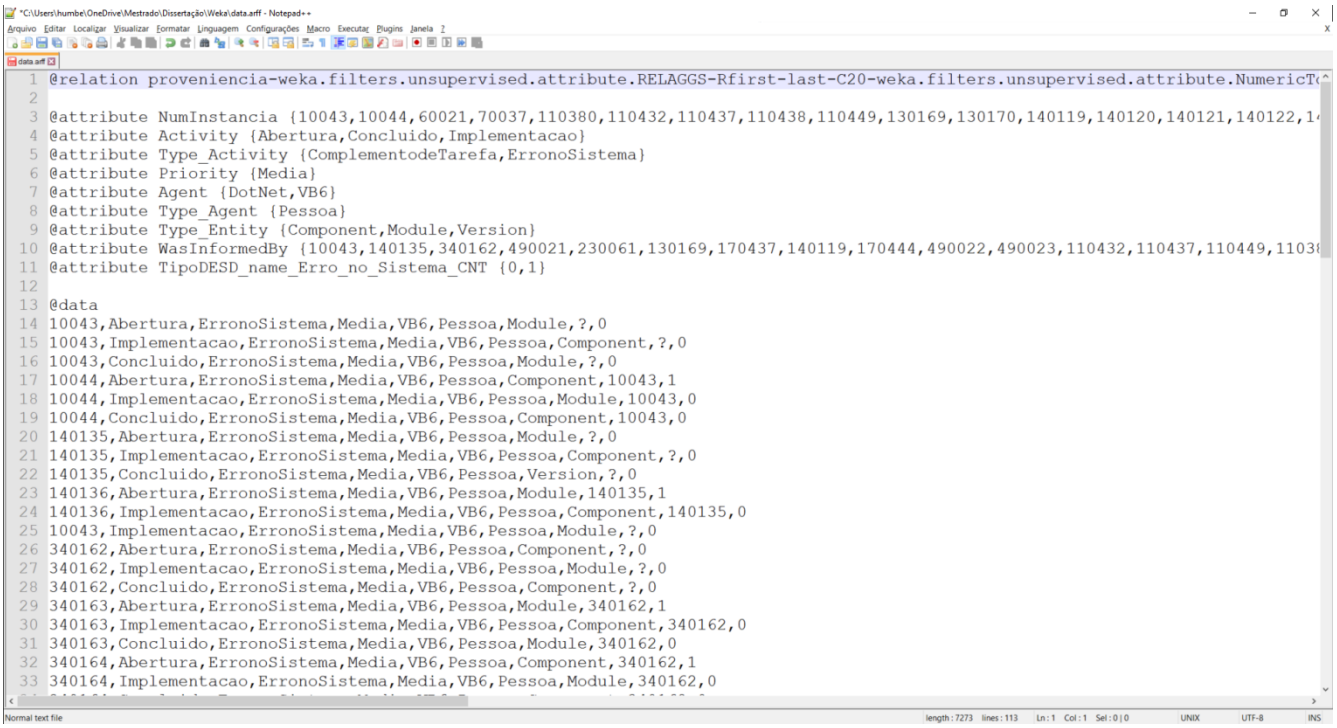


Figura 21: Arquivo data.arrf.

Em função do tipo de informação que deseja-se obter sobre projetos de desenvolvimento de software, optou-se pela utilização do algoritmo CBA (*Classification Based on Associations*) (LIU *et. al*, 1998) para identificação de padrões dos dados de proveniência. O arquivo gerado é interpretado por esse algoritmo através da API do Weka (HALL *et al.*, 2009), a qual foi implementada no PROV-Process. A escolha do algoritmo CBA deve-se ao fato de que o sistema PROV-Process visa prover informações estratégicas ao gerente de projetos, as quais podem ser

[2] Cada atributo constante no arquivo padrão de importação dos dados de execução de processos, dispostos na seção 4.2.3, é inserido, no arquivo com extensão '.arrf', precedido de "@attribute" e sucedido por todos os valores atribuídos ao mesmo, os quais são dispostos entre chaves ("}"). Já os dados preenchidos em cada linha da tabela do arquivo padrão de importação, são dispostos na ordem correspondente aos atributos definidos por "@attribute", sendo separados por vírgula e quando vazios representados por um ponto de interrogação (?). A especificação de que os valores constantes no arquivo correspondem aos dados referentes aos atributos, é dada por "@data", que precede os respectivos dados.

obtidas por meio de regras de associação de classes, foco do algoritmo CBA, conforme descrito na subseção 2.4.3.

Finalizado o processo de mineração dos dados, o sistema PROV-Process captura os dados das regras de classificação e apresenta ao usuário de forma mais simples, a fim de facilitar a interpretação dos resultados por parte do gerente de projetos. Segue abaixo exemplo de como a regra é apresentada após a mineração de dados e como é exibida pelo sistema:

- Forma apresentada após a mineração de dados:

  Activity=Implementacao 1 2 Agent=VB6

  ==>TipoDESD_name_Erro_no_Sistema_CNT=1 conf:(0.27), (98),

- Exibição no PROV-Process:

  27% of instances with Activity Implementacao and Agent=VB6 resulted in system error.

Na próxima seção apresentamos as possibilidades de visualização dos resultados, com base nas informações capturadas pela Ontologia, inferidas e geradas a partir das técnicas de mineração de dados.

4.2.6 Visualização dos Resultados

Uma aplicação web foi desenvolvida para facilitar a consulta aos dados por parte do gerente de projetos, além de possibilitar a visualização das informações advindas da análise dos dados de execução do processo pela arquitetura PROV-Process. Os dados são captados e exibidos na interface, a qual apresenta listagens distintas dos conteúdos, constantes nos registros de execução dos processos, e também permite a edição e/ou deleção destes.

Na Figura 22 é possível visualizar a listagem das instâncias de um processo específico. Vale ressaltar que a lista exibe os dados contidos no arquivo padrão importado pelo PROV-Process. A ferramenta possibilita ainda o detalhamento das instâncias, onde são apresentadas ao gerente de projetos informações referentes aos responsáveis pela execução das tarefas, (Agentes), especificação da respectiva tarefa (Atividades) e do módulo, componente ou versão (Entidades) envolvidos nesta. Nos detalhes, também são apresentadas ao gerente de projetos as inferências obtidas através do uso da Ontologia, o que auxilia na identificação de pontos de melhoria do processo.

Junto a opção *Data*, encontra-se disponível a opção *Load Data*, a qual é responsável por carregar os dados contidos no banco de dados do PROV-Process para a Ontologia, bem como o arquivo .arrf, o qual contém os dados necessários para aplicação da técnica de mineração de dados. Na Figura 22 é possível visualizar ainda a tela inicial da opção, a qual apresenta a listagem de todas as instâncias (fluxo de processo completo), com sua data de abertura e conclusão. Na coluna "*Option*" encontra-se a opção "*Details*", a qual exibe todas as atividades, agentes e entidades da instância, conforme já mencionado e também apresentado na Figura 23.

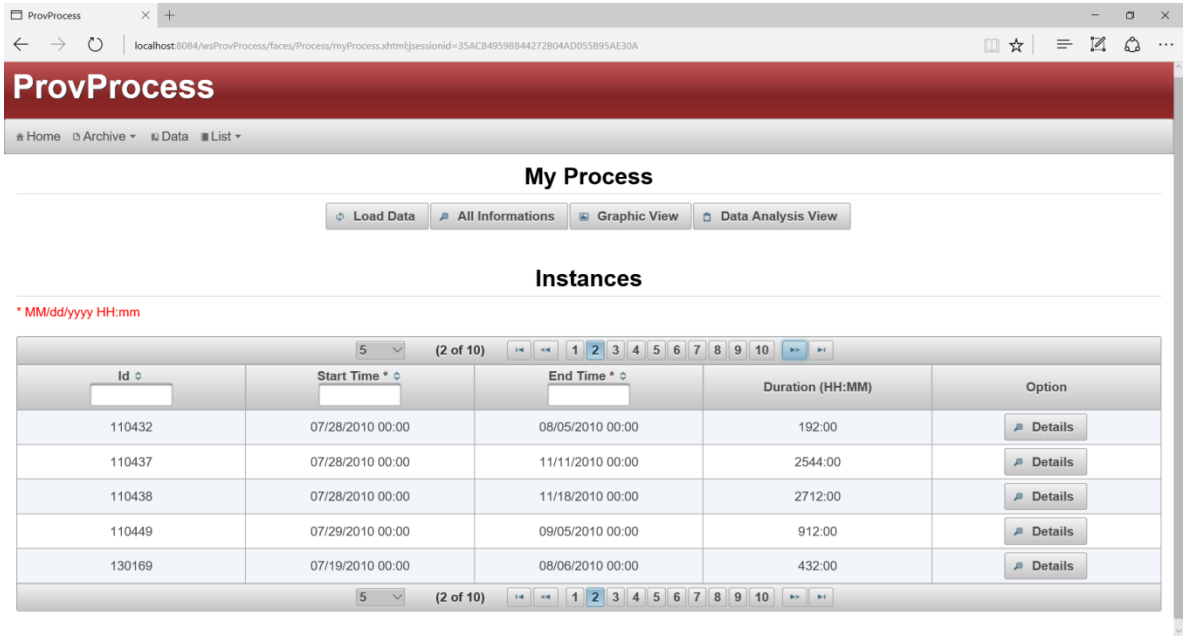


Figura 22: Lista de instâncias.

A Figura 22 exibe ao gerente de projetos a visualização de todas as instâncias que foram executadas em um determinado processo de software, permitindo a busca específica de instâncias por id, data de abertura ou data de finalização. Ainda é possível ao gerente de projetos a verificação do período de abertura ou finalização das instâncias, através da organização das guias em ordem crescente e/ou decrescente. A coluna *Duration* exibe o tempo em horas e minutos, formato HH:MM, entre a data de abertura e a data de finalização, possibilitando ao gerente de projetos um indicativo estratégico relativo a demora no tratamento de determinada instância. Para auxiliar o gerente no entendimento relativo ao período de execução da instância ou ainda detectar

alguma inconsistência no processo, pode-se utilizar a opção "*Details*", a qual apresenta informações sobre as atividades, agentes e entidades, podendo vir a auxiliar na compreensão do motivo da demora na execução de uma determinada instância, por exemplo.

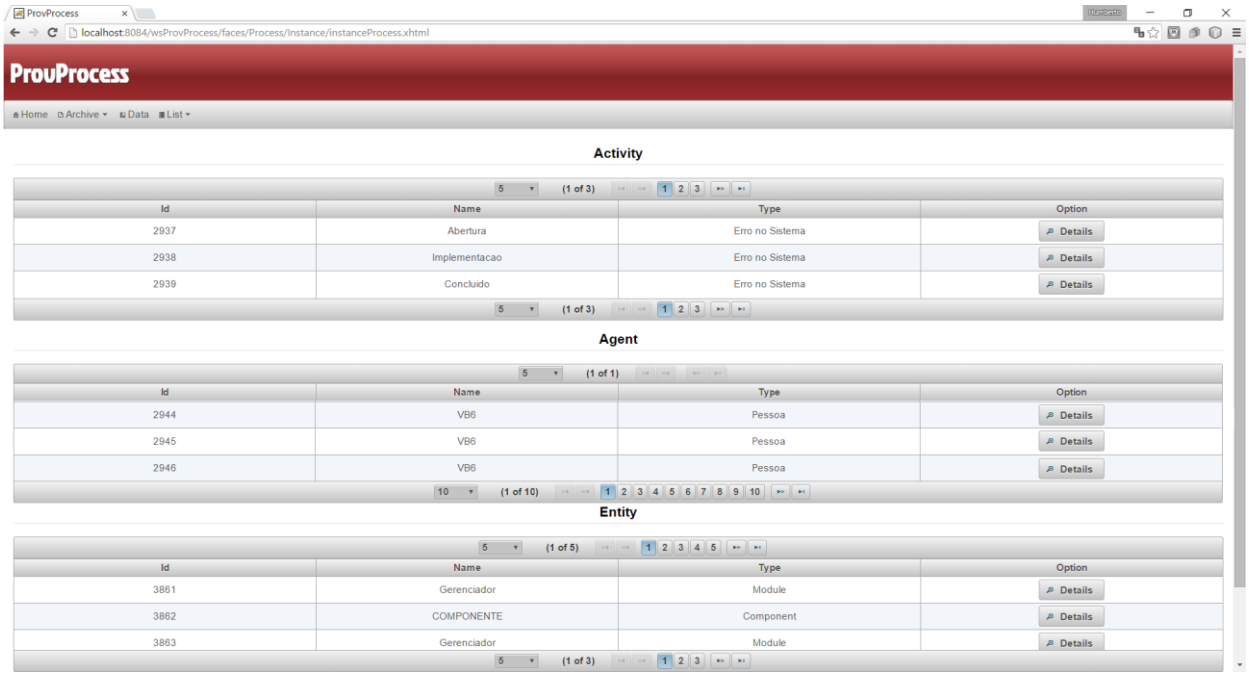


Figura 23: Lista de atividades, agentes e entidades de uma instância.

Conforme apresentado na Figura 23, também há a opção de detalhamento das atividades, agentes e entidades, as quais remetem o usuário a uma nova tela que exibe todas as informações da opção bem como as inferências obtidas junto a Ontologia, conforme apresentado na Figura 24.

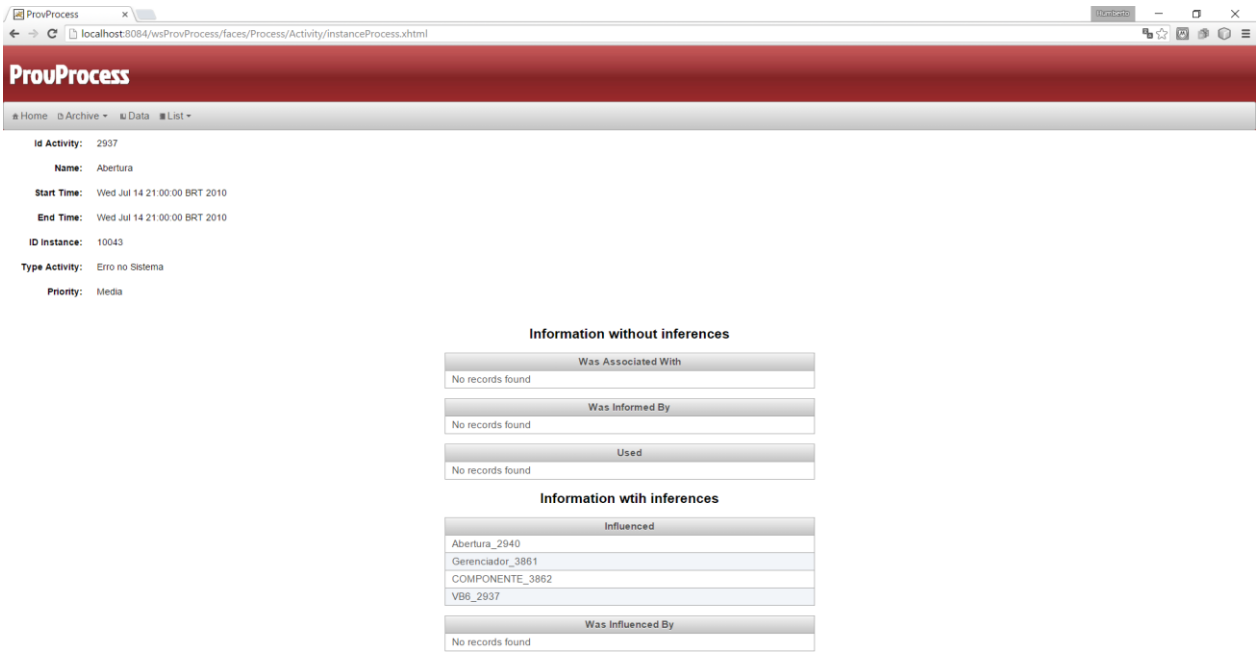


Figura 24: Detalhe atividade com inferências.

Caso não exista registro nas relações de uma atividade, agente ou entidade, tais como *WasAssociatedWith*, *WasInformedBy*, *Used*, *Influenced*, *WasInfluencedBy*, sejam estas inferidas ou não, esta informação será devidamente exibida ao gerente de projetos por meio da mensagem "*No records found*". Havendo registro de relação, esta será exibida ao gerente, apresentando-se de acordo com a forma de inclusão dos indivíduos na Ontologia, especificado na subseção 4.2.4, a qual será alocada na tabela correspondente à respectiva relação.

Conforme pode ser visto na figura 24, a atividade de id 2937, nome Abertura, influenciou a atividade Abertura, de id 2940. Esta informação estratégica indica ao gerente de projetos que a abertura da atividade 2940 originou-se após a conclusão da atividade 2937, ou seja, aponta-se um desdobramento da atividade de origem. Este desdobramento pode ocorrer devido a fatores tais como conclusão indevida da respectiva atividade, impacto da implementação em outras rotinas do sistema, fazendo com que as mesmas tenham seu funcionamento comprometido, e/ou ainda a necessidade de ajustar outras rotinas do sistema para adequação devido a mudança ocorrida mediante a conclusão da atividade de origem. Esta análise pode auxiliar o gerente de projetos na identificação de falhas no processo, onde de posse desta informação o mesmo pode adotar as medidas necessárias para sanar o problema, evitando reincidências e consequente retrabalho, o que pode vir a resultar na otimização da produtividade da equipe.

Visando proporcionar outra forma de visualização das inferências obtidas com o uso da Ontologia, a aplicação Web disponibiliza a visualização gráfica das inferências, com a opção "*Graphic View*" do menu "*Data*" (Figura 25). Por meio desta visualização, o gerente de projetos pode, através dos filtros, definir os símbolos utilizados na representação, podendo optar por círculos, pelo padrão de representação do modelo BPMN ou pelo padrão do modelo PROV. Também é possível aplicar zoom para melhor visualização de determinados vértices e/ou ligações, bem como arrastá-los para quaisquer posições da área do gráfico para melhorar a disposição ou identificação de uma parte específica do mesmo. São disponibilizados ainda filtros que permitem visualizar os arcos que representam as relações constantes no banco de dados relacional, representados por arcos na cor preta, e/ou as relações inferidas, representadas por setas na cor vermelha. Também é possível selecionar a exibição ou não dos arcos que representam as relações *Used*, *wasAssociatedWith*, *wasAttributedTo* e *Influenced*. Referente aos Agentes, Atividades e Entidades, é possível definir a exibição do nome e ícone referente a cada vértice do modelo PROV, sendo possível apresentar somente o nome, somente o vértice, o nome e o vértice juntos ou nenhum dos dois. A visualização, apresentada na Figura 25, auxilia, portanto, o gerente de projetos na identificação das inferências, agentes, atividades e entidades envolvidas no processo, apresentando de forma gráfica as respectivas relações e vértices.

Figura 25: Visualização Ontologia.

A aplicação Web exibe ainda uma visualização da análise dos dados, realizada por meio do uso da técnica de mineração de dados. Esta visualização encontra-se disponível através da opção "*Data Analysis View*", no menu "Data", a qual pode ser visualizada na Figura 26.

Através da análise dos dados com mineração de dados, realizada por meio do uso da técnica de regras de classificação de dados, o sistema apresenta ao gerente de projetos os resultados da mineração sobre os dados importados para o sistema PROV-Process, indicando o percentual de ocorrências de um determinado padrão que está relacionado a erro no sistema. Com base nestas informações, o gerente de projetos pode, por exemplo, identificar que toda atividade do tipo "implementação" executada pelo agente "VB6" resulta em um erro no sistema, conforme apresentado na figura 26. Com isso, o gerente pode verificar se o impacto deve-se ao tipo da atividade, a qual implementa um novo recurso no sistema, por exemplo, ou se o erro é oriundo do agente VB6, adotando as ações cabíveis com o objetivo de otimizar a produção e reduzir o número de erros no sistema.

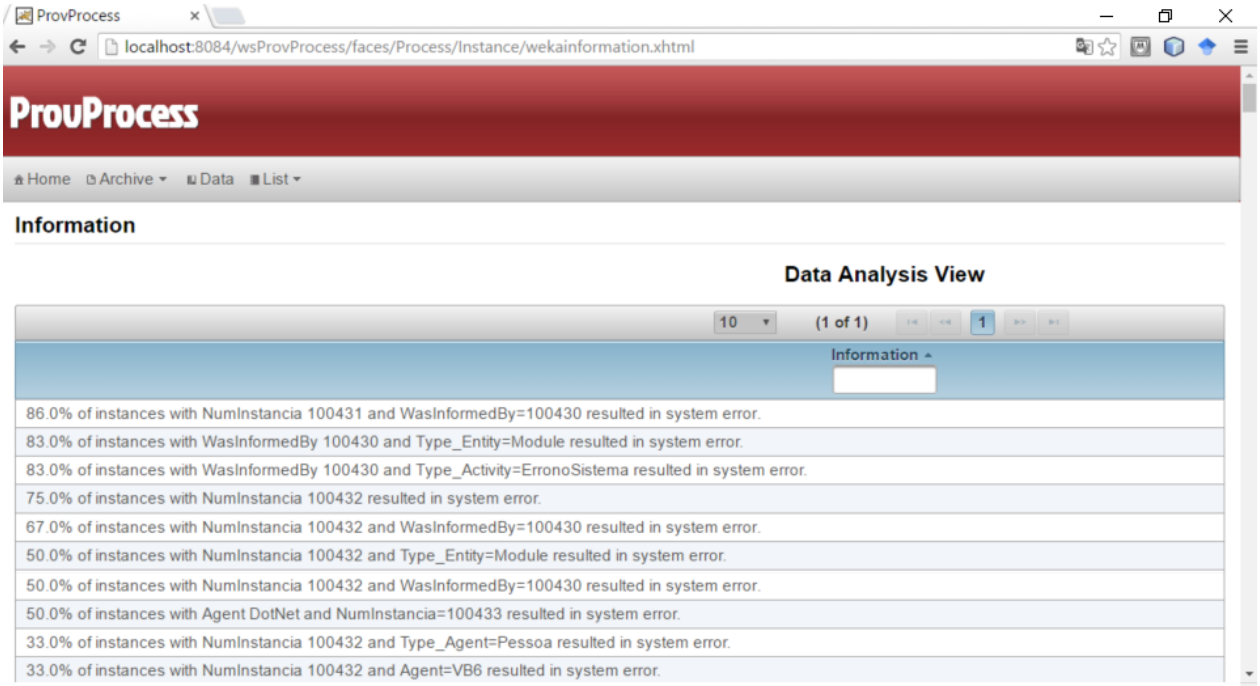


Figura 26: Exibição dos dados de mineração.

4.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou a arquitetura PROV-Process, desenvolvida para, através do uso da proveniência de dados aliado a Ontologias e técnicas de mineração de dados, indicar ao gerente de projetos possíveis pontos de melhoria no processo de desenvolvimento de software. Foram apresentadas todas as técnicas utilizadas e desenvolvidas para o funcionamento devido da arquitetura, bem como a forma de visualização dos resultados apresentados pelo mesmo.

O próximo capítulo apresenta a avaliação da arquitetura, contemplando a descrição das formas de avaliação bem como os resultados obtidos após a realização da mesma.

5 AVALIAÇÃO

Este capítulo descreve uma avaliação preliminar da abordagem PROV-Process, no intuito de verificar se esta oferece um mecanismo adequado para que gerentes de projeto possam analisar os dados de execução de um processo e, a partir dos mesmos, tomar decisões que permitam, futuramente, a melhoria das execuções de instâncias posteriores. Para que esta análise seja possível, os dados de execução do processo foram armazenados na forma de dados de proveniência, conforme apresentado nos capítulos anteriores.

A ferramenta desenvolvida para apoiar a abordagem PROV-Process apresenta três formas distintas para visualização dos dados de proveniência, sendo elas (i) Dados de execução de processos de desenvolvimento de software; (ii) Mineração de dados; e (iii) Inferências. Neste capítulo serão detalhados o planejamento, execução, resultados e conclusões obtidas com a avaliação dessa abordagem.

O capítulo está organizado nas seguintes seções: na Seção 5.1 é apresentado o estudo de caso realizado para avaliar a abordagem PROV-Process. A Seção 5.2 descreve o planejamento para realização do estudo de caso. Na sequência, a subseção 5.2.1 apresenta os objetivos do estudo. Os critérios para seleção dos participantes e a apresentação dos mesmos, são descritos na Seção 5.3. Os meios utilizados para realização da coleta de dados são explicitados na Seção 5.4. A Seção 5.5 apresenta as métricas utilizadas nas avaliações. A análise dos resultados é apresentada na Seção 5.6, detalhando a análise do PROV-Process, na subseção 5.6.1, e a análise do processo das empresas, na subseção 5.6.2. As ameaças a validade do estudo são descritas na Seção 5.7 e, por fim, a Seção 5.8 apresenta as considerações finais do capítulo.

5.1 ESTUDO DE CASO

Estudo de caso é um método padrão usado para estudos empíricos em diversas ciências, como a sociologia, medicina e psicologia (WOHLIN *et al.*, 2012). Na engenharia de software, estes estudos não devem ser utilizados apenas para avaliar como e por que ocorrem certos fenômenos, mas também para avaliar dentre os métodos, quais são mais adequados a uma determinada situação.

Os estudos de caso são muito adequados para a avaliação industrial de métodos e ferramentas de engenharia de software. A vantagem dos estudos de caso é que eles são mais fáceis de planejar e são mais realistas, porém, como desvantagem, tem-se que os resultados são difíceis de generalizar e interpretar, ou seja, é possível mostrar os efeitos de uma solução, mas é necessária maior análise para generalizar para outras situações (WOHLIN *et al.*, 2012).

Se o efeito de uma mudança de processo é muito difundido, um estudo de caso é mais adequado, haja vista que o efeito da mudança só pode ser avaliado em um alto nível de abstração, pois o processo de mudança inclui alterações menores e mais detalhadas de todo o processo de desenvolvimento. Além disso, os efeitos da mudança não podem ser identificados imediatamente. Por exemplo, se há interesse em saber se uma nova ferramenta de projeto aumenta a confiabilidade, pode ser necessário esperar até depois da entrega do produto desenvolvido para avaliar sobre falhas operacionais (WOHLIN *et al.*, 2012).

Com base nestas informações, identifica-se que o estudo de caso é a melhor forma de avaliar esta pesquisa, a qual é empírica e aplicada a um contexto real, visando a identificação quanto a aceitação e retorno, por parte de gerentes de projetos e desenvolvedores de software, em relação as informações obtidas através das análises dos dados de processos de desenvolvimento de software, junto a abordagem PROV-Process.

5.2 PLANEJAMENTO DO ESTUDO DE CASO

Inicialmente os participantes selecionados para realização deste estudo de caso preencheram um formulário de caracterização, disponível no Apêndice II. Na sequência, foram disponibilizados aos participantes dois formulários (Apêndices III e IV) contendo perguntas referentes as informações exibidas pela abordagem PROV-Process, bem como a utilização de seus recursos, tais como inferência sobre os dados de proveniência e mineração destes dados. A existência dos dois apêndices (III e IV) deve-se ao fato de que foi solicitado aos participantes a avaliação dos dados das duas empresas parceiras a fim de identificar se os mesmos conseguem localizar as informações independente do processo que está em avaliação. Os nomes das empresas foram omitidos e elas serão tratadas como empresa I e II.

Cabe ressaltar que as perguntas contidas nos respectivos formulários são específicas sobre o processo de cada uma das empresas parceiras. Neste aspecto é importante informar que

na empresa I não foi aplicada a técnica de mineração de dados, haja vista que a mesma não forneceu dados de desdobramentos de erro, os quais são analisados com o uso da referida técnica, conforme citado na subseção 2.4.3. Deve-se ressaltar que através das técnicas de mineração de dados pode-se obter outras informações, porém, no contexto desta dissertação, optou-se pelo uso apenas para detecção de padrões relativos a geração de desdobramentos de erro no sistema.

O estudo de caso conduzido nesta pesquisa foi especificado para avaliar os aspectos observados pelos participantes. As questões referentes a análise do processo das empresas foram usadas para observar a quantidade de participantes que responderam corretamente as questões previamente definidas sobre os dados exibidos pela ferramenta de apoio a abordagem PROV-Process. As perguntas utilizadas para esta avaliação encontram-se disponíveis nos Apêndices III e IV. As questões abertas foram formuladas para observar a concordância dos participantes em relação as informações apresentadas e/ou exibidas pelo sistema, a fim de verificar se estas são suficientes, relevantes e de fácil identificação por parte do gerente de projetos. Além disso, buscou-se também identificar se estas informações contribuem, de fato, para detecção de pontos de melhorias no processo de desenvolvimento de software. As perguntas utilizadas para esta avaliação estão disponíveis no Apêndice V.

5.2.1 Objetivos

Esta subseção apresenta o objetivo do estudo para avaliar a viabilidade da abordagem PROV-Process, conforme descrito na Tabela 9. Este foi estabelecido de acordo com a abordagem GQM (*Goal, Question, Metric*) (Basili *et al*. 1994).

Tabela 9: Objetivo da avaliação

Questões	Respostas
Objeto do estudo (o que será analisado?)	Abordagem PROV-Process
Objetivo (porquê / para o que o objeto vai ser analisado?)	Caracterizar
Foco da qualidade (que propriedades do objeto serão analisadas?)	Viabilidade para apoiar a tomada de decisão para a melhoria de processos de desenvolvimento de software a partir dos dados de proveniência de um processo.
Ponto de Vista / Perspectiva (quem irá utilizar os dados coletados)	Gerentes de processo e desenvolvedores de software.
Contexto (em qual contexto a análise será realizada?)	Projetos de desenvolvimento de software.

Assim, esta avaliação consiste em: *Analisar* ***a abordagem PROV-Process*** *com a finalidade de* ***caracterizá-la*** *com respeito* ***à viabilidade para apoiar a tomada de decisão para a melhoria de processos de desenvolvimento de software, a partir dos dados de proveniência de um processo,*** *do ponto de vista de* ***gerentes de processo e desenvolvedores de software,*** *no contexto de processos de software*.

A experiência do responsável pelo processo foi levada em consideração na análise realizada, para verificar se o tempo de experiência do indivíduo influencia na análise das informações.

5.3 SELEÇÃO DOS INDIVÍDUOS

A realização deste estudo de caso contou com a colaboração de 10 participantes voluntários. Dentre os participantes estão estudantes do programa de mestrado em Ciência da Computação da Universidade Federal de Juiz de Fora, uma doutoranda em Engenharia de Sistemas e Computação da Universidade Federal do Rio de Janeiro, com experiência em desenvolvimento de software. Dois gerentes de processos com experiência na área de gerência junto a empresas de desenvolvimento de software, onde exercem funções relativas a gerência de processos. Destes gerentes, ambos trabalham ou trabalharam[3] em uma das empresas parceiras utilizadas neste estudo, as quais cederam dados de execução de seus processos.

Cabe enfatizar que a escolha dos indivíduos considerou o conhecimento na área de desenvolvimento de software, haja vista que os dados utilizados pelo sistema se referem a dados de execução de processos de desenvolvimento de software. Sendo assim, todos os participantes possuem mestrado, ou estão em curso, na área de engenharia de software. Acredita-se que o conhecimento na área, aliado a experiência em gerência de processos de desenvolvimento de software, facilitará a compreensão das informações apresentadas pelo sistema, por parte dos participantes. Como dito anteriormente, os mesmos analisarão ambos os processos das empresas parceiras (Empresa I e Empresa II), a fim de identificar se o conhecimento do processo avaliado (no caso dos participantes que trabalham ou trabalharam em uma das empresas parceiras)

[3] Um dos participantes não trabalha mais na empresa.

interfere no resultado obtido, e também se é possível localizar as informações independente do processo que está em avaliação.

5.4 COLETA DE DADOS

A realização de um estudo de caso pode utilizar diferentes fontes de informação, tais como métodos diretos, indiretos e independente (LETHBRIDGE *et al.*, 2005). Nos métodos diretos, o pesquisador tem contato direto com os indivíduos participantes e a coleta é feita em tempo real, por meio de entrevistas, questionários, etc. Nos métodos indiretos, os pesquisadores coletam indiretamente os dados, através de interações dos indivíduos durante a coleta de dados, por meio de diários de trabalho, observação através de vídeos e áudio e *logs* do sistema. Já no independente, o pesquisador realiza a análise de artefatos de trabalho por meio de documentos.

De acordo com WOHLIN *et al.* (2012), as entrevistas e questionários podem ser divididas em não-estruturadas, semiestruturadas e totalmente estruturadas. As principais características de cada tipo de entrevista podem ser vistas na tabela 10.

Os registros para realização deste estudo foram coletados através de método direto. Esta coleta ocorreu por meio de entrevistas semiestruturadas, haja vista que o foco é a realização de análises do processo das empresas e do PROV-Process. Os formulários utilizados para captação dos dados a serem avaliados são compostos por perguntas abertas e fechadas, com o objetivo de identificar pontos referentes a experiência do usuário mediante o uso da ferramenta PROV-Process. As perguntas abertas visam detectar informações sobre o estudo de caso não previstas nas perguntas fechadas.

Tabela 10: Tipos de entrevistas [WOHLIN et al., 2012 *apud* SILVA, 2015]

	Não-estruturadas	Semiestruturadas	Totalmente estruturadas
Foco	Análise qualitativa a respeito da experiência dos indivíduos em relação a um	Análise qualitativa e quantitativa a respeito da experiência dos indivíduos em	Pesquisadores buscam encontrar relações entre dois fenômenos.

	fenômeno.	relação a um fenômeno.	
Questões	Um roteiro de entrevista englobando as áreas especificadas no foco da pesquisa.	Mistura entre questões abertas e fechadas.	Questões fechadas.
Objetivos	Exploratória	Descritiva e exploratória.	Descritiva e exploratória

5.5 MÉTRICAS

Visando atingir o objetivo descrito na subseção 5.2.1, foram definidas métricas a serem analisadas com base nas respostas obtidas, para cada questão previamente estabelecida. Nas questões 1 e 2, será utilizada a métrica *Precision* (ÁLVAREZ, 2007), a qual trata da precisão da informação retornada pelo participante, sendo calculada pela equação (1).

$$Precision = \frac{Número\ de\ Respostas\ Corretas}{Total\ de\ Respostas\ Dadas} \quad (1)$$

Já a questão 3 possui como métrica o tempo gasto pelo participante em cada tarefa, o qual é expresso em minutos.

Para os Formulários I e II (Apêndices III e IV), foram utilizadas as métricas apresentadas abaixo, de acordo com as questões a serem respondidas:

Questão 1: É possível identificar corretamente informações de proveniência de processos de software utilizando a abordagem PROV-Process?

- **Métrica:** Precisão da informação retornada pelo participante, equação (1).

Questão 2: É possível identificar corretamente informações para melhoria do processo de software a partir da utilização da abordagem PROV-Process?

- **Métrica:** Precisão da informação retornada pelo participante, equação (1).

Questão 3: Qual o esforço de cada participante para responder cada questão dos formulários I e II?

- **Métrica**: Tempo gasto em cada tarefa (expresso em minutos).

Para o Formulário II (Apêndice V), será avaliada a porcentagem de respostas obtidas em cada uma das questões para cada uma das cinco opções de resposta, sendo elas: Discordo totalmente, Discordo parcialmente, Indiferente, Concordo parcialmente, Concordo totalmente.

5.6 ANÁLISE DOS RESULTADOS

5.6.1 Análise do PROV-Process

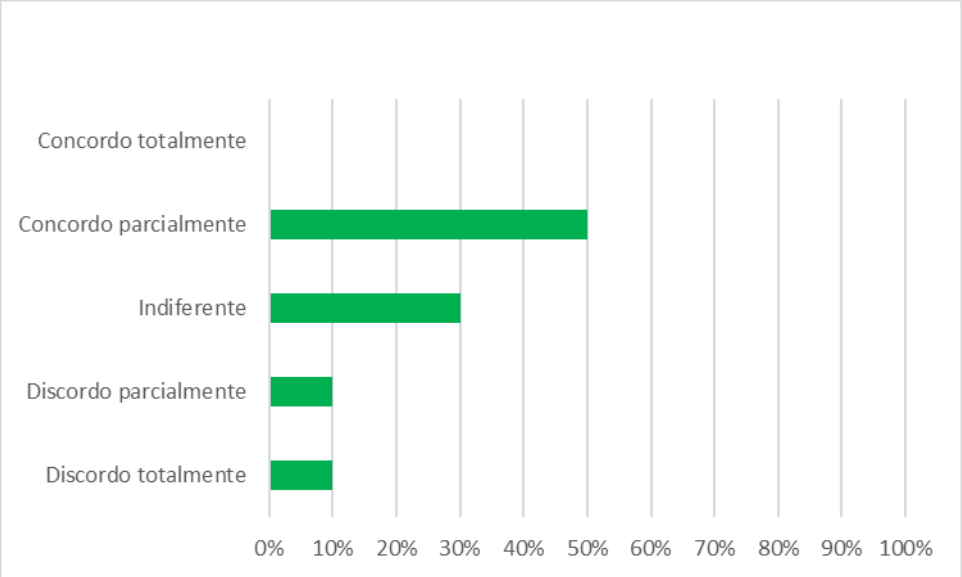
Por meio dos resultados obtidos mediante a aplicação da avaliação do PROV-Process (Formulário III – Apêndice V), foi possível identificar alguns aspectos acerca da utilidade da abordagem, os quais são descritos ao longo desta subseção.

Sobre o questionamento referente a *fácil identificação das informações relativas ao tempo de início, término e duração das instâncias*, todos os participantes indicaram que concordam totalmente com a afirmação. Esta concordância unânime indica que as respectivas informações encontram-se dispostas de forma adequada na ferramenta.

Em relação a afirmação "*As informações contidas no detalhamento da instância, auxiliam no entendimento do que ocorreu durante a execução do processo*", 60% dos participantes concordaram parcialmente e 40% concordaram totalmente. Este resultado indica que há uma convergência na concordância, porém, como o maior percentual dos participantes concordaram parcialmente, pode ser um indicativo de que nem sempre apenas estas informações serão suficientes para compreensão do processo como um todo. Cabe ressaltar que, nesta análise, os dois participantes que possuem experiência em gerência de processos de desenvolvimento de software concordam totalmente com a afirmação, o que denota que a experiência de ambos mostra que as informações contidas no detalhamento da instância são suficientes para compreensão do que ocorreu durante a execução do processo.

O gráfico 1 consolida as respostas para a afirmação de que "*As informações de ID, NOME e TIPO, de tarefas, pessoas envolvidas no processo e artefatos manipulados durante a execução de uma instância são suficientes para entendimento do processo*". Como apenas 50% dos participantes concordam parcialmente há indícios de que apenas estas informações talvez não sejam suficientes para entendimento do processo.

Gráfico 1 – Respostas da questão 3 (Formulário III – Apêndice V)



Quanto a questão sobre se "*as informações contidas no detalhamento de uma atividade auxiliam no entendimento do que ocorreu durante a execução do processo"*, 10% dos participantes discordaram parcialmente, 50% concordaram parcialmente e 40% concordaram totalmente. Os percentuais apontam para concordância em relação a afirmação, onde talvez sejam necessárias mais informações para compreensão do todo.

Em relação a questão "*As informações inferidas, contidas no detalhamento de uma atividade, apresentam novas informações acerca da atividade*", 10% dos participantes discordaram parcialmente, 10% indicaram indiferença, 40% concordaram parcialmente e 40% concordaram totalmente. O resultado indica que há uma concordância de que as inferências agregam informações para análise da atividade. Os 20% dos participantes que discordam parcialmente ou foram indiferentes, podem ser justificados por, nem sempre, ser possível apresentar novas informações com uso de inferências.

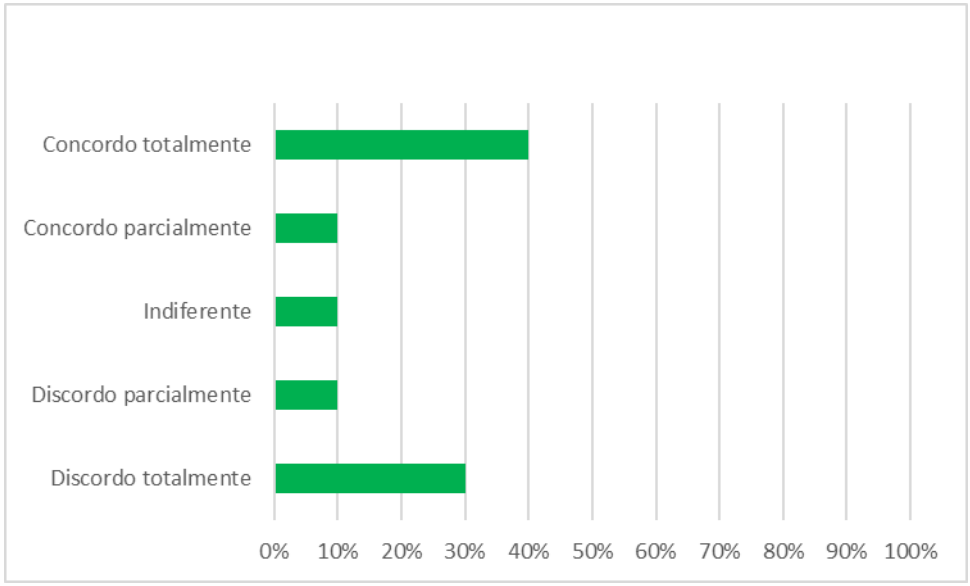
Mediante a afirmação de que as informações inferidas, contidas no detalhamento de um agente, apresentam novas informações acerca da participação do agente no processo, 10% dos

participantes discordaram parcialmente, 40% indicaram indiferença, 20% concordaram parcialmente e 30% concordaram totalmente. Assim como no resultado anterior, estes indicativos, principalmente o de indiferença, podem ser justificados pelo fato de nem sempre haverem informações novas apresentadas pelo uso de inferência.

Os resultados apresentados em relação a afirmação de que através da visualização gráfica é possível identificar, mais facilmente, as atividades, agente e entidades, mostram que 10% dos participantes discordam parcialmente, 40% concordam parcialmente e 50% concordam totalmente. Estes percentuais mostram que a visualização auxiliou os participantes. Nesta, cabe reiterar que, devido ao grande número de instâncias utilizadas foram exibidas muitas relações entre os nós, o que, de certa forma, retardou a localização das informações desejadas.

O gráfico 2 consolida as respostas à afirmação de que "*Através da visualização gráfica é possível identificar melhor as inferências obtidas por meio do uso da ferramenta PROV-Process*". Este resultado indica que a maior parte dos participantes teve dificuldade em visualizar as inferências. Isso deve-se ao fato de que, mediante ao grande número de instâncias utilizadas, a visualização não conseguiu apresentar, de forma clara, as inferências entre as relações dos nós do grafo gerado.

Gráfico 2 – Respostas da questão 8 (Formulário III – Apêndice V)



Em relação a afirmação de que a visualização gráfica possibilita uma análise mais rápida sobre os dados de execução de processos de desenvolvimento de software, 10% dos participantes discordaram parcialmente, 10% indicaram indiferença, 70% concordaram parcialmente e 10%

concordaram totalmente. O maior percentual indicando a concordância parcial, pode-se justificar pelo grande número de relações indicado anteriormente, o que dificulta a visualização do todo.

Com relação a afirmação de que a identificação de padrões relativos aos elementos que compõe o processo de desenvolvimento de software apresenta indícios significativos quanto a possíveis problemas do processo, 60% dos participantes concordaram parcialmente e 40% concordaram totalmente. Estes percentuais indicam que os padrões apresentados, de fato, podem auxiliar na identificação de problemas no processo.

Quanto a afirmação que indica que, por meio da identificação de padrões que culminam em tarefas de desdobramento de erros, é possível detectar a necessidade de melhoria para evitar novos erros, 30% dos participantes concordaram parcialmente e 70% concordaram totalmente. Este resultado novamente demonstra que a identificação de padrões que resultam em desdobramentos de erro facilita a identificação da necessidade de intervenção para melhoria do processo.

Em relação a afirmação de que quanto maior o percentual relativo ao número de vezes em que um conjunto de elementos resultou em um desdobramento de erro, mais forte o indício de problemas neste padrão, 10% dos participantes discordaram parcialmente, 40% concordaram parcialmente e 50% concordaram totalmente. Este resultado indica que os participantes compreenderam que quanto maior o percentual de ocorrências de determinado padrão, maior o indicativo de problema neste.

A Tabela 11 demonstra os resultados apresentados anteriormente de forma sintetizada.

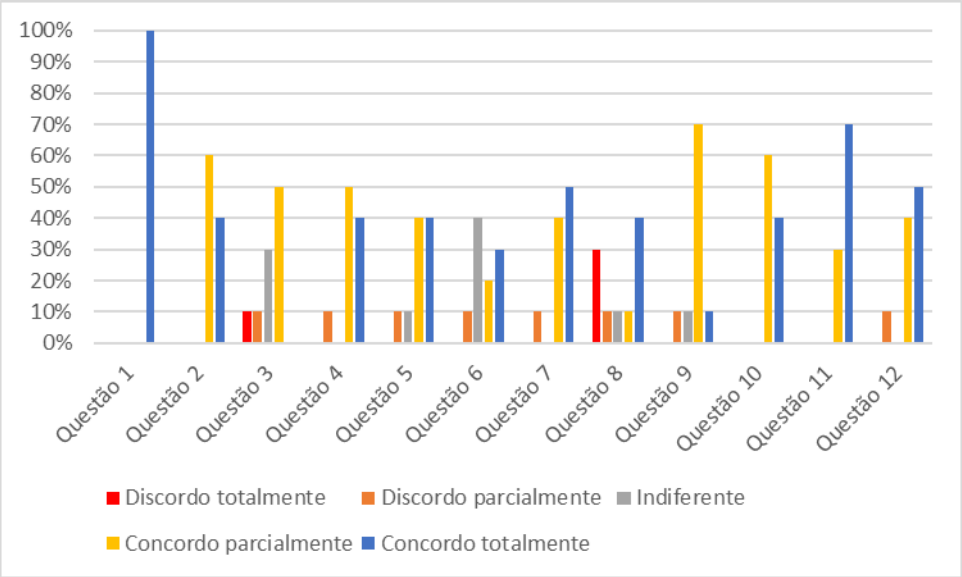
No Gráfico 3 é possível visualizar a comparação entre as respostas de cada questão, as quais encontram-se agrupadas em colunas. Neste é possível visualizar que há um maior percentual de participantes convergindo para concordância relativa as afirmações dispostas nas questões elaboradas para avaliação.

Tabela 11: Resultados da avaliação

	Discordo totalmente	**Discordo parcialmente**	**Indiferente**	**Concordo parcialmente**	**Concordo totalmente**
Questão 1	0%	0%	0%	0%	10%
Questão 2	0%	0%	0%	60%	40%
Questão 3	10%	10%	30%	50%	0%
Questão 4	0%	10%	0%	50%	40%
Questão 5	0%	10%	10%	40%	40%

Questão 6	0%	10%	40%	20%	30%
Questão 7	0%	10%	0%	40%	50%
Questão 8	30%	10%	10%	10%	40%
Questão 9	0%	10%	10%	70%	10%
Questão 10	0%	0%	0%	60%	40%
Questão 11	0%	0%	0%	30%	70%
Questão 12	0%	10%	0%	40%	50%

Gráfico 3 – Comparação de respostas da avaliação PROV-Process



Ao final da avaliação, conforme apêndice V, foram apresentadas perguntas dissertativas acerca da abordagem e da ferramenta PROV-Process. Analisando as respostas à pergunta "*O que mais gostou na abordagem PROV-Process*?" foi possível constatar que 50% dos participantes informaram que a identificação de padrões que culminam em desdobramentos de erro, por meio da mineração de dados, foi o que mais gostaram na abordagem PROV-Process, uma das respostas a essa pergunta foi: "A possibilidade de analisar desdobramentos de erros.".

Os outros 50% dos participantes, indicaram que a descoberta de novas informações por meio das inferências foi o que mais gostaram, conforme demonstra uma das respostas dadas por um dos participantes: "A descoberta por meio de inferências de informações que a análise manual não mostraria.". Estes 50% de participantes que relataram que as inferências foram o que mais gostaram, também informaram que gostaram da visualização gráfica, conforme pode-se verificar em uma das respostas que contempla este indicativo: "Possibilidade de realização de inferências e

possibilidade de observar graficamente os usuários associados a tarefas.”. Foi citado ainda a organização e disposição dos dados no sistema por 30% dos participantes, conforme pôde-se verificar em algumas respostas: “A organização dos dados em tabelas de forma clara.”, “A separação dos dados de forma a facilitar a busca...”.

Em relação à questão que indaga o que menos os participantes gostaram na abordagem PROV-Process, 40% relataram alguma dificuldade na identificação de dados, outros 40% informaram a falta de alguns filtros e buscas, 10% indicaram a exibição de padrões que culminaram em desdobramento de erro, ocorridos isoladamente (até 1 ocorrência) e outros 10% indicaram a questão de usabilidade, navegabilidade do sistema.

Por fim, ao serem indagados sobre o que mudariam na ferramenta PROV-Process, 90% dos participantes indicaram a adição de algum filtro e/ou busca, 10% apontaram a implementação de função de ordenação das colunas das tabelas *Activity*, *Agent* e *Entitty*, no detalhamento das instâncias, outros 10% também indicaram a inserção de legendas na parte da visualização gráfica, e 10% também indicaram a implementação da função de zoom na parte da visualização gráfica. Também houve a indicação do acréscimo de informações sobre os agentes, tais como número de projetos trabalhados, quem poderia substituí-los e quem executou mais tarefas, por 10% dos participantes.

5.6.2 Análise do processo das empresas

Com base nas respostas obtidas mediante ao preenchimento dos formulários I e II (Apêndices III e IV), foi possível responder as questões propostas na subseção 5.5.

Os resultados da análise quantitativa serão dispostos no decorrer desta subseção, onde estes serão iniciados pelas respostas obtidas pelo formulário I, referente a empresa I, e na sequência serão apresentados os resultados obtidos de acordo com as respostas do formulário II, referente a empresa II. Cabe salientar que o resultado da avaliação será indicado por meio da precisão (equação (1)), onde valores próximos a 1 (um) indicam maior número de acertos dos participantes.

Empresa I

A Empresa I não forneceu dados de desdobramentos das tarefas, logo, não foi possível fazer uso da técnica de mineração de dados, haja vista que a mesma tem o intuito de identificar padrões que culminam em desdobramentos de erro, conforme reiterado na subseção 5.2.

Na primeira questão, onde os participantes foram indagados quanto ao período em que uma instância foi iniciada, todos acertaram a resposta, resultando em precisão 1, com tempo médio de resposta de 24 segundos. A questão 2, onde a pergunta referia-se ao período de finalização de uma determinada instância, também teve precisão 1 e tempo médio de resposta de 36 segundos. Ainda em relação a tempo, a pergunta 3 questionou o tempo de duração da instância, considerando sua data e hora de início até sua conclusão, onde novamente todos os participantes acertaram a resposta, tendo como precisão 1, com tempo médio de resposta de 18 segundos.

A fim de verificar se os participantes conseguiam utilizar a funcionalidade de ordenação das colunas da tabela referente as instâncias, foram elaboradas as questões 4, 5 e 6. Observou-se que todos os participantes acertaram a resposta da questão 4, com tempo médio de resposta de 18 segundos. O mesmo ocorreu em relação a questão 5, a qual também resultou em precisão 1, porém com tempo médio de resposta de 24 segundos. Já na questão 6, onde indagou-se a instância iniciada por último, 8 participantes responderam corretamente e 2 incorretamente, resultando em 0,8 de precisão com média de 18 segundos para resposta.

As questões de número 7 a 10, tinham como objetivo a identificação de informações na tela de detalhes da instância. A questão 7 indaga qual(is) o(s) nome(s) do(s) módulo(s) que foram afetados pela execução de determinada instância, onde todos os participantes acertaram a resposta, indicando precisão 1 e tempo médio de resposta de 54 segundos. Na questão 8 a pergunta sobre as pessoas/equipes que participaram da execução de uma determinada instância foi respondida corretamente por 4 participantes, sendo que os outros 6 responderam errado, resultando em uma precisão de 0,4, com tempo médio de resposta em 3 minutos e 12 segundos. Nota-se que as respostas incorretas devem-se ao fato de que nem todos os participantes utilizaram o filtro de exibição de todos os registros, sendo assim, os participantes que responderam incorretamente indicaram exatamente as pessoas/equipes constantes na primeira página de registros, enquanto os demais, por meio da exibição de todas as páginas, conseguiram identificar todas as pessoas/equipes. A questão 9, a qual apresenta pergunta sobre a realização de alguma tarefa de determinada instância por mais de uma vez, foi respondida corretamente por todos os

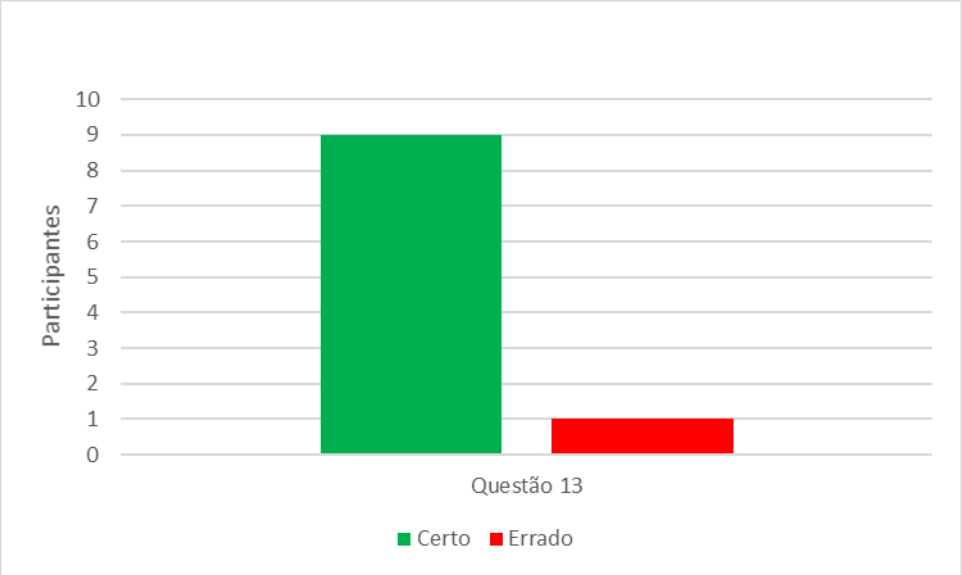
participantes, resultando em uma precisão de 1, com tempo médio de resposta de 1 minuto e 24 segundos. Finalizando as questões relativas aos detalhes da instância, a questão 10 indaga sobre o tipo de uma atividade, tendo sido respondida corretamente por todos os participantes, sendo assim, obtiveram precisão 1, com tempo médio de resposta de 18 segundos.

Dando prosseguimento a análise do formulário I, as perguntas 11 a 16 buscam verificar se os participantes conseguem identificar informações relativas as inferências. Na pergunta 11 questiona-se sobre quem executou uma dada tarefa de determinada instância, onde todos os participantes responderam corretamente a essa pergunta com tempo médio de resposta 1 minuto e 24 segundos.

Na pergunta 12 "*Quais módulos ou componentes foram manipulados durante a execução da atividade anterior?*" 7 participantes responderam corretamente e 3 incorretamente, precisão de 0,7, com tempo médio de resposta de 54 segundos. Os 3 participantes que erraram esta questão indicaram tanto o módulo quanto a versão, indicando que, possivelmente, não chegaram a verificar no sistema, na opção detalhes da instância, as entidades envolvidas nesta, as quais indicam o tipo das mesmas, o que auxiliaria na identificação da especificação das referidas.

A questão 13 "*Nesta mesma atividade, é possível identificar alguma inferência realizada pelo sistema que não conste nos dados que foram obtidos sem o mecanismo de inferência?*" obteve como resultado uma precisão de 0,9, em um tempo médio de resposta de 1 minuto e 18 segundos, conforme demonstrado no Gráfico 4.

Gráfico 4 - Respostas da questão 13 (Formulário I – Apêndice III)



A pergunta 14 refere-se a qual(is) módulo(s)/componente(s) foi influenciado por um determinado Agente, onde o baixo índice de 0,3 de precisão, em um tempo médio de 3 minutos e 24 segundos, apresenta indícios de desconhecimento dos participantes em relação ao que se refere a módulo, haja vista que 60% das respostas incorretas continham a descrição de todas as informações inferidas e as demais descartaram apenas a versão ou o cliente envolvido. Assim como ocorrido na questão 12, possivelmente, os participantes não chegaram a verificar, no detalhamento da instância, as entidades envolvidas nesta, as quais indicam o tipo das mesmas, o que auxiliaria na identificação da especificação das referidas.

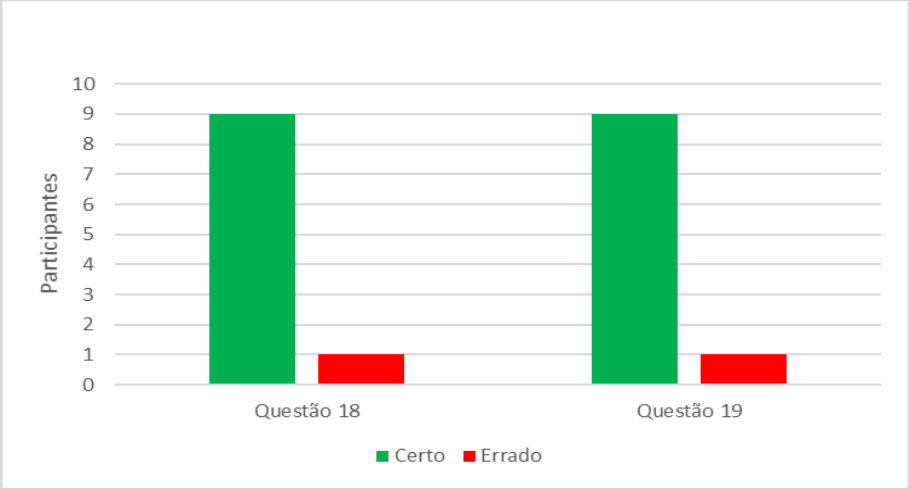
Na questão 15 pergunta-se se a informação da questão anterior pode ser obtida apenas consultando-se o detalhamento da instância, onde 6 participantes acertaram a resposta e o demais erraram, precisão de 0,6, com tempo médio de resposta de 6 segundos. Neste caso, as respostas incorretas podem ter ocorrido devido a exibição das informações corretas na tela de detalhamento das instâncias, o que não indica que estas foram influenciadas por um determinado agente, informação esta obtida apenas por meio do recurso de inferência no sistema.

Encerrando a verificação quanto a identificação das inferências pelos participantes, a pergunta 16 questiona o tipo de determinado agente e se haveria outro agente que pudesse substituí-lo durante a execução de determinada atividade. A precisão obtida de 0,9, com tempo médio gasto para resposta de 4 minutos e 18 segundos, indica que os participantes conseguiram identificar as informações, porém necessitaram de um tempo maior.

As perguntas finais do formulário I, que compreendem as questões 17 a 19, têm como objetivo verificar a utilização da visualização gráfica embutida no sistema PROV-Process. Desta forma, a questão 17 indaga qual pessoa ou equipe executou mais tarefas nas instâncias avaliadas, sendo obtida uma precisão de 0,4, em um tempo médio de resposta de 4 minutos e 48 segundos. As respostas incorretas, possivelmente, devem-se ao fato de que, devido ao elevado número de instâncias a exibição dos nós e suas respectivas relações não couberam na tela, sendo necessário arrastar um agente para o canto da tela para que outros agentes fossem visualizados. A pergunta 18 questiona sobre o agente que aparenta estar mais sobrecarregado, onde a precisão de 0,9, com tempo médio de resposta de 2 minutos e 6 segundos, indica que a maioria dos participantes conseguiu identificar este. Finalizando o formulário, a questão 19 indaga ao participante sobre as ações a serem adotadas mediante a identificação de um agente sobrecarregado, onde a precisão

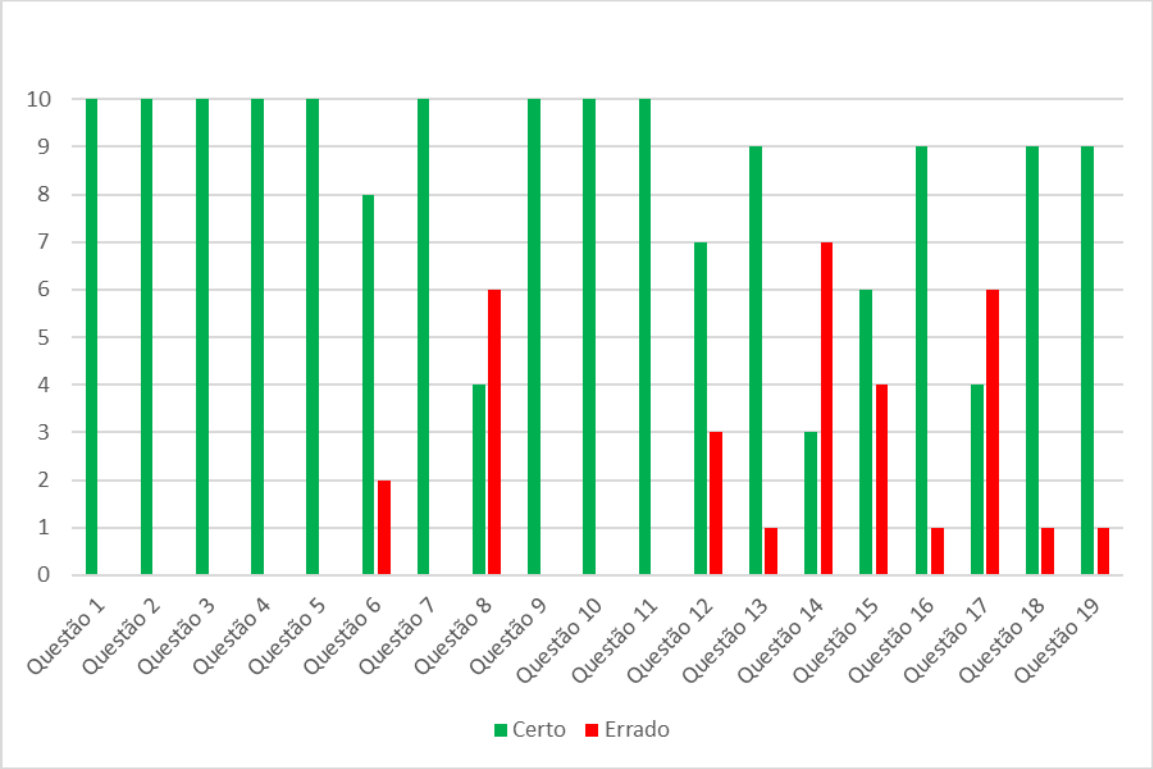
obtida foi a mesma da questão anterior, conforme demonstrado no Gráfico 5, com tempo médio de resposta de 2 minutos.

Gráfico 5 - Respostas das questões 18 e 19 (Formulário I – Apêndice III)



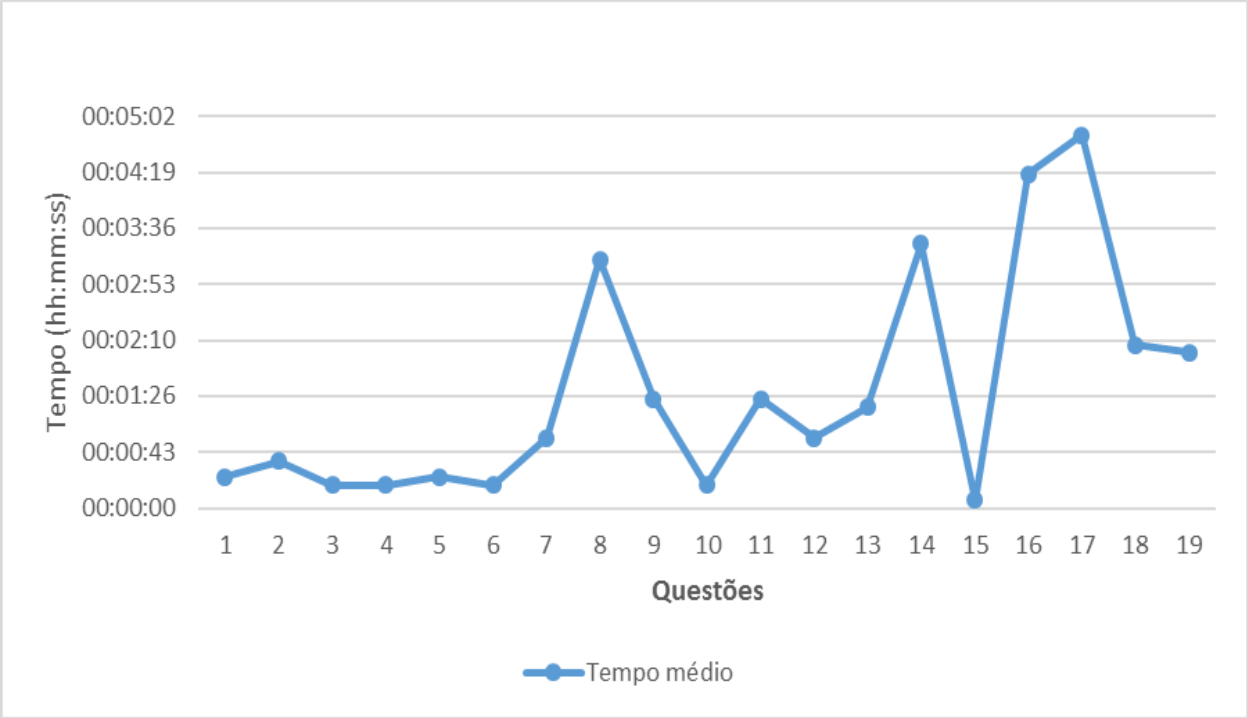
No Gráfico 6 é possível visualizar a comparação entre assertividade das respostas de cada questão, as quais encontram-se agrupadas em colunas. Neste é possível visualizar que há um maior percentual de acerto dos participantes.

Gráfico 6 - Assertividade das respostas para análise do processo da empresa I



O Gráfico 7 exibe o tempo médio gasto pelos participantes em cada questão, onde é possível notar que as questões em que a maior parte dos participantes errou a resposta, também coincide com a maiores médias de tempo gastos para respondê-las.

Gráfico 7 – Tempo médio gasto pelos participantes para responder as questões da análise do processo da empresa I



A Tabela 12 demonstra os resultados descritos anteriormente de forma sintetizada. A coluna *métricas* compreende as questões apresentadas na respectiva seção e a coluna tempo médio corresponde ao tempo médio gasto pelos participantes para responder à questão.

Tabela 12: Resultados da avaliação do formulário I

Empresa I						
Questões do Estudo de Caso	**Perguntas**	**Certo**	**Errado**	**Participantes**	**Precisão**	**Tempo médio**
Questão 1	Questão 1	10	0	10	1	00:00:24
	Questão 2	10	0	10	1	00:00:36
	Questão 3	10	0	10	1	00:00:18
	Questão 4	10	0	10	1	00:00:18
	Questão 5	10	0	10	1	00:00:24
	Questão 6	8	2	10	0,8	00:00:18
	Questão 7	10	0	10	1	00:00:54
	Questão 8	4	6	10	0,4	00:03:12
	Questão 9	10	0	10	1	00:01:24
	Questão 10	10	0	10	1	00:00:18
	Questão 11	10	0	10	1	00:01:24
	Questão 12	7	3	10	0,7	00:00:54
	Questão 13	9	1	10	0,9	00:01:18
	Questão 14	3	7	10	0,3	00:03:24
	Questão 15	6	4	10	0,6	00:00:06
Questão 2	Questão 16	9	1	10	0,9	00:04:18
	Questão 17	4	6	10	0,4	00:04:48
	Questão 18	9	1	10	0,9	00:02:06
	Questão 19	9	1	10	0,9	00:02:00

Empresa II

Os resultados referentes à Empresa II foram obtidos através das respostas dos participantes no formulário II (Apêndice IV), o qual segue o mesmo padrão de perguntas do formulário I, acrescido de questões relativas a mineração de dados, haja vista que a respectiva empresa forneceu dados referentes aos desdobramentos das atividades, possibilitando a utilização da função de mineração de dados, contida no sistema PROV-Process.

Com intuito de verificar se os participantes conseguiriam identificar as informações apresentadas na tela de instâncias, foram elaboradas as questões de 1 a 3. Na primeira questão, onde os participantes foram indagados quanto ao período em que uma instância foi iniciada, todos acertaram a resposta, resultando em precisão 1, com tempo médio de resposta de 36 segundos. A questão 2, onde a pergunta referia-se ao período de finalização de uma determinada

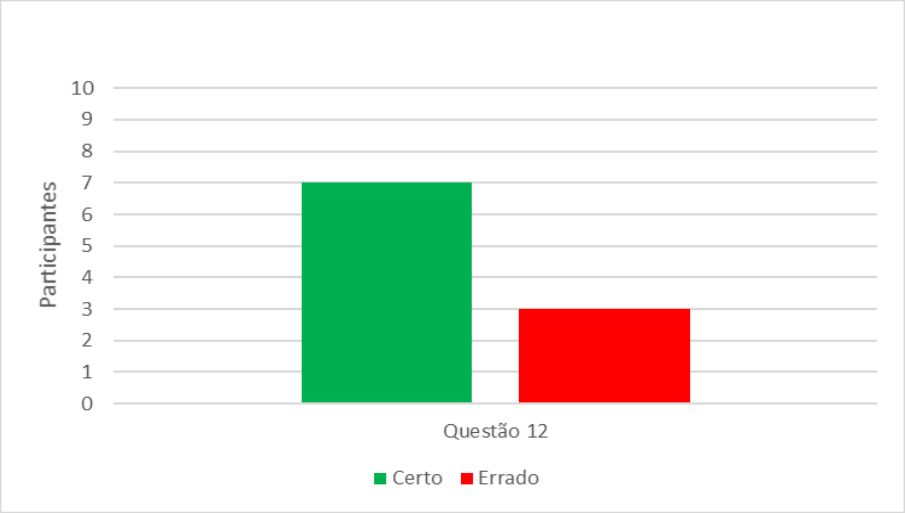
instância, também foi respondida assertivamente por todos os participantes, tendo precisão 1 e tempo médio de resposta de 6 segundos. Ainda em relação a tempo, a pergunta 3 questionou o tempo de duração da instância, considerando sua data e hora de início até sua conclusão, onde novamente todos os participantes acertaram a resposta, tendo como precisão 1, com tempo médio de resposta de 18 segundos.

A fim de verificar se os participantes conseguiam utilizar a funcionalidade de ordenação das colunas da tabela referente as instâncias, foram elaboradas as questões 4, 5 e 6. Observou-se que todos os participantes acertaram a resposta das questões 4 e 5, obtendo precisão 1, em um tempo médio de resposta de 12 segundos e 18 segundos, respectivamente. Na questão 6, onde indagou-se a instância iniciada por último, também todos os 10 participantes responderam corretamente, precisão de 1, sendo gastos em média 12 segundos para resposta.

As questões de número 7 a 10, tinham como objetivo a identificação de informações na tela de detalhes da instância, onde em todas obteve-se precisão de 1, o que apresenta indícios quanto a facilidade de identificação das informações na referida tela. Vale ressaltar que na questão 8, sobre as pessoas/equipes que participaram da execução de uma determinada instância, a assertividade das respostas deve-se ao fato de que todos os participantes, em algum momento da avaliação do Formulário I, utilizaram o filtro de exibição de todos os registros, sendo assim, por meio da exibição de todas as páginas, eles conseguiram identificar todas as pessoas/equipes. Os tempos médios gastos para responder cada questão encontram-se dispostos na Tabela 13.

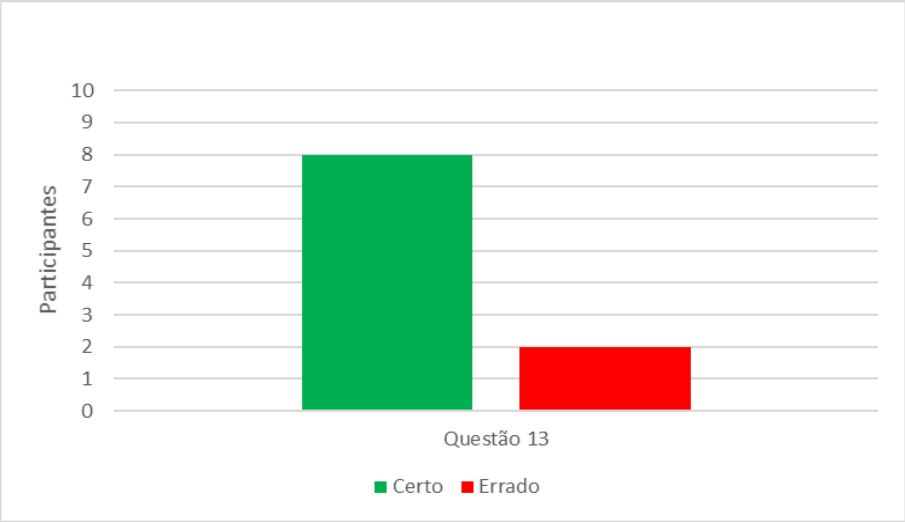
Dando prosseguimento a análise do formulário II, as perguntas 11 a 16 buscam verificar se os participantes conseguem identificar informações relativas as inferências. Na pergunta 11 questiona-se sobre quem executou uma dada tarefa de determinada instância, sendo obtida precisão de 1, com tempo médio de resposta 1 minuto e 18 segundos. A pergunta 12 indaga sobre quais módulos ou componentes foram manipulados durante a execução da tarefa anterior, onde obteve-se uma precisão de 0,7, conforme pode-se visualizar no Gráfico 8, com tempo médio de resposta de 1 minuto e 42 segundos.

Gráfico 8 - Respostas da questão 12 (Formulário II – Apêndice IV)



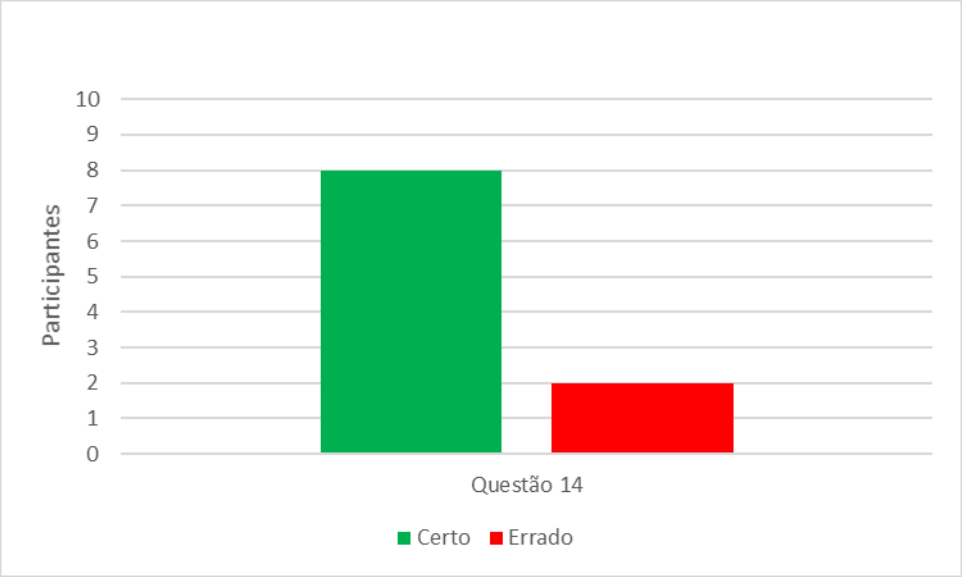
Os 3 participantes que erraram esta questão indicaram tanto os módulos quanto versões e atividades, o que indica que eles, possivelmente, não chegaram a verificar, no detalhamento da instância, as entidades envolvidas nesta, as quais indicam o tipo das mesmas, o que auxiliaria na identificação da especificação das referidas. A questão 13 indaga se na tarefa anterior é possível identificar alguma inferência realizada pelo sistema que não conste nos dados que foram obtidos sem o mecanismo de inferência, onde obteve-se uma precisão de 0,8, conforme demonstrado no Gráfico 9, com tempo médio de resposta de 42 segundos.

Gráfico 9 - Respostas da questão 13 (Formulário II – Apêndice IV)



A pergunta 14 refere-se a qual(is) módulo(s)/componente(s) foi influenciado por um determinado Agente em uma dada instância, a qual obteve precisão de 0,8, conforme demonstrado no Gráfico 10, em um tempo médio de 1 minuto e 54 segundos.

Gráfico 10 - Respostas da questão 14 (Formulário II – Apêndice IV)

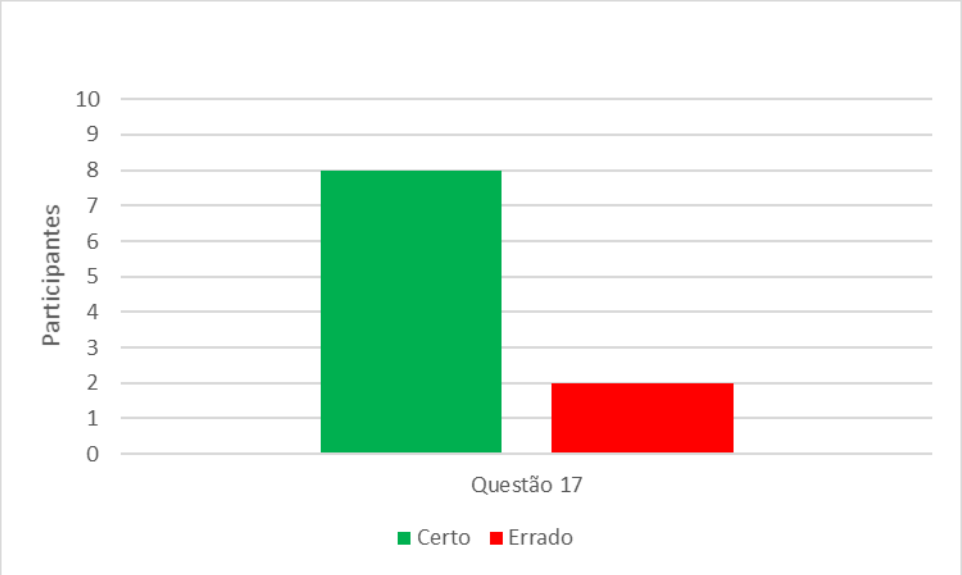


Este resultado demonstra que após a utilização e entendimento do sistema os participantes conseguiram responder à questão, verificando no detalhamento da instância, as entidades envolvidas nesta, as quais indicam o tipo das mesmas, possibilitando a identificação da especificação das referidas. Na questão 15 pergunta-se se a informação da questão anterior pode ser obtida apenas consultando-se o detalhamento da instância, onde 7 participantes acertaram a resposta e os demais erraram, indicando uma precisão de 0,7, com tempo médio de resposta de 36 segundos. Novamente nota-se um aumento da assertividade dos participantes, reafirmando indícios de que a utilização e entendimento do sistema auxilia os participantes na localização das informações desejadas. Encerrando a verificação quanto a identificação as inferências pelos participantes, a pergunta 16 questiona o tipo de determinado agente. A resposta a esta questão obteve precisão de 1, com tempo médio gasto para resposta de 6 segundos.

As perguntas que compreendem as questões 17 a 19, tem como objetivo verificar a utilização da visualização gráfica embutida no sistema PROV-Process. Desta forma, a questão 17 indaga qual pessoa ou equipe executou mais tarefas nas instâncias avaliadas, com resultado de

precisão 0,8, conforme pode-se verificar no Gráfico 11, em um tempo médio de resposta de 3 minutos e 36 segundos.

Gráfico 11 - Respostas da questão 17 (Formulário II – Apêndice IV)

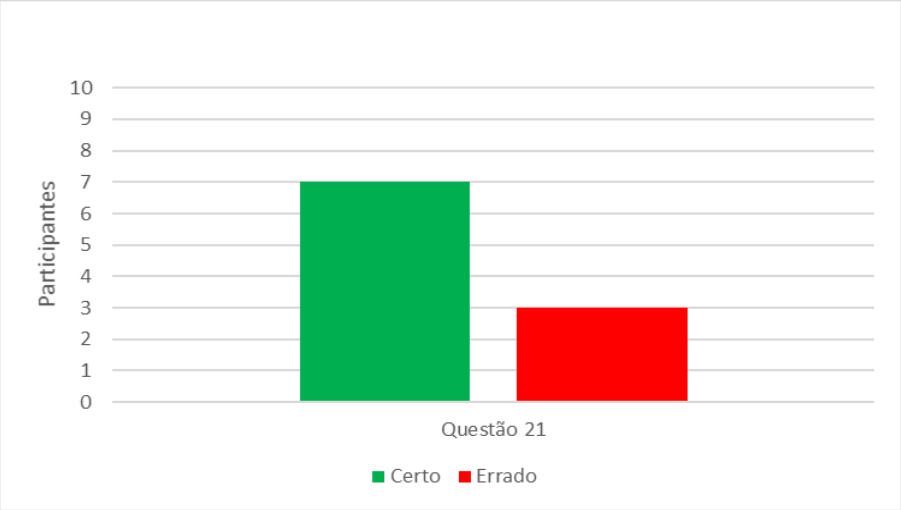


O aumento significativo da assertividade dos participantes nesta questão, justifica-se pelo fato de haverem apenas dois agentes. O número reduzido de agentes é devido aos dados fornecidos pela empresa II, a qual disponibilizou apenas o nome das equipes, sem a especificação dos integrantes destas, o que diminui drasticamente o número de agentes. Com um número menor de agentes, a visualização das atividades ligadas as estes foi mais fácil. A pergunta 18 questiona sobre o agente que aparenta estar mais sobrecarregado, onde todos os participantes responderam corretamente, indicando uma precisão de 1, com tempo médio de resposta de 30 segundos. A questão 19 indaga ao participante sobre as ações a serem adotadas mediante a identificação de um agente sobrecarregado, onde 9 participantes indicaram a mesma ação e apenas 1 divergiu desta. A precisão das respostas a essa questão foi de 0,9, com tempo médio de resposta de 36 segundos.

Finalizando o formulário II, as perguntas que compreendem as questões 20 a 23, visam avaliar a utilização da mineração de dados na descoberta de padrões que ocasionem o desdobramento de erros. A questão 20 indaga justamente sobre a possibilidade de identificação destes padrões, onde todos os participantes responderam corretamente, indicando uma precisão de 1, com tempo médio de resposta de 2 minutos e 18 segundos. Na questão 21 pergunta-se o percentual de um determinado padrão, formado por uma dada atividade e um determinado agente,

que resultaram em desdobramentos de erro, tendo como resultado uma precisão de 0,7, conforme demonstrado no Gráfico 12, com tempo médio gasto para responder à questão de 3 minutos.

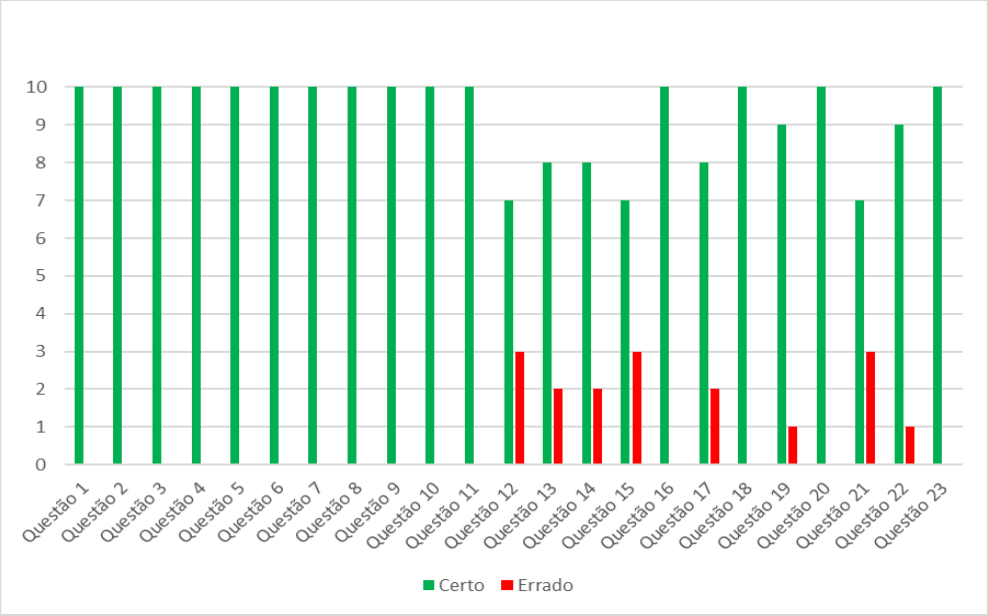
Gráfico 12 - Respostas da questão 21 (Formulário II – Apêndice IV)



A questão 22 indaga se, com base no percentual observado na questão anterior, é possível identificar a necessidade de adotar alguma ação, sendo respondida corretamente por 9 participantes. Isso indica uma precisão de 0,9, com um tempo médio gasto para responder a questão de 36 segundos. Por fim, a questão 23 indaga se através da identificação de um percentual alto de determinado padrão, é possível deduzir a necessidade de adoção de alguma medida para evitar novos desdobramentos de erro. Todos os participantes responderam corretamente a esta questão, resultando em uma precisão de 1, com um tempo médio de resposta de 54 segundos.

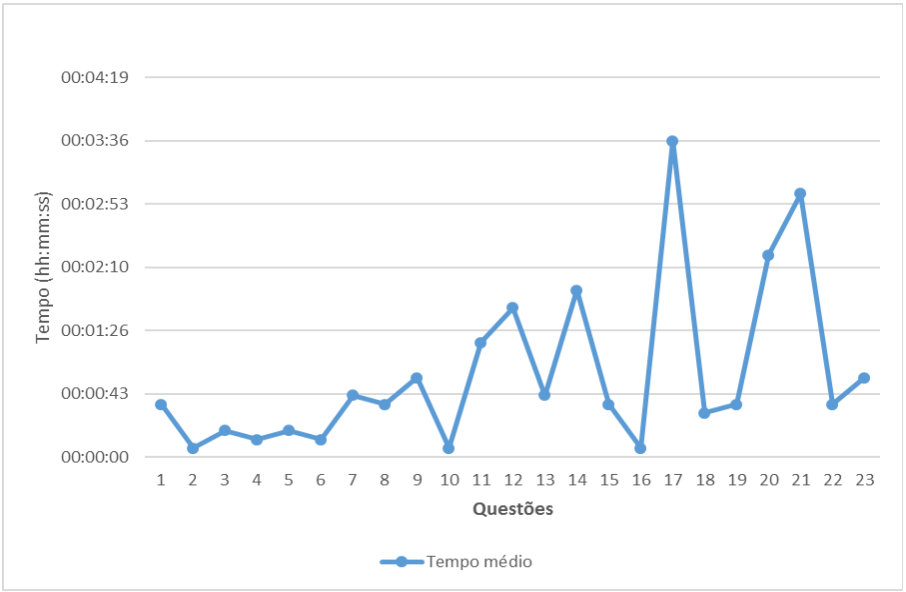
No Gráfico 13 é possível visualizar a comparação entre assertividade das respostas de cada questão, as quais encontram-se agrupadas em colunas. Neste é possível visualizar que há um maior percentual de acerto dos participantes em relação ao formulário I.

Gráfico 13 - Assertividade das respostas da análise do processo da Empresa II



O Gráfico 14 exibe o tempo médio gasto pelos participantes em cada questão, onde é possível notar a diminuição do tempo médio gasto para responder as questões do formulário II, em relação ao tempo médio registrado nas respostas do formulário I.

Gráfico 14 - Tempo médio gasto pelos participantes para responder as questões da análise do processo da empresa II



A Tabela 13 apresenta os resultados descritos anteriormente de forma sintetizada.

Tabela 13: Resultados da avaliação do formulário II

Empresa II						
Questões do Estudo de Caso	**Perguntas**	**Certo**	**Errado**	**Participantes**	**Precisão**	**Tempo médio**
Questão 1	Questão 1	10	0	10	1	00:00:36
	Questão 2	10	0	10	1	00:00:06
	Questão 3	10	0	10	1	00:00:18
	Questão 4	10	0	10	1	00:00:12
	Questão 5	10	0	10	1	00:00:18
	Questão 6	10	0	10	1	00:00:12
	Questão 7	10	0	10	1	00:00:42
	Questão 8	10	0	10	1	00:00:36
	Questão 9	10	0	10	1	00:00:54
	Questão 10	10	0	10	1	00:00:06
	Questão 11	10	0	10	1	00:01:18
	Questão 12	7	3	10	0,7	00:01:42
	Questão 13	8	2	10	0,8	00:00:42
	Questão 14	8	2	10	0,8	00:01:54
	Questão 15	7	3	10	0,7	00:00:36
Questão 2	Questão 16	10	0	10	1	00:00:06
	Questão 17	8	2	10	0,8	00:03:36
	Questão 18	10	0	10	1	00:00:30
	Questão 19	9	1	10	0,9	00:00:36
	Questão 20	10	0	10	1	00:02:18
	Questão 21	7	3	10	0,7	00:03:00
	Questão 22	9	1	10	0,9	00:00:36
	Questão 23	10	0	10	1	00:00:54

Os Gráficos 15 e 16 apresentam uma comparação entre as precisões das respostas obtidas junto a aplicação dos Formulários I e II, referentes as Empresas I e II, respectivamente.

Gráfico 15 – Comparativo de Precisão das perguntas semelhantes dos Formulários I e II, Questões 1 a 19

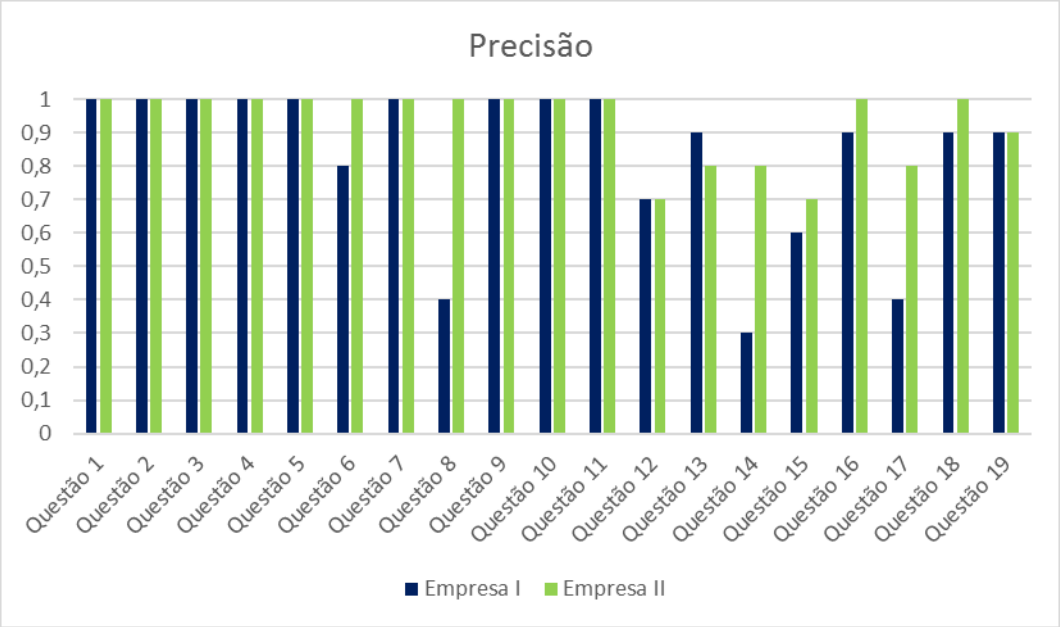
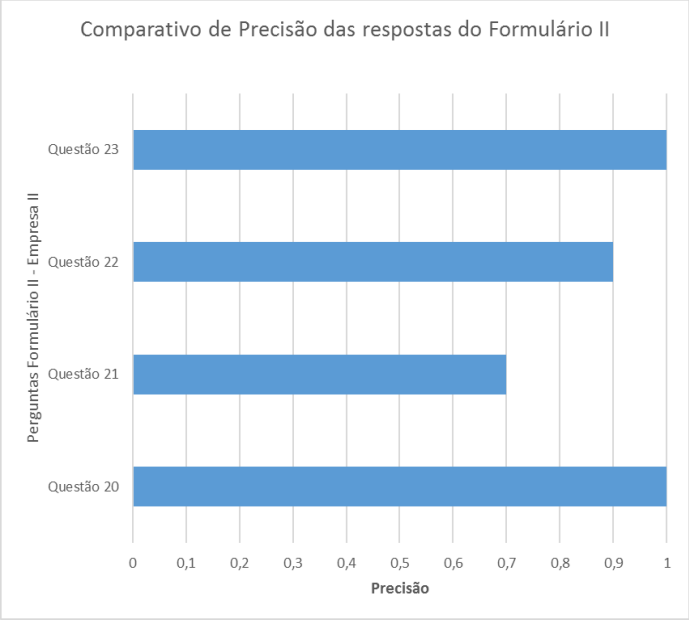


Gráfico 16 - Comparativo de Precisão das respostas do Formulário II, referentes a mineração de dados



5.7 AMEAÇAS A VALIDADE

Têm-se como ameaças a validade desse estudo o número de participantes e a participação de apenas 2 gerentes de processos, do total de 10 participantes. Destes, um trabalha na empresa parceira, sendo assim, conhece o processo atual da empresa. O outro gerente, já trabalhou na empresa parceira, continua atuando na área de desenvolvimento de software, porém em outra empresa. Do total de participantes, 4 possuem alguma experiência nas áreas afins do processo de desenvolvimento de software.

O desconhecimento dos processos de desenvolvimento de software e dos produtos da empresa parceira, por parte da maioria dos participantes, também pode ser considerado uma ameaça a validade deste estudo.

Embora nem todos os participantes tenham experiência no processo de desenvolvimento de software, todos formaram-se ou estão cursando mestrado em ciência da computação pela Universidade Federal de Juiz de Fora. Sendo assim, conhecem os conceitos de ontologia e mineração de dados, o que pode ter facilitado a identificação das informações inferidas e/ou a compreensão das mesmas.

A não divulgação de todos os dados de execução dos processos de desenvolvimento de software das empresas parceiras, também é uma ameaça, haja vista que a falta de informações, tais como o nome dos agentes (caso da empresa I), pode culminar em uma análise indevida com relação ao processo.

A grande quantidade de instâncias utilizadas ocasionou um problema na ferramenta de visualização gráfica, a qual não exibiu as inferências. A não exibição destas informações pode ter ocasionado algumas respostas incorretas em relação as inferências. Esta inconsistência foi reportada ao aluno responsável pelo desenvolvimento da ferramenta, para ajuste devido.

Como análise final é importante ressaltar que estes estudos de caso são válidos apenas neste contexto específico. Assim, não é possível generalizar os resultados para outros contextos ou mesmo para outros problemas relacionados a outras áreas do desenvolvimento de software.

Desta forma, é necessário planejar e conduzir avaliações adicionais, em outros contextos reais, para estender a validade das análises realizadas ao longo dos estudos de caso. Estudos adicionais, considerando questões não abordadas no presente estudo, podem trazer novas

evidências não observadas. Neste contexto, um maior número de participantes pode revelar novos aspectos não tratados.

É também importante ressaltar ainda que as conclusões aqui apresentadas significam que os resultados podem ser tratados como evidências preliminares dos benefícios de se usar a abordagem PROV-Process, em conjunto com seu ferramental de apoio. Assim, devemos destacar que a validade das conclusões deste estudo depende da replicação do experimento em outros contextos para assegurar a viabilidade da abordagem e aumentar a validade do estudo. No entanto, apesar dos resultados não poderem ser generalizados, é possível identificar situações em que resultados similares podem ser alcançados.

5.8 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou a avaliação da abordagem PROV-Process. Através dos resultados obtidos com a aplicação dos formulários é possível responder as questões definidas como métricas, as quais encontram-se dispostas na sequência.

Em resposta a questão 1 pode-se afirmar que, de acordo com os resultados obtidos na aplicação dos formulários I e II, sim, é possível identificar corretamente informações de proveniência de processos de software utilizando a abordagem PROV-Process. Essa questão foi contemplada nas 16 primeiras perguntas dos formulários I e II, onde obteve-se como resultado uma precisão média de 0,89, o que indica um alto índice de assertividade dos participantes em relação a identificação das informações de proveniência.

Em resposta a questão 2, também pode-se afirmar que, de acordo com os resultados obtidos na aplicação dos formulários I e II, sim, é possível identificar corretamente informações para melhoria do processo de software a partir da utilização da abordagem PROV-Process. Essa questão foi contemplada nas 3 últimas questões do formulário I (questões 17, 18 e 19) e nas 7 últimas questões do formulário II (questões 17, 18, 19, 20, 21, 22 e 23), onde obteve-se como resultado uma precisão média de 0,85, o que indica um alto índice de assertividade dos participantes em relação a identificação de informações que podem auxiliar na melhoria do processo.

Em relação ao esforço de cada participante para responder cada questão dos formulários I e II, constatou-se que a maior parte dos participantes responderam as questões em menos de 1

minuto e 30 segundos. Trata-se de um tempo baixo considerando-se a necessidade de análise de muitos dados para obtenção das respectivas informações.

Também foi possível constatar que a experiência do indivíduo não influencia na análise realizada junto às informações que este conseguiu extrair da ferramenta, haja vista que o índice médio de precisão dos participantes que possuem alguma experiência em processos de desenvolvimento de software foi de 0,82, correspondente aos demais participantes que também obtiveram 0,82 de precisão.

Através da análise das respostas obtidas na avaliação dos processos das empresas parceiras (Empresa I e Empresa II), foi possível identificar que o conhecimento do processo avaliado (no caso dos participantes que trabalham ou trabalharam em uma das empresas parceiras) não interfere, significativamente, no resultado obtido, haja vista que o índice médio de assertividade dos participantes que conhecem o processo de uma das empresas foi de 84%, enquanto os demais participantes obtiveram índice médio de 83%.

Por fim, também foi possível concluir, de acordo com o resultado do estudo realizado, que a utilização e conhecimento da ferramenta impacta na identificação das informações, haja vista que o índice médio de assertividade do formulário I, realizado primeiramente, foi de 83%, enquanto o índice médio do formulário II, realizado após a utilização da ferramenta para resposta ao formulário I, foi de 93%. Este resultado também indica que, nem sempre, a identificação das informações independerá do processo que está em avaliação, haja vista que um processo pode conter informações mais claras que outro.

6 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as considerações finais desta dissertação, considerando a arquitetura PROV-Process e seus benefícios. Também são apresentadas as principais contribuições deste trabalho, suas limitações e os trabalhos futuros.

6.1 VISÃO GERAL

Esta pesquisa apresentou uma proposta de arquitetura, denominada PROV-Process, a qual é capaz de importar dados de execução de processos e, por meio de ontologias e técnicas de mineração de dados, apresentar, ao gerente de projetos, indicativos de melhorias nos processos.

Os resultados obtidos mediante a realização do estudo de caso, apresentado na seção 5, mostram a viabilidade de aplicação da proposta apresentada. Estes resultados também apontam indícios de que "*O armazenamento e posterior análise dos dados de proveniência de processos de software, utilizando um modelo de proveniência de dados, Ontologias e técnicas de mineração de dados, são capazes de oferecer informações adicionais sobre o processo analisado, facilitando a tomada de decisão por parte do gerente de projeto sobre ações de melhoria para a execução das próximas instâncias do mesmo"*, embasando a hipótese dessa dissertação.

Considerando os objetivos deste trabalho e os estudos de caso realizados, é relevante destacar que a arquitetura PROV-Process, que contempla o armazenamento dos dados de proveniência de processos e a posterior análise dos mesmos através do uso de uma Ontologia e técnica de mineração de dados apresenta indícios de que estes objetivos foram atendidos. Assim, algumas análises acerca destes objetivos podem ser feitas:

- Mediante a análise das propostas existentes na literatura, por meio da realização da revisão sistemática descrita no apêndice I, não foram encontrados trabalhos que englobem os conceitos e aplicações de proveniência de dados, ontologias e mineração de dados.
- O modelo PROV, utilizado como base para criação da camada de proveniência de dados desta proposta, possui uma vasta família de documentos, o que possibilitou a criação de um banco de dados relacional específico para processos de

desenvolvimento de software e que atende as especificações e restrições do referido modelo. Por meio deste, é possível obter informações relativas a proveniência de dados, as quais podem auxiliar no entendimento do processo, bem como indicar pontos críticos e/ou pontos onde pode-se melhorar o processo como um todo.

- O uso da ontologia e da mineração de dados auxilia a análise dos dados de proveniência capturados sobre o processo de desenvolvimento de software. Através destes, é possível obter novos conhecimentos sobre o processo, o que pode vir a auxiliar o gerente de projetos na identificação da origem de determinados problemas ou ainda em pontos passíveis de melhoria para execução das próximas instâncias deste.
- A ferramenta PROV-Process, desenvolvida para apoiar a abordagem, apresenta ao gerente de projetos os resultados da análise dos dados de execução de processos, de forma simples e clara. Estes resultados apresentam informações importantes sobre o processo, não sendo necessário o conhecimento do gerente nas áreas de proveniência de dados, ontologia ou mineração de dados.
- A avaliação da abordagem mostra que é possível identificar informações de proveniência de dados de processo de desenvolvimento de software, bem como identificar informações para melhoria deste processo. A identificação destas informações independe da experiência do indivíduo em processos de desenvolvimento de software ou no processo específico analisado. A avaliação indicou ainda que o conhecimento da ferramenta impacta na identificação das informações, as quais podem ser mais claras dependendo do processo.

6.2 CONTRIBUIÇÕES

A abordagem apresentada neste trabalho traz aspectos interessantes, tal como a utilização conjunta de um modelo de proveniência de dados, uma ontologia e técnica de mineração de dados, com o objetivo de capturar proveniência de dados de execução de processos e analisá-los por meio da ontologia e mineração de dados, gerando para os gerentes de projetos, indicativos para melhoria das próximas execuções de instâncias destes processos.

O desenvolvimento da arquitetura PROV-Process, da ferramenta e do estudo de caso mostraram a viabilidade técnica e a aceitação dos usuários em relação as novas informações e

conhecimento gerados sobre os processos analisados. A obtenção de informações resultantes da análise dos dados de execução de processos independe do conhecimento nas áreas de proveniência de dados, ontologia ou mineração de dados, por parte do gerente de projetos.

Os resultados das análises dos dados de execução de processos podem auxiliar os gerentes de projetos na tarefa de encontrar pontos de melhoria destes em um tempo menor, mais facilmente e com maior eficácia, o que pode proporcionar a otimização dos processos, possivelmente resultando em maior produtividade e melhor qualidade do produto desenvolvido.

A forma de captura e análise de dados de proveniência deste trabalho, apresenta uma nova perspectiva de uso de proveniência de dados, mostrando a possibilidade de utilizar proveniência de dados, ontologia e mineração de dados, em um contexto real de processos de desenvolvimento de software.

A apresentação das informações dos processos de desenvolvimento de software na ferramenta PROV-Process, de forma mais detalhada e organizada, contribui para o melhor entendimento e visualização das partes que formam o processo, facilitando a identificação e a tomada de decisão para melhoria deste.

6.3 LIMITAÇÕES

A seguir são apresentadas algumas limitações tanto da ferramenta desenvolvida quanto da abordagem PROV-Process como um todo.

Considerando a ferramenta desenvolvida, a biblioteca utilizada para apoio ao uso da ontologia, denominada JENA, apresentou bons resultados, porém possui limitações, principalmente em relação ao volume de dados utilizados pelo sistema. Considerando o uso de um grande volume de dados, a ferramenta apresentou uma demora exacerbada para o carregamento da ontologia, culminando, em alguns momentos, no travamento da mesma.

A quantidade de dados também impacta na visualização gráfica disponível na ferramenta, haja vista que mediante a exibição de uma grande quantidade de vértices e relações, as inferências não são exibidas devidamente.

A necessidade de um hardware mais robusto, também pode ser considerada uma limitação, pois o carregamento da ontologia por meio da referida biblioteca necessita de um

maior processamento e memória, caso contrário, mesmo um pequeno volume de dados pode levar um tempo demasiado para carregamento da ontologia.

Na visualização da análise dos dados por meio da mineração de dados, a ocorrência única de um determinado padrão que culmina em desdobramento de erro pode exibir uma confiança (probabilidade condicional de que as combinações de atributos resultem no parâmetro a ser analisado) que não é interessante para análise, haja vista que, a frequência única de um dado padrão não indica a necessidade de averiguação deste para melhoria do processo. Neste caso, o algoritmo utilizado na mineração de dados indicará que 100% das vezes em que se identifica este dado padrão, tem-se um desdobramento de erro. Porém este alto percentual, na verdade, refere-se a uma única ocorrência.

Considerando a abordagem PROV-Process como um todo, a restrição de informações disponibilizadas pelas empresas parceiras pode ser também considerada uma limitação para análise dos dados de execução de processos de desenvolvimento de software. Acredita-se que quanto mais informações sobre os dados de execução do processo alimentarem o repositório de proveniência, melhores serão os resultados apresentados considerando a análise destes.

Além disso, o presente trabalho limitou-se a analisar dados de execução de processos de desenvolvimento de software. Sendo assim, o arquivo definido como padrão para importação pela ferramenta PROV-Process, baseou-se na modelagem dos respectivos processos, realizada juntamente aos gerentes destes processos. Portanto, não foram explorados outros modelos e processos.

6.4 TRABALHOS FUTUROS

A ampliação da proposta apresentada nesta dissertação para outros processos é uma sugestão de trabalhos futuros. A abordagem foi desenvolvida com possibilidade de adaptações em caso de expansão para outros tipos de processos. Será necessário considerar as especificidades destes outros processos, tais como, as informações que deseja-se obter por meio da abordagem.

Conforme apontado na avaliação realizada, seria de grande valia a criação de filtros na opção de visualização gráfica, haja vista que considerando um grande volume de vértices e relações, a visualização fica comprometida devido a falta de espaço na tela para exibir todas as informações.

Na visualização da análise dos dados por meio da mineração de dados, seria importante a implementação de filtros para consulta das informações apresentadas, tendo em vista que quanto maior a quantidade de dados armazenados, maior o volume de informações geradas mediante a análise destes.

A construção de novas Ontologias poderia resultar em novo conhecimento que englobe diferentes partes dos processos. Outra possibilidade para obtenção de novo conhecimento seria a exploração de outras técnicas de mineração de dados, que aplicadas ao processo, poderiam gerar novas informações sobre o mesmo, haja vista que no contexto deste trabalho, a mineração de dados foi utilizada para descoberta de uma informação específica sobre o processo, pois as demais informações desejadas foram obtidas por meio da ontologia.

Por fim, a própria abordagem usada neste trabalho pode ser refinada e melhorada, considerando, por exemplo, os resultados e respostas obtidas nas avaliações realizadas junto ao PROV-Process.

REFERÊNCIAS

ACUNA, S. T., DE ANTONIO, A., FERRE, X., *et al.*, 2000, "The Software Process: Modelling, Evaluation and Improvement", Handbook of Software Engineering and Knowledge Engineering, pp. 1–35.

AGRAWAL, R. and SRIKANT, R. 1994. Fast algorithms for mining association rules. VLDB-94, 1994.

ÁLVAREZ, Alberto Cáceres. Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem. USP - Universidade de São Paulo, 2007. Disponível em: <http://lakh.unm.edu/handle/10229/35211>.

ALVES, A. S., Regras de Associação e Classificação em Ambiente de Computação Paralela Aplicadas a Sistemas Militares. Diss. UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, 2007.

ALVES, T. O. M., 2013, SciProvMiner: Captura e Consulta de Proveniência utilizando Recursos Web Semânticos para Ampliação do Conhecimento Gerado e Otimização do Processo de Coleta. Dissertação de Mestrado, PGCC/UFJF, Juiz de Fora, MG, Brasil.

BARRETO, A. S., 2011, "Uma Abordagem para Definição de Processos baseada em Reutilização Visando à Alta Maturidade em Processos". Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

BASILI, V., CALDIERA, G., ROMBACH, H. (1994). "Goal Question Metric Paradigm", Encyclopedia of Software Engineering, v. 1, edited by John J. Marciniak, John Wiley & Sons, pp. 528-532.

BEEFERMAN D., BERGER A., & LAFFERTY J. Statistical models for text segmentation. Machine Learning, 34(1-3):177.210, 1999.

BELHAJJAME, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J. (2013) "PROV-O: The PROV Ontology". Available in <http://www.w3.org/TR/prov-o/>. Accessed in July 2015.

BELHAJJAME, K.; DEUS, H.; GARIJO, D.; KLYNE, G.; MISSIER, P.; SOILANDREYES, S.; ZEDNIK, S. Prov model primer. URL: http://www.w3.org/TR/provprimer, 2012.

BIVAR, B.; SANTOS, L.; KOHWALTER, T. C.; MARINHO, A.; MATTOSO, M.; BRAGANHOLO, V. Uma Comparação entre os Modelos de Proveniência OPM e PROV, In Proocedings of BRESCi 2013, Maceió, Brazil, 2013.

BIZAGI, 2015. Disponível em http://www.bizagi.com/. Acesso em: 23 jan 2015.

BONITASOFT, 2014. Bonita BPM, BPM de código aberto. Disponível em: http://br.bonitasoft.com/produtos/bonita-bpm-bpm-de-codigo-aberto. Acesso em: 23 jan 2015.

BRAUN, U., GARFINKEL, S., HOLLAND, D. A., MUNISWAMY-REDDY, K. K., SELTZER, M. I., "Issues in automatic provenance collection". In Proceedings of the 2006 international conference on Provenance and Annotation of Data (IPAW'06), Luc Moreau and Ian Foster (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 171-183, 2006.

BREIMAN, L, FRIEDMAN, J., OLSHEN, R. and STONE C. 1984. Classification and regression trees. Belmont: Wadsworth.

BRUSSO, M. J. Access Miner: Uma proposta para a Extração de Regras de Associação Aplicada à Mineração do Uso da Web. Master's thesis, PPGC da UFRGS, Porto Alegre - RS, 2000.

BUDGEN, D., TURNER, M., BRERETON, P., *et al.*, 2008, "Using Mapping Studies in Software Engineering". In: PPIG, pp. 195-204, Lancaster, UK.

BUNEMAN, P., Khanna, S. and Tan, W.C.. Why and where: A characterization of data provenance. In: 8th International Conference on Database Theory, London. p. 4-6, 2001.

CABENA, P; HADJINIAN, P.AND STADLER, R; VERHEES, J; ZANASI, A. Discovering Data Mining: From Concept to Implementation. Prentice Hall, 1997.

COSTA, G. C. B.; BRAGA, R. ; DAVID, J. M. N. ; CAMPOS, F. A Scientific Software Product Line for the Bioinformatics Domain. Journal of Biomedical Informatics, v. 56, p. 239-264, 2015.

DELEN, D.. "Real-World Data Mining: Applied Business Analytics and Decision Making". FT Press.

DEUTCH, D., MOSKOVITCH, Y., TANNEN, V., 2014, “A Provenance Framework for Data-Dependent Process Analysis”, 40th International Conference onVery Large Data Bases, September 1st - 5th 2014, Hangzhou, China.

DONG G., ZHANG X., WONG L., & LI J.. Classification by aggregating emerging patterns. In Discovery Science, December 1999.

ECLARUS, 2015. eClarus Business Process Modeler for Business Analysts. Disponível em http://eclarus-business-process-modeler-for-bus.software.informer.com/. Acesso em: 23 jan 2015.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, 39(11):27–34, November 1996.

FILHO, L. A. S., FAVERO, E. L.; DIAS, M. M.; MENDONCA, C. K. L. . Mineração de Regras de Associação em Dados e Textos: Uma Aplicação em Segurança Pública. In: Congresso Internacional de Gestão de Tecnologia e Sistemas de Informação, 2010, São Paulo. 7º CONTECSI. São Paulo : TECSI EAC FEA USP, 2010. v. 7. p. 227-227.

FLUIG, 2015. Disponível em http://www.fluig.com/. Acesso em: 23 jan 2015.

FOSTER, I., VÖCKLER, J., WILDE, M. and ZHAO, Y., "The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration," in CIDR, 2003.

FUGGETTA, A., 2000, “Software process: a roadmap”. In: Proceedings of the Conference on The Future of Software Engineering, pp. 25–34, Limerick, Ireland.

GARIJO, D., GIL, Y., 2014, “The OPMW-PROV Ontology”. Disponível em <http://www.opmw.org/model/OPMW/>. Acessado em set. 2014.

GHOSHAL, D., PLALE, B, 2013, “Provenance from log files: a BigData problem”. In Proceedings of the Joint EDBT/ICDT 2013 Workshops (EDBT '13). ACM, New York, NY, USA, 290-297. DOI=10.1145/2457317.2457366 http://doi.acm.org/10.1145/2457317.2457366

GREENWOOD, M., GOBLE, C., STEVENS, R., ZHAO, J., ADDIS, M., MARVIN, D., MOREAU, L., and OINN, T., "Provenance of e-Science Experiments - experience from Bioinformatics," in Proceedings of the UK OST e-Science second All Hands Meeting, 2003.

GROTH, P., MILES, S., MOREAU, L., 2009, “A model of process documentation to determine provenance in mash-ups”. ACM Trans. Internet Technol. 9, 1, Article 3 (February 2009). DOI=10.1145/1462159.1462162 http://doi.acm.org/10.1145/1462159.1462162

GROTH, P., MOREAU, L., PROV- Overview: An Overview of the PROV Family of Documents. Disponível em: <http://travesia.mcu.es/portalnb/jspui/bitstream/10421/7487/1/PROV-Overview.pdf>. Acessado em nov. 2015.

GUNTHER, C. W., RINDERLE, S., REICHERT, M., VAN DER AALST, W.: “Change mining in adaptive process management systems”. In On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE. Springer Berlin Heidelberg. 309--326 (2006)

GUNTHER, C. W., RINDERLE-MA, S., REICHERT, M., VAN DER AALST, W. M., RECKER, J.: “Using process mining to learn from process changes in evolutionary systems. International Journal of Business Process Integration and Management”, 3(1). 61--78 (2008)

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., & WITTEN, I. H. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, v. 11, n. 1, p. 10-18, 2009.

HAN, J; KAMBER, M., and PEI, J.,Data mining: concepts and techniques. Elsevier, 2011..

HOLMES, G., DONKIN, A., &WITTEN, I. H. (1994, December). Weka: A machine learning workbench. In Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on (pp. 357-361). IEEE.

INTALIO, 2014. Disponível em http://www.intalio.com/. Acesso em: 23 jan 2015.

JABREF, 2014, “JabRef reference manager”. Disponível em: <http://jabref.sourceforge.net/>.Acessado em ago. 2014.

JENA, Apache. "Apache jena." jena. apache. org [Online]. Available: http://jena. apache. org [Accessed: Jun. 02, 2016].

JOGLEKAR, G. S., GIRIDHAR, A., REKLAITIS, G., 2014, “A workflow modeling system for capturing data provenance”, Computers & Chemical Engineering, V. 67, 4 August 2014, pp. 148-158, ISSN 0098-1354. DOI=http://dx.doi.org/10.1016/j.compchemeng.2014.04.006.

JUNAID, M. M., BERGER, M., VITVAR, T., PLANKENSTEINER, K., FAHRINGER, T., 2009, “Workflow composition through design suggestions using design-time provenance information”, In Proceedings of the 5th IEEE International Conference on e-Science, December 09-11, 2009, Oxford, UK.

LANTER, D., "Design Of A Lineage-Based Meta-Data Base For GIS," in Cartography and Geographic Information Systems, vol. 18, 1991, pp. 255-261.

LEBO, T., SAHOO, S., MCGUINNESS, D., BELHAJJAME, K., CHENEY, J., CORSAR, D., ... &ZHAO, J. (2013). Prov-o: The prov ontology. W3C Recommendation,30.

LETHBRIDGE, Timothy C.; SIM, Susan Elliott; SINGER, Janice. Studying software engineers: Data collection techniques for software field studies. Empirical Software Engineering v. 10, p. 311–341 , 2005.1382-3256.

LIM T.-S.,. LOH W.-Y, & SHIH Y.-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning, 40(3):203. 228, September 2000.

LIM, C., LU, S., CHEBOTKO, A., FOTOUHI, F, “Prospective and Retrospective Provenance Collection in Scientific Workflow Environments”. In Proceedings of the 2010 IEEE International Conference on Services Computing (SCC '10). IEEE Computer Society, Washington, DC, USA, pp. 449-456, 2010.

LIU B., HSU W., & MA Y. Integrating classification and association rule mining. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), pp 27–31, 1998.

MAGDALENO, A. M., 2013, *COMPOOTIM: Planejamento, Acompanhamento e Otimização da Colaboração na Composição de Processos de Software*. Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

MARINHO, A. S. PROVMANAGER: Uma Abordagem para Gerenciamento de Proveniência de Experimentos Científicos, dissertação de mestrado, Engenharia de Sistemas e Computação, UFRJ/COPPE, Rio de Janeiro, RJ, Brasil, 2011.

MERETAKIS D., HONGJUN L., & WUTHRICH B. A study on the performance of large bayes classifier. In ECML, pp 271.279. LNAI, 2000.

MILES, S., GROTH, P., MUNROE, S., MOREAU, L., 2011, “PrIMe: A methodology for developing provenance-aware applications”. ACM Trans. Softw. Eng. Methodol. 20, 3, Article 8 (August 2011), 42 pages. DOI=10.1145/2000791.2000792 http://doi.acm.org/10.1145/2000791.2000792

MISSIER, P., DEY, S., BELHAJJAME, K., VICENTTÍN, V. C., LUDÄSCHER, B., 2013, “D-PROV: extending the PROV provenance model with workflow structure”, In *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance* (TaPP '13). USENIX Association, Berkeley, CA, USA.

MONTONI, M., 2007, “Uma Abordagem para Condução de Iniciativas de Melhoria de Processos de Software”, Exame de Qualificação para o Doutorado, Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ, Rio de Janeiro, Brasil.

MOREAU, L., FREIRE, J., FUTRELLE, J., MCGRATH, R. E., MYERS, J., PAULSON, P., “The Open Provenance Model: An Overview”. In *Provenance and Annotation of Data and Processes*, Juliana Freire, David Koop, and Luc Moreau (Eds.). Lecture Notes In Computer Science, Vol. 5272. Springer-Verlag, Berlin, Heidelberg pp. 323-326, 2008.

MOREAU, L.; CLIFFORD, B.; FREIRE, J.; FUTRELLE, J.; GIL, Y.; GROTH, P.; KWASNIKOWSKA, N.; MILES, S.; MISSIER, P.; MYERS, J. *et al.* The Open Provenance Model core specification (v1.1). Future Generation Computer Systems, v. 27 n. 6, p.743 –756, 2011.

MOREAU, L.; MISSIER, P. PROV-DM: The prov data model, 2013.

OMG, 2005, “Unified Modeling Language: Superstructure”, UML Superstructure Specification v2.0, formal/05-07-04, Object Management Group.

OMG, 2008, “Software & Systems Process Engineering Metamodel (SPEM)”. Disponível em: http://www.omg.org/spec/SPEM/. Acesso em: mar. 2015.

OMG, 2011, “Business Process Model and Notation (BPMN) Version 2.0”. Disponível em: http://www.omg.org/spec/BPMN/2.0/. Acesso em: mar. 2015.

OSTERWEIL, L. , 1987; "Software processes are software too", Prac. 9th 1m/. Conf Software Engíneering, IEEE Computer Society Press.

PAULK, M. C., 2009, A History of the Capability Maturity Model for Software, Technical Report, American Society for Quality (ASQ).

PEDRINACI, C., LAMBERT, D., WETZSTEIN, B., Van LESSEN, T., CEKOV, L., DIMITROV, M.: “SENTINEL: A Semantic Business Process Monitoring Tool”. In Proceedings of the First International Workshop on Ontology-supported Business Intelligence, New York, USA. 1--12. (2008)

PETERSEN, K., FELDT, R., MUJTABA, S., *et al.*, 2008, “Systematic Mapping Studies in Software Engineering”. In: 12th International Conference on Evaluation and Assessment in Software Engineering, Bari, Italy.

PETRINJA, E., STANKOVSKI, V., TURK, Z., 2007, “A Provenance Data Management System for Improving the Product Modelling Process”, Automation in Construction, v. 16, pp. 485-497.

PLALE, B., ALAMEDA, J., WILHELMSON, B., GANNON, D., HAMPTON, S., ROSSI, A., and DROEGEMEIER, K., "Active Management of Scientific Data," Internet Computing, 2005.

PROV-DM: O modelo de dados prov . Recomendação W3C, Abril de 2013. PROV-O Inverses. Disponível em http://www.w3.org/ns/prov-o-inverses. Acesso em: 20 mar 2016.

QUINLAN J. R.. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

RED HAT, 2015. jBPM - Open Source Business Process Management - Process Engine. Disponível em: http://www.jbpm.org/ Acesso em: jan. 2015.

REIS, C. A. L. Uma Abordagem Flexível para Execução de Processos de Software Evolutivos. [s.l.] Universidade Federal do Rio Grande do Sul, 2003.

REIS, C.A., 2003, Uma Abordagem Flexível para Execução de Processos de Software Evolutivos. Tese de Doutorado, PPGC/UFRGS, Porto Alegre, RS, Brasil.

SILVA, C. K. P., 2015, Extração de Características de Perfil e de Contexto em Redes Sociais para Recomendação de Recursos. Dissertação de Mestrado, PGCC/UFJF, Juiz de Fora, MG, Brasil.

SIMMHAN, Y. L. Provenance framework in support of data quality estimation. 2007. 350 f. Doctoral thesis, Indiana University, Indianapolis, USA, 2007.

SIMMHAN, Y. L., PLALE, B., GANNON, D., “A survey of data provenance in e-science”. SIGMOD Rec. 34, 3 (September 2005), pp. 31-36, 2005a.

SIMMHAN, Y. L., PLALE, B., GANNON, D., “A Survey of Data Provenance Techniques". Technical Report IUB-CS-TR618, Indiana University, Bloomington, 2005b.

SIMMHAN, Y. L., PLALE, B., GANNON, D., 2008, “Karma2: Provenance management for data driven workflows”, International Journal of Web Services Research, Idea Group Publishing, v. 5, 2008.

SOMMERVILLE, I., 2004, Software Engineering. 7 ed. Addison Wesley.

SOUZA, R. R., and ALVARENGA, L. (2004). “A Web Semântica e suas contribuições para a ciência da informação.” Ciência da Informação, Brasília,33(1), 132-141.

STEINMACHER, I., CHAVES, A. P., GEROSA, M. A., 2013, “Awareness Support in Distributed Software Development: A Systematic Review and Mapping of the Literature”. *Comput. Supported Coop. Work* 22, 2-3 (April 2013), pp. 113-158. DOI=10.1007/s10606-012-9164-4 http://dx.doi.org/10.1007/s10606-012-9164-4

TRAVASSOS, G.H., DOS SANTOS, P.S.M., NETO, P.G.M., et al., 2008, "An Environment to Support Large Scale Experimentation in Software Engineering". In: Engineering of Complex Computer Systems, 2008. ICECCS 2008. 13th IEEE International Conference on, pp. 193-202, Belfast, March.

VASCONCELOS, L. M. R. and Carvalho, C. L. "Aplicação de Regras de Associação para Mineração de Dados na Web."Brasil, Universidade Federal do Rio Grande do Sul(2004): 11-14.

VIEIRA, M. A., FORMAGGIO, A. R., RENNÓ, C. D., ATZBERGER, C., AGUIAR, D. A., & MELLO, M. P. (2012). “Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas”. Remote Sensing of Environment, 123, 553-562.

W3C, 2015 http://www.w3.org/blog/SW/2013/03/12/call-for-review-prov-family-of-documents-published-as-proposed-recommendations/. Acesso em: 23 jan 2015.

WENDEL, H., KUNDE, M., SCHREIBER, A., 2010, “Provenance of software development processes”, In Deborah McGuinness, James Michaelis, and Luc Moreau, editors, Provenance and Annotation of Data and Processes, volume 6378 of Lecture Notes in Computer Science, pp. 59–63. Springer Berlin / Heidelberg, 2010.

WFMC, 2002, Workflow Management Coalition Workflow Standard: Workflow Process Definition Interface – XML Process Definition Language (XPDL), Technical Report, Workflow Management Coalition, Lighthouse Point, Florida, USA.

WOHLIN, C et al. Experimentation in Software Engineering. [S.l.]: Springer Berlin Heidelberg, 2012.

WU, X., ZHU, X., WU, G. Q., and DING, W.. “Data mining with big data”. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

XIAOXIN Y. & JIAWEI H.. CPAR: Classification based on Predictive Association Rules. In SDM, 2003.

ZHAO, J., GOBLE, C., STEVENS, R. and BECHHOFER, S. "Semantically Linking and Browsing Provenance Logs for E-science," in ICSNW, 2004, pp. 158-176.

APÊNDICE I – REVISÃO SISTEMÁTICA

1. Protocolo de Pesquisa

Tem-se como objetivo deste estudo baseado em revisão sistemática: *Através do* ***objeto de estudo*** *proveniência de dados, a* ***intenção/propósito*** *é identificar técnicas, abordagens, métodos, metodologias e ferramentas que tenham como* ***efeito*** *uma melhoria de processos através do uso de proveniência de dados,* ***do ponto de vista de*** *pesquisadores, no* ***contexto*** *empresarial e acadêmico.*

Partindo-se do objetivo supracitado, foram elaboradas as questões de pesquisa apresentadas a seguir:

Questão 1: Como a Proveniência de Dados pode ser utilizada na melhoria de processos, independente do domínio de aplicação?

Questão 2: Como as vantagens obtidas no uso da proveniência auxiliam a melhoria de processos?

Questões 3:

- Como a confiabilidade dos dados de proveniência pode ser avaliada?
- Como o volume de dados de Proveniência em Processos pode ser controlado?

O foco nas vantagens obtidas no uso de proveniência

Para realização de buscas foram utilizadas as seguintes máquinas de busca: ACM, IEEE, ISI e Scopus. A base IEEE, mesmo não retornando os artigos de controle, foi utilizada no estudo pois, mediante a execução da *string* de busca, retornou alguns artigos considerados, a princípio, relevantes. O não retorno dos artigos de controle pela base IEEE justifica-se pelo fato dos mesmos não estarem indexados junto a mesma.

1.1 Método de Busca

Nesta *quasi* revisão sistemática utilizou-se a abordagem PICO (*Population, Intervention, Comparsion, Outcome Measure*) para organizar e estruturar a busca a ser realizada. Com base nesta abordagem, definiu-se:

- **População:** processos;

- **Intervenção:** proveniência de dados;
- **Comparação:** não se aplica;
- **Resultados:** técnicas, abordagens, métodos, metodologias e ferramentas que venham a auxiliar na melhoria de processos através do uso de proveniência de dados, apontando suas vantagens e apresentando formas de controle para avaliar a confiabilidade dos dados de proveniência.

Mediante a organização obtida com o uso da abordagem PICO, foram definidas palavras chave que embasaram a montagem da *string* de busca, conforme exibido na Tabela 14.

Tabela 14: Definição de palavras chave

CATEGORIA		PALAVRAS-CHAVE
P	**PROCESSOS**	**PROCESS DEFINITION, PROCESS COMPOSITION, PROCESS SELECTION, PROCESS ADAPTATION, PROCESS TAILORING, PROCESS CUSTOMIZATION, PROCESS DEVELOPMENT, PROCESS ENGINEERING, PROCESS IMPROVEMENT, PROCESS DESIGN, SOFTWARE PROCESS MODELING, PROCESS REENGINEERING, PROCESS IMPLEMENTATION, MANAGING PROCESSES**
I	**PROVENIÊNCIA DE DADOS**	**PROVENANCE, DATA PROVENANCE, PROVENANCE FOR DATA, PROVENANCE OF DATA**
C	-	-
O	**MELHORIA, VANTAGENS, CONTROLE, CONFIABILIDADE**	**IMPROVEMENT, QUALITY, ADVANTAGE, TRUST, CONTROL**

Através da junção das palavras chave apresentadas na tabela anterior foi definida uma *string* de busca inicial, a qual foi alterada e refinada gradativamente. Este refinamento, por conseguinte, diminuiu o número de artigos retornados, passando a atender o contexto desejado no respectivo trabalho. Após os refinamentos, a *string* adotada para esta revisão foi:

"processes" AND ("provenance" OR "data provenance" OR "provenance for data" OR "provenance of data") AND ("improvement" OR "quality" OR "advantage" OR "trust" OR "control")

Também foi considerado como critério para definição da string de busca a exibição dos artigos de controle, sendo estes listados a seguir:

- MILES, S., GROTH, P., MUNROE, S., MOREAU, L. PrIMe: A methodology for developing provenance-aware applications, ACM Transactions on Software Engineering and Methodology (TOSEM), v.20 n.3, p.1-42, August 2011.
- WENDEL, H., KUNDE, M., SCHREIBER, A. Provenance of software development processes. In: Deborah McGuinness, James Michaelis, and Luc Moreau, editors, Provenance and Annotation of Data and Processes, v. 6378 of Lecture Notes in Computer Science, p.59-63. Springer Berlin / Heidelberg, 2010.
- GHOSHAL, D., PLALE, B. Provenance from log files: a BigData problem, Proceedings of the Joint EDBT/ICDT 2013 Workshops, Genoa, Italy, 2013.

1.2 Procedimento de Seleção e Critérios

O procedimento para seleção dos artigos foi realizado em seis passos:

1. **Busca e Catalogação**: nesta etapa foram realizadas as buscas utilizando a *string* e as bases apresentadas na seção 2.1, os resultados retornados foram catalogados para análise posterior.
2. **Eliminação de artigos repetidos na mesma base (1° filtro):** para gerenciamento dos artigos retornados mediante a execução da *string* de busca, foi utilizada a ferramenta JabRef (JABREF, 2014), a qual foi escolhida por atender a critérios como: ser uma ferramenta de código aberto, funcionar em todos os sistemas operacionais, possibilitar importação de arquivos com extensões ‘bibtex’ e ‘txt’ e possuir exportação de bases nos formatos HTML, latex e csv. Para execução deste primeiro filtro foi realizada a importação dos resultados de cada base de forma separada, a fim de identificar repetições dentro do retorno apresentado pela própria base.
3. **Eliminação de artigos repetidos entre bases (2° filtro):** para diminuir a possibilidade de detectar artigos repetidos, mediante a aplicação dos filtros subsequentes, foi realizada a importação dos resultados de todas as bases de forma unificada, com o intuito de eliminar os artigos repetidos entre as bases.
4. **Seleção de artigos com base nos títulos (3° Filtro):** os trabalhos retornados após a aplicação do 3º filtro foram verificados com base no título, onde os que não

enquadravam-se no contexto da pesquisa foram eliminados. Como forma de evitar que a inclusão ou exclusão de um determinado artigo com base apenas no título fosse realizada apenas sobre a visão do autor deste trabalho, participaram deste 3º filtro 3 pesquisadores (dois estudantes de mestrado e um de doutorado). Após avaliação dos títulos dos artigos retornados, dois dos pesquisadores posicionaram-se em relação à inclusão ou exclusão dos mesmos nas referências a serem analisadas. Em caso de empate, o terceiro pesquisador efetivava a avaliação do respectivo título e compartilhava sua opinião, contribuindo com uma espécie de voto de minerva, que definia a inclusão ou exclusão do artigo da seleção realizada. O fato de haver um número ímpar de pesquisadores envolvidos foi propositalmente determinado para evitar a situação de empate na avaliação realizada.

5. **Seleção de artigos com base na leitura dos resumos (4º Filtro):** como a busca realizada através das *strings* de busca criadas são restritas ao aspecto sintático, isto não garante que todas as publicações selecionadas no passo anterior são úteis para o propósito deste estudo. Partindo deste princípio, os resumos dos artigos selecionados através da análise do título foram lidos e analisados por dois dos três pesquisadores participantes da seleção anterior. Assim como na seleção anterior, o terceiro pesquisador só foi acionado para definição dos casos de empate, onde os pesquisadores participantes tenham opinado de forma divergente sobre a inserção ou deleção de determinado artigo. Para a aplicação deste filtro foram utilizados os seguintes critérios de exclusão (CE):

- **CE1** – Publicações que não tratavam proveniência de dados voltadas a processos;
- **CE2** – Publicações voltadas especificamente à fluxos de trabalho científicos;
- **CE3** – Publicações não disponíveis para download, em sua forma completa, nas bibliotecas digitais, nem através de nenhum outro meio sem custos para o pesquisador;

De acordo com os critérios acima, as publicações obtidas nas máquinas de busca foram selecionadas nesta etapa apenas se NÃO se enquadrassem nestes critérios.

6. **Seleção das publicações relevantes com base na leitura completa do artigo (5º Filtro)**: Neste quinto e último filtro, as publicações selecionadas no passo anterior

foram lidas por completo e analisadas usando os mesmos critérios de exclusão do passo 2, haja vista que a análise apenas dos resumos não garante que a publicação será útil para o estudo apresentado neste artigo.

1.3 Procedimento para Extração e Armazenamento dos Dados

Conforme mencionado na Seção 2, foram utilizadas 4 bases de pesquisa para realização deste estudo, sendo respectivamente ACM, ISI, Scopus e IEEE. Em cada uma destas bases foi realizada a pesquisa utilizando a *string* de busca apresentada na Subseção 2.1 (sendo feitas as devidas adaptações de acordo com a base em uso). Os artigos retornados, mediante a execução da pesquisa, foram exportados, para formatos compatíveis com a ferramenta de gerenciamento de publicações utilizada, e em seguida importados junto ao JabRef. Cada uma das bases utilizadas neste estudo possui suas particularidades, as quais podem tornar o processo mais ágil ou então acarretar em morosidade na exportação dos resultados. A seguir são listados os procedimentos de busca realizados em cada base, bem como os contratempos ocorridos em meio a utilização das mesmas.

- ACM: para realização da busca foi utilizada a opção “Busca avançada”. Junto ao campo exibido foi inserida a *string* de busca, conforme é possível visualizar na Figura 1. Ao verificar as opções para exportação dos resultados foi identificada a necessidade de exportar individualmente os artigos retornados. Devido ao volume de artigos, tal processo foi demorado, tanto para exportar da base quanto para importar no JabRef, tendo em vista que em ambos os casos o processo foi realizado um a um.

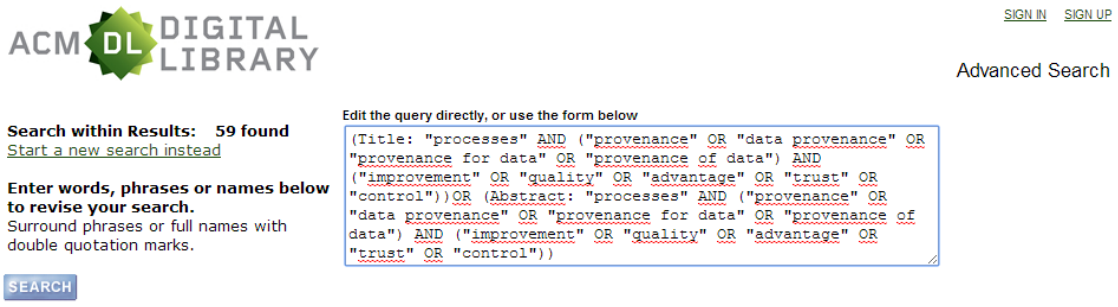


Figura 27: **Pesquisa ACM**

- ISI: nesta ferramenta foram preenchidos e adicionados campos, conforme exibido na Figura 2. Dentre as opções de extensões de arquivos possíveis à exportação dos resultados, optou-se pela extensão '.txt', a qual é compatível para importação junto ao JabRef. Neste ponto, foi constatada uma limitação da base, a qual realiza a exportação de no máximo 50 artigos por vez, obrigando o usuário a gerar vários arquivos, no caso de uma quantidade elevada de artigos retornados.

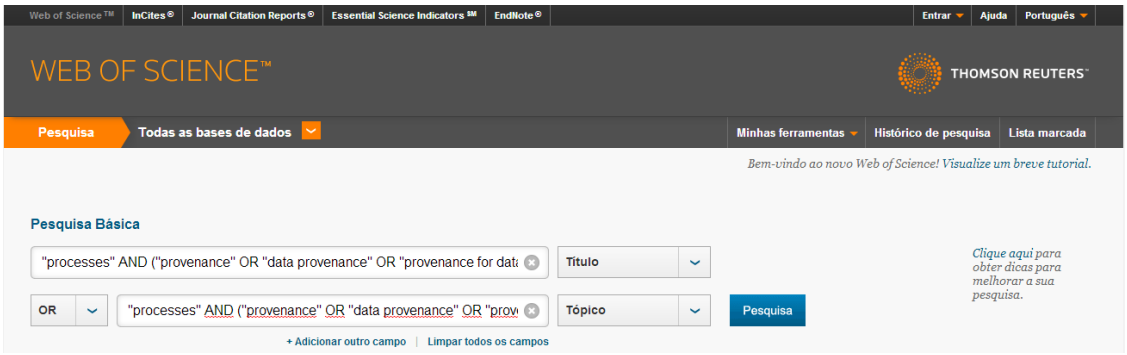


Figura 28: **Pesquisa ISI**

- Scopus: assim como na ACM, nesta ferramenta também utilizou-se a opção de "Busca Avançada", sendo preenchidos os campos de pesquisa conforme Figura 3.

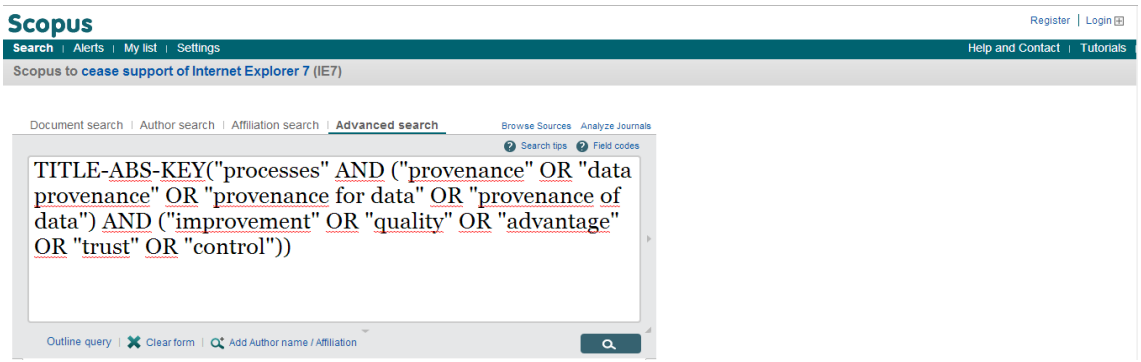


Figura 29: **Pesquisa Scopus**

A base Scopus possibilitou a exportação de todos os artigos retornados em um único arquivo, em formato bibtex, o que facilitou consideravelmente o processo de importação junto ao JabRef.

- IEEE: sem fugir ao padrão das demais bases, a pesquisa da *string* na respectiva também utilizou a opção de "Busca Avançada", onde nesta se fez necessário um pouco

mais de conhecimento para utilização da opção devida, a fim de obter o melhor retorno mediante a busca. As opções utilizadas e o preenchimento dos campos podem ser visualizados na Figura 4. Esta base permite a exportação dos resultados apenas em formatos não reconhecidos pelo JabRef, sendo assim, como o volume de artigos retornados era pequeno, optou-se pela exportação dos resultados para um formato '.csv', reconhecido pelo Libre Office. A verificação de repetições foi realizada de forma manual, onde foram realizadas comparações com os resultados das demais bases.

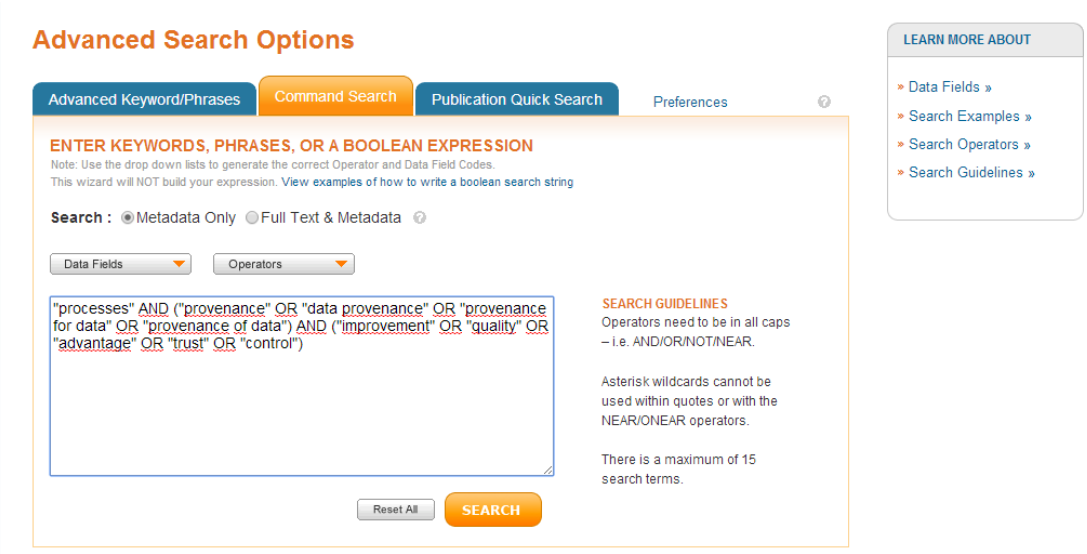


Figura 30: **Pesquisa IEEE**

Todos os resultados obtidos mediante a execução da *string* de busca junto as bases, com exceção da IEEE, foram armazenados na ferramenta JabRef.

Os totais de publicações retornadas pelas *strings* de busca são apresentados na próxima Seção. Além disto, estes totais, eliminando-se as duplicatas, foram listados na Tabela 2 da Seção 3. Os totais de publicações eliminadas de acordo com os filtros aplicados e critérios de exclusão adotados, foram marcados nesta mesma tabela. Para todas as outras publicações que não foram excluídas, foram registradas sua referência, um resumo e a resposta para cada uma das questões de pesquisa, conforme exibido na subseção 2.6.

2. Execução da Pesquisa

Mediante a definição do protocolo apresentado anteriormente, iniciou-se o processo de execução do mesmo. Assim, esta revisão foi realizada entre os meses de julho e agosto de 2014, como parte dos requisitos para aprovação na disciplina de Engenharia de Software Experimental. Nesta Seção são apresentados os resultados obtidos em cada um dos seis passos definidos na Subseção 2.2 (Procedimento de seleção dos artigos) do Protocolo de Pesquisa, sendo eles: (1) Busca e Catalogação; (2) Eliminação de artigos repetidos na mesma base (1° filtro); (3) Eliminação de artigos repetidos entre bases (2° filtro); (4) Seleção de artigos com base nos títulos (3° Filtro); (5) Seleção de artigos com base na leitura dos resumos (4º Filtro) e (6) Seleção das publicações relevantes com base na leitura completa do artigo (5º Filtro).

1. **Busca e Catalogação**

Inicialmente foi definida uma *string* de busca para efetivação da pesquisa junto as bases. Abaixo são dispostas as respectivas *strings* utilizadas nas buscas realizadas, devidamente refinadas, contendo informações de link, número de artigos retornados e artigos de controle listados, de acordo com a base em uso:

- **Base:** Scopus (http://www.scopus.com/)

Resultados: 577

Artigos de Controle Indexados:

- PrIMe: A Methodology for Developing Provenance-Aware Applications (AC1)
- Provenance of Software Development Processes (AC2)
- Provenance from log files: A BigData problem (AC3)

***String*:** TITLE-ABS-KEY("processes" AND ("provenance" OR "data provenance" OR "provenance for data" OR "provenance of data") AND ("improvement" OR "quality" OR "advantage" OR "trust" OR "control"))

- **Base:** ISI –Web of Science (http://www.isiknowledge.com/)

Resultados: 178

Artigos de Controle Indexados:

- AC1
- AC2

***String*:** Título: ("processes" AND ("provenance" OR "data provenance" OR "provenance for data" OR "provenance of data") AND ("improvement" OR "quality" OR "advantage" OR

"trust" OR "control")) OR Tópico: ("processes" AND ("provenance" OR "data provenance" OR "provenance for data" OR "provenance of data") AND ("improvement" OR "quality" OR "advantage" OR "trust" OR "control"))

- **Base:** ACM – (http://dl.acm.org/)

Resultados: 130

Artigos de Controle Indexados:

- AC1
- AC3

***String*:** (Title: "processes" AND ("provenance" OR "data provenance" OR "provenance for data" OR "provenance of data") AND ("improvement" OR "quality" OR "advantage" OR "trust" OR "control")) OR (Abstract: "processes" AND ("provenance" OR "data provenance" OR "provenance for data" OR "provenance of data") AND ("improvement" OR "quality" OR "advantage" OR "trust" OR "control"))

- **Base: IEEE** – (http://ieeexplore.ieee.org/Xplore/home.jsp)

Resultados: 15

Artigos de Controle Indexados: Não foram retornados os artigos de controle, haja vista que os mesmos não estão indexados junto a respectiva base.

String: *("processes" AND ("provenance" OR "data provenance" OR "provenance for data" OR "provenance of data") AND ("improvement" OR "quality" OR "advantage" OR "trust" OR "control"))*

Tabela 15: Total de resultados obtidos mediante a aplicação de cada filtro

Fonte	Data	Total	Filtro 1 Repetições mesma base	Filtro 2 Repetições entre bases	Filtro 3 Título	Filtro 4 Resumo	Filtro 5 Texto
ACM	**14/08/14**	**130**	**0**	**0**	**112**	**13**	**0**
ISI	**14/08/14**	**178**	**0**	**4**	**163**	**8**	**0**
Scopus	**14/08/14**	**577**	**11**	**80**	**464**	**18**	**0**
IEEE	14/08/14	15	0	12	3	-	-

Os valores exibidos informam o total de artigos descartados ao término de cada filtro

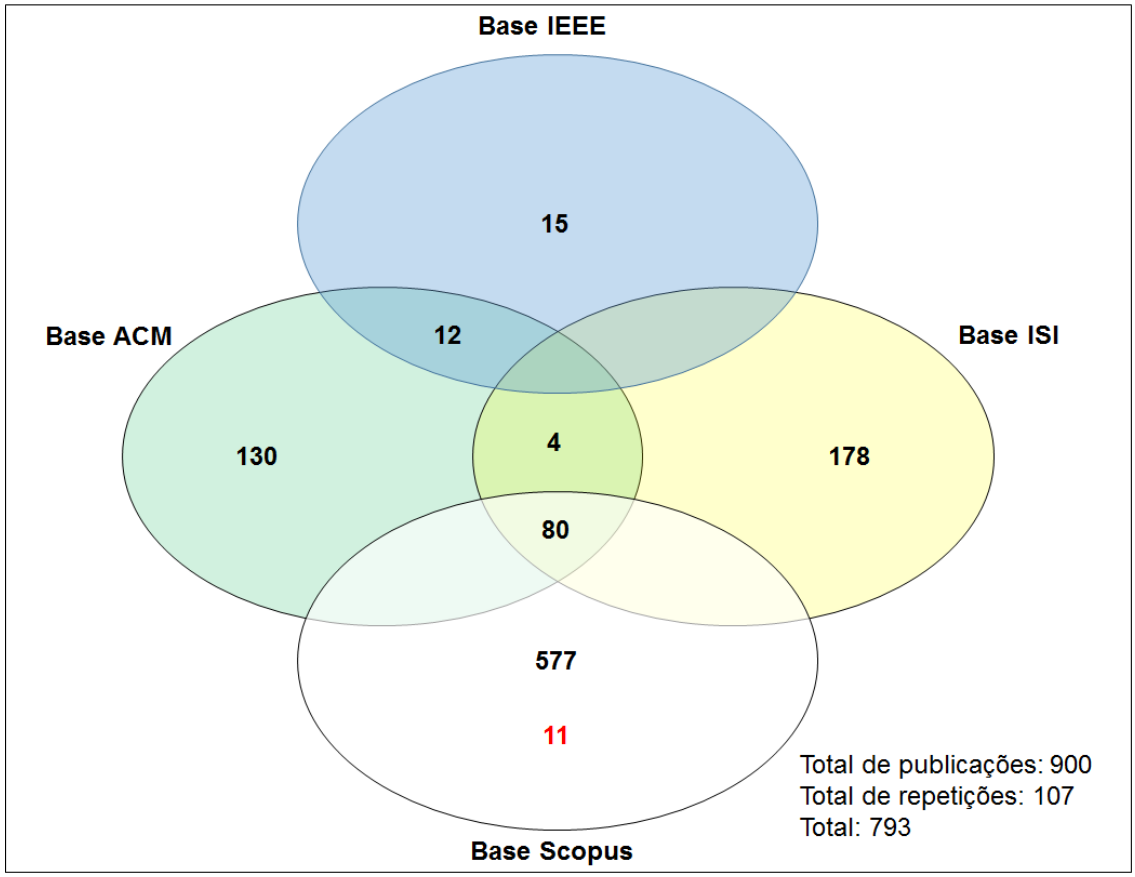


Figura 31: **Artigos retornados pela *string* de busca**

2.1 Eliminação de artigos repetidos na mesma base (1° filtro)

Nesta etapa, após realizada a importação dos resultados de cada base de forma separada no JabRef, foram identificadas e eliminadas as repetições dentro do retorno apresentado pela própria base. Conforme exibido na Tabela 10, a base Scopus trouxe um total de 11 artigos duplicados.

2.2 Eliminação de artigos repetidos entre bases (2° filtro)

Nesta etapa, com o intuito de eliminar os artigos repetidos entre as bases, foram importados os resultados obtidos em cada base no JabRef, na mesma ordem disposta na Tabela 2. As repetições deste filtro foram excluídas e contabilizadas para a base que está sendo importada no momento. A seguir é apresentado um exemplo para melhor entendimento de como foi

aplicado este filtro e como foi realizado o preenchimento da respectiva tabela: A primeira base a ser importada foi a ACM, sendo assim a mesma não apresentou repetições, haja vista que não havia outra base para comparações. Na sequência foi realizada a importação da ISI, a qual apresentou 4 artigos repetidos, comparados aos artigos retornados pela ACM. Os artigos da ISI repetidos foram desmarcados e a importação foi realizada somente dos artigos não repetidos.

As buscas em todas as quatro bases mencionadas na retornaram 900 publicações no total, onde mediante a eliminação das repetições, em duas etapas (repetições de mesma base e entre bases), restaram 793 publicações, conforme ilustrado na Figura 31.

2.3 Seleção de artigos com base nos títulos (3° Filtro)

Dando continuidade ao processo de filtragem para seleção das publicações, iniciou-se a aplicação dos filtros específicos. Foi realizada uma leitura dos títulos dos artigos selecionados após a eliminação das repetições, onde os títulos que não continham nenhuma das palavras chave definidas e/ou não remetiam aos objetos de estudo foram excluídos. Na Figura 32 é possível visualizar a quantidade de artigos restantes após a eliminação dos artigos repetidos e dos artigos em que o título não enquadrava-se nos critérios válidos.

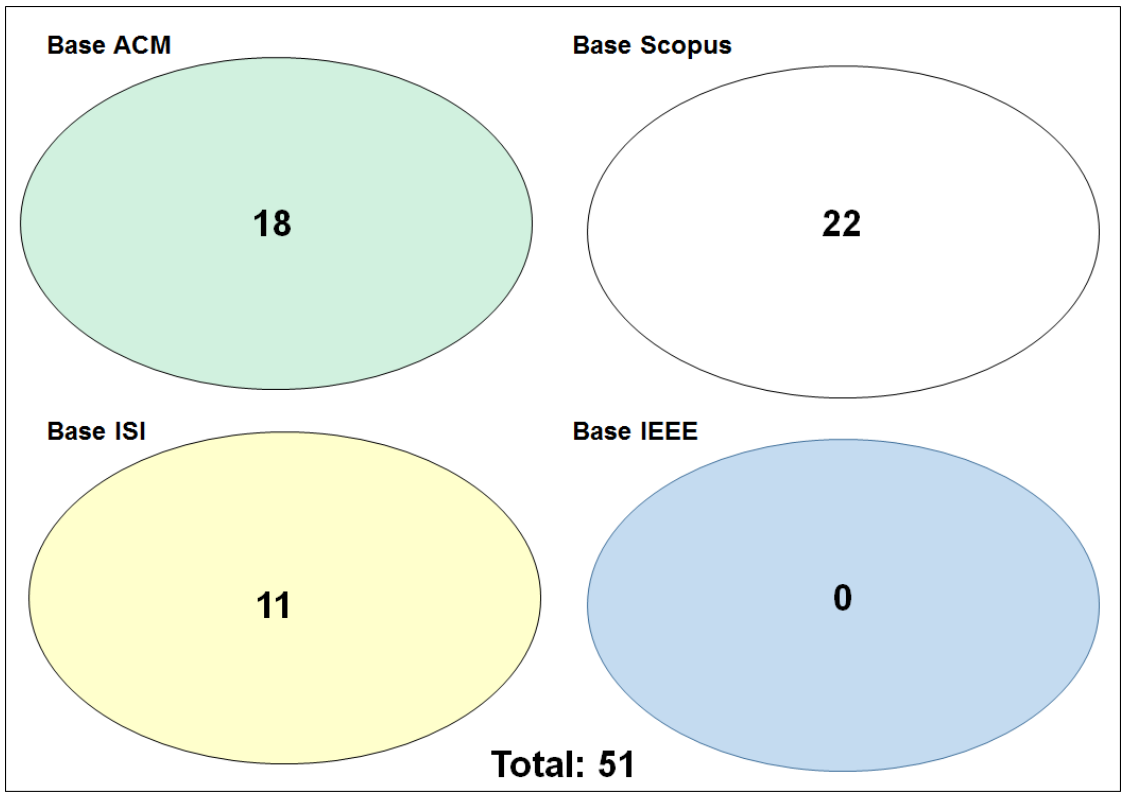


Figura 32: **Artigos selecionados após a eliminação de repetições e análise do título**

2.4 Seleção de artigos com base na leitura dos resumos (4º Filtro)

Foi realizada a leitura do resumo dos artigos selecionados após a execução dos três filtros anteriores, onde foram adotados os critérios de exclusão descritos no passo 5, da Seção 2.2. Os resultados após a execução desta etapa podem ser observados na Figura 33.

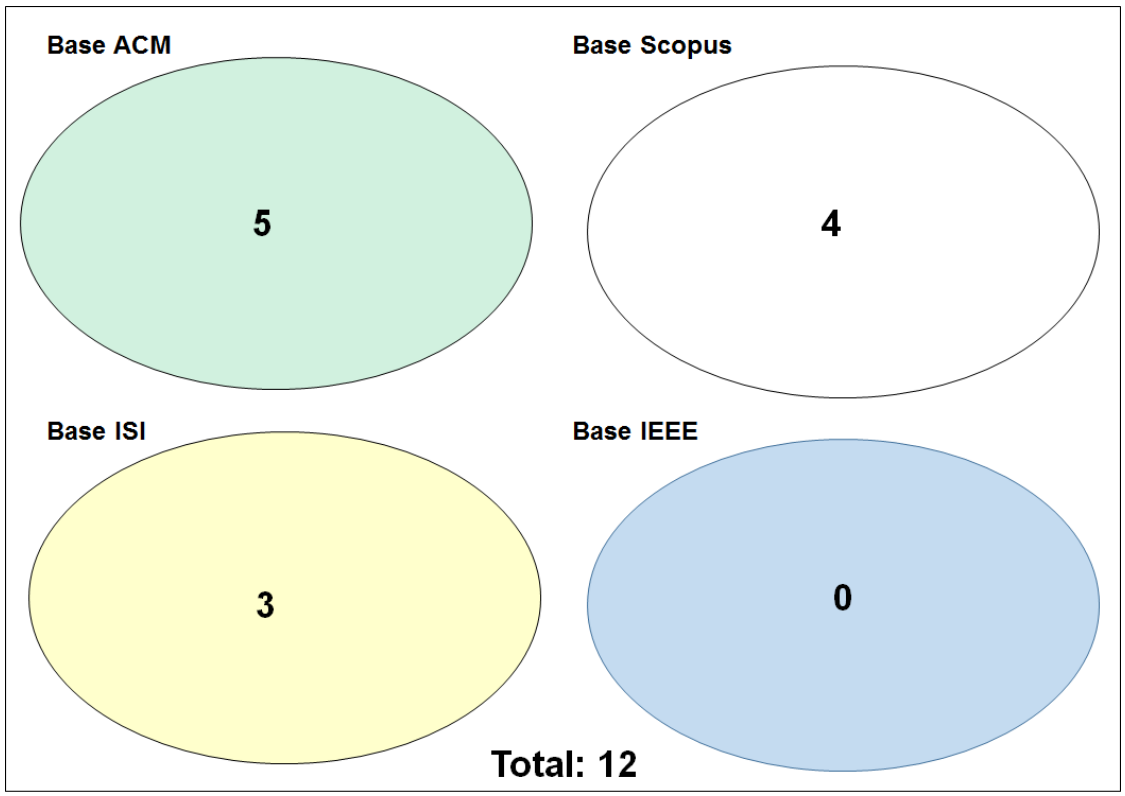


Figura 33: **Artigos selecionados após leitura do resumo dos artigos selecionados**

2.5 Seleção das publicações relevantes com base na leitura completa (5º Filtro)

Após a execução dos passos supramencionados, 1 dos 12 artigos selecionados não estava disponível nas bibliotecas eletrônicas, de forma gratuita. Foi enviado e-mail para o principal autor solicitando a verificação de viabilidade de disponibilizar o respectivo artigo, porém até a presente data não houve retorno por parte do mesmo.

Todas as 11 publicações selecionadas foram completamente lidas e reclassificadas em relação aos critérios de exclusão definidos na Seção 2.2. Mediante a execução do 4º filtro dos procedimentos de seleção de publicações, as 11 publicações restantes foram mantidas. As mesmas foram resumidas e tiveram respondidas as questões de pesquisa.

2.6 Dados Coletados das Publicações Selecionadas

Na sequência, por meio de formulários, são dispostas informações referentes aos artigos selecionados após a execução dos filtros relatados no decorrer deste artigo.

Dados da publicação	
Título	**A Model of Process Documentation to Determine Provenance in Mash-Ups**
Autor(es)	**GROTH, P., MILES, S., MOREAU, L.**
Ano de publicação	**2009**
Resumo	
Autores definem um modelo conceitual de dados genéricos, que suporta a criação autônoma de atribuições da documentação do processo e fatores para aplicações multi-institucionais dinâmicas. O modelo de dados é instanciado usando dois formatos de Internet, OWL e XML, é avaliado em relação a questões sobre a proveniência dos resultados gerados por uma bioinformática complexa com uso de *mash-up*.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **As análises finais, especificamente a determinação da origem de um resultado (ou seja, o processo que levou a ele), são ativados por documentação do processo, que refere-se a documentação do processo passado de um aplicativo criado pelos componentes desse aplicativo em tempo de execução.** • **Não mencionado no artigo.**	

Dados da publicação	
Título	**A provenance data management system for improving the product modelling process**
Autor(es)	**PETRINJA, E., STANKOVSKI, V., TURK, Z.**
Ano de publicação	**2007**
Resumo	
Os autores mostram que a gestão de modelos de produtos complexos pode ser melhorada com o uso de metadados relacionados ao tempo. Este objetivo é alcançado através da concepção e implementação de um sistema de gerenciamento de dados semânticos de proveniência. Uma parte do sistema é um serviço de *grid* de origem, sendo o mesmo acessível através de um portal. Os autores mostram que o serviço de origem, e em particular, as extrações de dados poderão ser utilizados para melhorar a gestão de modelos de produtos complexos, sem atrasos significativos de tempo de processamento. Além disso, a utilização de infraestrutura da rede pode aumentar a confiança no processo de modelação do produto e a sua segurança.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Através do projeto e a implementação de um sistema de gestão de dados semânticos de proveniência.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado no artigo.** • **Através do uso da estrutura da especificação IFC.**	

Dados da publicação	
Título	**A Provenance Framework for Data-Dependent Process Analysis**
Autor(es)	**DEUTCH, D., MOSKOVITCH, Y., TANNEN, V.**
Ano de publicação	**2014**
Resumo	
Os autores consideram novas construções que generalizam uma abordagem chamada *semiring* ao contexto de análise de processo dependente de dados (DDP). Estas construções abordam dois novos desafios, sendo respectivamente: (1) combinar as anotações de proveniência, tanto de informação que reside no banco de dados quanto de informações sobre as entradas externas (por exemplo, as escolhas do usuário), e (2) a finita captura de execução de processo infinito. Uma solução é proposta e analisada a partir de perspectivas teóricas e experimentais, sendo comprovada sua eficácia.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado neste artigo	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Utilizando a especificação PROV, de padrão de modelagem.** • **Não mencionado no artigo.**	

Dados da publicação	
Título	**A workflow modeling system for capturing data provenance**
Autor(es)	**JOGLEKAR, G., GIRIDHAR, A., REKLAITIS, G.**
Ano de publicação	**2014**
Resumo	
Os autores descrevem um quadro geral para a construção de novos fluxos de trabalho e implementam as ações associadas, o que visa facilitar a compreensão dos processos de trabalho em várias disciplinas. São descritos os principais blocos de construção no quadro, as suas funcionalidades e é ilustrada a integração de fluxos de trabalho entre um experimento e um processo científico.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado no artigo** • **Não mencionado no artigo**	

Dados da publicação	
Título	**D-PROV: Extending the PROV Provenance Model with Workflow Structure**
Autor(es)	**MISSIER, P., DEY, S., BELHAJJAME, K., CUEVAS-VICENTTÍN, V., LUDASCHER, B.**
Ano de publicação	**2013**
Resumo	
Os autores apresentam uma extensão do modelo de proveniência W3C PROV[1], que visa representar uma estrutura do processo. Mediante a introdução de novas relações de proveniência para a estrutura de processo de modelagem, juntamente com os seus padrões de uso, obtém-se benefícios demonstrados em pesquisas apresentadas.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado no artigo.** • **Não mencionado no artigo.**	

Dados da publicação	
Título	**Issues in Automatic Provenance Collection**
Autor(es)	**BRAUN, U., GARFINKEL, S., HOLLAND, D., MUNISWAMY-REDDY, K., SELTZER, M.**
Ano de publicação	**2006**
Resumo	
Os autores debatem os desafios encontrados e as questões expostas mediante ao desenvolvimento de um coletor de proveniência automático, o qual é executado no nível do sistema operacional.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Coleta de informações que geram novos dados que, por conseguinte, tronam-se novos dados de proveniência.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado neste artigo.** • **Através de dois modelos independentes, onde um controla o acesso convencional sobre proveniência e o outro fornece o controle de acesso sobre os ramos da árvore de ascendência.**	

Dados da publicação	
Título	**Karma2: Provenance Management for Data Driven Workflows**
Autor(es)	**SIMMHAN, Y., PLALE, B., GANNON, D.**
Ano de publicação	**2009**
Resumo	
Os autores focam o trabalho na coleta de proveniência para os fluxos de trabalho, necessários para validar o fluxo de trabalho e para determinar a qualidade de produtos de dados gerados. O desafio proposto pelos autores refere-se a gravação de metadados de proveniência uniforme e utilizável que atenda às necessidades de domínio, minimizando a carga sobre os autores de serviço e a sobrecarga de desempenho no mecanismo de workflow e serviços.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado no artigo.** • **Não mencionado no artigo.**	

Dados da publicação	
Título	**PrIMe: A Methodology for Developing Provenance-Aware Applications**
Autor(es)	**MILES, S., GROTH, P., MUNROE, S., MOREAU, L.**
Ano de publicação	**2011**
Resumo	
Os autores propõem uma técnica de engenharia de software, denominada Prime, para adaptar projetos de aplicações para que possam interagir com uma camada mediadora de proveniência, tornando-os, assim, voltados à proveniência. Os autores especificam as etapas envolvidas na aplicação de Prime, analisam a sua eficácia e ilustram seu uso com dois estudos de caso, bioinformática e medicina.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Prime é aplicado a um determinado tipo de caso de uso, questões de proveniência e tecnologias que ajudam a satisfazer os casos de uso, caso o projeto tenha uma forma particular.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Utilização de sistema separado de sua função primária, o qual trata processos e dados na aplicação. Respondendo a cada tipo de questão, proveniência pode ser vista como um caso de uso, ao invés de uma mudança no projeto para o benefício dos desenvolvedores.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Possibilitando a compreensão de como os dados foram derivados até seu estado atual.** • **Não mencionado no artigo.**	

Dados da publicação	
Título	**Provenance from Log Files: a BigData Problem**
Autor(es)	**GHOSHAL, D., PLALE, B.**
Ano de publicação	**2013**
Resumo	
Os autores exploram a opção de obter proveniência dos arquivos de log existentes, uma abordagem que reduz a tarefa de instrumentação substancialmente, mas levanta questões sobre refinar enormes quantidades de informação para o que pode ou não ser proveniência. Os autores estudam a troca de facilidades de captura, integralidade de proveniência e mostram que em algumas circunstâncias, a captura através de registros pode resultar em proveniência de alta qualidade.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Na identificação de anomalias e erros.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
• **Não atrapalha a execução dos processos mediante a sua aplicação** • **Permite ao usuário refinar regras de filtragem** • **A captura de proveniência através de registros pode resultar em proveniência de alta qualidade**	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Captando nuances de ambiente de execução que poderiam influenciar os resultados de grande escala de análise de dados.** • **Utilizando um sistema baseado em regras, o qual facilita as tarefas de extração de proveniência, pois permite uma flexibilidade na maneira de selecionar informações relevantes e também dá o controle na gestão da granularidade de informações de proveniência.**	

Dados da publicação	
Título	**Provenance of Software Development Processes**
Autor(es)	**WENDEL, H., KUNDE, M., SCHREIBER, A.**
Ano de publicação	**2010**
Resumo	
Os autores propõem uma solução para problemas relacionados a falhas nos processos de desenvolvimento de software, com base em tecnologias de proveniência.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Não mencionado no artigo.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado no artigo.** • **Não mencionado no artigo.**	

Dados da publicação	
Título	**Workflow Composition through Design Suggestions using Design-Time Provenance Information**
Autor(es)	**JUNAID, M., BERGER, M., VITVAR, T., PLANKENSTEINER, K., FAHRINGER, T.**
Ano de publicação	**2010**
Resumo	
Os autores fazem uma abordagem onde o sistema de proveniência intercepta as ações dos usuários, processa, armazena essas ações e fornece sugestões sobre possíveis ações futuras para o projeto de fluxo de trabalho. Essas ações sugeridas são baseadas nas ações do usuário atual e são calculados com base nas informações de proveniência disponíveis.	
Questão primária: Como a Proveniência de Dados pode ser utilizada na melhoria de processos?	
Através da aplicação de proveniência em um fluxo de trabalho é possível criar novos fluxos com base nas sugestões dadas pelos sistemas de proveniência. Neste estudo foi identificado que a criação de novos fluxos de trabalho, utilizando sugestões dadas pelos sistemas de proveniência, tem ganho significativo em eficiência.	
Questão secundária: Quais as vantagens da aplicação de Proveniência de Dados em processos?	
Não mencionado no artigo.	
Questões terciárias: • **Como avaliar a confiabilidade dos dados de proveniência?** • **Como controlar o volume de dados de Proveniência em Processos?**	
• **Não mencionado no artigo.** • **Não mencionado no artigo.**	

APÊNDICE II – FORMULÁRIO DE CARACTERIZAÇÃO DO PARTICIPANTE

Nome: _____

1) Formação Acadêmica:

☐Pós-Doutorado ☐Doutorado ☐Mestrado ☐Especialização ☐Graduação ☐Técnico

Outra: _____

2) Atualmente trabalha na:

☐Academia ☐Indústria ☐Academia e Industria

Tempo na academia (em anos): _____

Tempo da indústria (em anos):_____

3) Conhece o processo analisado?

☐Não ☐Sim - Tempo (em anos):_____- Papel desempenhado:_____

Experiência em Processos de Desenvolvimento de Software

Indique o grau de sua experiência para cada item nesta seção, seguindo a escala de 5 pontos abaixo:

1 = nenhum

2 = conhecimento teórico (estudei em aula ou em livro ou apliquei em projeto em sala de aula)

3 = participei em um ou mais projetos na indústria, mas sem ser o responsável

4 = participei em até 03 oportunidades na indústria como responsável

5 = participei em mais de 03 oportunidades na indústria como responsável incluindo diferentes organizações e/ou diferente tipos de processos.

Além disto, inclua o tempo em anos e o número de projetos que participou.

4) Experiência implantando processos de desenvolvimento de software: 1 2 3 4 5

Tempo: _____ Número de Projetos: _____

5) Experiência definindo processos de desenvolvimento de software: 1 2 3 4 5

Tempo: _____ Número de Projetos: _____

6) Experiência revisando (atividades de validação/verificação) processos de desenvolvimento de software: 1 2 3 4 5

Tempo: _____ Número de Projetos: _____

7) Experiência gerenciando processos de desenvolvimento de software: 1 2 3 4 5

Tempo: _____ Número de Projetos: _____

8) Experiência com melhoria de processos de desenvolvimento de software:1 2 3 4 5

Tempo: _____ Número de Projetos: _____

APÊNDICE III – FORMULÁRIO I - AVALIAÇÃO EMPRESA I

Nome: _____

Data: ___/___/___

Questão 1 - Hora de Início: ___:___ Hora de Término: ___:___
Quando foi iniciada a instância 11570?

Questão 2 - Hora de Início: ___:___ Hora de Término: ___:___
Quando foi finalizada a instância 11570?

Questão 3 - Hora de Início: ___:___ Hora de Término: ___:___
Qual foi o tempo de duração da instância 11570, considerando sua data e hora de início até sua conclusão?

Questão 4 - Hora de Início: ___:___ Hora de Término: ___:___
Quantas horas foram gastas na instância que teve o maior tempo de duração?

Questão 5 - Hora de Início: ___:___ Hora de Término: ___:___
Quantas horas foram gastas na instância que teve o menor tempo de duração?

Questão 6 - Hora de Início: ___:___ Hora de Término: ___:___
Qual o ID da instância que foi iniciada por último?

Questão 7 - Hora de Início: ___:___ Hora de Término: ___:___
Qual(is) o(s) nome(s) do(s) módulo(s) que foram afetados pela execução da instância 11570?

Questão 8 - Hora de Início: ___:___ Hora de Término: ___:___
Qual(is) pessoas/equipes participaram da execução da instância 11570?

Questão 9 - Hora de Início: ___:___ Hora de Término: ___:___

Alguma tarefa da instância 11570 foi realizada mais de uma vez? Qual(is) é(são) o(s) nome(s) da(s) tarefa(s)?

Questão 10 - Hora de Início: ___:___ Hora de Término: ___:___

Qual é o tipo da atividade Reportar Erro no Sistema da instância 11570?

Questão 11 - Hora de Início: ___:___ Hora de Término: ___:___

Quem executou a tarefa de Resolução do Caso da instância 11570?

Questão 12 - Hora de Início: ___:___ Hora de Término: ___:___

Quais módulos ou componentes foram manipulados durante a execução da tarefa anterior?

Questão 13 - Hora de Início: ___:___ Hora de Término: ___:___

Na tarefa anterior é possível identificar alguma inferência realizada pelo sistema que não conste nos dados que foram obtidos sem o mecanismo de inferência?

Questão 14 - Hora de Início: ___:___ Hora de Término: ___:___

Qual(is) módulo(s)/componente(s) foi influenciado pelo Agente Marcos Miguel?

Questão 15 - Hora de Início: ___:___ Hora de Término: ___:___

A informação da questão anterior pode ser obtida apenas consultando-se o detalhamento da instância?

Questão 16 - Hora de Início: ___:___ Hora de Término: ___:___

Qual o tipo do agente de Nome Marcos Miguel? Qual agente poderia substituir este agente, durante a execução da atividade de Resolução do Caso da instância?

Questão 17 - Hora de Início: ___:___ Hora de Término: ___:___

Qual pessoa ou equipe executou mais tarefas nas instâncias avaliadas?

Questão 18 - Hora de Início: ___:___ Hora de Término: ___:___

Qual agente aparenta estar mais sobrecarregado?

Questão 19 - Hora de Início: ___:___ Hora de Término: ___:___

Sabendo que um agente está sobrecarregado, que ações você poderia tomar?

APÊNDICE IV – FORMULÁRIO II - AVALIAÇÃO EMPRESA II

Nome: _____

Data: ___/___/___

Questão 1 - Hora de Início: ___:___ Hora de Término: ___:___
Quando foi iniciada a instância 100430?

Questão 2 - Hora de Início: ___:___ Hora de Término: ___:___
Quando foi finalizada a instância 100430?

Questão 3 - Hora de Início: ___:___ Hora de Término: ___:___
Qual foi o tempo de duração da instância 100430, considerando sua data e hora de início até sua conclusão?

Questão 4 - Hora de Início: ___:___ Hora de Término: ___:___
Quantas horas foram gastas na instância que teve o maior tempo de duração?

Questão 5 - Hora de Início: ___:___ Hora de Término: ___:___
Quantas horas foram gastas na instância que teve o menor tempo de duração?

Questão 6 - Hora de Início: ___:___ Hora de Término: ___:___
Qual o ID da instância que foi iniciada por último?

Questão 7 - Hora de Início: ___:___ Hora de Término: ___:___
Qual(is) o(s) nome(s) do(s) módulo(s) que foram afetados pela execução da instância 100430?

Questão 8 - Hora de Início: ___:___ Hora de Término: ___:___
Qual(is) pessoas/equipes participaram da execução da instância 100430?

Questão 9 - Hora de Início: ___:___ Hora de Término: ___:___
Alguma tarefa da instância 100430 foi realizada mais de uma vez? Qual(is) é(são) o(s) nome(s) da(s) tarefa(s)?

Questão 10 - Hora de Início: ___:___ Hora de Término: ___:___

Qual é o tipo da atividade Abertura da Requisição de Mudança no Software da Instância de ID 100430?

Questão 11 - Hora de Início: ___:___ Hora de Término: ___:___

Quem executou a primeira atividade de Implementação da Solução da Instância de ID 100430?

Questão 12 - Hora de Início: ___:___ Hora de Término: ___:___

Quais módulos ou componentes foram manipulados durante a execução da atividade anterior?

Questão 13 - Hora de Início: ___:___ Hora de Término: ___:___

Nesta mesma atividade, é possível identificar alguma inferência realizada pelo sistema que não conste nos dados que foram obtidos sem o mecanismo de inferência?

Questão 14 - Hora de Início: ___:___ Hora de Término: ___:___

Qual(is) módulo(s)/componente(s) foi influenciado pelo Agente DotNet, na Instância de ID 100430?

Questão 15 - Hora de Início: ___:___ Hora de Término: ___:___

A informação da questão anterior pode ser obtida apenas consultando-se o detalhamento da instância?

Questão 16 - Hora de Início: ___:___ Hora de Término: ___:___

Qual o tipo do agente de Nome DotNet?

Questão 17 - Hora de Início: ___:___ Hora de Término: ___:___

Qual pessoa ou equipe executou mais tarefas nas instâncias avaliadas?

Questão 18 - Hora de Início: ___:___ Hora de Término: ___:___

Qual agente aparenta estar mais sobrecarregado?

Questão 19 - Hora de Início: ___:___ Hora de Término: ___:___
Sabendo que um agente está sobrecarregado, que ações você poderia tomar?

Questão 20 - Hora de Início: ___:___ Hora de Término: ___:___
É possível identificar padrões que culminam em desdobramentos de erro?

Questão 21 - Hora de Início: ___:___ Hora de Término: ___:___
Qual o percentual de atividade Erronosistema e agente DotNet que resultaram em desdobramentos de erro?

Questão 22 - Hora de Início: ___:___ Hora de Término: ___:___
Com base no percentual observado na questão anterior, é possível identificar a necessidade de adotar alguma ação?

Questão 23 - Hora de Início: ___:___ Hora de Término: ___:___
Sabendo-se que o percentual de determinado padrão é alto, pode-se deduzir a necessidade de adoção de alguma medida para evitar novos desdobramentos de erro?

APÊNDICE V – FORMULÁRIO III - AVALIAÇÃO PROV-PROCESS

1) As informações relativas ao tempo de início, término e duração, foram facilmente identificadas na ferramenta PROV-Process.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

2) As informações contidas no detalhamento da instância, auxiliam no entendimento do que ocorreu durante a execução do processo.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

3) As informações de ID, NOME e TIPO, de tarefas, pessoas envolvidas no processo e artefatos manipulados durante a execução de uma instância são suficientes para entendimento do processo.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

4) As informações contidas no detalhamento de uma atividade, auxiliam no entendimento do que ocorreu durante a execução do processo.

a. Discordo totalmente

b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

5) As informações inferidas, contidas no detalhamento de uma atividade, apresentam novas informações acerca da atividade.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

6) As informações inferidas, contidas no detalhamento de um agente, apresentam novas informações acerca da participação do agente no processo.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

7) Através da visualização gráfica é possível identificar, mais facilmente, as atividades, agentes e entidades.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

8) Através da visualização gráfica é possível identificar melhor as inferências obtidas por meio do uso da ferramentas PROV-Process.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

9) A visualização gráfica possibilita uma análise mais rápida sobre os dados de execução de processos de desenvolvimento de software.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

10) A identificação de padrões relativos aos elementos que compõe o processo de desenvolvimento de software, apresentam indícios, significativos, quanto possíveis problemas do processo.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

11) Por meio da identificação de padrões que culminam em tarefas de desdobramento de erros, é possível detectar a necessidade de melhoria para evitar novos erros.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

12) Quanto maior o percentual, relativo ao número de vezes em que um conjunto de elementos, resultou em um desdobramento de erro, mais forte o indício de problemas neste padrão.

a. Discordo totalmente
b. Discordo parcialmente
c. Indiferente
d. Concordo parcialmente
e. Concordo totalmente

13) O que mais gostou na abordagem PROV-Process?

_____

_____

_____

14) O que menos gostou na ferramenta PROV-Process?

_____

_____

_____

15) O que mudaria na ferramenta PROV-Process?

_____

_____

_____

APÊNDICE VI – DETALHAMENTO DA BASE DE DADOS

A tabela *Entity* possui um identificador (*idEntity*), conforme especificação do PROV-DM, o qual indica ainda a possibilidade de atributos opcionais. Para este modelo, foram criados os atributos *Name* e *Type*, os quais representam, respectivamente, o nome e o tipo da entidade, sendo que o tipo da entidade deve ser preenchido com *Plan, Bundle, Collection* ou *EmptyCollection*, por tratarem-se de tipos de entidades.

A tabela *Activity* é composta por um identificador (*idActivity*), um atributo *startTime* para registro do início da atividade e um atributo *endTime* para registro do fim da atividade, conforme especificação do PROV-DM, o qual indica ainda a possibilidade de atributos opcionais. Para este modelo foram criados os atributos *Name*, para informação do nome da atividade, *idProcessInstance*, para indicação do número da instância, *Type_Activity*, para especificação do tipo da atividade, e *Priority*, para especificação quanto a prioridade da respectiva atividade, devendo ser preenchido com “alta”, “media” ou “baixa”.

A tabela *wasGeneratedBy* possui um identificador (*idWasGeneratedBy*) e um atributo *Entity_idEntity* para identificar uma entidade criada, conforme especificado pelo modelo PROV-DM. Os atributos *Activity_idActivity* e *Time*, indicados como opcionais no PROV, também foram inseridos, representando, respectivamente, um identificador para atividade que cria entidade e o tempo de criação da entidade.

A tabela *Used* possui um atributo *Activity_idActivity* para identificação da atividade consumida e um atributo *Entity_idEntity* para identificação da entidade consumida, conforme definido no PROV. Os atributos *idUsed* e *Time*, especificados como opcionais no PROV, constam na tabela e indicam, respectivamente, o identificador e o tempo em que a entidade começou a ser utilizada.

A tabela *wasStartedBy* possui um atributo *Activity_idActivity*, para identificação da atividade iniciada, conforme definido no PROV. Os atributos *idWasStartedBy*, *Entity_idEntity_Trigger*, *Activity_idActivity_Started* e *Time*, especificados como opcionais no PROV, constam na tabela e indicam, respectivamente, um identificador para o início da atividade, um identificador para entidade desencadeando uma atividade, um identificador para atividade que gerou a entidade e o tempo em que a atividade foi iniciada.

A tabela *wasEndedBy* possui um atributo *Activity_idActivity*, para identificação da atividade finalizada, conforme definido no PROV. Os atributos *idWasEndedBy*, *Entity_idEntity_Trigger*, *Activity_idActivity_Ended* e *Time*, especificados como opcionais no PROV, constam na tabela e indicam, respectivamente, um identificador para o final da atividade, um identificador para entidade desencadeando uma atividade final, um identificador para atividade que gerou a entidade e o tempo em que a atividade foi terminada.

A tabela w*asInvalidatedBy* possui um atributo *Entity_idEntity*, para identificação de entidades inválidas, conforme definido no PROV. Os atributos *idWasInvalidatedBy*, *Activity_idActivity* e *Time*, especificados como opcionais no PROV, constam na tabela e indicam, respectivamente, identificador para uma invalidação, um identificador para a atividade que invalidou a entidade e o tempo em que a entidade começou a ser anulada.

A tabela *wasDerivedFrom* possui um atributo *Entity_idEntity_GeneratedEntity*, para identificação da entidade gerada pela derivação e um atributo *Entity_idEntity_UsedEntity*, para identificação da entidade utilizada pela derivação, conforme definição do PROV. Os atributos *idWasDerivedFrom*, *Activity_idActivity*, *WasGeneratedBy_idWasGeneratedBy_Generation* e *Used*, especificados como opcionais no PROV, constam na tabela e indicam, respectivamente, um identificador para um derivação, um identificador para atividade usando e gerando as entidades acima, um identificador para a geração, envolvendo a entidade gerada e a atividade, e um identificador para o uso que envolve a entidade usada e a atividade. O atributo *Type_Derived* deve ser preenchido com *WasRevisionOf*, *WasQuotedOf* ou *HasPrimarySource*, indicando o tipo de derivação, sendo este não obrigatório.

A tabela *Agent* possui um atributo *idAgent*, para identificação de um agente, um atributo *Name*, para especificação do nome do mesmo, e um atributo *Type_Agent*, o qual deve ser preenchido com *Person*, *Organization* ou *SoftwareAgent*, para definição de categoria de agentes, em uma perspectiva de interoperabilidade.

A tabela *wasAttributedTo* possui um atributo *Entity_idEntity*, para identificação de entidade e um atributo *Agent_idAgent*, que se refere ao identificador do agente a quem a entidade é atribuída, indicando responsabilidade por sua existência, conforme definição do PROV. O atributo opcional, indicado no PROV, foi criado como *idWasAttributedTo*, o qual trata-se de um identificador para a relação.

A tabela *wasAssociatedWith*, apresenta o atributo *Activity_idActivity*, um identificador para a atividade, conforme definido no PROV. Os atributos opcionais, criados nesta tabela, correspondem a *idWasAssociatedWith*, *Agent_idAgent* e *Entity_idEntity_Plan*, indicando respectivamente, um identificador para a associação entre uma atividade e um agente, um identificador para o agente associado a atividade e um identificador para o plano do agente invocado no âmbito desta atividade.

A tabela *ActedOnBehalfOf* possui um atributo *Agent_idAgent_Delegate*, para identificação do agente associado a uma atividade e um atributo *Agent_idAgent_Responsabile*, para identificação do agente em nome do qual o agente delegado agiu, conforme definição do PROV. Os atributos opcionais constantes no PROV e criados na tabela, correspondem a *idActedOnBehalfOf* e *Activity_idActivity*, indicando, respectivamente, um identificador para o link entre delegado e responsável e um identificador de uma atividade, a qual detém o link delegado.

Na sequência dos componentes do PROV encontram-se os *Bundles*, os quais, no modelo desenvolvido, são especificados na tabela *Entity*, atributo *Type_Entity*, haja vista que se tratam de tipos de entidade.

A tabela *AlternateOf* é formada por um auto relacionamento com a tabela *Entity*, possuindo o atributo *Entity_idEntity_Alternate1*, que se trata de um identificador da primeira das duas entidades, e o atributo idEntity*_Alternate2*, que indica um identificador da segunda das duas entidades.

Assim como a tabela supracitada, a tabela *SpecializationOF* também se trata de um auto relacionamento com a tabela *Entity*, onde o atributo *Entity_idEntity_SpecificEntity* indica um identificador da entidade que é uma especialização da entidade geral, e o atributo *Entity_idEntity_GeneralEntity* indica um identificador da entidade que está sendo especializada.

Por fim, a tabela *HadMember*, é formada pelo atributo *idCollection*, que se refere a um identificador para coleção de membros, e pelo atributo *Entity_idEntity*, que indica o identificador de uma entidade que é membro da coleção.