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

Weiner Esmério Batista de Oliveira

Um Framework para Análise e Visualização de Dados de Proveniência

Juiz de Fora

2017

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

Weiner Esmério Batista de Oliveira

Um Framework para Análise e Visualização de Dados de Proveniência

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

Juiz de Fora

2017

Weiner Esmério Batista de Oliveira

Um Framework para Análise e Visualização de Dados de Proveniência

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 1 de Setembro de 2017.

BANCA EXAMINADORA

Profa. D.Sc. Regina Maria Maciel Braga - Orientadora
Universidade Federal de Juiz de Fora

Prof. D.Sc. Victor Stroële de Andrade Menezes
Universidade Federal de Juiz de Fora

Prof. Ph.D. Mario Antônio Ribeiro Dantas
Universidade Federal de Santa Catarina

A minha esposa, gentileza e dedicação. Aos meus pais, segurança e esforço.

AGRADECIMENTOS

Agradeço à Deus por todas as potencialidades que recebi e as bençãos que recebo até hoje. Obrigado pais que trabalharam muito por mim e acreditaram que os resultados chegariam. Obrigado mãe pelo exemplo de dedicação e esforço incansável. Obrigado pai pelo exemplo de ética e as lições de vida, mesmo quando precisaram chegar bem tarde. Obrigado cara esposa por me apoiar nos momentos difíceis, pela companhia e força. Agradeço também pela nossa filha que iluminou nossas vidas e pelo novo bebê que ainda não chegou. Obrigado meu irmão e todos os meus familiares que me incentivaram. Obrigado professores e orientadora. Reconheço a importância de todos nas minhas conquistas. Obrigado.

"Estou interessado nas ideias,
não apenas em produtos visuais."
Marcel Duchamp

RESUMO

A proveniência é reconhecida hoje como um desafio central para estabelecer confiabilidade e prover segurança em sistemas computacionais. Em workflows científicos, a proveniência é considerada essencial para apoiar a reprodutibilidade dos experimentos, a interpretação dos resultados e o diagnóstico de problemas. Estes benefícios podem também ser utilizados em outros contextos, como, por exemplo, em processos de software. No entanto, para sua melhor compreensão e utilização, são necessários mecanismos eficientes e amigáveis. Pesquisas em visualização de software, ontologias e redes complexas podem ajudar neste processo, gerando novo conhecimento sobre os dados e informações estratégicas para tomada de decisão. Esta dissertação apresenta um framework chamado Visionary, para auxiliar na compreensão e uso dos dados de proveniência através de técnicas de visualização de software, ontologias e análise de redes complexas. O framework captura os dados de proveniência e gera novas informações usando ontologias e análise do grafo de proveniência. A visualização apresenta e destaca as inferências e os resultados obtidos com a análise. O Visionary é um framework livre de contexto que pode ser adaptado para qualquer sistema que utiliza o modelo PROV de proveniência. Com o objetivo de avaliar a proposta, foi realizado um estudo experimental que encontrou indícios que o framework auxilia na compreensão e análise dos dados de proveniência, dando suporte à tomada de decisão.

Palavras-chave: Proveniência de dados, Visualização de software, Redes complexas.

ABSTRACT

Provenance is recognized today as a central challenge to establish reliability and provide security in computational systems. In scientific workflows, provenance is considered essential to support the reproducibility of experiments, interpretation of results and diagnosis of problems. We consider that these benefits can be used in new contexts, like software process. However, for a better understanding and use, efficient and friendly mechanisms are needed. Software visualization, ontology, and complex networks can help in this process by generating new data insights and strategic information for decision making. This dissertation presents a framework named Visionary, to assist in the understanding and use of provenance data through software visualization techniques, ontologies and analysis of complex networks. The framework captures the provenance data and generates new information using ontologies and analysis of provenance graph. The visualization presents and highlights the inferences and the results obtained with the analysis. Visionary is a context-free framework that can be adapted to any system that uses the PROV provenance model. In order to evaluate the proposal, an experimental study was carried out, which found indications that the framework assists in the understanding and analysis of provenance data, supporting decision making.

Keywords: Provenance Data. Software Visualization. Complex Networks.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BPMN Business Process Model and Notation

CE Critério de Exclusão

CI Critério de Inclusão

JSON JavaScript Objetc Notation

MDF Distância Finita Máxima

OPM Open Prevenance Model

OWL Web Ontology Language

PICOC População, Intervensão, Comparação, Saída e Contexto

QM Questões de Mapeamento

QP Questões de Pesquisa

QR Questões de Revisão

QS Questões de Pesquisa Secundária

SGWfC Sistemas de Gerência de Workflows Científicos

WfC Workflow Científico

SUMÁRIO

1 INTRODUÇÃO

Proveniência de dados pode ser conceituada como sendo a informação sobre as partes envolvidas na produção de um objeto (GROTH; MOREAU, 2013). Hoje a proveniência é reconhecida como um desafio central para estabelecer confiabilidade e prover segurança em sistemas computacionais, principalmente na Web (ACAR et al., 2012). Para os workflows científicos, a proveniência é considerada essencial para apoiar a reprodutibilidade dos experimentos, a interpretação dos resultados e o diagnóstico de problemas (SIMMHAN et al., 2005; GIL et al., 2007). Pela mesma razão, diversas áreas estudam a proveniência e seus benefícios como sistemas de arquivos (MUNISWAMY-REDDY et al., 2006), processos de software (COSTA et al., 2016), a web semântica (MARGO; SMOGOR, 2010), computação nas nuvens (MUNISWAMY-REDDY; SELTZER, 2010) e banco de dados (CHENEY et al., 2009). Além disso, com a utilização de modelos de proveniência, como o OPM (MOREAU et al., 2011) ou PROV (GROTH; MOREAU, 2013), a interoperabilidade desses dados é facilitada, auxiliando no processo de reutilização e cooperação entre grupos de trabalho.

Compreender a proveniência e os processos pelos quais ela é capturada é uma tarefa difícil de se realizar (BUNEMAN et al., 2001; PACKER; MOREAU, 2014). Mas a visualização de software, auxiliando na compreensão do software (DIEHL, 2007) pode facilitar o uso da proveniência, comunicando intuitivamente os principais aspectos de dados complexos para o usuário (ARSHAD et al., 2015). No entanto, apenas o uso dos mecanismos de visualização não garante uma adequada compreensão dos dados de proveniência. Os dados de proveniência são abundantes e valiosos e precisam de técnicas para auxiliar em sua exploração e compreensão (HOEKSTRA; GROTH, 2014). Mecanismos de inferência, juntamente com ontologias e análise de redes complexas podem ajudar no processo de identificar informações importantes contidas nos dados. Dessa forma, usando essas técnicas em conjunto, é possível melhorar a compreensão das informações e a confiabilidade dos resultados, bem como sua reutilização. Assim, os dados podem ser compreendidos de forma mais rápida e segura.

A fim de estimular o uso da proveniência e auxiliar na sua compreensão e análise, este trabalho propõe um framework para capturar os dados de proveniência, analisar

os dados através de ontologias e técnicas de redes complexas para gerar conhecimento novo e apresentar os resultados utilizando mecanismos de visualização, favorecendo a compreensão dos dados e da análise. O framework, denominado Visionary, foi projetado para ser flexível e ser capaz de se adaptar a diferentes contextos.

1.1 PROBLEMA

Conforme já mencionado, a proveniência tem importância fundamental nos sistemas computacionais, mas compreender as informações capturadas pela proveniência e compreender os mecanismos que realizam essa captura, não é uma tarefa trivial de ser executada por causa do seu conteúdo técnico (BUNEMAN et al., 2001; PACKER; MOREAU, 2014). Como afirma Hoekstra and Groth (2014) são necessários técnicas para auxiliar na navegação e investigação dessas informações devido à riqueza disponível das informações de proveniência.

Além disso, apesar de muitos trabalhos tratarem de proveniência de dados, a maior parte desses trabalhos foca na coleta e na análise semântica das informações de proveniência (DAI et al., 2008). Hoje, muitas ferramentas utilizam a proveniência de dados mas poucas se preocupam com a facilidade de utilização e compreensão para um público não especializado ou usuários casuais (RICHARDSON; MOREAU, 2016).

1.2 OBJETIVOS

Este trabalho tem como objetivo principal definir e implementar uma abordagem capaz de utilizar os dados de proveniência, realizando análises semânticas e estruturais e apresentar de forma intuitiva as informações coletadas e analisadas para um público não especializado. Essa proposta deve ser de fácil adoção para sistemas que já utilizam a proveniência sem se deter à um contexto específico.

Além disso, esse trabalho possui os seguintes objetivos específicos:

- Encontrar e investigar as propostas existentes na literatura que analisam e visualizam dados de proveniência.
- Propor uma abordagem que captura, analisa e visualiza as informações de proveniência.

- Gerar conhecimento novo a partir de dados de proveniência armazenados.
- Simplificar a compreensão da proveniência dos dados.

1.3 ESTRUTURA DO TRABALHO

Essa dissertação está dividida em cinco capítulos, incluindo este capítulo de introdução. O Capítulo 2 apresenta os principais conceitos relacionado à proposta e as tecnologias envolvidas. O Capítulo 3 apresenta o resultado de duas revisões sistemáticas conduzidas para dar verificar as principais propostas existentes na literatura relacionadas aos temas de pesquisa desta dissertação. O Capítulo 4 apresenta o framework Visionary e descreve seus passos. O Capítulo 6 conclui a dissertação, apresentando as considerações finais e trabalhos futuros.

2 PRESSUPOSTOS TEÓRICOS

Esse capítulo apresenta os principais conceitos relacionados à proposta deste trabalho. Neste sentido, a Seção 2.1 detalha proveniência de dados e o modelo de proveniência padrão utilizado nesse trabalho; a Seção 2.2 discute redes complexas e a Seção 2.3 apresenta os principais conceitos relacionados a visualização de software relevantes para o contexto deste trabalho.

2.1 PROVENIÊNCIA DE DADOS

Um dos primeiros autores a definir proveniência de dados foi Buneman et al. (2001) como a descrição da origem de um objeto de dados ou o processo pelo qual esse chegou em um banco de dados. Uma definição mais recente e precisa foi apresentada por Moreau et al. (2009) como a documentação dos processos do ciclo de vida de um objeto digital. Moreau também enfatiza a importância do uso de proveniência no contexto de experimentos científicos (MOREAU et al., 2009), com destaque para seu uso em *workflows* científicos [1] e Sistemas de Gerenciamento de Workflows Científicos (SGWfC).

A proveniência de dados é considerada de fundamental importância para os SGWfCs, principalmente para garantir a reprodutibilidade dos resultados e processos obtidos na execução de *workflow* científicos (MOREAU et al., 2009; LIM et al., 2010) bem como para determinar a autoria e a qualidade dos dados (DAVIDSON; FREIRE, 2008).

Pode-se distinguir dois tipos de proveniência de dados: a proveniência retrospectiva e a proveniência prospectiva (LIM et al., 2010). A proveniência retrospectiva modela a execução dos workflows e as informações da derivação dos dados, como as tarefas que foram executadas e como os artefatos de dados foram derivados. A proveniência prospectiva modela uma especificação abstrata dos workflows como uma receita para a derivação futura dos dados.

Muitos modelos de proveniência foram propostos na literatura (BOWERS et al., 2006; BUNEMAN et al., 2006; CAO et al., 2009; GROTH et al., 2009), mas a interoperabilidade entre esses modelos ainda é muito deficiente (DAVIDSON; FREIRE, 2008). Dois modelos

[1] *Workflow* científico é um modelo ou template que representa uma sequência de atividades científicas implementadas por ferramentas, a fim de alcançar um determinado objetivo (DEELMAN et al., 2009)

se destacaram, tornando-se padrão para a captura de dados de proveniência e são largamente utilizados: o modelo OPM (MOREAU et al., 2011) e o modelo PROV (GROTH; MOREAU, 2013). O modelo OPM foi concebido para a captura da proveniência retrospectiva, já o modelo PROV, além de ser um modelo mais recente e possuir recursos para a captura da proveniência retrospectiva, é considerado um modelo mais abrangente.

O modelo OPM (MOREAU et al., 2011) surgiu como resultado de Desafios de Proveniência (Provenance Challenges) propostos no contexto da conferência IPAW (MATTOSO; GLAVIC, 2016). Os desafios de Proveniência tiveram quatro edições e o modelo OPM foi resultado dos dois primeiros desafios. O modelo PROV foi especificado a partir do quarto Desafio de Proveniência.

Tanto o OPM quanto o PROV são modelos genéricos. O OPM é mais simples, com foco nos fluxos de execução e em como capturar a proveniência neste contexto, indicando, por exemplo, que um dado processo foi iniciado por um outro. Por outro lado, o modelo PROV foca nas responsabilidades e no histórico dos dados, tendo várias relações entre agentes e os outros tipos (atividades e entidades), como será descrito na subseção seguinte. Assim, o modelo PROV permite a captura de mais informações sobre os agentes, focando nas suas responsabilidades, ao passo que o OPM é mais focado no controle do fluxo de execução. Neste contexto, o PROV é considerado mais completo, porque possui mais relações que o OPM e provê uma documentação extensa, além de uma ontologia associada – PROV-O – e regras e restrições específicas.

Por essa razão o modelo PROV foi selecionado para ser utilizado nesse trabalho.

2.1.1 MODELO PROV

O modelo PROV foi criado com o objetivo de divulgar e facilitar a troca de informações de proveniência em ambientes heterogêneos como a Web. Conforme já dito, o modelo é especificado em um conjunto de doze documentos, onze deles representados na Figura 2.1. O último documento é o PROV-Overview (GROTH; MOREAU, 2013) que fornece uma visão geral de toda a documentação do PROV.

O PROV-Primer é um texto básico que instrui sobre o modelo de dados da proveniência para usuários que querem compreender o PROV e utilizar aplicações que o suportam. Os outros documentos são divididos em duas cores: verde para desenvolvedores que querem criar aplicações que geram ou consomem proveniência utilizando o PROV; rosa para

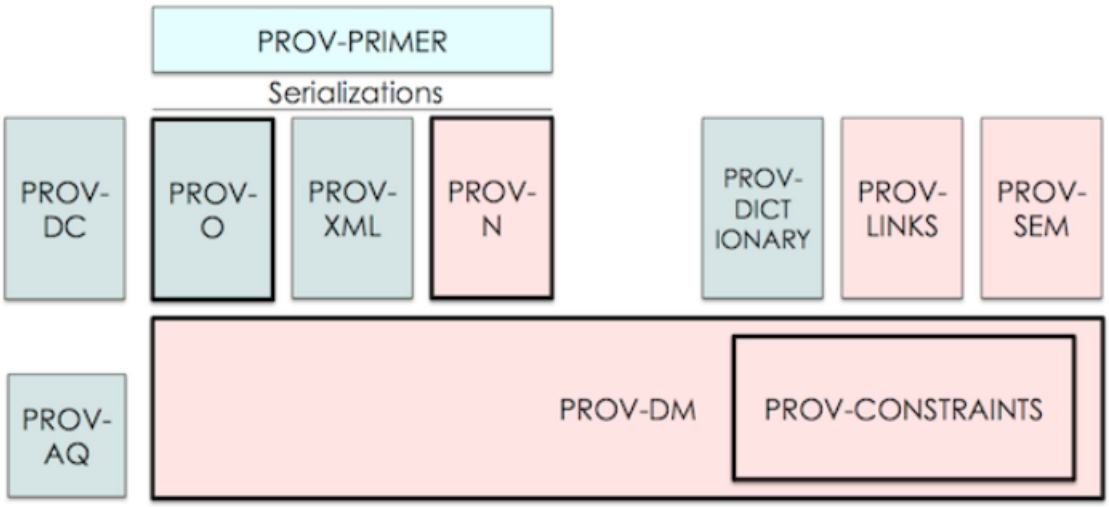


Figura 2.1: Conjunto de documentos que compõem o modelo PROV de proveniência (GROTH; MOREAU, 2013).

desenvolvedores avançados que desejam criar validações, outras serializações do PROV ou sistemas avançados de proveniência. Esses documentos são descritos na Tabela 2.1 (GROTH; MOREAU, 2013).

O PROV define três elementos fundamentais e as relações entre eles. Os elementos ou tipos são: Entidades, Atividades e Agentes. As entidades são elementos físicos, digitais ou conceituais que possuem características fixas. Documentos, arquivos, sistemas, um carro ou uma ideia são exemplos de entidade. As atividades são ocorrências dentro de um período de tempo que atuam sobre as entidades. Como exemplo de atividades pode-se citar a cópia e duplicação de objetos digitais ou a direção de um carro e a impressão de um livro.

Os agentes são caracterizados pela responsabilidade de iniciar uma atividade, pela existência de uma entidade ou pela atividade de outro agente. Pessoas são bons exemplos de agentes, mas alguns tipos particulares de entidades e atividades também podem ser agentes (GIL; MILES, 2012; MOREAU; MISSIER, 2013). A Figura 2.2 apresenta os símbolos utilizados para representar cada um dos três elementos fundamentais definidos pelo PROV.

Além dos três tipos fundamentais o PROV define sete relações básicas que representam a geração, o uso, a comunicação, a derivação, a atribuição, a associação e a delegação entre elementos. A Tabela 2.2 apresenta o nome dessas relações especificando o domínio e o contradomínio de cada uma delas. A Figura 2.3 representa graficamente.

Tabela 2.1: Descrição dos documentos que constituem o PROV

Documento	Tipo	Descrição
PROV-O	Desenvolvedores	Define uma ontologia OWL2 para o modelo de dados de proveniência.
PROV-XML	Desenvolvedores	Define um esquema XML para o modelo de dados de proveniência.
PROV-DM	Avançado	Define um modelo conceitual de dados de proveniência, incluindo diagramas UML. PROV-O, PROV-XML e PROV-N são serializações desse modelo conceitual.
PROV-N	Avançado	Define uma notação mais amigável (human-readable) para o modelo de proveniência. Também é utilizado para fornecer exemplos com o modelo conceitual, como os utilizados na definição de PROV-Constraints.
PROV-Constraints	Avançado	Define um conjunto de restrições sobre o modelo de dados do PROV que especifica a noção de proveniência válida. Essas definições são principalmente direcionadas aos desenvolvedores para criar validações.
PROV-AQ	Desenvolvedores	Define como utilizar mecanismos baseados na Web para localizar e recuperar informações de proveniência.
PROV-DC	Desenvolvedores	Este documento define um mapeamento entre Dublin Core e PROV-O.
PROV-Dictionary	Desenvolvedores	Descreve extensões do PROV para facilitar a modelagem da proveniência para a estrutura de dados de dicionário.
PROV-SEM	Avançado	Define uma especificação declarativa em termos de lógica de primeira ordem do modelo de dados PROV.
PROV-Links	Avançado	Define extensões para o PROV para permitir a ligação de informações de proveniência através de faixas de descrição de proveniência.

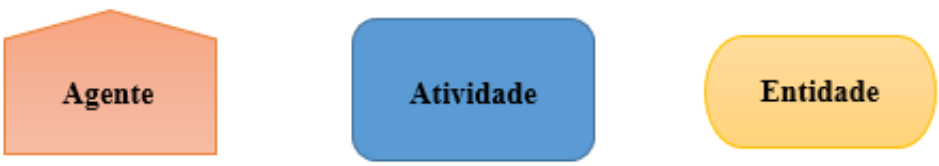


Figura 2.2: Os três símbolos definidos para os três elementos fundamentais no modelo PROV: agente, atividade e entidade.

Tabela 2.2: As relações básicas do PROV e seus domínios

Relação	Nome	Domínio	Contradomínio
Geração	WasGeneratedBy	Entidade	Atividade
Uso	Used	Atividade	Entidade
Comunicação	WasInformedBy	Atividade	Atividade
Derivação	WasDerivedFrom	Entidade	Entidade
Atribuição	WasAttributedTo	Entidade	Agente
Associação	WasAssociatedWith	Atividade	Agente
Delegação	ActedOnBehalfOf	Agente	Agente

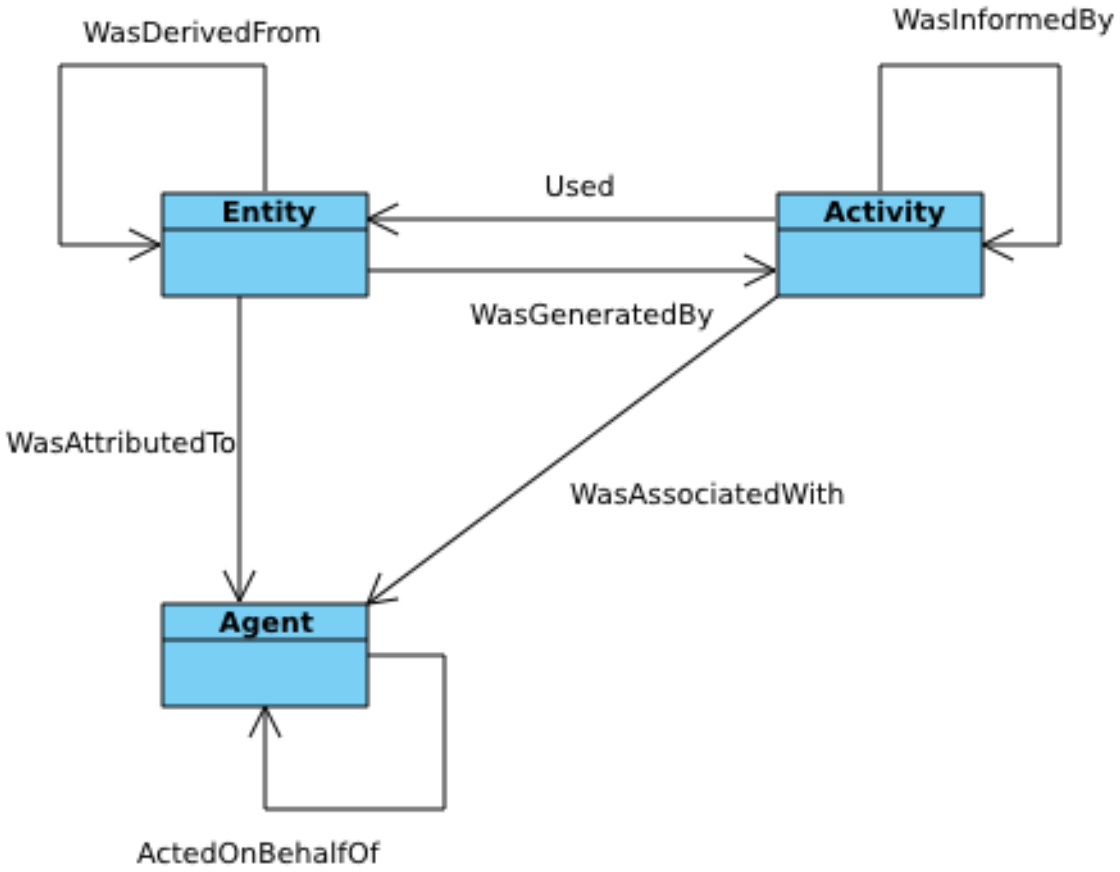


Figura 2.3: Estrutura básica do PROV com a representação dos tipos básicos e suas relações (MOREAU; MISSIER, 2013).

O PROV é um modelo livre de domínio que pode receber extensões para se adequar as especificidades do usuário ou desenvolvedor. Isso fez com que o PROV fosse bastante utilizado e adaptado para contextos específicos como o de workflow científicos (STITZ et al., 2016; SIRQUEIRA et al., 2016) e processos de software (DALPRA, 2016), entre outros.

No contexto deste trabalho, outro componente importante da especificação do PROV é o documento que trata de uso de ontologias, PROV-O. Uma ontologia, de modo geral, se preocupa com a identificação dos tipos de objetos e como descrevê-los (ANTONIOU; HARMELEN, 2004). Em computação, uma ontologia é descrita como uma especificação formal e explicita de uma conceituação compartilhada (GUARINO et al., 1998). A utilização de ontologias possibilita o compartilhamento de conhecimento sobre os conceitos de um determinado domínio, reutilização do conhecimento e processamento de máquina (YU, 2011).

A Ontologia PROV (PROV-O) expressa o Modelo de Dados PROV (PROV-DM) usando OWL2 (Web Ontology Language) (LEBO et al., 2013). Fornece um conjunto de classes, propriedades e restrições que podem ser usadas para representar e trocar informações de procedência geradas em diferentes sistemas e em diferentes contextos. A Figura 2.4 apresenta algumas especializações das classes básicas, suas relações e atributos.

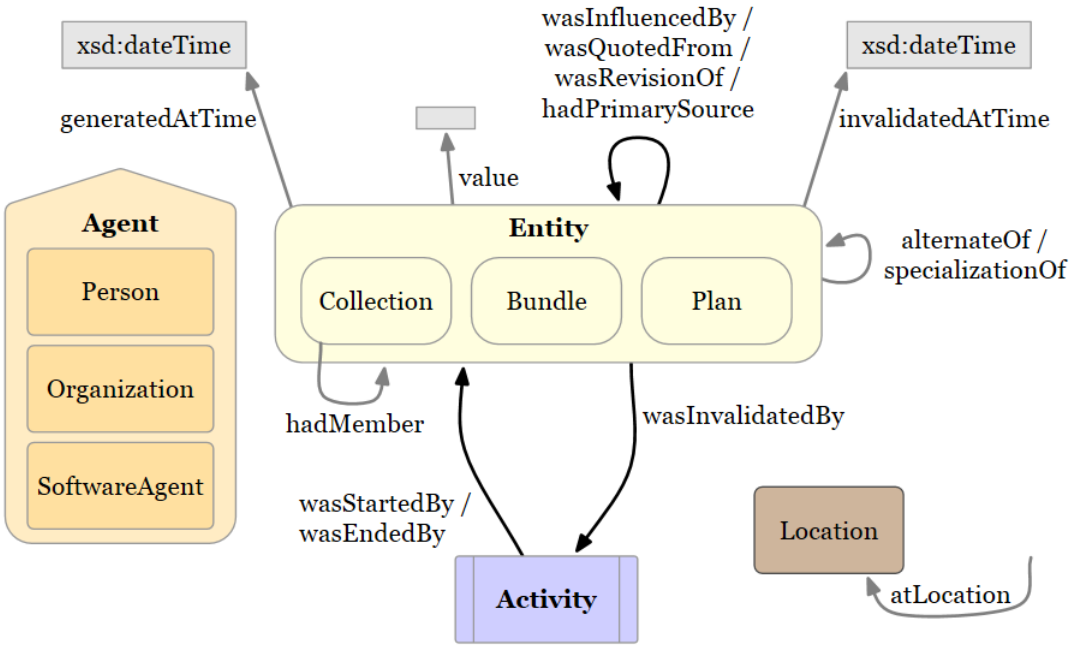


Figura 2.4: Representação das classes e suas propriedades do PROV-O (LEBO et al., 2013).

2.2 REDES COMPLEXAS

Um sistema complexo pode ser definido como um sistema com grande quantidade de elementos que são capazes de interagir entre si e com o ambiente, além de evoluir com o tempo. A principal característica encontrada em todos os sistemas complexos é a organização que formada sem qualquer intervenção externa. Nesses sistemas a soma é maior do que as partes (AMARAL; OTTINO, 2004).

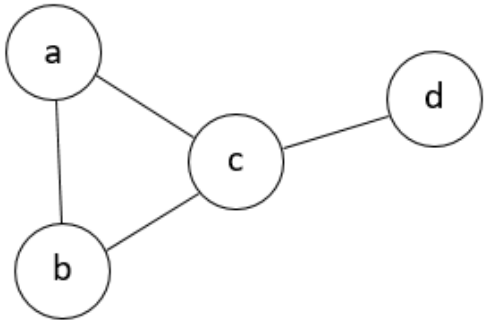
Três grandes áreas suportam os estudos sobre os sistemas complexos: a dinâmica não-linear, a física estatística e a teoria de redes complexas (AMARAL; OTTINO, 2004), sendo essa última a mais recente e um dos pontos principais desse trabalho. As redes, também conhecidas como grafos, são estruturas formadas por um conjunto de nós, também chamados de vértices, e um conjunto de ligações entre esses nós, chamados de arestas (HARARY, 1969). Os estudos sobre grafos iniciaram na matemática e encontraram aplicações em várias áreas como física, química, ciência da comunicação, tecnologia da computação, engenharia civil e elétrica, arquitetura, genética, psicologia, sociologia, economia, antropologia e linguística (HARARY, 1969). É comum essas áreas modelarem os sistemas complexos em estudo através de estruturas de redes, mapeando os elementos como vértices e as interações entre eles como arestas.

Deste ponto em diante, como convenção, será utilizado a palavra "grafo" ao se referir às redes complexas e a palavra "nó" no lugar de vértice.

Mais formalmente um grafo G pode ser definido como uma estrutura $G = (V, E)$ onde V é um conjunto discreto e E é uma família de elementos não vazios, definidos em função dos elementos de V (WASSERMAN; FAUST, 1994; BOAVENTURA, 2012). Os grafos podem ainda ser representados de forma gráfica, pela notação de conjunto e através da matriz de adjacência. A Figura 2.5 apresenta um mesmo grafo sendo representado de forma gráfica (a), onde os nós são os círculos e as arestas as linhas conectando os nós; matricial (b), a existência da conexão é representado pelo valor 1; e por conjunto (c).

As restrições sobre o conjunto V definem alguns conceitos sobre os grafos. Um grafo é considerado direcionado quando as ligações possuem o sentido especificado. Em outras palavras, considere $e \in E$ e $f \in E$ onde $e = v_1, v_2$ e $f = v_2, v_1$ logo $e \neq f$. No mesmo caso se $e = f$ o grafo é considerado não direcionado. A Figura 2.6 ilustra um grafo direcionado no formato gráfico (a) e de matriz de adjacência (b).

Os grafos são valorados quando as arestas possuem valores ou peso. Muito utilizado



(a)

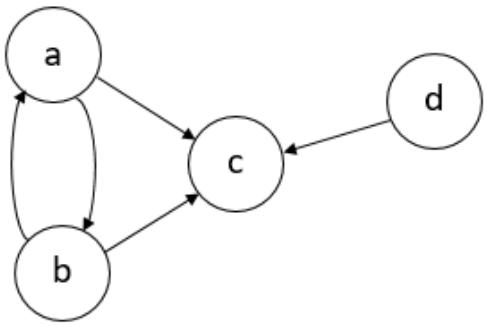
	a	b	c	d
a	-	1	1	0
b	1	-	1	0
c	1	1	-	1
d	0	0	1	-

(b)

$G = \{V, E\}$ sendo: $V = \{a, b, c, d\}$ e $E = \{ab, ac, bc, cd\}$

(c)

Figura 2.5: Três diferentes representações do mesmo grafo. Representação gráfica (a), matriz de adjacência (b) e representação por conjunto (c).



(a)

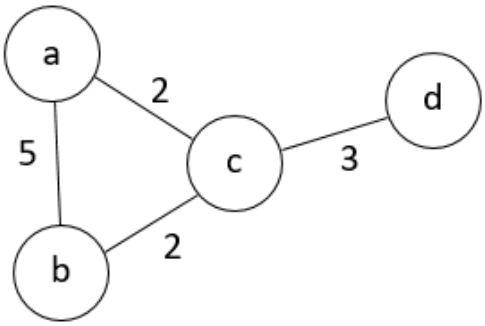
	a	b	c	d
a	-	1	1	0
b	1	-	1	0
c	0	0	-	0
d	0	0	1	-

(b)

Figura 2.6: Representação de um grafo direcionado no formato gráfico (a) e por matriz de adjacência (b).

para definir a distância entre dois pontos ou a força da conectividade entre os nós. Se as arestas não possuírem peso diferenciado o grafo é dito não valorado (ilustrado na Figura 2.6). Os valores das arestas geralmente são representados perto da linha que une os nós ou dentro do elemento da matriz, como apresenta a Figura 2.7. Na figura estão presentes a representação gráfica (a) e matricial de um grafo valorado (b).

Ao invés de pesos as arestas podem ser definas com diferentes qualidades ou tipos, assim dois nós podem ser conectados por arestas de diferentes tipos, isso caracteriza grafos multi-relacionais ou com múltiplas arestas. A Figura 2.8 representa graficamente um grafo multi-relacional.



(a)

	a	b	c	d
a	-	5	2	0
b	5	-	2	0
c	2	2	-	3
d	0	0	3	-

(b)

Figura 2.7: Representação de um grafo valorado no formato gráfico (a) e por matriz de adjacência (b).

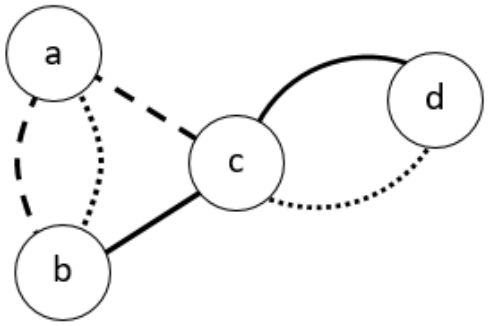


Figura 2.8: Representação de um grafo multi-relacional. As linhas seccionadas e contínuas representam diferentes ligações.

Como as análises apresentadas na Seção 4.1.3 são realizas no contexto de proveniência de dados, os grafos gerados são direcionados, não valorados e multi-relacionais, já que as relações não possuem peso mas qualidades distintas.

2.2.1 CARACTERÍSTICAS E MÉTRICAS

Com o passar dos anos, cientistas de várias áreas desenvolveram um conjunto extenso de ferramentas matemáticas, estatísticas e computacionais para analisar, modelar e compreender as redes. Com a representação de um sistema em forma de rede, essas ferramentas podem ser utilizadas para, com alguns cálculos, apresentar informações sobre a rede que podem ser muito úteis. (NEWMAN, 2010). Neste contexto, os cálculos realizados nas redes são obtidos a partir de algumas medidas e métricas.

Um conceito importante é o de caminho ou percurso em grafos. O caminho é a família de arestas adjacentes que conectam dois nós. O caminho pode ser fechado se o

último e o primeiro nó conectado são iguais ou aberto, caso contrário. Para determinar o comprimento do caminho, basta contar o número de arestas presentes nele, caso o grafo for não valorado. Em grafos valorados o comprimento do caminho é calculado com a soma dos valores de cada aresta utilizada. Desse conceito surge a definição de caminho mínimo ou distância entre dois vértices que é definido como o caminho com o menor cumprimento conectando os dois vértices.

O diâmetro de um grafo é calculado pela média de todos os caminhos mínimos ou através do maior caminho mínimo encontrado no grafo (NEWMAN, 2010). Quando o conceito de diâmetro é utilizado nesse trabalho refere-se ao maior dos caminhos mínimos.

Uma das métricas mais estudadas na análise de nós isoladamente é o grau. O grau é determinado pelo número total de conexões de um vértice. Caso o grafo seja direcionado, os nós possuem um grau de entrada, para as arestas que chegam; e um grau de saída para as arestas que saem do nó. O grau do nó é muitas vezes definido como a centralidade local. A centralidade tenta definir o vértice mais importante da rede. Outras formas de calcular a centralidade são chamadas de *closeness* e o *betweenness*. O *closeness* de um nó é calculado como a média das distâncias entre o vértice e todos os outros, demonstrado na Equação 2.1. O *betweenness* é definido após todos os caminhos mínimos serem definidos. O *betweenness* de determinado vértice é definido pelo número de caminhos mínimos que passa pelo vértice, como apresenta a Equação 2.2 (NEWMAN, 2010).

$$C_a = \frac{|V|}{\sum_b D_{a,b}} \tag{2.1}$$

Na Equação 2.1, o *closeness* do nó a é notado como C_a. No cálculo é considerado o número de nós $|V|$ e as distâncias entre o nó a para todos os outros nós (b) em $D_{a,b}$.

$$B_a = \sum_{b \neq a \neq c} \frac{\sigma_{bc}(a)}{\sigma_{bc}} \tag{2.2}$$

Na Equação 2.2, o *betweenness* do nó a é notado como B_a, onde σ_{bc} é o número total de menores caminhos do nó b para o nó c e $\sigma_{bc}(a)$ é o número de menores caminhos que passa por a.

Outro conceito central na análise de redes sociais é o da semelhança entre vértices.

Existem duas abordagens fundamentais para a construção de medidas de similaridade de rede, denominadas equivalência estrutural e equivalência regular. Dois vértices em uma rede são estruturalmente equivalentes se compartilham muitos dos mesmos vizinhos da rede. A equivalência regular é mais difícil de ser determinada. Dois vértices que possuem equivalência regular não compartilham necessariamente os mesmos vizinhos, mas eles têm vizinhos que são semelhantes, são do mesmo tipo, ou possuem outras características que os assemelham (NEWMAN, 2010).

Todas essas características são utilizadas na análise do grafo de proveniência. A forma com que essas métricas e conceitos são aplicadas e o que elas representam está apresentado no Capítulo 4.

2.3 VISUALIZAÇÃO DE SOFTWARE

A visualização é um importante meio de compreensão e é fundamental para apoiar a construção de um modelo mental a respeito de uma determinada situação ou realidade (SPENCE, 2007). Visualização é o processo de transformar dados em uma forma visual, permitindo usuários obterem informações. A visualização possui um grande papel na computação, auxiliando na compreensão humana (DIEHL, 2007).

Hoje existem duas grandes áreas de visualização: a visualização científica que processa dados físicos e a visualização de informação que processa dados abstratos (DIEHL, 2007). Pode-se entender a visualização de informação como parte de um processo de compreensão com o objetivo de atingir um conhecimento aprofundado a respeito de um tema a partir de um conjunto de dados (JACOBSON et al., 1999).

2.3.1 ASPECTOS DA VISUALIZAÇÃO DE SOFTWARE

Segundo Diehl (2007) a visualização de software é uma subárea da visualização da informação, que tem o objetivo de representar um software em três diferentes aspectos: a estrutura, o comportamento e a evolução.

- A estrutura refere-se às partes estáticas e às relações do sistema, isto é, aquelas que podem ser computadas ou inferidas sem executar o programa. Isso inclui o código do programa e as estruturas de dados, o gráfico de chamadas estáticas e a organização do programa em módulos (DIEHL, 2007). Este aspecto busca ilustrar as estruturas,

relacionamentos e propriedades das entidades do software. Os dados de entrada para o modelo são obtidos, na maioria das vezes, do código fonte dos aplicativos, sem a execução do programa (CASERTA; ZENDRA, 2011; WARE, 2012).

- O comportamento refere-se à execução do programa com dados reais e abstratos. A execução pode ser vista como uma sequência de estados de programa, onde um estado de programa contém o código atual e os dados do programa. Dependendo da linguagem de programação, a execução pode ser visualizada em um nível superior de abstração como funções que chamam outras funções ou objetos que se comunicam com outros objetos (DIEHL, 2007). Este tipo de visualização busca auxiliar a compreensão do comportamento do software (WARE, 2012).

- A evolução refere-se ao processo de desenvolvimento do sistema de software e, em particular, enfatiza o fato de que o código do programa é alterado ao longo do tempo para ampliar a funcionalidade do sistema ou simplesmente para remover erros (DIEHL, 2007). Esse tipo de visualização incrementa a visualização estática utilizando informações temporais do modelo (CASERTA; ZENDRA, 2011).

A visualização de proveniência abrange os metadados gerados com a execução de aplicações. Os dados de proveniência, dependendo de como são capturados e modelados, podem refletir as três áreas da visualização. A estrutura, por exemplo, pode ser mapeada quando os dados de proveniência são retirados do código fonte das aplicações. Na maior parte das vezes os dados são capturados com a execução dos programas, ou com registro dessa execução, refletindo o comportamento do sistema. Por fim, a evolução do sistema pode ser representada se a captura da proveniência é duradoura. Dessa forma, a visualização dos dados de proveniência atinge indiretamente as três áreas da visualização de software.

2.3.2 PERCEPÇÃO HUMANA

A percepção visual é o principal sistema cognitivo humano, considerando que 75% de todas as informações do mundo real são captadas visualmente (DIEHL, 2007). Algumas características da visão humana são importantes para serem consideradas:

- A cor é a percepção humana da luz. A tonalidade de uma cor está relacionada ao seu

comprimento de onda dominante, enquanto o brilho está relacionado à intensidade ou amplitude da onda.

- A percepção de padrões é a tarefa de decidir se os elementos visuais, como linhas e áreas, pertencem ao mesmo objeto.
- A identificação imediata de objetos distintos ocorre quando esses estão em um grupo homogêneo e diferenciados pela orientação, comprimento, largura das linhas, tamanho, cor, forma ou tamanho por exemplo.
- A percepção do movimento é a tarefa de decidir se os elementos visuais, como linhas e áreas, realizam o mesmo movimento em imagens subsequentes. É baseado no reconhecimento de padrões.

A importância da visualização de software é ligada a dependência que o ser humano possui da visão. Além disso, essas características percebidas pela visão humana são amplamente exploradas na geração de sistemas de visualização como descrito a seguir.

2.3.3 VISUALIZAÇÃO

O processo de visualização inclui quatro estágios básicos (DIEHL, 2007) ilustrados na Figura 2.9. Os estágios são combinados em um número de loops de feedback. Os quatro estágios são:

1. A coleta e armazenamento de dados em si.
2. O pré-processamento projetado para transformar os dados em algo que pode-se entender.
3. O hardware de exibição e os algoritmos gráficos que produzem uma imagem na tela.
4. O sistema perceptivo e cognitivo humano (o receptor).

É possível identificar claramente esses quatro estágios de visualização nos sistemas de visualização de proveniência. A coleta e armazenamento de dados é executado junto a aplicação ou aos dados históricos; o pré-processamento fica a cargo da ontologia PROV-O e/ou análises do grafo de proveniência; o hardware de exibição variam bastante, desde aplicações stand alone aos sistemas web; finalmente o receptor realiza o último estágio para compreender os dados.

Figura 2.9: Representação dos quatro estágios do processo de visualização: coleta e armazenamento de dados; pré-processamento; hardware e algorítimos e receptor.

2.3.4 REQUISITOS DE VISUALIZAÇÃO DE SOFTWARE

A identificação dos requisitos é uma parte importante para todo projeto de desenvolvimento de software. Em um survey, Kienle and Muller (2007) identificaram os requisitos básicos para a visualização de software. Os autores dividiram os requisitos funcionais em sete grupos: múltiplas visões; abstrações; pesquisa; filtros; proximidade com o código fonte; layouts automáticos e histórico.

Múltiplas visualizações são importantes para satisfazer a necessidade de diferentes usuários e podem facilitar a compreensão. As múltiplas visões também enfatizam as diferentes dimensões dos dados como a dimensão temporal ou os níveis de abstração. As abstrações também são importantes, já que os resultados podem se tornar muito complexos para serem visualizados. Um mecanismo de pesquisa é uma das funcionalidades mais úteis em visualização de software. A falta dela pode impedir o progresso do usuário. Filtrar informação em uma ferramenta de visualização pode ser visto como uma forma rudimentar de consulta (*query*). Isso permite ao usuário reduzir a quantidade de dados visualizados e limitar a análise. A proximidade com o código fonte é a capacidade da visualização de prover um fácil e rápido acesso ao código fonte. A disposição da informação em layouts automáticos é um dos aspectos funcionais das ferramentas de visualização. Um bom algoritmo de layout pode fazer os dados muito mais legíveis, por exemplo, minimizando a sobreposição de informações. Como usuários realizam interações e manipulações na visualização, um mecanismo histórico que permite reverter para os estados anteriores é importante. Outros requisitos também foram identificados, ainda que menos importantes que os anteriores, são eles: a utilização de cores; registro de anotações na visualização; o zoom; a navegação (*panning*); a deleção de elementos e a gravação da visualização.

Esses são requisitos importantes para guiar a construção de qualquer ferramenta de visualização. Mesmo que determinados contextos não sejam necessários a implementação de todos os requisitos, a maior parte deles deve estar presente. Esses requisitos são utilizados para comparar os trabalhos encontrados na área no Capítulo 3.

2.4 CONSIDERAÇÕES FINAIS DOS CAPÍTULO

Como mencionado anteriormente, a proveniência armazena informações importantes dos sistemas computacionais. As regras e restrições da ontologia geram novas informações sobre os dados de proveniência. Além disso, esses dados são modelados como um grafo e analisados, utilizando os conceitos apresentados nesse capítulo. Finalmente, todas as informações geradas nas etapas de análise são apresentadas através dos recursos de visualização. Portanto, todos esses conceitos são revisitados no decorrer desse trabalho, e são apresentadas as razões de suas aplicações.

3 TRABALHOS RELACIONADOS

Esse capítulo apresenta os trabalhos relacionados encontrados a partir da execução de duas revisões sistemáticas realizadas para dar suporte à abordagem descrita nessa dissertação. A primeira revisão tem como foco a visualização de dados de proveniência enquanto a segunda se detém na análise de dados de proveniência, o que abrange as duas principais áreas de pesquisa relacionadas a esta dissertação.

3.1 INTRODUÇÃO AO CAPÍTULO

Uma revisão sistemática e um mapeamento sistemático tem como objetivo identificar e avaliar os estudos relevantes em uma área de pesquisa particular. A revisão sistemática é um método de pesquisa para conduzir uma revisão da literatura (WOHLIN et al., 2012). Um mapeamento sistemático é uma revisão mais abrangente sobre uma área específica afim de identificar as evidências disponíveis sobre a área (KITCHENHAM; CHARTERS, 2007).

Qualquer pesquisa com o objetivo de compreender o estado da arte de uma área particular gera a necessidade de uma revisão sistemática. O mesmo ocorre com a escolha de praticantes que querem ou precisam usar evidências empíricas em sua tomada de decisão ou aprimorar atividades estratégicas (WOHLIN et al., 2012).

Foram realizadas duas diferentes revisões e mapeamentos sistemáticos, na Seção 3.2 são apresentadas as questões de pesquisa que servem de guia para as revisões. Na Seção 3.6 os resultados do mapeamento e da revisão sobre visualização de proveniência são apresentados, na Seção 3.7 são apresentados os resultados da revisão de análise de proveniência.

3.2 QUESTÕES DE PESQUISA

As questões de pesquisa precisam ser bem formuladas, porque, juntamente com o campo, definem o foco para a identificação de estudos primários, a extração de dados e a análise das publicações (WOHLIN et al., 2012). As questões de pesquisa foram divididas em questões de mapeamento (QM), para questões abrangentes que buscam compreender a

área de estudo; e questões de revisão (QR), que são questões mais específicas para identificar os trabalhos e conhecer as propostas da área. Todas as questões serão respondidas separadamente para cada revisão.

Segue-se as três questões de mapeamento para a condução do mapeamento sistemático.

- QM 1: Como as publicações são distribuídas ao longo dos anos?

A resposta dessa questão busca apontar se a área de pesquisa está em crescimento ao longo dos anos ou em retração.

- QM 2: Quem são os principais autores da área?

Identificar os principais autores da área é uma maneira de conhecer os grupos de pesquisa mais influentes que trabalham com o mesmo objeto de pesquisa e fornecer referências para futuras publicações.

- QM 3: Quais veículos de publicação estão mais interessados na área?

Com a resposta a essa pergunta, é possível identificar oportunidades para novas publicações na área, bem como os locais onde pode-se encontrar futuras pesquisas relacionadas ao objeto de estudo.

Duas questões de revisão foram formuladas.

- QR 1: Quais são as propostas encontradas?

A resposta a esta questão inclui todas as publicações que apresentam uma abordagem ou software que inclui o objetivo da pesquisa (visualização ou análise de dados de proveniência).

- QR 2: Quais técnicas são utilizadas na área?

Dentro das várias opções e técnicas existentes, essa questão busca identificar as mais apropriadas para serem aplicadas à proveniência de dados.

- QR 3: Quais modelos de proveniência são utilizados?

Esta questão pode indicar uma tendência no campo ou destacar modelos de origem ainda pouco explorados.

3.3 CRITÉRIOS DE INCLUSÃO E EXCLUSÃO

Todos os documentos candidatos devem ser analisados para verificar a relevância para a pesquisa. O processo usado para incluir ou excluir um trabalho foi baseado em Kitchenham and Charters (2007) e Prikladnicki and Audy (2010), e compreende os seguintes critérios de inclusão (CI):

- CI 1a: Publicações que apresentam abordagens para visualizar dados de proveniência.

- CI 1b: Publicações que apresentam abordagens para analisar dados de proveniência.

Este é o principal critério de inclusão que pretende ser alcançado com a elaboração da *string* de busca. Esse critério foi dividido para se adequar a cada escopo. Na questão CI 1a é aceita qualquer publicação que ofereça uma abordagem ou software para visualizar dados de proveniência. Na questão CI 1b são aceitos abordagens e aplicações que analisam os dados de proveniência.

- CI 2: Publicações dentro da área da ciência da computação.

A palavra visualização é comumente usada na visualização científica, que difere do foco desta pesquisa que é a visualização do software (DIEHL, 2007). Da mesma forma, a proveniência significa a origem de algo e é usada em outras áreas sem se referir à proveniência dos dados. Para melhorar a precisão da pesquisa, apenas as publicações no campo da ciência da computação foram aceitas.

- CI 3: Publicações do ano de 2007 e posteriores.

O primeiro modelo de proveniência de dados amplamente utilizado foi proposto em 2007, por isso apenas publicações desse ano, e posteriores, são aceitas nessa pesquisa.

De forma similar, os critérios de exclusão (CE) foram elaborados e aplicados com a identificação de todas as publicações candidatas. Os critérios de exclusão são os seguintes:

- CE 1: Documentos duplicados.

É desnecessário a manutenção de documentos idênticos. Esse processo é simples de ser automatizado e diminui o número de publicações facilitando as etapas seguintes.

- CE 2: Publicações não escritas em inglês.

Publicações em outras línguas fogem do conhecimento do pesquisador. Como a maioria das pesquisas na área são publicadas em inglês, esta foi escolhida como idioma padrão para este trabalho.

- CE 3: Documentos sem informações básicas de autor ou título.

A falta de informação nas publicações, mesmo quando encontradas pela pesquisa, impede a identificação correta.

- CE 4: Estudos secundários, chamadas de artigos e anais de eventos.

Esses documentos referem-se a várias publicações importantes, facilitando o resultado positivo com a string de busca, mas não apresentam o conteúdo da pesquisa e seus detalhes para serem selecionados. Assim, essas obras foram excluídas aguardando a identificação direta do artigo citado nessas obras.

- CE 5: Título, resumo e palavras-chave claramente irrelevantes.

Os títulos, resumos e palavras-chave dos artigos são verificados se correspondem às questões de pesquisa. As publicações que se apresentam irrelevantes são excluídas.

3.4 ESTRATÉGIA DE BUSCA

As questões de pesquisa são a base para formar os termos e / ou palavras-chave que formam a string de busca a ser aplicada em bibliotecas ou bases digitais. Após a definição das palavras-chave, a string é enriquecida com sinônimos dos termos para ampliar os resultados e evitar que publicações relevantes fiquem de fora do grupo de publicações candidatas (PETERSEN, 2011).

Os termos de pesquisa são organizados através do PICOC (População, intervenção, comparação, saída e contexto) (KITCHENHAM, 2004). A população é onde a evidência é coletada. A intervenção é a mesma aplicada no estudo empírico. A comparação, se existir, é algo que deve ser comparado com a intervenção. A saída é o resultado do experimento e não deve ser significativo apenas estatisticamente, mas também de forma prática. O contexto, quando definido, é uma extensão da população (WOHLIN et al., 2012).

Para ambas as pesquisas, não houve comparação e o contexto foi considerado irrelevante, por isso as strings de busca foram montadas sem esses dois termos. As palavras-chave encontradas através deste processo e os termos de pesquisa finais, incluindo sinônimos, são apresentados na tabela 3.1 para a visualização de dados de proveniência e na tabela 3.2 para análise de dados de proveniência.

Tabela 3.1: Termos de pesquisa identificados através do PICOC para visualização de dados de proveniência

PICOC	Palavras Chave	Termos de Busca
Population	*proveniência*	*provenance*
Intervention	*visualização, exibição*	*visual*, display*, exibi**
Comparison		
Outcomes	*ferramenta, software, programa, sistema, modelo, método, técnica, abordagem*	*tool*, software*, system*, model*, method*, technique*, approach**
Context		

Tabela 3.2: Termos de pesquisa identificados através do PICOC para análise de dados de proveniência

PICOC	Palavras Chave	Termos de Busca
Population	*proveniência*	*provenance*
Intervention	*análise, inferência, avaliação*	*analy*, inference*, assessment**
Comparison		
Outcomes	*ferramenta, software, programa, sistema, modelo, método, técnica, abordagem*	*tool*, software*, system*, model*, method*, technique*, approach**
Context		

Os termos identificados em cada parte do PICOC são agrupados com o operador OR, os grupos gerados são agrupados com o operador AND. Os símbolos de asterisco (*) são usados como um curinga, ajudando na busca de várias palavras semelhantes com o mesmo radical. A tabela 3.3 apresenta a *string* de busca genérica gerada com a concatenação dos termos de busca.

Para encontrar as publicações candidatas, as strings de busca são aplicadas às bibliotecas digitais que foram selecionadas anteriormente por Prikladnicki and Audy (2010) para pesquisas na área da ciência da computação:

- ACM Digital Library (portal.acm.org)
- IEEE Digital Library (ieeexplore.ieee.org)

- EI Compendex (www.engineeringvillage.com)
- Science @ Direct (www.sciencedirect.com)
- Scopus (www.scopus.com)
- Springer Link (link.springer.com)

Tabela 3.3: Strings genéricas de busca geradas com o PICOC

String básica de busca para a visualização de dados de proveniência
(“provenance”) AND (“visual*” OR “display” OR “exibi*”) AND (“tool*” OR “software*” OR “program*” OR “system*” OR “model*” OR “process*” OR “framework*” OR “method*” OR “technique*” OR “approach*”)
String básica de busca para a análise de dados de proveniência
(“provenance”) AND (“analy*” OR “inference*” OR “assessment*”) AND (“tool*” OR “software*” OR “program*” OR “system*” OR “model*” OR “process*” OR “framework*” OR “method*” OR “technique*” OR “approach*”)

A pesquisa booleana em cada biblioteca digital funciona de forma diferente, sendo comum a necessidade de adaptar a string para pesquisar algumas das bases. Neste caso, apenas a base IEEE exigiu mudanças na string de busca, por causa da sua limitação de uso de caracteres curinga (*). As strings usadas na base IEEE são apresentadas na tabela 3.4:

Três artigos de controle foram selecionados para a revisão sobre visualização. Os artigos de controle servem para verificar a qualidade dos resultados obtidos pela *string* de busca. Os três artigos são:

- BOYD, Madelaine D. Inprov: visualizing provenance graphs with radial layouts and time-based hierarchical grouping. **Harvard College Cambridge, Massachusetts**, 2012.
- HOEKSTRA, Rinke; GROTH, Paul. PROV-O-Viz-understanding the role of activities in provenance. In: **International Provenance and Annotation Workshop**. Springer, Cham, 2014. p. 215-220.

- KOHWALTER, Troy et al. Prov viewer: a graph-based visualization tool for interactive exploration of provenance data. In: **International Provenance and Annotation Workshop**. Springer International Publishing, 2016. p. 71-82.

Tabela 3.4: String de busca adaptada para a base digital IEEE

String de busca adaptada para a base digital IEEE para visualização de dados de proveniência
(provenance) AND (visual* OR display OR exibi*) AND (tool OR software OR program OR system OR model OR process OR framework OR method OR technique OR approach)
String de busca adaptada para a base digital IEEE para análise de dados de proveniência
(provenance) AND (analy* OR inference OR assessment*) AND (tool OR software OR program OR system OR model OR process OR framework OR method OR technique OR approach)

Na revisão sobre análise de dados de proveniência dois artigos foram selecionados como artigos de controle, são eles:

- CHEAH, You-Wei; PLALE, Beth. Provenance analysis: Towards quality provenance. In: **E-Science (e-Science), 2012 IEEE 8th International Conference on**. IEEE, 2012. p. 1-8.

- OLIVEIRA, Weiner et al. A Framework for Provenance Analysis and Visualization. **Procedia Computer Science**, v. 108, p. 1592-1601, 2017.

3.5 CONDUÇÃO

A condução das revisões e mapeamentos sistemáticos foi realizada em julho de 2017 seguindo 4 passos

1. Execução da string de busca

2. Preparação das publicações candidatas

3. Aplicação dos critérios de seleção

4. Avaliação da qualidade

A primeira etapa da condução consiste em executar a *string* de busca diretamente nas bases selecionadas. A pesquisa foi realizada de acordo com os critérios CI2 e CI3 usando as ferramentas disponíveis em cada base. As buscas em todas as bases resultaram nas publicações chamadas de publicações candidatas.

Na segunda etapa, os resultados foram exportados para um arquivo no formato bibtex, contendo a informação dos artigos candidatos. A ferramenta Zotero[1] foi usada quando as bases exportavam os resultados em um formado indesejado (Springer Link) ou com informações importantes ausentes (ACM Digital Library). Zotero pode coletar as informações básicas de todas as publicações resultantes da pesquisa e exportá-las no formato bibtex.

Com a informação dos artigos candidatos no formato bibtex, estes foram importados para a ferramenta Parsifal[2] para auxiliar na seleção de artigos. O Parsifal é uma ferramenta on-line que auxilia durante a implementação da revisão sistemática, incluindo a identificação automática de publicações duplicadas. Esta ferramenta permite a marcação de todas as publicações como "aceitas" ou "excluídas" e registra o motivo de cada decisão. O Parsifal foi usado para todas as etapas subsequentes da revisão.

Das publicações candidatas, o terceiro passo de condução foi feito com a aplicação dos critérios de exclusão. A Tabela 3.5 apresenta a quantidade de publicações eliminadas em cada critério de exclusão da revisão sobre visualização, enquanto a Tabela 3.6 mostra os resultados da revisão sobre análise de proveniência.

Tabela 3.5: Número de artigos encontrados nas bases digitais e os números de publicações eliminadas por critério de exclusão da revisão de **visualização de proveniência**

Bases	**Publicações Candidatas**	**CE1**	**CE2**	**CE3**	**CE4**	**CE5**	**Aceitos**
ACM Digital Library	163	26	0	0	5	132	0
IEEE Digital Library	80	2	0	0	1	77	0
EI Compendex	112	26	0	0	8	70	8
Science@Direct	972	80	0	6	24	860	2
Scopus	1534	191	0	0	35	1.290	18
Springer Link	881	72	1	29	7	770	2
Total	**3.742**	**397**	**1**	**35**	**80**	**3.199**	**30**

[1]https://www.zotero.org/
[2]https://parsif.al/

Tabela 3.6: Número de artigos encontrados nas bases digitais e os números de publicações eliminadas por critério de exclusão da revisão de **análise de proveniência**

Bases	**Publicações Candidatas**	**CE1**	**CE2**	**CE3**	**CE4**	**CE5**	**Aceitos**
ACM Digital Library	237	55	0	0	2	177	0
IEEE Digital Library	227	4	0	0	3	220	0
EI Compendex	254	85	0	0	22	138	9
Science@Direct	1326	70	0	10	58	1196	2
Scopus	1196	137	0	0	105	943	11
Springer Link	399	49	0	0	0	348	2
Total	**3.713**	**400**	**0**	**10**	**190**	**3025**	**24**

O critério CE1 foi executado primeiro e automaticamente pela Parsifal, em seguida, os títulos e resumos das publicações foram avaliados e os critérios CE2, CE3 e CE4 foram aplicados imediatamente. O CE5 exigiu mais atenção na análise do resumo e às vezes a publicação completa foi consultada para a aplicação do critério.

O quarto passo é a avaliação da qualidade que buscou através de uma lista de verificação (tabela 3.7) avaliar e eliminar publicações de baixa qualidade. A lista de verificação contém oito perguntas, apresentadas na tabela 3.7, essas perguntas foram respondidas para avaliação de cada publicação. As respostas foram definidas entre Sim, parcial ou não, e tiveram um valor agregado de 1, 0,5 e 0, respectivamente. A soma dos valores de cada resposta para uma determinada publicação é a nota final do artigo. Para evitar publicações de baixa qualidade, todas as publicações que obtiveram pontuação igual ou inferior ao último quartil (8/4 = 2) do total de pontos possíveis (8) foram eliminados nesta fase. A segunda questão foi especificada para cada revisão, dividida em 2a para visualização e 2b para análise.

Tabela 3.7: Questões de qualidade aplicadas às publicações aceitas.

Questão	**Pergunta**
1	Os objetivos da pesquisa são especificados claramente?
2a	A pesquisa possui a visualização de dados de proveniência como foco?
2b	A pesquisa possui a análise de dados de proveniência como foco?
3	As técnicas utilizadas são claramente descritas?
4	A seleção das técnicas utilizadas é justificada?
5	A proposta foi avaliada de forma adequada?
6	Os resultados negativos (se existem) são apresentados?
7	Os pesquisadores discutem alguma ameaça a validade dos resultados?
8	Os resultados são obtidos por mais de uma pesquisa?

Na primeira revisão, das 30 publicações restantes após os critérios de exclusão, 14

apresentaram pontuação maior do que a 2 na avaliação de qualidade e foram aceitas. Na revisão de análise, das 24 publicações selecionadas, 10 foram aceitas na avaliação de qualidade.

3.6 VISUALIZAÇÃO DE DADOS DE PROVENIÊNCIA

As 30 publicações encontradas na revisão e mapeamento sistemático sobre visualização de dados de proveniência foram utilizadas para responder as questões de mapeamento, apresentados na Subseção 3.6.1. Para responder as questões de revisão foram utilizados os artigos restantes após a avaliação de qualidade, 14 no total. O resultado da revisão sobre visualização está apresentado na Subseção 3.6.2.

3.6.1 RESULTADOS DO MAPEAMENTO SISTEMÁTICO

A Figura 3.1 apresenta a distribuição das publicações entre 2007 e 2017 para responder a QM 1 (como as publicações são distribuídas ao longo dos anos?). Em destaque o ano de 2016 com a maior quantidade de publicações, quatro no total. A linha tracejada vermelha mostra a tendência do gráfico.

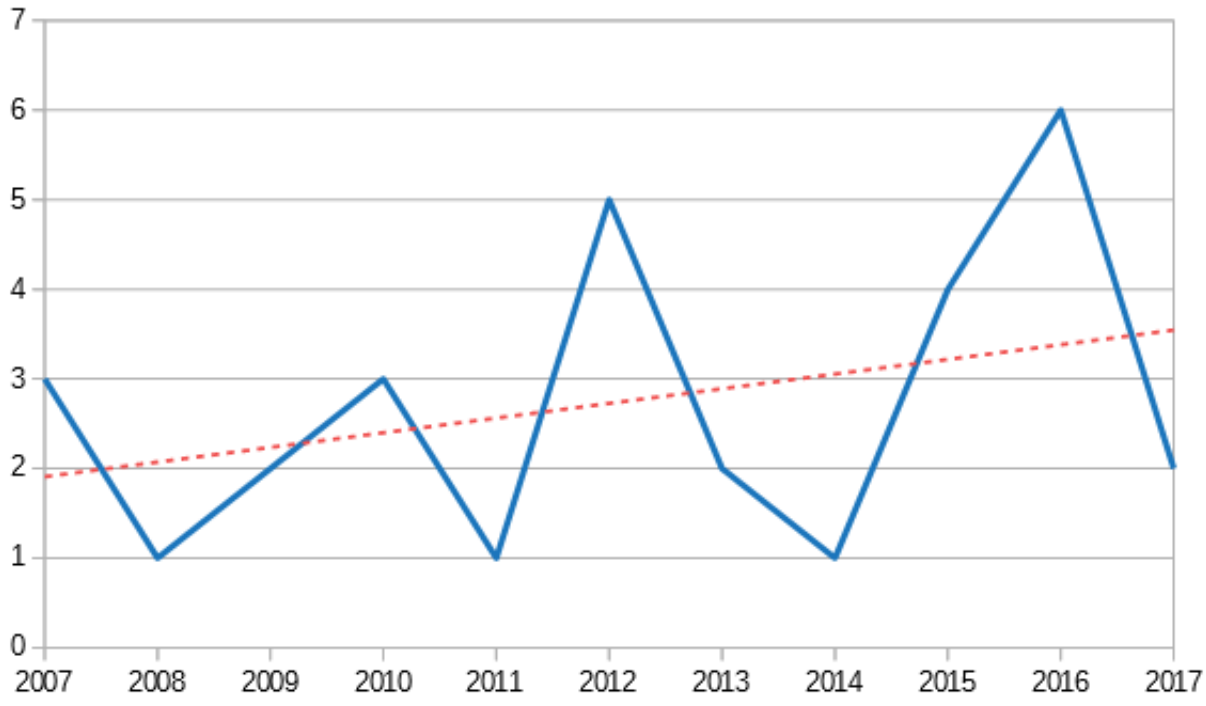


Figura 3.1: Distribuição pelos anos das publicações aceitas no mapeamento de visualização de dados de proveniência.

É possível verificar uma tendência crescente no número de publicações, isso é um indício do aumento do interesse de pesquisadores na área de visualização de dados de proveniência.

Todos os pesquisadores foram identificados e relacionados para responder a QM 2 (quem são os autores principais da área?). Um total de 100 pesquisadores são autores das 30 publicações do mapeamento. Dez deles se destacam com mais de uma publicação cada. A Tabela 3.8 relaciona os dez autores que mais publicaram na área.

Tabela 3.8: Número total de publicações dos dez pesquisadores que mais publicaram na área.

Pesquisador	**Número de Publicações**
Ludäscher, B. A. C.	5
Anand, M. K. A.	4
Bowers, S. B.	4
Bouttaz, T.	2
Eckhardt, A.	2
Eckhardt, A.	2
Edwards, P.	2
Mellish, C.	2
Missier, P. B.	2
Chen, P.	2
Plale, B. A.	2

As publicações encontradas também foram distribuídas pelos seus veículos de publicação. Assim pode-se identificar os principais veículos que possuem interesse na área. Dessa forma chega-se a resposta da QM 3 (quais veículos de publicação estão mais interessados na área?). O IPAW (*International Provenance and Annotation Workshop*) claramente se destaca dos outros veículos com dez publicações, todos os outros veicularam apenas uma publicação. Como o IPAW é um workshop específico sobre proveniência, era esperado que trouxe-se mais publicações que outros veículos.

Finalizando o mapeamento sistemático, as respostas das QMs foram resumidas na Tabela 3.9. É importante ressaltar que esses resultados são específicos do conjunto de publicações aceitas para esse mapeamento e que as decisões tomadas durante toda sua execução estão ligadas fortemente as questões de pesquisas apresentadas.

Tabela 3.9: Resumo das respostas das questões de mapeamento sistemático da revisão de visualização de dados de proveniência

Pergunta	Respostas
QM 1	As publicações estão bem distribuídas pelos anos mas demonstram um crescimento.
QM 2	Ludäscher, B. A. C. se destaca dos demais pesquisadores com cinco publicações.
QM 3	IPAW se destaca fortemente com dez publicações.

3.6.2 RESULTADOS DA REVISÃO SISTEMÁTICA

Com a intenção de responder a QR 1 (quais são as propostas encontradas?) e a QR 2 (quais técnicas são utilizadas na área?) as publicações aceitas tiveram sua proposta resumida nos parágrafos seguintes. Juntamente com a descrição de cada publicação foi realizada uma comparação com a proposta apresentada nessa dissertação.

A ferramenta Provenance Browser é apresentada em (ANAND et al., 2010) e gera visualizações a partir do resultado de uma pesquisa nos dados realizada pelo usuário. A ferramenta foi desenvolvida para ler arquivos XML com o registro das execuções de workflows (workflow traces) e armazenar em um banco de dados relacional. Os autores desenvolveram técnicas de otimização para reduzir o grafo de proveniência antes de armazenar os resultados da busca no banco relacional. Os autores também desenvolveram uma linguagem de busca de auto nível (query language – QLP) que é reescrita em SQL para ser executada na base relacional. O resultado da query é apresentado em três visualizações diferentes: o dependency history view, que combina dependência de dados e grafos de invocação de processos; o collection history view, apresenta os dados de entrada e saída por invocação; e o graph view, mostra as dependências dos processos. O Provenance Browser foi integrado com o Sistema de Gerenciamento de Workflow Cientifico (SGWfC) Kepler mas também pode ser utilizado como uma aplicação independente.

A proposta de (ANAND et al., 2010) fornece visualizações para uma parte dos dados selecionada a partir da pesquisa. Além disso, o resultado sofre um processo de sumarização para ser gerada a visualização. A proposta dessa dissertação apresenta uma visualização integral dos dados e possui recursos para filtrar as visualizações a fim de permitir ao usuário navegar e compreender os dados. Além disso, nossa proposta gera visualizações a partir de um modelo de proveniência padrão, ao contrário da proposta de Anand et al. (2010) que faz a leitura de arquivos XML.

O InProv, apresentada por Borkin et al. (2013), utiliza um zoom semântico para

filtrar as informações visualizadas. Nesta ferramenta, as informações de proveniência são dispostas em um layout radial, chamado de anel. Cada setor do anel representa um nó (ou grupo de nós) e suas relações em determinado contexto. Informações sobre o setor são apresentadas com o passar do mouse em um tooltip. Cada grupo de nó pode ser explorado gerando um novo anel com seus nós e/ou subgrupos. Uma miniatura dos anéis visitados é mantida à direita da visualização e seus nomes acima, permitindo o retorno mais fácil aos anéis passados. Os grupos são formados a partir da análise temporal dos dados e uma linha do tempo é disposta abaixo da visualização, onde é destacado o grupo atual explorado. O usuário também pode escolher o agrupamento pelo método de "árvore de processo". As relações são representadas por linhas dentro do anel, conectando os segmentos (nós) correspondentes. O sentido de cada conexão é determinado pelo fluxo dos dados. A Figura 3.2 mostra uma visualização gerada pelo InProv, exemplificando suas áreas.

Figura 3.2: Tela da ferramenta InProv (BORKIN et al., 2013). O círculo central apresenta a relação dos nós dentro do contexto. A área de contexto à direita mostra os círculos visitados. A linha do tempo à baixo destaca o grupo visualizado.

Muitos recursos de visualização são utilizados no InProv, as conexões utilizam cores para determinar ligações de chegada ou saída, azul ou vermelho respectivamente, do setor selecionado. As conexões são mais claras (transparentes) se poucas ligações existem entre dois setores, já com um grande número de ligações entre dois setores, as cores são mais

intensas e opacas. O tamanho do setor é proporcional ao número de nós que o compõe. Um esquema de cores foi adotado para os setores: cinza escura para processos, branco para arquivos e cinza claro para todos os outros tipos, onde o setor selecionado é apresentado em roxo.

O InProv possui a aplicação de bons recursos de visualização que auxilia o usuário na navegação e compreensão dos dados. Diferente da nossa proposta, um layout radial foi apresentado, onde é mais usual a representação de grafo. Além disso a seleção das opções visuais, como as cores por exemplo, não é justificada no InProv,, ao contrário do que ocorre na proposta dessa dissertação onde até os símbolos são padronizados e baseados no modelo de proveniência.

A proposta do artigo (KADIVAR et al., 2009) é apresentar e descrever o sistema CZ-Saw. O CZSaw possui um modelo simbólico representado como um grafo de dependências que é utilizado para gerar as visualizações. As interações do usuário com o sistema são capturadas, processadas por scripts que atualizam o modelo simbólico, mas o usuário pode alterar o modelo diretamente. No modelo, as cores são utilizadas para destacar tipos deferentes de nós.

O CZSaw possui ainda outros tipos de visualização que auxiliam o usuário na compreensão dos dados, mas nenhum padrão visual foi estabelecido para gerar as visualizações. As visualizações criadas pela nossa proposta possuem símbolos padronizados pelo PROV e outras opções para facilitar a compreensão do usuário sobre os dados apresentados. Assim, considera-se que a proposta dessa dissertação utiliza melhor os recursos visuais para facilitar a interação e tornar o sistema mais amigável.

Para visualizar e analisar grandes grafos de proveniência, Chen et al. (2014) utiliza a ferramenta Cytoscape (LOPES et al., 2010) que é um grande projeto de código aberto com o objetivo de visualizar a rede de interação molecular adicionando anotações e outras informações extras. Na proposta de Chen et al. (2014) os dados de proveniência são apresentados como grafos e o processamento dos dados é realizado ao mesmo tempo que eles são gerados. No processo, alguns dados são filtrados pelo usuário para a utilização de técnicas de mineração de dados, estatística e redução de grafo. Além disso, estas técnicas são utilizadas para diminuir as informações que devem ser exibidas para dar foco nas informações mais importantes.

Na proposta dessa dissertação, o grafo de proveniência não é alterado e recursos visuais

são utilizados para reduzir a quantidade de informação disposta nas visualizações. O resultado das análises da proposta desta dissertação,é apresentado para o usuário que toma a decisão final sem alterar os dados.

A abordagem Probe-It! é apresentada em Rio and Silva (2007) como uma ferramenta interativa de visualização de dados de proveniência. Essa ferramenta possui três diferentes visualizações: de resultados, de justificativas e de proveniência. A primeira visualização (*results view*) mostra os resultados finais ou intermediários da execução do workflow científico, que são obtidos através de um plug-in de integração com o gerenciador de workflow utilizado. Essa visualização não trabalha com os dados de proveniência em si. A segunda visualização (*justification view*) tem a proveniência de processos como foco e suas informações associadas. Essas informações são apresentadas em um grafo direcionado. A última visualização (*provenance view*) apresenta informações das fontes (documentos, dados etc.) utilizadas e algumas informações mais detalhadas (tempo de acesso, duração etc.). O usuário pode interagir com a ferramenta, navegando entre os três tipos de visualização. O Probe-It! utiliza a Proof Markup Language – PML (SILVA et al., 2006) para codificar os dados de proveniência e gerar inferências. PML é uma linguagem baseada em OWL e possui regras já estabelecidas que levam a inferências.

O framework Visionary, proposto no contexto desta dissertação, utiliza um padrão de proveniência e pode ser facilmente integrado em outras ferramentas que utilizam o mesmo padrão. O foco do Visionary é a proveniência, ao contrário do Probe-It!, que gera visualizações científicas. As inferências do framework Visionary são realizadas em dois níveis: nas regras e restrições da ontologia (OWL2) e também nas análises do grafo de proveniência.

O PROV-O-Viz (HOEKSTRA; GROTH, 2014) é uma ferramenta de visualização disponível na web , para auxiliar na compreensão de proveniências baseadas no padrão PROV. Esta ferramenta utiliza um diagrama de fluxo *(sankey diagram*) com o objetivo de identificar as atividades mais importantes dentro do fluxo de dados e compreender como os dados percorrem atividades selecionadas e adiciona características específicas de proveniência. O PROV-O-Viz pode receber o arquivo RDF do PROV-O ou se conectar através de SPARQL para receber os dados. A Figura 3.3 mostra uma visualização gerada pelo PROV-O-Viz.

O usuário pode selecionar uma atividade específica que ganha o foco da ferramenta

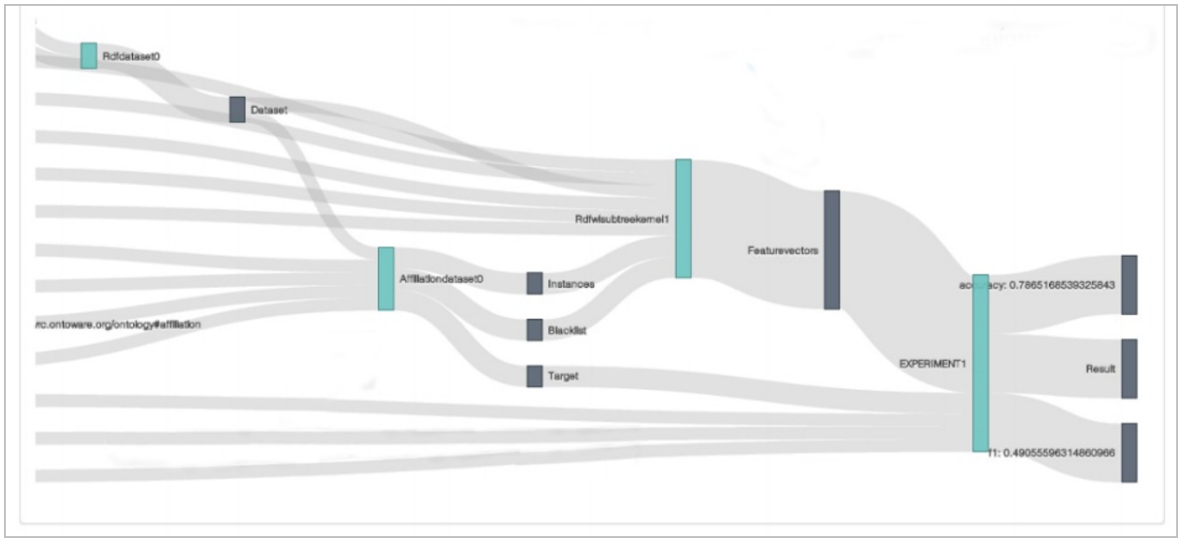


Figura 3.3: Sankey diagram gerado pelo PROV-O-Viz a partir dos dados de proveniência (HOEKSTRA; GROTH, 2014).

para gerar a visualização. A altura das caixas é proporcional ao fluxo de dados que passa por cada atividade. Com a ajuda da ontologia do PROV (PROV-O), o sistema também infere informações não declaradas e apresenta na visualização.

O foco do PROV-O-Viz é trabalhar com o fluxo dos dados coletados, o que pode restringir a análise dos usuários do sistema. Os recursos de visualização também são pouco explorados nesta ferramenta, poucas cores distintas, nenhum ícone e nenhuma diferença de formato nos processos apresentados (caixas). O framework Visionary, além de apresentar itens de visualização específicos, ainda apresenta uma etapa extra de análise, a análise de grafo, detalhada no capítulo 4.

O Provenance Explorer é apresentado por Hunter and Cheung (2007), com o objetivo de apresentar informações de proveniência simples para explicar a metodologia e validar os resultados de experimentos científicos, sem comprometer a propriedade intelectual dos pesquisadores. O Provenance Explorer possui três componentes básicos: uma base de conhecimento, composta por arquivos SWRL/OWL com dados e metadados das instâncias de proveniência e das regras de inferência; o visualizador de proveniência; e a máquina de inferência que utiliza as regras dos arquivos SWRL/OWL para gerar conhecimento novo sobre os dados. O visualizador possui três áreas ou painéis: a primeira área apresenta graficamente os processos de proveniência modelados usando grafos RDF; na segunda área o usuário pode arrastar os nós do primeiro painel para gerar um grafo menor para publicação; e a última área é utilizada para apresentar detalhes dos nós selecionados na

primeira visualização.

Nossa proposta possui um foco diferente do Provenance Explorer que se concentra na publicação do grafo de proveniência. Talvez por isso os recursos de exploração e análise da visualização apresentadas são muito limitados. Os tipos dos nós não são diferenciados e a visualização, mesmo para um grafo limitado, possui muitas informações, sobrecarregando a visualização.

Em Karsai (2016) é apresentado um sistema web para visualizar proveniências baseadas no modelo PROV. Neste trabalho o grafo de proveniência é agrupado para auxiliar na visualização e os nomes dos agrupamentos são gerados com o intuito de dar mais significado ao agrupamento. Técnicas também são aplicadas para evitar a falsa dependência entre os grupos ou referências circulares. Na visualização do grafo, o usuário pode mostrar detalhes dos nós, navegar (*panning*), dar *zoom*, reorganizar nós e criar grupos manualmente. O sistema permite renomear nós, refazer e desfazer ações e exportar parte ou todo grafo. Os tipos dos nós utilizam a mesma simbologia apresentada pelo PROV (pentágonos para agentes, elipses para entidades e retângulos para as atividades).

O framework Visionary também propõe uma ferramenta web e utiliza os símbolos do modelo PROV para gerar as visualizações, mas permite apresentar um número muito maior de informação que a proposta de Karsai (2016).

A proposta de Kohwalter et al. (2016), denominada Prov Viewer, utiliza o padrão PROV para modelagem de proveniência. Esta abordagem gera um grafo de proveniência interativo para prover visualizações e utiliza várias técnicas desenvolvidas pelos autores para tratar características de diferentes cenários. Alguns dos recursos empregados são: collapsing, que evidencia as informações relevantes dentro de um grafo; filtering, que remove informações não relevantes para determinada análise; graph merge, que integra a análise de diferentes execuções; specialized layout, que reorganiza o grafo para auxiliar na compreensão; domain configuration, que customiza a visualização para necessidades específicas; além de explorar técnicas de visualização para auxiliar na distinção de informações e importar dados com o formato do PROV-N. A ferramenta apresenta muitas opções de manipulação, o que auxilia o usuário a evidenciar as características que precisa analisar.

Considera-se que os recursos presentes no framework Visionary são suficientes para a exploração do grafo de proveniência em nível satisfatório. E as análises do grafo auxiliam o usuário na tomada de decisão.

Em Khan et al. (2016) é apresentado um sistema de visualização de grafos de proveniência que foca em auxiliar a busca de arquivos. Os autores definiram uma estrutura de grafo chamada Search Provenance Graph (SPG) que armazena as pesquisas e permite a visualização dos dados. O sistema utiliza cores para identificar o autor de cada busca e apresenta o SPG de duas formas: com um layout radial para apresentar a similaridade das buscas e um layout temporal que é utilizado para explorar o grafo. Os nós selecionados apresentam, com um conjunto de ícones estabelecidos pelos autores, mais detalhes da busca, como: falsos positivos e sua probabilidade, tamanho dos resultados, data da busca etc.

Com o foco na análise dos dados de proveniência, o Visionary consegue apresentar de forma mais amigável os dados para a compreensão e exploração.

A ferramenta AVOCADO é apresentada em Stitz et al. (2016) como um visualizador de proveniência de workflow de pesquisas biomédicas. O sistema utiliza duas estratégias de agrupamento para reduzir o tamanho do grafo: agrupamento hierárquico e agrupamento temático (*motif-based aggregation*). Uma heurística foi criada pelos autores para calcular o nível de interesse de cada nó, levando em conta seus atributos e o comportamento do usuário. O nível de interesse é utilizado para gerar os agrupamentos. Vários recursos visuais são utilizados para auxiliar na geração da visualização como os símbolos de cada tipo de nó, a claridade dos nós é utilizada para codificar o tempo, e filtros são aplicados as legendas de acordo com o zoom atual etc.

O Visionary também possui muitos recursos de visualização e não é focado na utilização de proveniência derivada de workflows como o AVOCADO.

Após a publicação dos modelos de proveniência OPM (MOREAU et al., 2011) em 2007 e PROV (GROTH; MOREAU, 2013) em 2013 era esperada a utilização desses modelos na maior parte dos trabalhos encontrados. Apesar disso, muitos dos trabalhos não apresentaram estes modelos de proveniência ou trabalham com modelos próprios. Apenas Chen et al. (2012) utiliza o modelo OPM e as publicações de Karsai (2016), Hoekstra and Groth (2014), Kohwalter et al. (2016) e Oliveira et al. (2017) trabalham com o modelo PROV. Assim, respondendo a QR 3 (quais modelos de proveniência são utilizados?), o PROV é o modelo mais utilizado com 4 das 14 publicações. A Tabela

3.6.3 ANÁLISE DOS RESULTADOS

A Tabela 3.10 apresenta um resumo dos recursos de visualização encontrado nas propostas descritas nessa revisão. Nessa tabela os requisitos funcionais (RF) apresentados no capítulo 2 estão numerados, sendo: RF1 múltiplas visões; RF2 abstrações; RF3 busca; RF4 filtros; RF5 proximidade do código; RF6 layout automático; RF7 histórico; RF8 outros. Quando o requisito foi encontrado a célula está marcada com um 'S' verde, destacando o resultado positivo. Quando não foi encontrado, a resposta é um 'N' vermelho. A cor amarela é utilizada para requisitos atendidos parcialmente, marcados com 'P' e quando não é claro a presença do requisito, representado com um '?'.

Tabela 3.10: Publicações encontradas na revisão sistemática e requisitos de visualização utilizados

Proposta	**RF1**	**RF2**	**RF3**	**RF4**	**RF5**	**RF6**	**RF7**	**RF8**
Provenance Browser	S	N	N	S	N	S	S	P
InProv	S	N	S	N	P	S	S	P
CZSaw	S	N	S	N	S	S	?	N
(CHEN et al., 2014)	?	N	S	?	?	S	?	S
Probe-It!	N	N	?	N	S	S	?	N
PROV-O-Viz	N	P	N	N	P	S	?	P
Provenance Explorer	N	N	P	N	P	S	?	N
(KARSAI, 2016)	N	N	S	?	P	S	S	P
Prov Viewer	S	S	?	?	?	S	?	S
(KHAN et al., 2016)	S	P	S	P	N	S	?	S
AVOCADO	P	S	S	S	?	S	S	S
Framework Visionary	S	P	S	S	S	S	N	S

Como é possível conferir pela tablela 3.10, não foi encontrado um sistema de visualização que atende à todos os requisitos funcionais. Mesmo alguns requisitos foram atendidos em outro nível da aplicação. Alguns filtros, por exemplo, são realizados na execução de queries que antecipam a visualização, já no Visionary o usuário não precisa sair da visualização para filtrar os elementos.

Outros elementos, além dos requisitos funcionais, devem ser considerados, como a visão temporal gerada pelo PROV-O-Viz e AVOCADO, que é específica para análises do fluxo da informação, o que restringe as análises e contextos de aplicações. A utilização do modelo PROV também é fundamental para a flexibilidade do sistema de visualização, o que ocorreu em poucos trabalhos. Como o modelo já estabeleceu símbolos para a codificação dos elementos, é desnecessário a codificação em outros padrões visuais. No

caso do AVOCADO, e Prov Viewer, por exemplo, o usuário deve conhecer o modelo PROV, os símbolos utilizados em cada e fazer a ligação entre eles.

Por essas razões, o framework Visionary foi desenvolvido, para facilitar a adoção de usuários que já utilizam ou que querem começar a utilizar a proveniência através do PROV. O framework contempla todos os requisitos funcionais de visualização, exceto os recursos de modificação e manutenção de visualizações anteriores. Essas modificações não são o objetivo do Visionary que foi construído para auxiliar na análise dos dados e indicar modificações nos dados da aplicação.

3.7 ANÁLISE DE DADOS DE PROVENIÊNCIA

As 24 publicações encontradas na revisão e mapeamento sistemático sobre análise de dados de proveniência foram utilizadas para responder as questões de mapeamento, apresentados na Subseção 3.7.1. Para responder as questões de revisão foram utilizados os artigos restantes após a avaliação de qualidade, 9 no total. O resultado da revisão sobre visualização está apresentado na Subseção 3.7.2.

3.7.1 RESULTADOS DO MAPEAMENTO SISTEMÁTICO

A Figura 3.4 apresenta a distribuição das publicações entre 2007 e 2017 para responder a QM 1 (como as publicações são distribuídas ao longo dos anos?). Em destaque o ano de 2012 com a maior quantidade de publicações, quatro no total. A linha tracejada vermelha mostra a tendência do gráfico.

Não é possível verificar uma tendência clara de crescimento ou redução do número de publicações, podemos considerar a área estável.

Todos os pesquisadores foram identificados e relacionados para responder a QM 2 (quem são os autores principais da área?). Apenas Beth Plale possui três publicações, os outros autores com mais de uma publicação são: Archana Nottamkandath, Davide Ceolin, Paul Groth, Wan Fokkink e Willem R. Van Hage.

As publicações encontradas também foram distribuídas pelos seus veículos de publicação o que destacou dois veículos com mais de uma publicação: IPAW (*International Provenance and Annotation Workshop*) com quatro publicações; e o *IEEE International Conference on E-Science* com duas publicações.

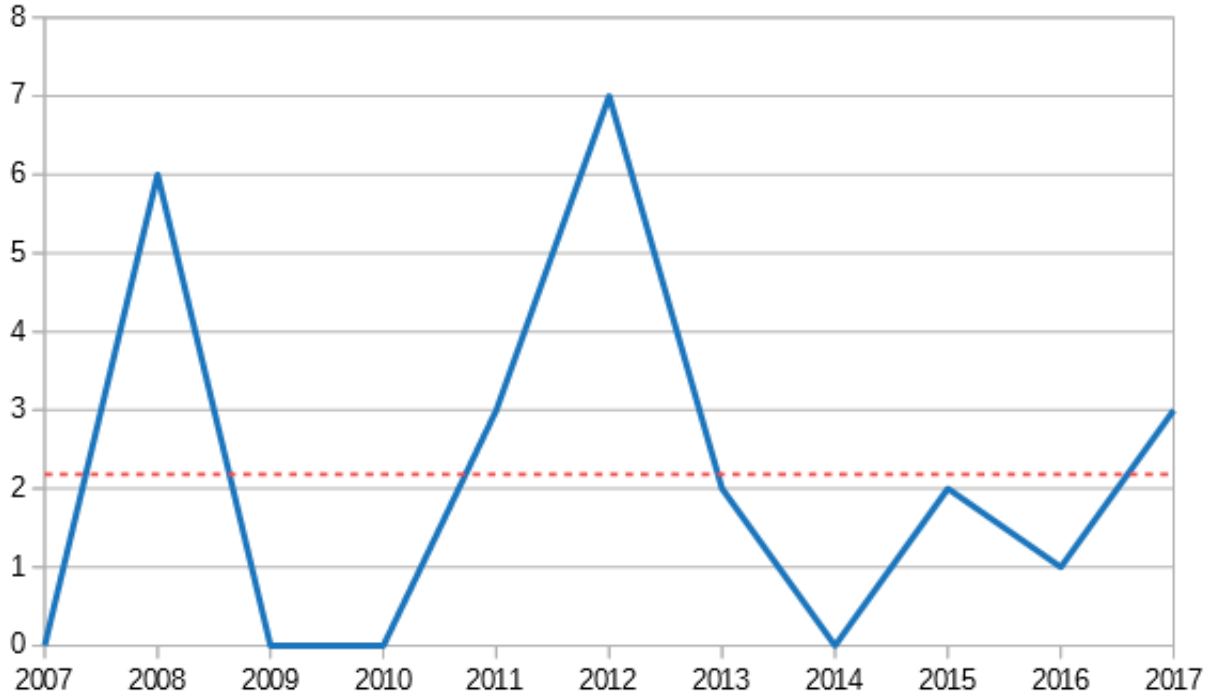


Figura 3.4: Distribuição pelos anos das publicações aceitas na revisão de visualização de dados de proveniência.

Finalizando o mapeamento sistemático, as respostas das QMs foram resumidas na Tabela 3.11.

Tabela 3.11: Resumo das respostas das questões de mapeamento sistemático da revisão de nálise de dados de proveniência

Pergunta	**Respostas**
QM 1	As publicações na área estão estáveis.
QM 2	Beth Plale se destaca com três publicações.
QM 3	IPAW se destaca com quatro publicações.

3.7.2 RESULTADOS DA REVISÃO SISTEMÁTICA

Com a intenção de responder a QR 1 (quais são as propostas encontradas?) e a QR 2 (quais técnicas são utilizadas na área?), as publicações aceitas tiveram sua proposta resumida nos parágrafos seguintes. Juntamente com a descrição de cada publicação, foi realizada uma comparação com a proposta apresentada nessa dissertação.

Em Ceolin et al. (2012) os autores apresentam uma combinação de reputação e proveniência para determinar valores de confiabilidade. Os valores de confiabilidade de alguns

artefatos são utilizados para calcular a reputação dos usuários do jogo Waisda?[3]. Esses valores são combinados com análises realizadas sobre as relações do modelo PROV, através de técnicas de aprendizagem de máquina para predizer o nível de confiabilidade de outros artefatos. Os autores apresentam resultados estatisticamente relevantes e positivos na combinação das técnicas. O trabalho se desenvolve em um pipeline completo para avaliar a confiabilidade de artefatos em Ceolin et al. (2016). Nessa sequência do trabalho, os autores descrevem o pipeline que realiza a classificação a partir da proveniência e um conjunto de exemplos.

As análises realizadas pelo framework Visionary não consideram dados específicos de domínio, o que permite a aplicação em qualquer sistema que utiliza o PROV. Além disso, as análises abrangem mais aspectos do que a confiabilidade dos dados como a importância dos elementos no grafo, o que oferece mais informações para a tomada de decisão.

A proposta de Cheah and Plale (2012) apresenta uma metodologia para avaliar a qualidade dos grafos de proveniência gerados com o modelo OPM. Primeiramente, o modelo verifica a correção dos dados através de análise contextual: as anotações são verificadas para encontrar a duplicidade de artefatos e os timestamps também são analisados para encontrar inconsistências na sequência dos eventos. Após isso, a completude do grafo é verificada através da análise estrutural. A entrada e a saída de cada ligação são verificadas e a ocorrência temporal é analisada para encontrar inconsistências. O número dos nós também é analisado junto a outros grafos para verificar a ocorrência de outliers.

Essa proposta é complementar ao nosso trabalho já que atua na qualidade da proveniência em si, e o framework Visionary atua com a qualidade dos artefatos representados pela proveniência.

A proposta de McGrath and Futrelle (2008) é combinar regras escritas em SWRL com uma representação em OWL da proveniência modelada com OPM. Os autores afirmam que a combinação de ambas as linguagens permite melhorar as inferências encontradas com a análise dos dados. Trabalho semelhante foi detalhado em Missier and Belhajjame (2012), que além de apresentar uma codificação do grafo do PROV, apresenta regras e restrições que podem ser validadas por uma máquina DLV (inteligência artificial baseada em disjunção lógica).

Muitas formas de análise podem se juntar aos grafos de proveniência em nível onto-

[3]www.waisda.nl

lógico ou em grafo. O framework Visionary apresenta essas duas abordagens, explorando as inferências da ontologia (OWL2) e das análises de grafo.

Em Prat and Madnick (2008) é apresentado um modelo de proveniência e uma abordagem para calcular a credibilidade baseada em metadados de proveniência. Os autores calculam a credibilidade da origem de dados, do resultado dos processos e uma credibilidade geral. A credibilidade da origem dos dados é calculada com métricas simples de distância e similaridade que são utilizadas para gerar as outras métricas e resultados.

A abordagem Visionary é baseada em um padrão de proveniência, o que encoraja a reutilização e permite a integração com outras técnicas. A análise de grafo também é baseada em métricas de redes complexas e pode ser explorada fora de um contexto específico.

O workflow apresentado em Strubulis et al. (2012) possui regras de inferência baseadas em proveniência para reduzir a quantidade de informações de proveniência e auxiliar o controle de qualidade. Os autores buscam reduzir o espaço necessário para o armazenamento das informações de proveniência com regras de inferências intuitivas que podem ser calculadas dinamicamente. São apresentadas três regras de inferência mas os autores afirmam que a proposta pode ser estendida para outras regras e outros modelos.

Ao poupar espaço de armazenamento essa proposta onera o processamento dos dados, o que concorre com as análises e inferências sobre os dados. Além disso, o foco do Visionary é gerar conhecimento novo a partir dos dados de proveniência.

Para concluir com a resposta da QR 3 (quais modelos de proveniência são utilizados?), destaca-se o PROV como o modelo mais utilizado, 4 das 9 publicações, sendo o OPM o segundo mais utilizado com 2 das 9 publicações.

3.7.3 ANÁLISE DOS RESULTADOS

Das abordagens encontradas algumas análises são específicas para determinados domínios e outras para alguns modelos. Algumas análises tratam da qualidade dos meta dados de proveniência mas não fornecem ao usuário informações capazes de auxiliar na melhoria da aplicação que gera os meta dados.

Devido às limitações apresentadas nas análises listadas, este trabalho propõe a abordagem Visionary que cumpre os requisitos de análise listados abaixo:

- Atuar sobre o modelo PROV que é um modelo popular e extensível.

- Fornecer resultados sem a dependência de dados para treinamento.
- Operar em diversos contextos.
- Estabelecer informações estratégicas para a melhoria dos processos analisados.

3.8 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Esse capítulo apresentou os resultados de duas revisões sistemáticas realizadas com o objetivo de fazer um levantamento das principais abordagens existentes na literatura relacionadas à proposta dessa dissertação. A primeira revisão abrangeu o tema de visualização de dados de proveniência e a segunda foi relacionada a análise de dados de proveniência. Foram descritas as questões de pesquisa que guiaram as revisões e mapeamentos sistemáticos e os resultados foram sumarizados nesse capítulo. Cada questão foi respondida junto à sua revisão específica.

4 FRAMEWORK VISIONARY

A fim de estimular o uso da proveniência e auxiliar na sua compreensão e análise, este trabalho propõe um framework, denominado Visionary. O Visionary foi projetado para ser flexível e ser capaz de se adaptar aos diferentes contextos. Os principais objetivos do Visionary são: i) simplificar a compreensão da proveniência dos dados; ii) melhorar a compreensão dos dados de proveniência e iii) dar suporte a tomada de decisão através da análise dos dados de proveniência.

O framework Visionary captura os dados de proveniência de execução de processos digitais, manipula e analisa os dados para apresentar visualmente informações para o usuário. Através das visualizações, o usuário pode explorar os dados para compreender e analisar os processos que originaram os dados históricos. Assim, usuário ganha mais conhecimento e suporte para ampliar, corrigir e modificar os processos originais. A Figura 4.1 ilustra esse processo de aquisição de conhecimento através dos dados históricos.

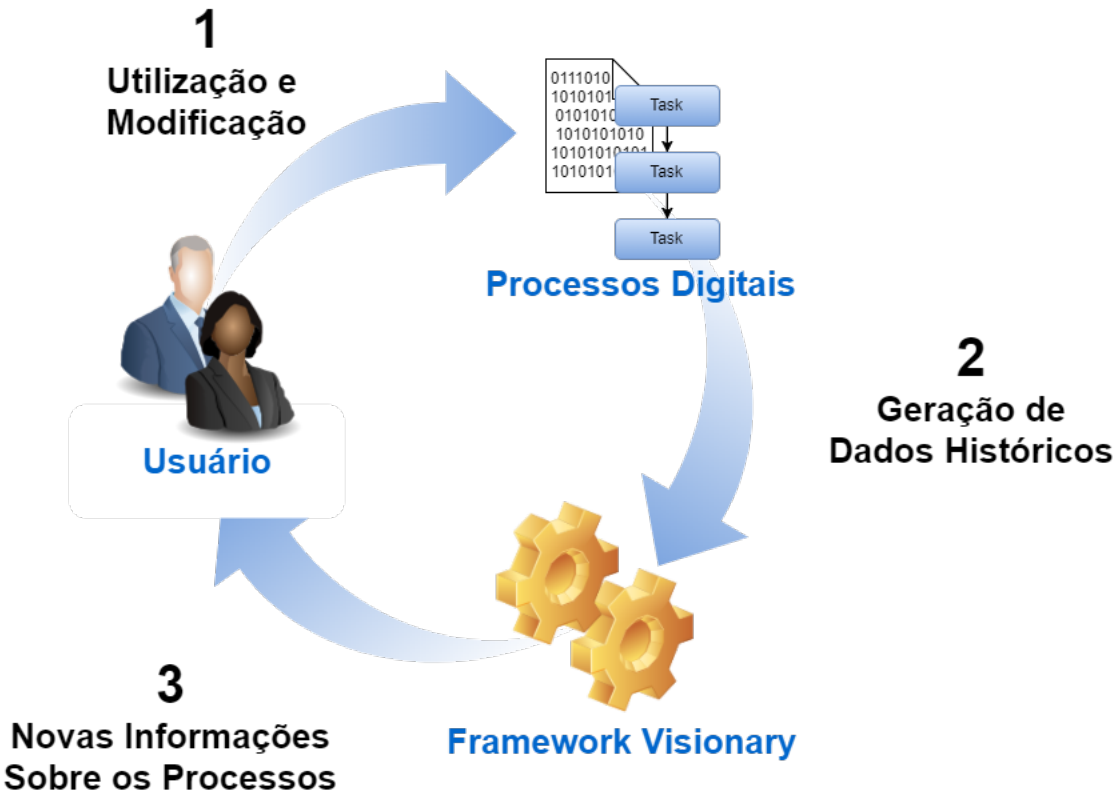


Figura 4.1: Ciclo de aprimoramento de processos digitais através do framework Visionary. O usuário utiliza os processos que fornecem dados para o framework que gera informações para os usuários melhorarem os processos.

A Figura 4.1 apresenta no item 1 a utilização de processos digitais; no item 2, os processos executados geram dados de proveniência que alimentam o framework; no item 3 da figura, o Visionary gera novas informações e apresenta para os usuários; novamente no item 1 os usuários podem modificar os processos digitais utilizando as informações geradas pelo framework.

Na Seção 4.1.1 todas as etapas do framework são apresentadas. Essas etapas são descritas com mais detalhes nas seções seguintes: a Seção 4.1.1 apresenta a captura dos dados de proveniência; a Seção 4.1.2 e Seção 4.1.3 apresenta as análises dos dados através de ontologias e técnicas de redes complexas para gerar novas informações; na Seção 4.1.4 é apresentado os mecanismos de visualização utilizados para favorecer a compreensão dos dados.

4.1 ETAPAS DO FRAMEWORK

O framework possui cinco etapas distintas. Ele foi desenvolvido para ser aplicado a partir de qualquer implementação do modelo PROV. As cinco etapas estão ilustradas na Figura 4.2, são elas: (1) Captura, (2) Inferência, (3) Transformação, (4) Análise e (5) Visualização.

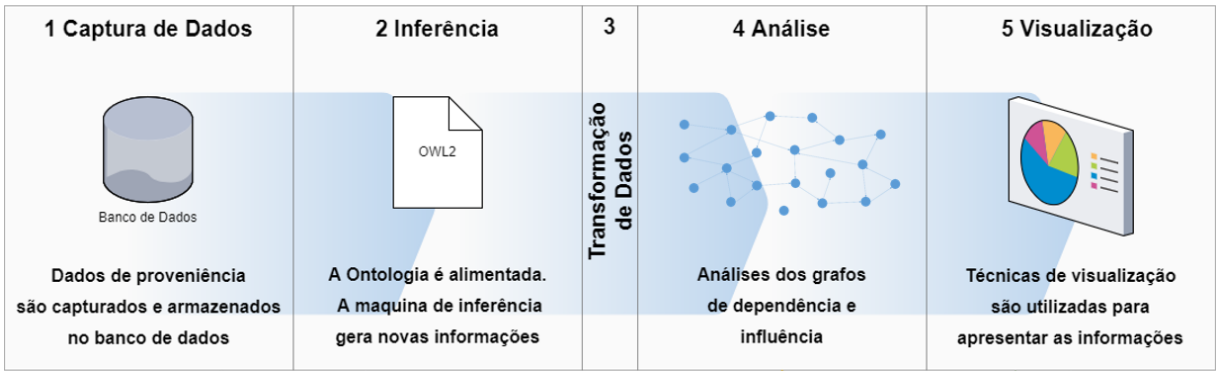


Figura 4.2: Representação das etapas do framework Visionary e suas atividades: (1) Captura, (2) Inferência, (3) Transformação, (4) Análise e (5) Visualização.

Na etapa 1 são capturadas informações importantes referente aos processos digitais modelados. Essas informações compõe os dados de proveniência e são armazenadas em um banco relacional. Na etapa 2 os dados do banco relacional alimentam a ontologia PROV-O (um arquivo OWL2) que possui regras e restrições para gerar novas informações para o usuário. A etapa 3 faz uma transformação dos dados da ontologia para o formato

de grafo, codificado em JSON. A etapa 4 Analisa os dados em forma de grafo e gera conhecimento novo para o usuário, preparando também vários recursos de visualização da próxima etapa. Na etapa 5, os dados são codificados visualmente e uma série de recursos são fornecidos para o usuário explorar e compreender os dados e as novas informações.

4.1.1 CAPTURA E ARMAZENAMENTO DE DADOS (ETAPA 1)

O Visionary utiliza o PROV como modelo de proveniência. A escolha deste modelo de proveniência define aspectos importantes das etapas posteriores. O PROV é um modelo genérico e permite especificações para diferentes contextos, como já apresentado. O Visionary, por essa razão, explora os recursos centrais do PROV, e pode ser adaptado para contextos mais específicos, como por exemplo, processos de software e experimentação científica. Mudanças e adaptações no PROV são refletidas nas outras etapas do framework com poucas ou nenhuma modificação necessária.

Na captura dos dados, o contexto analisado (processos de gestão, processos de software, workflow científico, etc.) deve ser previamente adaptado para capturar informações de proveniência. A captura pode ser realizada na execução das tarefas onde os próprios processos gravam informações de proveniência no banco de dados. Por exemplo, um usuário ao iniciar uma nova tarefa teria seu identificador registrado (como agente no modelo PROV) em um banco relacional, o identificador da tarefa também seria registrado (como uma atividade) e uma terceira informação seria registrada, a que associa os dois registros (essa relação é *WasAssociatedWith* segundo a modelagem do PROV). Se essa mesma tarefa utiliza um documento digital (um formulário por exemplo), o identificador desse documento também é registrado (entidade) e mais uma ligação é registrada entre a tarefa e o documento (*used*). Dessa forma o registro da proveniência é realizado semelhante ao registro de um log do sistema, mas, capturando dados específicas para fornecer informações sobre a origem e qualidade dos objetos de dados.

Dessa forma, a aplicação a ser utilizada para a execução do processo deve ser 'instrumentalizada', ou seja, componentes para captura de dados de proveniência específicos devem ser desenvolvidos e acoplados a aplicação de forma a capturar e armazenar os dados de proveniência em um repositório de proveniência específico.

Em (DALPRA, 2016; SIRQUEIRA et al., 2016) foram desenvolvidas aplicações específicas para esta captura. Além destes, pode-se citar a captura dos dados de proveniência em

sistemas de gerenciamento de workflows, principalmente nos SGWfC como (OINN et al., 2007; LUDÄSCHER et al., 2006; CALLAHAN et al., 2006). Esses sistemas são capazes de registrar eventos durante a execução do workflow (BOWERS et al., 2012; TOWNEND et al., 2013) e as dependências podem ser geradas diretamente com as etapas de execução do workflow ou computadas través de padrões pré-estabelecidos (BOWERS et al., 2012). No entanto, a utilização do modelo PROV não é nativa destes sistemas.

Outra forma de capturar os dados de proveniência é através dos registros de execução dos sistemas ou tarefas. Dessa forma os dados são capturados em fontes secundárias que estão relacionadas com o sistema. Registros de execução como logs ou outros documentos de gestão de tarefas podem ser utilizados para gerar os dados de proveniência e enfim alimentar a base relacional. Um exemplo dessa captura é mostrada em Dalpra (2016) que exporta os dados da execução de processos de desenvolvimento de software para um arquivo padronizado XML, esse arquivo permite a importação dos dados para o banco de dados.

Essa armazenagem dos dados no banco de dados relacional padronizado pelo PROV-DM facilita a interoperabilidade com outros sistemas. Desta forma, o Visionary pode ser adotado por uma aplicação que captura a proveniência segundo o modelo PROV. No entanto, algumas adaptações podem ser necessárias.

As extensões do PROV que especificam o domínio das aplicações são variadas e, portanto, variam de acordo com o tipo de aplicação utilizada. O autor da extensão deve fazer a adequação das etapas seguintes sempre que for necessário. Dessa forma o framework fornece suporte para os dados definidos no PROV-DM e é aberto para receber outras extensões específicas da aplicação. Se o modelo de proveniência for modificado para um domínio específico, salvo modificações estruturais dos dados de proveniência, o Visionary vai continuar obtendo resultados da análise, mesmo que de forma mais genérica. No entanto, a adaptação do Visionary vai permitir resultados também específicos para o domínio utilizado.

4.1.2 ONTOLOGIA E INFERÊNCIA (ETAPA 2)

Nesta fase, os dados são analisados para auxiliar o usuário na descoberta de novas informações sobre a aplicação. Os dados de proveniência são processados em duas etapas de análises: através da ontologia PROV-O, detalhada nesta seção, e através de algoritmos

de análise de redes complexas, que será detalhada na seção 4.1.3.

Assim, a primeira análise dos dados realizada pelo framework é através do uso de ontologia e regras de inferência. Os dados modelados com PROV-DM podem ser carregados na ontologia PROV-O que fornece um conjunto de classes, propriedades e restrições para auxiliar na análise de dados de proveniência. A PROV-O segue as características do modelo PROV, sendo livre de domínio, mas permitindo extensões para domínios específicos. A definição e utilização da PROV-O é detalhada por Lebo et al. (2013). Os dados armazenados no banco de dados modelado pelo PROV-DM são carregados na ontologia e a máquina de inferência é executada, e as novas informações são salvas.

Como afirmam Moreau and Missier (2013), todas as relações contidas no PROV são relações que geram algum nível de influência. A influência entre os elementos é fundamental durante a análise de grafos de proveniência, por isso é importante mapear a cadeia de influência gerada a partir das relações entre os elementos definidos no PROV. Além das relações básicas do PROV-O, o framework utiliza a relação *influenced* que define a influência das relações. Essa relação rastreia a influência durante a fase de análise. O domínio e o contradomínio dessa relação são qualquer entidade, atividade ou agente. A relação *influenced* possui a relação inversa chamada *wasInfluencedBy*. A relação, ou propriedade, *wasInfluencedBy* é uma relação abrangente e muitas vezes substituída por uma de suas sub propriedades, logo, todas as sub propriedades de *wasInfluencedBy* devem ser relacionadas como relações inversas a *influenced*. Da mesma forma, *influenced* também deve ser relacionada à todas as relações básicas, como uma relação inversa. Essas associações na ontologia são ilustradas na Tabela 4.1.

Tabela 4.1: Geração da relação *influenced* a partir das relações básicas do PROV com especificação de domínio e contradomínio da relação.

Relações básicas	**Relação de influência gerada**
Entidade → *WasGeneratedBy* → Atividade	Atividade→ *Influenced* → Entidade
Atividade → *Used* → Entidade	Entidade → *Influenced* → Atividade
Atividade → *WasInformedBy* → Atividade	Atividade → *Influenced* → Atividade
Entidade → *WasDerivedFrom* → Entidade	Entidade → *Influenced* → Entidade
Entidade → *WasAttributedTo* → Agente	Agente → *Influenced* → Entidade
Atividade → *WasAssociatedWith* → Agente	Agente → *Influenced* → Atividade
Agente → *ActedOnBehalfOf* → Agente	Agente → *Influenced* → Agente

Além da importância da relação *influenced* na fase de análise, ela permite a qualquer utilizador do PROV, que realizou adaptações para um domínio específico, utilizar essa

relação sempre que necessário na criação de novas relações de influência. Por exemplo, caso a relação de gerenciamento (*wasManagedBy*) seja criada para mapear a gestão de um grupo ou organização por determinada pessoa, deve-se associar a relação *influenced* como inversa à *wasManagedBy*. Assim as análises sobre o grafo de proveniência poderá rastrear a relação *wasManagedBy* e conseguir resultados mais precisos no contexto utilizado.

Um exemplo de uma ontologia adaptada a um contexto específico é a ontologia PROV-Process (DALPRA, 2016). Essa ontologia é uma extenção da PROV-O, adaptada ao domínio de processos de software e que utiliza o framework Visionary para análises sobre o grafo de proveniência. Na PROV-Process são definidas três property chains sobre as relações básicas que podem ser utilizadas dentro do Visionary. As três *property chains* geram a relação *wasAssociatedWith*, definidas como:

- *used* **o** *wasAttributedTo* **SubPropertyOf:** *wasAssociatedWith*
- *wasStartedBy* **o** *wasAttributedTo* **SubPropertyOf:** *wasAssociatedWith*
- *wasEndedBy* **o** *wasAttributedTo* **SubPropertyOf:** *wasAssociatedWith*

A relação *wasAssociatedWith* é uma das sub propriedades da relação *wasInfluencedBy* sendo portanto uma relação oposta a relação influenced como apresenta a Tabela 1. Portanto a PROV-Process pode ser utilizada junto do framework Visionary sem modificações.

O mesmo ocorre com a ontologia apresentada em Sirqueira et al. (2016), que é uma versão da PROV-O estendida e adaptada para o contexto de manutenção de experimentos e workflows científicos. Na adaptação, além da criação de novas classes, duas relações foram criadas (*evolutionOf* e *evolutionTo*) e são inferidas a partir de outras relações já existentes (*wasDerivedFrom*, *specializationOf* e *alternateOf*). A relação *wasDerivedFrom* é uma sub propriedade de *wasInfluencedBy*, logo inversa à *influenced*. A relação *specializationOf* é definida como sub propriedade de *alternateOf*, e para definir a relação inversas à *influenced* basta relacionar *alternateOf* como inversa. Assim as duas relações inferidas estariam ligadas as duas relações de influência. O framework Visionary pode ser utilizado neste contexto com o objetivo de ampliar as análises realizadas para a manutenção e evolução de workflows científicos e experimentos associados. No entanto, neste contexto, algumas adaptações são necessárias.

A adaptação das classes não altera o funcionamento do Visionary, desde que sejam mantidas as três classes básicas definidas no modelo PROV (entidade, atividade e agente),

sobre as quais o framework trabalha.

Todas as inferências geradas com a ontologia, são utilizadas na etapa 5, sendo destacadas na visualização. Esta etapa além de gerar novo conhecimento, prepara também os dados para a próxima etapa de análise.

4.1.3 TRANSFORMAÇÃO E ANÁLISE (ETAPA 3 E 4)

Após a segunda etapa, os dados da ontologia são carregados e armazenados em formato de grafo para prosseguir com a segunda análise do framework. A leitura diretamente do arquivo OWL (ontologia) permite uma transformação simples para o modelo de grafo e prepara os dados para análises posteriores. Essa transformação dos dados constitui a etapa 3 do framework.

Com os dados de proveniência em formato de grafo é possível extrair características geradas pelo modelo PROV, para descrever sua estrutura. Esta descrição pode ser utilizada para identificar semelhanças entre nós e a faixa de influência de cada nó. Ebden et al. (2012) apresenta várias características dos grafos de proveniência com o objetivo de analisar a evolução dos grafos com o tempo de uso. Algumas dessas características são métricas comuns usadas para análise de rede, como diâmetro ou número de nós, outras são métricas adaptadas aos grafos de proveniência e suas particularidades.

Quatro métricas diferentes foram selecionadas por Huynh et al. (2013) ao analisar subgrafos de proveniência. As métricas foram utilizadas com sucesso pelos autores inferindo a qualidade de cada nó analisado. As métricas são (i) número de nós, (ii) número de arestas, (iii) diâmetro e (iv) distância finita máxima (MFD) (EBDEN et al., 2012). Huynh et al. (2013) utiliza as métricas para caracterizar o subgrafo de influência (definido a seguir) de cada nó e inferir a qualidade de outros nós a partir de informações de qualidade pré-determinadas. Influenciados por este trabalho, o framework Visionary conta com um algoritmo que utiliza as quatro métricas citadas, mas não precisa de informações prévias para determinar a qualidade dos nós. Este algoritmo determina a similaridade regular entre os nós através da análise do subgrafo que influencia o nó. Este recurso pode ser ou não utilizado e, portanto, é ativado a partir da seleção do usuário.

A seguir, é apresentado as métricas utilizadas e algumas definições importantes para o processamento das análises.

Métricas

Considerando o grafo de proveniência como um grafo direcionado $G = (V, E)$, onde V é o conjunto de nós e E o conjunto de arestas, o número de nós é representado por $|V|$ e o número de arestas por $|E|$. O diâmetro do grafo é a maior distância encontrada dentro do grafo, onde a distância é o caminho mais curto entre dois nós. Em outras palavras, depois de definir o caminho mais curto entre todos os pares de nós, o maior desses valores é considerado o diâmetro do grafo. Como os grafos de proveniência são direcionados, é possível encontrar distâncias infinitas, porque pode não existir um caminho entre os pares de nós. Portanto, ao determinar as distâncias, o grafo é considerado temporariamente não-direcionado.

O MFD é uma métrica específica para grafos de proveniência e é definida como a maior distância finita de um tipo de nó para outro dentro de um grafo direcionado G. Considerando os três tipos de nó presentes em PROV, são 6 MFDs a serem calculados (agente e agente; entidade e entidade; atividade e atividade; agente e entidade; entidade e atividade; atividade e agente), atingindo um total de 9 valores.

Subgrafos

Todas as métricas são usadas para caracterizar partes do grafo de proveniência. O grafo é dividido em vários subgrafos relacionados a cada nó, e através das suas características, é possível relacionar os subgrafos gerados e, consequentemente, os nós relacionados. Esta é uma fórmula livre de contexto que permite uma análise dos grafos e de cada nó sem a necessidade de ajustes adequados às especificidades de cada aplicação. Mesmo com a adaptação da proveniência para um contexto, a análise dos grafos continua funcionando e tem resultados mais específicos, de acordo com a adaptação realizada.

Cada aresta de um grafo do PROV representa formas de influência entre o nó de origem e o nó de destino (MOREAU; MISSIER, 2013). Em outras palavras, se houver um caminho entre v_0. e v_i, notado como $v_0 \rightarrow *v_i$, pode-se considerar que v_i foi potencialmente influenciado por v_0. Desta forma, pode-se construir um subgrafo de $D_{G,a}$ de dependência, no qual contém apenas vértices que influenciam o nó a e suas respectivas arestas, como descreve Huynh et al. (2013):

$$D_{G,a} = (V_{G,a}, E_{G,a}) \tag{4.1}$$

$$V_{G,a} = \{v \in V : v \to *a\} \tag{4.2}$$

$$E_{G,a} = \{e \in E : (\exists\, v_s, v_t \in V_{G,a})(e = (v_s, v_t))\} \tag{4.3}$$

Cada métrica possui igual peso para determinar a similaridade entre os nós. As métricas são normalizadas entre 0 e 1 de acordo com a Equação (4.4), onde M_m é o maior valor encontrado entre as métricas analisadas, M_0 é o menor valor e M o valor obtido no nó analisado. VP é o resultado do cálculo.

$$VP = \frac{M - M_0}{M_m - M_0} \tag{4.4}$$

O usuário pode então, identificar um elemento ou processo que merece destaque, positivo ou negativo, e assim relacionar o nó identificado com a mesma estrutura de dependência através da análise do grafo $D_{G,a}$. Como exemplo, para entender a importância de tal análise, pode-se considerar que ao identificar um elemento ou processo na aplicação sujeito a um erro, esta análise permite encontrar nós que representam os mesmos tipos de elementos ou processos que podem ter o mesmo erro devido à similaridade do gráfico de dependência, já que a proveniência trata de qualidade e confiabilidade dos dados (MOREAU; MISSIER, 2013). O mesmo pode acontecer com elementos com a alta qualidade confirmada pelo usuário, uma vez que o framework encontra elementos semelhantes, e o usuário pode aumentar a importância desses elementos na aplicação, consequentemente, aumentando a importância dos respectivos nós no grafo de proveniência.

Com o mesmo princípio do subgrafo de dependência $D_{G,a}$, pode-se determinar um subgrafo de influência $I_{G,a}$ que contém apenas os nós que foram potencialmente influenciados por um determinado nó a. Essa análise vai além das propostas de Huynh et al. (2013); Ebden et al. (2012) e apresenta uma nova forma de determinar a importância de cada nó dentro do grafo de proveniência. A importância do nó a é determinada pela capacidade do nó influenciar os outros nós do grafo e as métricas são utilizadas para calcular o grau da influência. Foi definido $I_{G,a}$ da seguinte forma:

$$I_{G,a} = (V_{G,a}, E_{G,a}) \tag{4.5}$$

$$V_{G,a} = \{v \in V : a \rightarrow *v\} \tag{4.6}$$

$$E_{G,a} = \{e \in E : (\exists\, v_s, v_t \in V_{G,a})(e = (v_s, v_t))\} \tag{4.7}$$

Para auxiliar os usuários na compreensão da aplicação, a análise do subgrafo é destacada na visualização. Com o grafo $I_{G,a}$ representando a rede de influência do nó a, o usuário pode identificar na rede o impacto de excluir ou modificar o elemento ou processo representado por esse nó. Uma opção na visualização permite destacar os nós (alterando seu tamanho) de acordo com a influência que ele tem na rede. Isso permite ao usuário identificar rapidamente os nós mais influentes na rede e tomar decisões, se necessário.

O cálculo da importância do nó a no grafo de proveniência G, notado como $Imp_{G,a}$, é baseado nas métricas retiradas do subgrafo de influência do nó como demonstram as Equações 4.8 e 4.9.

$$E_M = \frac{|V_{G,a}| * (|V_{G,a}| - 1)}{2} \tag{4.8}$$

$$Imp_a = \frac{|V_{G,a}|}{|V|} * \frac{\frac{|E_{G,a}|}{E_M} + \frac{1}{D} + \frac{1}{MDF_1} + ... + \frac{1}{MDF_6}}{8} \tag{4.9}$$

O cálculo considera o número de nós do subgrafo $|V_{G,a}|$ em relação ao número de nós total do grafo de proveniência $|V|$. Esse é o valor base da influência que é modificado pelas outras métricas: o número de arestas em relação ao total de arestas do subgrafo completo (E_M); o diâmetro (D); e o inverso das seis métricas MFD. O valor final de $Imp_{G,a}$ será 0 para influência nula e 1 para influência máxima. A montagem da Equação 4.9 considera os pontos abaixo:

- O número de nós é o valor mais importante a ser considerado, por isso foi colocado em um termo separado e considerado em relação ao número total de nós do grafo.
- Quanto menor a distância entre dois nós, maior é a influência exercida entre eles.
- Quanto maior o número de arestas em um grafo, maior o número de caminhos

que podem ser traçados do nó analisado para os outros nós do grafo. Logo, o nó analisado influencia os outros nós de formas diferentes.

- Quanto menor o diâmetro do grafo, menor as distâncias entre os nós do grafo, portanto, maior a influência entre eles. O mesmo ocorre com o MFD que considera as distâncias entre os tipos de nó. Por isso foi considerado o inverso desses valores.

Assim sendo, para calcular a importância do nó no grafo de proveniência são utilizados o inverso dos valores de diâmetro e MFDs. O resultado do cálculo afeta o tamanho do nó na visualização do grafo de proveniência.

A aplicação destas análises varia de acordo com o domínio específico. Assim, enquanto em uma aplicação o usuário pode procurar aumentar a influência de um nó no grafo para melhores resultados, em outra aplicação pode procurar diminuir a influência dos nós sobre o grafo, deixando o grafo com menos conexões para também obter melhores resultados. Desta forma, o framework e seus recursos podem melhorar a compreensão da aplicação, o que possibilita uma melhor reutilização, cooperação entre parceiros e confiabilidade. No entanto, não substitui o papel do usuário que deve conduzir a análise e tomar a decisão final. Para destacar essas funcionalidades, são apresentados três exemplos de uso das análises.

Exemplo 1: em um extenso workflow científico os dados parecem ter sido modificados de forma incorreta. O cientista, analisando alguns dos serviços utilizados, descobre um erro em um serviço, e tem razões para acreditar que outros serviços também não são confiáveis. Analisando os dados de proveniência, o cientista identifica o serviço mapeado como uma atividade, e através do framework Visionary, encontra atividades de qualidade semelhante. Finalmente o cientista concentra seu esforço na verificação de erros das atividades destacadas pelo framework para encontrar outros serviços de má qualidade.

Exemplo 2: Para a otimização dos trabalhos, um dos processos de uma empresa deve ser substituído. O gerente responsável verifica os dados de proveniência para identificar qual o impacto da mudança nos outros processos. Utilizando a análise do grafo de influência, o gerente descobre que muitos processos e artefatos de dados são impactados com a mudança e decide rever os planos ou ponderar a mudança com os demais gerentes.

Exemplo 3: Impossibilitado de verificar todo o sistema devido ao prazo estipulado, um analista decide utilizar o princípio de Pareto (principio 80 / 20) e verificar apenas 20% do sistema que é mais importante. Para identificar esse grupo de elementos, o analista

vai até os dados de proveniência e descobre, através da análise de influência dos nós, os elementos mais importantes envolvidos no sistema e faz sua verificação.

4.1.4 VISUALIZAÇÃO DOS DADOS (ETAPA 5)

Os dados de proveniência são apresentados em formato de grafo com vários recursos visuais para navegação, exploração e compreensão. Essa última etapa tem o objetivo de apresentar as informações geradas pela etapa de análise de forma amigável e compreensível para o usuário.

A Figura 4.3 mostra uma visualização gerada pelo framework. Na parte superior da imagem é destacada a opção de visualização, indicado pelo número 1; do lado esquerdo estão presentes os controles de símbolos e análise, indicado pelo número 2; controles de filtro, indicados pelo número 3; e busca por nós, indicado pelo número 4. Do lado direito, na área maior, estão os dados de proveniência em formato de grafo, indicado pelo número 5.

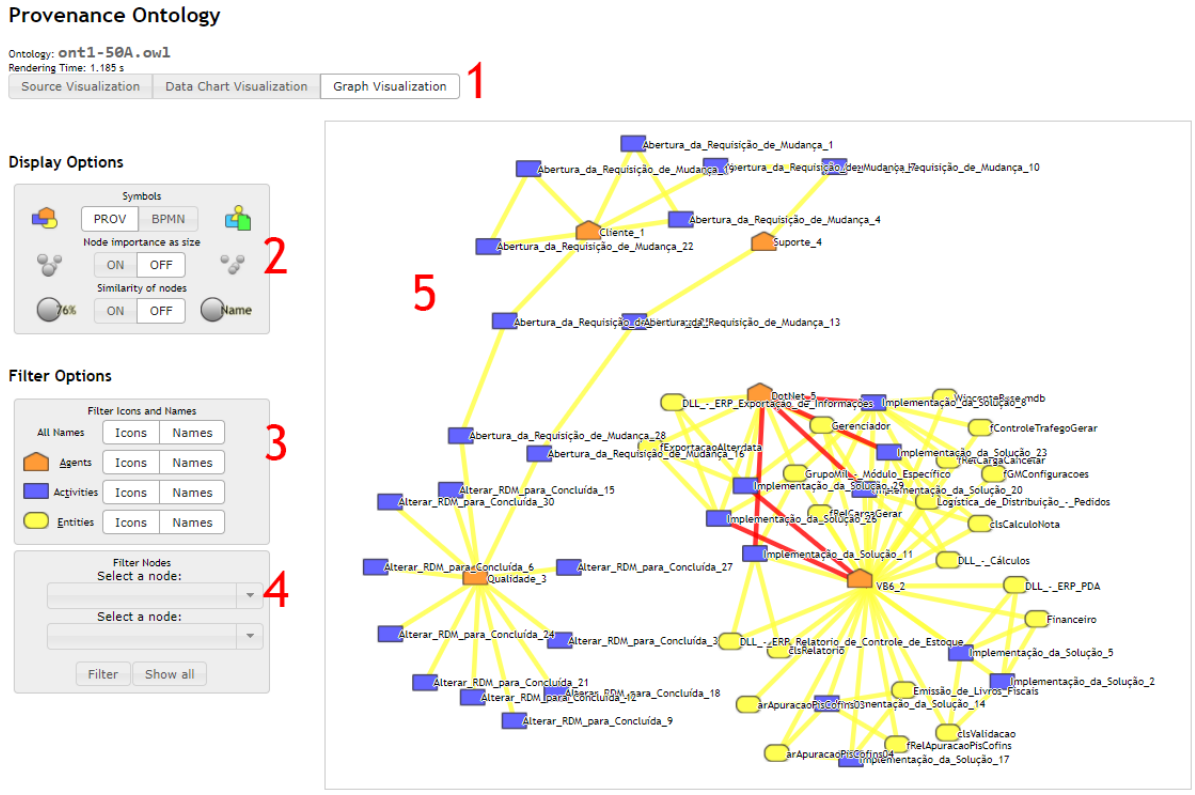
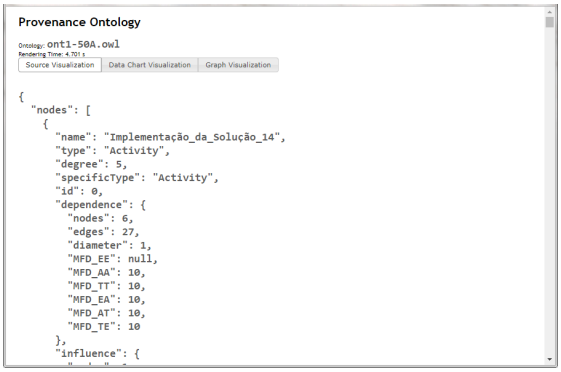


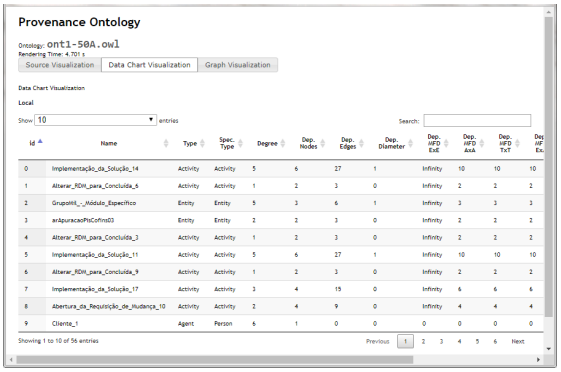
Figura 4.3: Visualização gerada pelo framework. Em destaque as áreas de opção de visualização (1), controle de símbolos e análises (2), controle de filtros (3), busca por nós (4) e a área de visualização (5).

1. Opção de visualização

Esta opção permite ao usuário alternar entre o código fonte (*Source Visualization*), métricas gerais dos dados de proveniência (*Data Chart Visualization*) e a visualização do grafo (*Graph Visualization*). A Figura 4.4 apresenta as visões de código fonte (a) e métricas (b) fornecidas pelo framework.



(a) *Source Visualization* ou visão do código fonte



(b) *Data Chart Visualization* ou visão das métricas gerais

Figura 4.4: Duas opções de visualização presentes no framework.

A opção de código fonte exibe os dados codificados em JSON como lista de nós e lista de aresta. Essa opção permite uma busca direta pelo nome dos elementos e auxilia na solução de problemas de sintaxe e codificação de caracteres. A opção de métricas exibe o número total de nós e de arestas discriminados por tipo. Dessa forma o usuário possui uma dimensão geral do grafo através de números diretos. A opção de visualização do grafo exibe os dados codificados visualmente e é onde se encontra a maior parte dos recursos de visualização.

2. Controle de símbolos

Dois conjuntos de símbolos são utilizados para codificar o tipo de cada nó. O primeiro conjunto é apresentado por Moreau and Missier (2013) e retrata diretamente os tipos do modelo PROV que são codificados como classes na ontologia PROV-O. O segundo conjunto é uma correspondência direta com os símbolos usados no BPMN (OMG, 2011) de acordo com a Tabela 4.2. O símbolo do Ator do BPMN foi adaptado para representar um nó do grafo. Os conjuntos de símbolos podem ser alterados com a demanda do usuário. Outros conjuntos de símbolos podem ser adicionados. Esse recurso permite

Tabela 4.2: Correlação entre os conjuntos de símbolos (PROV e BPMN) utilizados na fase de visualização do framework.

Modelo PROV		**Modelo BPMN**	
Nome	**Símbolos**	**Nome**	**Símbolos**
Entidade		Objeto de dados	
Agente		Ator	
Atividade		Tarefa	

que a visualização seja compreendida mais rapidamente pelo usuário, sem necessidade de conhecer ou aprender novos símbolos para interpretar os dados.

3. Controle de filtros

O controle de filtros permite ao usuário filtrar as informações exibidas na tela. As legendas podem ser subtraídas da visualização, bem como os nós de determinado tipo. Essa função auxilia na compreensão da visualização, especialmente em grafos muito densos, com muita informação em um pequeno espaço. Esse espaço do controle também funciona como uma legenda, vinculando os nomes de cada tipo com o símbolo utilizado.

4. Busca por nós

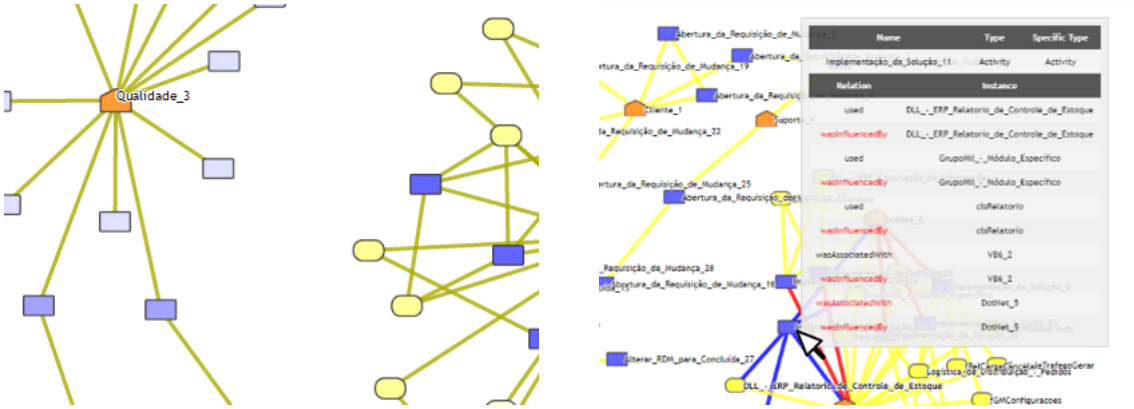
Esta opção destaca um nó escolhido pelo nome. Até dois nós e suas relações imediatas podem ser destacadas com esta opção. A busca, filtra todos os outros nós e relações, sendo utilizado para comparar dois elementos ou analisar visualmente a relação entre dois nós.

5. Visualização

Muitos recursos foram empregados para codificar as informações contidas no grafo ou na proveniência. A área de visualização permite o zoom, para focar a exploração na área de interesse e o *panning* para auxiliar na navegação do grafo de proveniência.

Os ícones possuem níveis diferentes de cores, de acordo com o grau de cada nó. A Figura 4.5(a) apresenta três diferentes tons para representar tarefas com diferentes graus, quanto menor o grau, mais fraco é o tom do nó. O clique no nó apresenta um painel com as informações disponíveis do nó e suas conexões. A Figura 4.5(b) apresenta esse

recurso, mostrando mais informações sobre uma tarefa indicada pelo cursosr do mouse na visualização.



(a) Níveis de cor diferentes para diferentes graus de cada nó

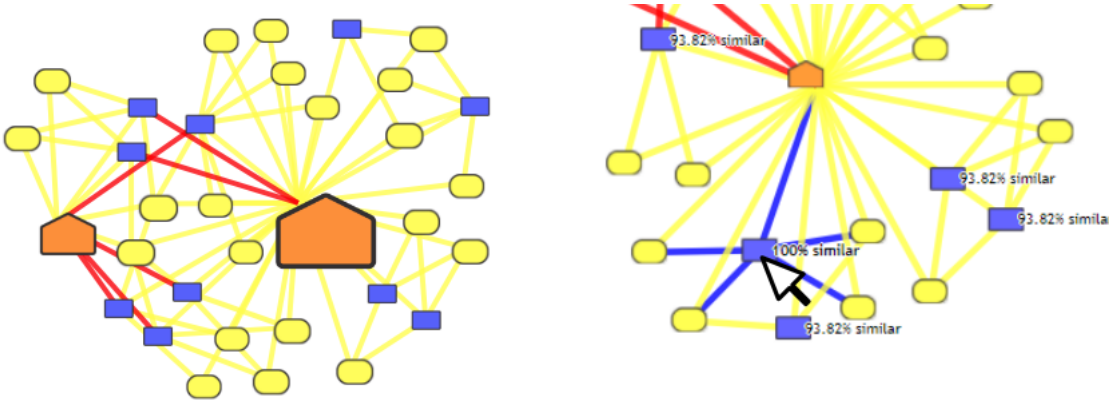
(b) Quadro de informações do nó, apresentado com o clique do mouse sobre o nó.

Figura 4.5: Recursos de visualização utilizados para auxiliar na compreensão e na exploração dos dados.

Os links entre os nós também são codificados visualmente. Para evitar o excesso de informação, a visualização exibe um grafo não direcionado, diferente do que define o modelo PROV, mas essa informação está codificada. Cada aresta do grafo contém todas as conexões entre os dois nós comunicantes e guarda internamente as informações de cada ligação. Essas informações, junto com o sentido de cada ligação, são exibidas com a demanda do usuário. Em outras palavras, cada aresta da visualização representa todas as conexões com a mesma direção, enquanto as conexões discriminadas por sentido são apresentadas na exploração de cada aresta. A espessura da aresta representa o número de ligações internas que possui e a cor da aresta representa os tipos de ligações, sendo verde, amarelo ou vermelho para os links afirmados, afirmados e inferidos e apenas inferidos, respectivamente. Desta forma, as inferências geradas com a utilização da ontologia são destacadas para atrair a atenção do usuário para novas informações e possíveis descobertas importantes sobre os dados.

O segundo passo de análise é apresentado de duas maneiras diferentes na visualização. O subgrafo da influência de $I_{G,a}$ é apresentado modificando o tamanho dos nós. Quanto maior o nó, maior sua influência no grafo de proveniência, como apresenta a Figura 4.6(a). Esta opção pode ser ativada pelo usuário para identificar pontos críticos da aplicação e também exibe o subgrafo de influência do nó permitindo ao usuário analisar os impactos de modificação, manutenção e substituição dos objetos representados pelo nó.

A análise do subgrafo $D_{G,a}$ identifica nós com estrutura similar e apresenta valores percentuais na visualização. Esses valores representam similaridade e auxiliam o usuário na identificação de nós semelhantes ao analisado. Como a proveniência representa a qualidade e a confiabilidade dos dados, esse recurso apresenta nós com qualidade e confiabilidade semelhantes. A Figura 4.6(b) apresenta o recursos de similaridade sendo utilizado na visualização.



(a) Quanto maior o tamanho do nó, mais ele influencia outros nós no grafo de proveniência

(b) Recurso utilizado para encontrar nós com qualidade semelhante à um nó escolhido.

Figura 4.6: Recursos de visualização desenvolvidos a partir das análises dos subgrafos de influência e dependência de cada nó.

Uma forma de redução do grafo foi criada e é opcional no framework. A redução foi criada para auxiliar na exposição de informações durante a visualização, sem sobrecarregá-la. O algorítimo busca nós de determinado tipo que estão ligados aos mesmos vizinhos pelas mesmas arestas. Ao encontrar mais de um nó com essa característica, eles são eliminados da visualização e um novo nó é criado, indicado como grupo de nós e contendo os nós subtraídos.

Algumas decisões de projeto foram tomadas durante a implementação do framework. As decisões mais importantes estão relacionadas aqui com suas justificativas.

- O padrão de cores tipo semáforo foi selecionado para a exibição de inferências nas arestas por ser um padrão de fácil compressão, multicultural e presente no cotidiano de todos.

- As arestas entre os nós foram reduzidas a apenas uma ligação não direcionada. Como as relações presentes no PROV-O possuem relações inversas, esta opção diminuiu

para menos da metade o número de ligações exibidas. A visualização se tornou mais simples para a facilitar a compreensão e possui informação sobre demanda.

- As legendas de todas as arestas foram eliminadas e apenas o nome dos nós é exibido no início da visualização.

- O layout do grafo é construído automaticamente para melhor aproveitar o espaço da visualização. O layout do grafo é reajustado todas as vezes que o posicionamento dos nós é modificado.

4.2 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou o framework Visionary, descrevendo seu objetivo de analisar e visualizar dados de proveniência para auxiliar na compreensão, desenvolvimento e manutenção de processos digitais. As etapas do modelo foram descritas e foi detalhado como cada etapa influencia na visualização gerada.

O resultado buscado foi uma abordagem capaz de auxiliar pesquisadores, cientistas e desenvolvedores que utilizam dados de proveniência. O Visionary pode ser adaptado à vários contextos com a utilização do modelo PROV. A abordagem também tem o objetivo de ser de fácil utilização para usuários que não possuem conhecimento profundo em proveniência.

5 AVALIAÇÃO DA PROPOSTA

Este capítulo descreve uma avaliação inicial da abordagem Visionary com o objetivo de verificar se a abordagem oferece um mecanismo adequado para a análise e compreensão dos dados de proveniência e se oferece apoio a tomada de decisão.

A Seção 5.1 apresenta os resultados de um estudo piloto que avaliou a abordagem Visionary e auxiliou no seu desenvolvimento e construção do estudo regular. A Seção 5.2 apresenta o estudo regular da abordagem, discriminando seu planejamento na Subseção 5.2.1, os participantes do estudo na Subseção 5.2.2, como foi realizada a coleta de dados na Subseção 5.2.3 e a análise dos resultados na Subseção 5.2.4 com as conclusões na Subseção 5.2.5. Finalmente na Subseção 5.2.6 são apresentadas as ameaças a validade do estudo e as considerações finais do capítulo na Seção 5.3.

5.1 ESTUDO PILOTO

Uma estudo piloto da abordagem foi realizado por Dalpra (2016). Nesta avaliação a abordagem Visionary foi adaptada e integrada com a arquitetura PROV-Process, foco do estudo de Dalpra (2016).

O PROV-Process é uma arquitetura que visa identificar melhorias nos processos de desenvolvimento de software e apresentá-las ao gerente de projetos por meio de uma aplicação orientada a serviços. Uma das formas das interfaces das abordagens da arquitetura utiliza a arquitetura Visionary para processamento, análise e visualização dos dados de proveniência gerados.

A realização deste estudo 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 em uma das empresas parceiras utilizadas neste estudo, as quais cederam dados de execução de seus processos.

Por meio dos resultados obtidos mediante a aplicação da avaliação do PROV-Process (Formulário de Avaliação PROV-Process – Apêndice A) realizada por Dalpra (2016), foi possível identificar alguns aspectos específicos considerando o uso da abordagem Visionary.

Considerando o formulário de avaliação (Apêndice A) e as respostas dadas pelos participantes em relação as inferências e visualização, foram compilados os resultados a seguir. Em relação a questão 6 ("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.

A Figura 5.1 (DALPRA, 2016) consolida as respostas da questão 9 do formulário ("Através da visualização gráfica é possível identificar melhor as inferências obtidas por meio do uso da ferramenta PROV-Process"). Este resultado indica que grande 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.

Em relação a afirmação de que a visualização gráfica possibilita uma análise mais rá-

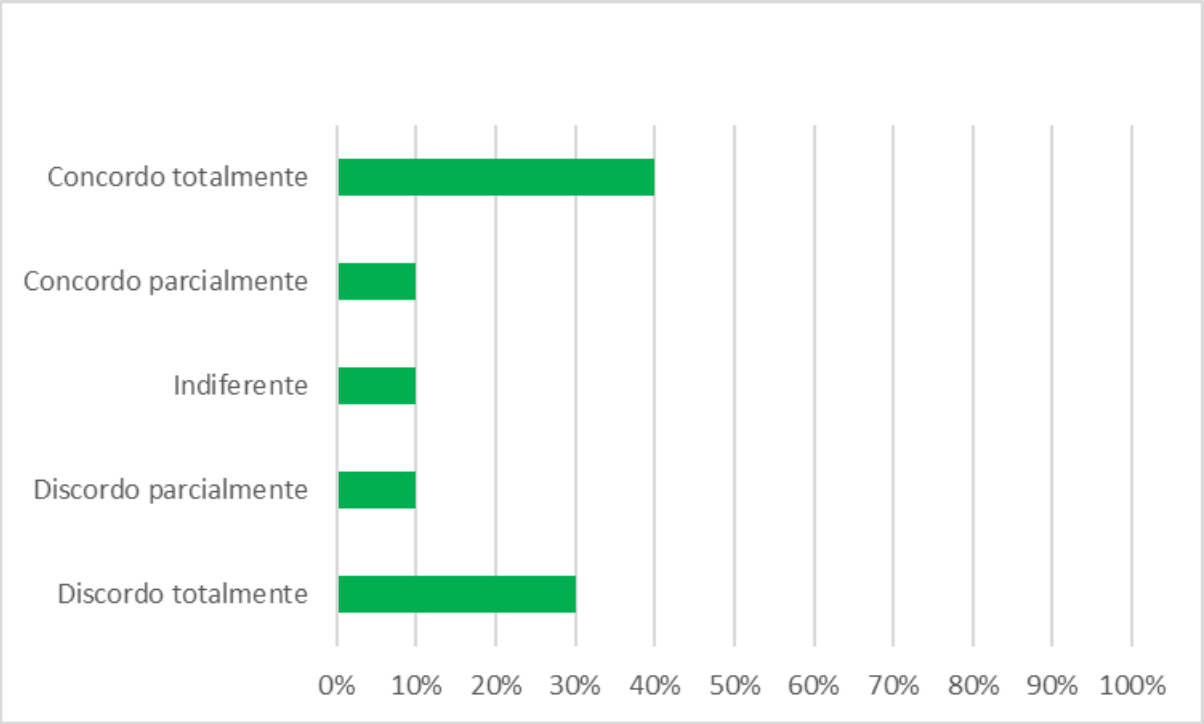


Figura 5.1: Respostas da questão 8 do formulário de avaliação do PROV-Process (Apêndice A) (DALPRA, 2016)

pida 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.

Ao final da avaliação, conforme Apêndice A, foram apresentadas perguntas dissertativas acerca da abordagem e da ferramenta PROV-Process. Analisando as respostas à pergunta 14 (O que mais gostou na abordagem PROV-Process?), segundo Dalpra (2016) 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.

Por fim, ao serem indagados sobre o que mudariam na ferramenta PROV-Process, considerando somente as questões de visualização, 90% dos participantes indicaram a adição de algum filtro e/ou busca, 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.

Com base nestes resultados, os pontos destacados pelos participantes como falhos na parte de visualização foram considerados para aprimorar a abordagem e para serem avaliados no estudo de caso regular, apresentado na próxima seção.

Assim, este estudo piloto serviu para demonstrar a viabilidade técnica do modelo, conceitos e tecnologias envolvidas no projeto. Desta forma, ele ajudou a ser um estudo teste para que pudéssemos descobrir melhorias a serem realizadas e falhas que deveriam ser corrigidas na condução do estudo de caso regular.

5.2 ESTUDO REGULAR

Após o estudo piloto, conforme já dito, as respostas dos participantes foram utilizadas para melhorar a proposta e ajustar possíveis deficiências da abordagem e da arquitetura. Assim, alguns aspectos da abordagem foram modificados e outros recursos foram adicionados. O estudo regular foi utilizado para avaliar aspectos específicos do framework Visionary, como o auxílio na compreensão e análise dos dados de proveniência, a geração de novas informações, a facilidade de utilização da interface e a utilização dos recursos de visualização.

5.2.1 PLANEJAMENTO DO ESTUDO

Como afirma Lethbridge et al. (2005) o primeiro passo de uma avaliação é a definição clara dos objetivos do estudo, já que muitas decisões posteriores são consequências deles. Dessa forma, o escopo da avaliação foi construído com base na estrutura do método

GQM (BASILI et al., 1994). Essa estrutura fornece um *template* para definir o escopo da avaliação, segue o modelo:

"**Analisar o** <objetivo de estudo> **com a finalidade de** <objetivo> **com respeito à** <foco da qualidade> **do ponto de vista de** <perspectiva> **no contexto de** <contexto>".

Dessa forma foi criado o escopo da avaliação, que é: "**Analisar** a abordagem Visionary **com a finalidade de** verificar sua capacidade de auxiliar na compreensão e análise dos dados de proveniência a partir do uso de mecanismos de visualização, redes complexas e ontologias **com respeito ao** auxílio na tomada de decisão **do ponto de vista de** usuários e desenvolvedores **no contexto de** aplicações científicas".

Atendendo ao escopo especificado, a seguinte questão de pesquisa (QP) foi elaborada:

> **QP:** O framework Visionary suporta a tomada de decisão através da análise e compreensão dos dados de proveniência?

Essa questão de pesquisa principal gerou três questões de pesquisa secundárias (QS) de apoio:

> **QS 1:** O framework Visionary é de fácil utilização?
>
> **QS 2:** O framework Visionary auxilia na compreensão de dados de proveniência?
>
> **QS 3:** O framework Visionary auxilia na análise de dados de proveniência?

A hipótese nula (H0) e a hipótese alternativa (H1) foram derivadas da questão de pesquisa principal. As hipóteses são:

> **H0:** O framework Visionary NÃO suporta a tomada de decisão através da análise e compreensão dos dados de proveniência.
>
> **H1:** O framework Visionary suporta a tomada de decisão através da análise e compreensão dos dados de proveniência.

A implementação do framework foi integrada com a arquitetura E-SECO (SIRQUEIRA et al., 2016). O E-SECO é um ecossistema de software científico (SOUZA et al., 2015) que permite a prototipação, manutenção e execução de workflows científicos. Durante a

execução dos workflows e em suas modificações o E-SECO captura os dados de proveniência que são utilizados para alimentar o framework Visionary. Essa integração foi utilizada para realizar o estudo presente.

Os dados de proveniência foram disponibilizados para os participantes que foram instruídos para executar roteiros de atividades pré-estabelecidos (Apêndices D, E, F e G) utilizando o framework. Quatro diferentes conjuntos de dados com seus respectivos roteiros foram disponibilizados para cada participante.

5.2.2 SELEÇÃO DOS INDIVÍDUOS

Participaram do estudo 5 voluntários que iniciaram sua colaboração com a assinatura do Termo de Consentimento Livre e Esclarecimento apresentado no Apêndice B e o questionário de caracterização apresentado no Apêndice C. Os voluntários são cientístas com experiência no uso de workflows científicos e mestrandos, com experiência com aplicações científicas e ecossistemas.

O objetivo da participação deste grupo é avaliar a compreensão e análise de dados de proveniência através do Visionary. Os resultados do questionário do Apêndice C mostram participantes com diferentes níveis de conhecimento nas áreas envolvidas, o que caracteriza diferentes cientistas com diferentes níveis de conhecimento sobre dados de proveniência. O questionário de auto caracterização coleta o conhecimento dos participantes em três áreas importantes: (i) proveniência de dados, (ii) visualização de software e (iii) redes complexas. Os resultados estão apresentados na Figura 5.2.

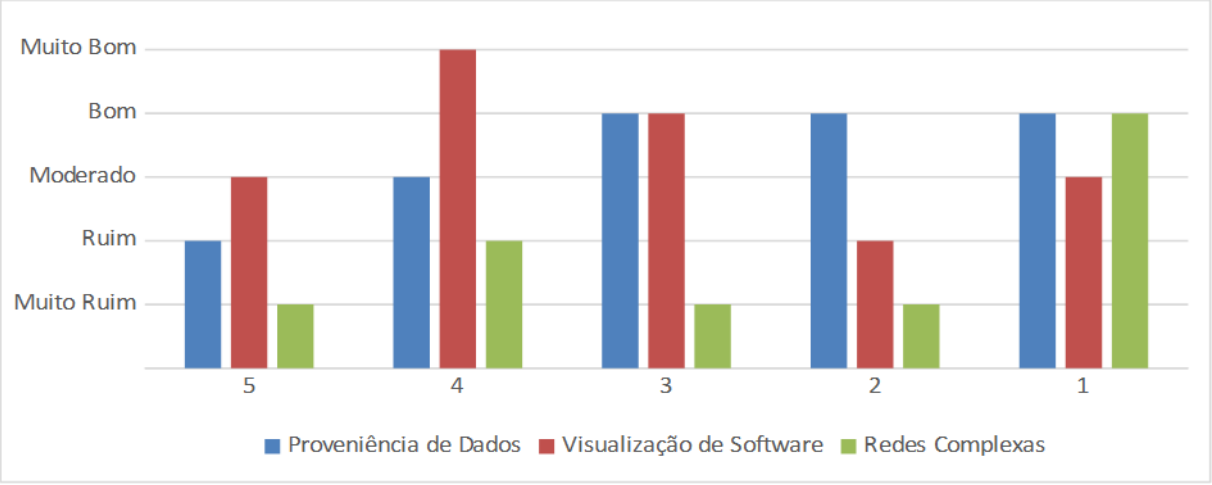


Figura 5.2: Compilação dos resultados do questionário de caracterização, mostrando o conhecimento dos participantes nas diferentes áreas envolvidas.

Essas três áreas foram analisadas para verificar uma possível tendência na avaliação. Um alto conhecimento em proveniência de dados auxilia na compreensão dos dados, no entanto verificamos conhecimentos medianos, entre ruim e bom. Participantes que possuem experiência em visualização de software podem contribuir mais para a melhoria da proposta. Com conhecimento em redes complexas, os participantes poderiam executar análises visuais através de métricas simples de serem calculadas (como o grau do nó por exemplo), como 3 dos 5 participantes possuem conhecimento ruim em redes complexas, eles vão depender mais das análises apresentadas pelo framework.

5.2.3 COLETA DE DADOS

Para melhor compreender o objeto de estudo, é importante a utilização de diferentes métodos de coleta de dados (LETHBRIDGE et al., 2005). Cada método de coleta é caracterizado por Lethbridge et al. (2005) em três diferentes ordens: os métodos de primeira ordem envolvem o pesquisador em contato direto com os indivíduos participantes onde a coleta é feita em tempo real, por meio de entrevistas, questionários, etc; nos métodos de segunda ordem o pesquisador coleta indiretamente os dados, através de interações dos indivíduos durante o estudo, por meio de diários de trabalho, observação através de vídeos e áudio e logs do sistema; nos métodos de terceira ordem o pesquisador não entra em contato com os participantes, apenas realiza a análise de artefatos de trabalho por meio de documentos.

Segundo Wohlin et al. (2012) as entrevistas e questionários são divididas em três categorias: não-estruturado, semiestruturado e totalmente estruturado. As entrevistas não-estruturadas são realizadas através de roteiros de entrevistas e realizam uma análise qualitativa com o objetivo de explorar o fenômeno analisado. As entrevistas semiestruturadas são realizadas com questões abertas e fechadas com o objetivo de explorar e descrever o fenomeno. Já as entrevistas totalmente estruturadas busca a relação entre dois fenômenos através de questões fechadas com o objetivo de descrição e exploração.

Foram estabelecidas três métodos de coleta de dados durante o estudo experimental. O primeiro método é de primeira ordem e semiestruturado, onde os participantes completam os quatro roteiros de atividades (Apêndices D, E, F e G) durante o estudo experimental. Cada roteiro possui 12 atividades diferentes e específicas para um conjunto de dados. Os conjuntos de dados são dados de proveniência relacionados a software científico, o primeiro

conjunto possui 56 nós e o segundo possui 64 nós. O terceiro e quarto conjunto de dados são maiores com 410 e 327 nós respectivamente. Os roteiros foram apresentados em ordem para todos os participantes. Esse método de coleta de dados pretende verificar a precisão das respostas dos participantes em cada tipo de atividade proposta. Cada roteiro possui 12 atividades específicas de um conjunto de dados diferentes.

O segundo método é de segunda ordem e não estruturado, onde o pesquisador faz uma análise observacional do participante tomando nota de acontecimentos relevantes. Esse método pretende verificar a qualidade da interface proposta, a aplicação dos recursos de visualização e identificar dificuldades na localização de funções e informações.

O terceiro método é de primeira ordem e semiestruturado onde o participante responde um questionário de avaliação (Apêndice H) após a realização das atividades. Esse método é utilizado para verificar a percepção do participante sobre o suporte fornecido pelo framework nas atividades propostas. Além disso, o questionário aceita também respostas discursivas.

A Tabela 5.1 apresenta um resumo dos três métodos de coleta do estudo experimental.

Tabela 5.1: Classificação e objetivo dos métodos de coleta de dados utilizados na avaliação da proposta.

Método de Coleta	**Ordem**	**Tipo**	**Objetivo**
Roteiros de Atividade (Apêndices D, E, F e G)	Primeira	Semiestruturado	Verificar a precisão das atividades
Observação do Participante	Segunda	Não-estruturado	Avalia a qualidade da interface e a dificuldade dos participantes
Questionário de Avaliação (Apêndice H)	Primeira	Semiestruturado	Verificar o apoio à tomada de decisão

Os quatro roteiros de atividades possuem três grupos, onde cada atividade foi especificada com base no objetivo proposto. As atividades 1, 2, e 3 de cada roteiro são de informações gerais, utilizadas para analisar a interface e os recursos de visualização. As atividades 4, 5, 6 e 12 são de exploração e compreensão dos dados e servem para verificar se os resultados são apresentados de forma clara e objetiva. Já as atividades 7, 8, 9, 10 e 11 são de suporte à decisão e análise dos dados e verificam se as novas informações apresentadas são úteis e se auxiliam na tomada de decisão. Os roteiros de atividade estão disponíveis nos Apêndices D, E, F e G.

5.2.4 ANÁLISE DOS RESULTADOS

Nesta sessão os resultados dos três métodos de coleta são apresentados separadamente. A métrica "precisão" (ÁLVAREZ, 2007) foi utilizada para avaliar as atividades realizadas, juntamente com o tempo de execução dos roteiros.

A Subseção 5.2.4.1 analisa os roteiros de atividades; a Subseção 5.2.4.2 apresenta os resultados colhidos com a observação dos participantes; já a Subseção 5.2.4.3 analisa os resultados do questionário de autoavaliação.

5.2.4.1 Roteiros de Atividades

Todas as respostas foram compiladas e a precisão de todos os participantes está resumida na Tabela 5.2.

Tabela 5.2: Precisão das respostas dos participantes em cada atividade dos roteiros propostos.

Atividade	**Roteiros**				**Total**
	1 Apêndice D	**2 Apêndice E**	**3 Apêndice F**	**4 Apêndice G**	
1	1,0	1,0	1,0	1,0	1,00
2	0,6	1,0	0,8	0,8	0,80
3	1,0	1,0	1,0	1,0	1,00
4	0,8	1,0	0,8	1,0	0,90
5	0,8	0,8	0,8	1,0	0,85
6	0,8	0,8	0,8	0,8	0,80
7	1,0	1,0	1,0	0,8	0,95
8	1,0	0,6	0,8	1,0	0,85
9	0,8	0,8	1,0	1,0	0,90
10	1,0	1,0	1,0	1,0	1,00
11	1,0	1,0	0,8	0,8	0,90
12	0,6	0,4	0,4	0,4	0,45
Total	0,87	0,87	0,85	0,88	0,87

Inicialmente pode-se destacar a precisão das atividades de 0,87. Comparando esse valor com a precisão do estudo piloto (0,8167) (DALPRA, 2016) verifica-se um pequeno aumento.

As atividades que se destacam positivamente pela precisão são as atividades 1, 3 e 10 (Apêndices D, E, F e G). As atividades 1 e 2 (Apêndices D, E, F e G) possuem a resposta literal apresentada diretamente na visualização o que justifica a precisão alta da atividade 1 mas se contrapõe com a precisão baixa e não esperada da atividade 2. Já as atividades 3

e 10 podem ser encontradas com uma análise visual dos dados (observando a visualização gerada pelo framework, a resposta pode ser encontrada), mesmo sem treinamento no uso do framework, os participantes realizaram as atividades com sucesso.

A Figura 5.3 apresenta as precisões das respostas dos participantes discriminadas por atividade. Em destaque a atividade 12 que apresenta precisão além do desvio padrão e foi analisada separadamente. Após a análise da questão e dos resultados, o pesquisador considerou a questão mal formulada o que gerou dúvidas nos participantes e a baixa precisão.

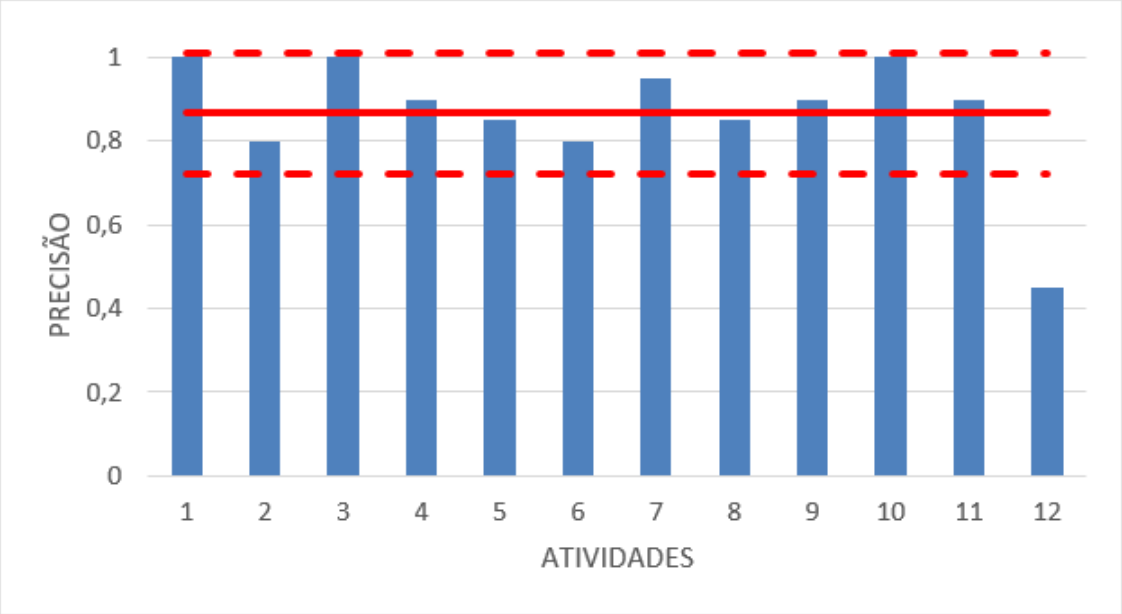


Figura 5.3: Precisão das respostas das atividades propostas. A linha vermelha apresenta a média dos valores, a linha tracejada o desvio padrão.

As atividades também foram analisadas segundo o tipo de questão: informações gerais (atividades 1, 2 e 3), exploração e compreensão (atividades 4, 5, 6 e 12) e análise e suporte à decisão (atividades 7, 8, 9, 10 e 11). A Tabela 5.3 apresenta a precisão obtida para cada tipo de questão.

Tabela 5.3: Precisão das respostas dos participantes em cada **tipo** de atividade dos roteiros propostos.

Atividade	**Roteiros**				**Total**
	1	**2**	**3**	**4**	
Informações Gerais	0,87	1,0	0,93	0,93	0,93
Exploração e Compreensão	0,75	0,75	0,70	0,80	0,75
Análise e Suporte à Decisão	0,96	0,88	0,92	0,92	0,92
Total	0,87	0,87	0,85	0,88	0,87

As atividades de informações gerais são as mais simples de responder, as respostas são mais diretas e facilmente encontradas nas visões disponíveis pelo framework. A maior precisão encontrada está nessas atividades com o valor de 0,93. Esse resultado reflete a interface e os recursos de visualização, indicando um bom resultado mas a necessidade de pequenos ajustes, tais como a apresentação de mais informações sobre os dados visualizados e a utilização de legendas, sobre alguns dos recursos de visualização.

As atividades de exploração e compreensão são parte importantes no escopo da avaliação e apresentaram o pior resultado com uma precisão de 0,75. A atividade 12 teve um resultado inesperado, como apresenta a Figura 5.3 o resultado da atividade é um *outlier*, ou seja, seu valor está além de um desvio padrão (destacado pela linha tracejada na Figura 5.3) da média de todos os resultados. Isso indica uma interferência no resultado da atividade, como um erro na elaboração que permite dupla interpretação da atividade, levando à respostas diversas dos participantes. Mesmo desconsiderando o resultado da atividade 12, as atividades de exploração e compreensão teriam 0,85 de precisão. Apesar de ser a menor precisão no escopo deste tipo de atividade, o resultado é bem maior do que uma resposta aleatória com duas opções (precisão 0,5), sem deixar de considerar ainda que as respostas eram abertas. Portanto, podemos considerar que existe indícios que o framework auxilia na exploração e compreensão dos dados de proveniência.

As atividades de análise e suporte à decisão tiveram uma precisão de 0,92. Esse valor, muito perto do primeiro grupo de atividades, demonstra um bom resultado das principais atividades da avaliação. Uma das possibilidades da alta precisão é o fato que essas atividades possuem respostas mais abrangentes que as outras. De qualquer forma, os resultados das questões trazem uma indicação que o framework auxilia na análise e suporte à tomada de decisão.

O tempo na realização dos roteiros também foi registrado. A Tabela 5.4 apresenta um resumo dos tempos obtidos nos quatro roteiros de atividades propostas.

Dentre todos os participantes o tempo gasto no roteiro 2 é expressivamente menor que o roteiro 1. A redução varia entre 37,5% e 64,29%. Essa variação reflete a curva de aprendizado do framework, já que nenhuma instrução foi dada para os participantes compreenderem os recursos do Visionary. O tempo gasto aumenta entre o segundo e o terceiro roteiro, refletindo a diferença do tamanho dos dados de proveniência de 64 nós do roteiro 2 para 410 nós do roteiro 3. Outra redução ocorre do roteiro 3 para o roteiro

Tabela 5.4: Tempo de realização dos roteiros de atividade por cada participante.

	Roteiros				
Participante	**1**	**2**	**3**	**4**	**Total**
1	16min	8min	12min	12min	48min
2	17min	10min	13min	12min	52min
3	16min	10min	18min	13min	57min
4	28min	10min	23min	13min	74min
5	17min	8min	18min	14min	57min
Média	18min 48s	9min 12s	16min 48s	12min 48s	57min 36s

4, dessa vez menos expressiva, a redução varia entre 0% e 43,48%. Após lidar com um número grande de nós era previsível a redução do tempo gasto nas atividades com o roteiro 4.

Apesar de alguns resultados positivos na precisão das atividades também foram identificadas adaptações para serem realizadas no framework. As principais adaptações estão na apresentação das informações para o usuário

A avaliação preliminar da precisão das atividades fornece indícios que o framework auxilia na compreensão e análise dos dados de proveniência e oferece suporte à tomada de decisão. Mesmo assim algumas necessidades de adaptação do framework foram identificadas, como a apresentação de mais informação ao usuário sob demanda e modificações na interface do sistema.

5.2.4.2 Observação dos Participantes

A observação dos participantes revelou alguns comportamentos recorrentes. Antes de começar a realização das atividades, 3 dos participantes exploraram os recursos disponíveis para conhecer o framework e os outros 2 iniciaram as atividades juntamente com a exploração dos recursos. Alguns dos recursos não foram utilizados desde o primeiro roteiro, os recursos só foram amplamente utilizados a partir do segundo ou terceiro roteiro. Não foi encontrada relação entre a utilização de recursos e a precisão já registrada das atividades.

Dois dos participantes utilizaram amplamente a visão de *Data Chart*, sugerindo uma ampliação dos recursos aplicados nessa visão, como foi sugerido pelos participantes posteriormente. Uma combinação de recursos na visualização apresentou um conflito que gerava mal funcionamento dos mesmos. Essa falha na implementação atrapalhou 4 dos participantes e um deles sofreu a interferência do pesquisador para retomar a normalidade no uso framework.

Outras sugestões e necessidades de melhorias observadas consideram a correção e ampliação de algumas funcionalidades além da criação de um sistema de ajuda para facilitar a compreensão dos recursos. Ampliar o sistema de similaridade dos nós considerando a equivalência estrutural pode melhorar os resultados dessa funcionalidade, e precisa ser testado. A apresentação de valores numéricos para o cálculo de importância também foi sugerido pelos participantes e pode ser facilmente adicionado na visão de *Data Chart*.

5.2.4.3 Questionário de Avaliação

O questionário de avaliação buscou registrar a percepção dos participantes sobre a importância do framework na realização das atividades propostas. O questionário possui 31 questões, onde 28 são fechadas e 3 abertas. As respostas das questões fechadas utilizam um item *Likert* com cinco opções. O questionário aplicado pode ser consultado no Apêndice H.

As primeiras perguntas do questionário são perguntas diretas sobre a percepção dos participantes sobre a utilização do framework. A questão número 1 pergunta ao participante *"Você considera o Visionary uma ferramenta de fácil utilização?"*, 40% das respostas concordaram parcialmente e 60% concordaram totalmente, 0 respostas foram indiferentes, descordaram parcialmente ou totalmente.

Com o foco no objetivo definido para a avaliação, o questionário pergunta se o framework auxilia na compreensão dos dados de proveniência e indaga também se o framework auxilia na realização de cada atividade. Sendo assim podemos comparar a pergunta direta sobre compreensão e o resumo das questões que tratam da compreensão dos dados. O mesmo pode ser feito com o foco de análise e suporte à decisão. A Figura 5.4 traz a comparação das respostas diretas e indiretas sobre a compreensão dos dados de proveniência.

Perguntados diretamente 80% dos participantes afirmaram que concordam totalmente sobre o framework auxiliar na compreensão dos dados de proveniência e 20% concordam parcialmente, conforme apresentado na Figura 5.4. Um resultado semelhante ocorre quando a questão é sobre o framework auxiliar nas atividades correspondentes às questões de compreensão e exploração dos dados. Nesse caso 75% concordaram totalmente e 25% concordaram parcialmente. Podemos considerar portanto que existe forte indício que a abordagem auxilia em atividades de compreensão dos dados de proveniência.

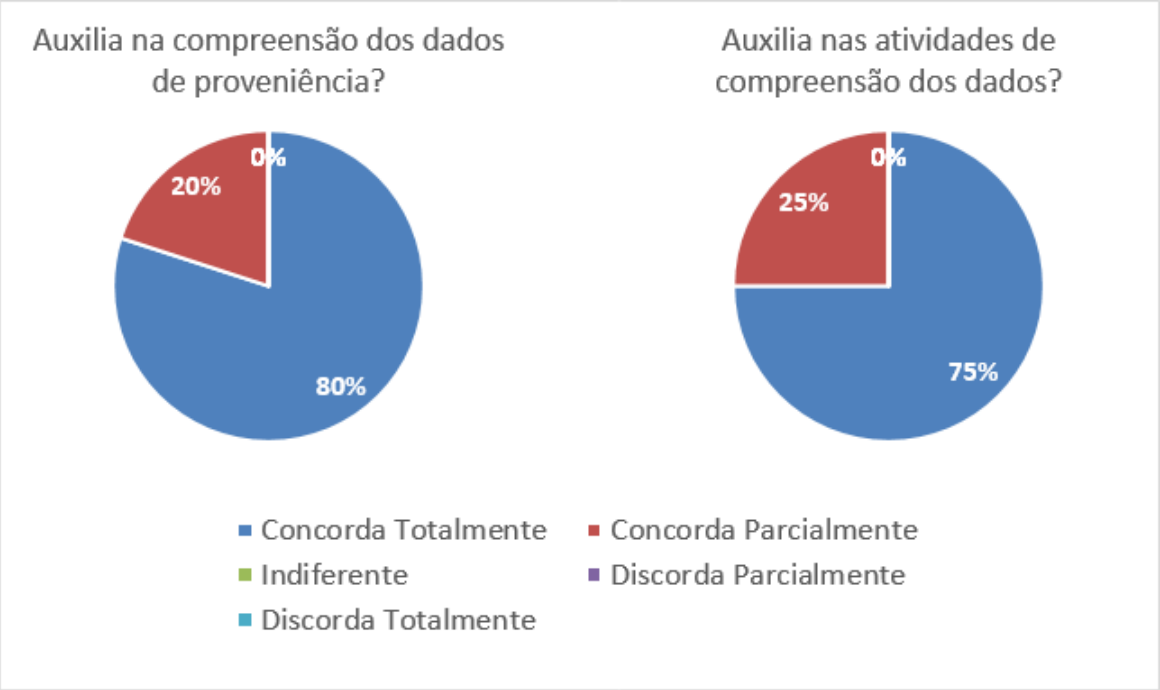


Figura 5.4: Comparação entre as respostas diretas e indiretas do questionário de avaliação sobre a compreensão dos dados de proveniência.

As questões de análise e suporte a decisão possuem uma diferença maior dos resultados da questão direta e das questões indiretas, como apresentado na Figura 5.5.

Perguntados diretamente 80% dos participantes afirmaram que concordam totalmente sobre o framework auxiliar na análise dos dados de proveniência e 20% concordam parcialmente, conforme apresentado na Figura 5.5. Nas questões indiretas, 4% descordaram totalmente, 8% acharam que o framework é indiferente, 28% concordaram parcialmente e a grande maioria, 60% concordaram totalmente. Isso significa que 1 participante considerou que o framework não contribuiu para a realização de 1 questão (4%). O que pode ser devido à dificuldade de utilização ou não compreensão dos recursos do framework. O mesmo ocorre com as respostas indiferentes, 2 atividades (8%) foram consideradas por 1 ou 2 participantes como não recebendo auxílio adequado do framework para a solução. Apesar disso os participantes consideraram que, de forma geral o framework auxiliou na análise e suporte à decisão. Isso traz indícios que o framework suporta de maneira adequada a análise dos dados de proveniência e suporta a decisão sobre os dados, apesar de modificações serem necessárias para melhor suportar essas atividades.

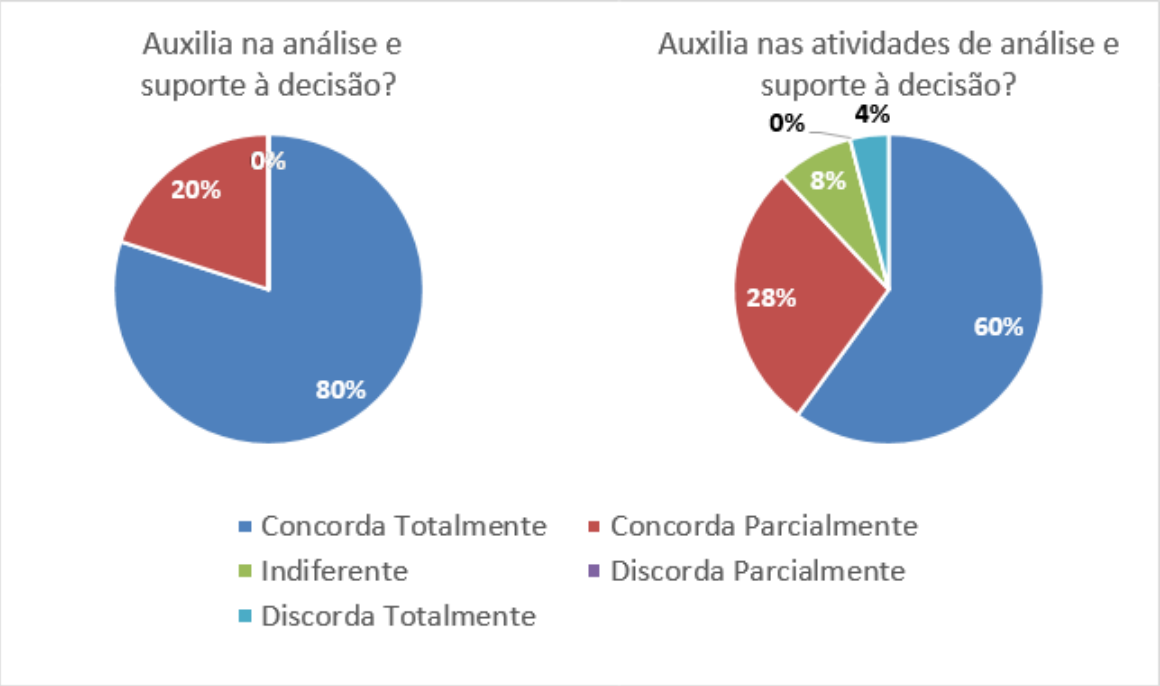


Figura 5.5: Comparação entre as respostas diretas e indiretas do questionário de avaliação sobre a análise dos dados e suporte à decisão.

5.2.5 CONCLUSÕES PRELIMINARES

O estudo experimental foi construído para responder as questões de pesquisas apresentadas na Seção 5.2.1. Com base nos resultados obtidos com as três coletas de dados pode-se realizar as seguintes afirmações, considerando as questões de pesquisa secundárias:

Pode-se responder positivamente a **QS 1** (O Framework Visionary é de fácil utilização?), considerando os indícios encontrados como: (i) apesar da falta de treinamento dos participantes com o framework, eles responderam a maior parte das atividades com sucesso, como mostra a Tabela 5.2 e (ii) na resposta da questão 1 do questionário de avaliação presente no Apêndice H (Você considera o Visionary uma ferramenta de fácil utilização?) com um total de 60% de respostas concordando totalmente com a questão, como apresentado na Seção 5.2.4.3.

A **QS 2** (O framework Visionary auxilia na compreensão de dados de proveniência?) também pode ser respondida positivamente com base na: (i) realização bem sucedida das atividades de compreensão, como apresenta a Tabela 5.3 com precisão de 0,75 nessas atividades e (ii) nas perguntas diretas e indiretas do questionário de avaliação, como mostra a Figura 5.4 com mais de 75% das respostas concordando totalmente com a resposta afirmativa para a questão de pesquisa.

A última questão de pesquisa secundária (**QS 3**) pode ser respondida afirmativamente com as mesmas referências da **QS 2**: (i) a realização bem sucedida das atividades de análise, como apresenta a Tabela 5.3 com precisão de 0,92 nessas atividades e (ii) nas perguntas diretas e indiretas do questionário de avaliação, como mostra a Figura 5.5 com mais de 88% das respostas concordando parcialmente ou totalmente com a afirmativa que o framework Visionary auxilia na análise e suporte à decisão.

Dessa forma temos vários indícios para responder a questão de pesquisa principal, rejeitar a hipótese nula (**H0**) e aceitar a hipótese alternativa (**H1**), afirmando que o framework Visionary suporta a tomada de decisão através da análise e compreensão dos dados de proveniência.

5.2.6 AMEAÇAS A VALIDADE

A validade do estudo é relacionado à confiabilidade dos resultados, considerando se os resultados são tendenciosos ou não, a partir do ponto de vista subjetivo dos pesquisadores (RUNESON et al., 2012). Para avaliar a qualidade da avaliação, são comumente utilizados quatro testes (YIN, 2015): (i) a validade do construto, que verifica se as medidas operacionais foram corretamente estabelecidas para os conceitos que estão sob estudo; (ii) validade interna, que permite estabelecer corretamente relação de causa e consequência entre duas condições; (iii) validade externa, para estabelecer o domínio ao qual as descobertas podem ser generalizadas e (iv) confiabilidade, que demonstra a reprodutibilidade do estudo.

Validade de construto: as atividades propostas podem não ter sido abrangentes o suficiente para abarcar todas as necessidades de usuários e desenvolvedores que precisam compreender e analisar os dados de proveniência. Um estudo em cenário real auxiliaria na evolução do framework e forneceria mais sustentabilidade para as relações estabelecidas no estudo. O estudo também precisa ser ampliado em número de participantes para auxiliar sua validade e generalização dos resultados.

Validade Interna: os resultados obtidos são ainda preliminares e apesar de indicarem uma conclusão positiva, um estudo mais detalhado sobre as respostas, incluindo métodos estatísticos, é importante para apresentar conclusões mais concretas.

Validade externa: o estudo foi apresentado sob o contexto de SGWfC e seus resultados são limitados para o contexto empregado e o grupo de dados utilizado.

Confiabilidade: detalhes minuciosos da execução do estudo não foram apresentados, mas a documentação garante relativa reprodutibilidade da avaliação.

5.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Nesse capítulo foi apresentado a avaliação da abordagem Visionary. A abordagem sofreu duas avaliações, uma avaliação piloto, com uma integração com uma arquitetura de análise de processos de software e uma avaliação regular, integrada à um SGWfC. A avaliação regular teve como foco o auxílio fornecido pelo Visionary para a compreensão e análise dos dados de proveniência e suporte à decisão. O estudo obteve indícios que o Visionary cumpre com seu objetivo de auxiliar na análise e compreensão dos dados e fornecer informações importantes para dar suporte à decisão. Outras avaliações ainda são necessárias para garantir a eficácia do Visionary.

6 CONSIDERAÇÕES FINAIS

Essa dissertação apresentou o framework Visionary, que tem como objetivo auxiliar na compreensão e análise dos dados de proveniência. Além disso, o desenvolvimento desta dissertação ressalta a importância do uso de dados de proveniência e como estes podem auxiliar na tomada de decisão considerando informações estratégicas das empresas. Neste contexto, duas revisões sistemáticas foram realizadas e seus resultados demostram a importância do tema atualmente. A primeira revisão foi relacionada ao tema "Visualização de Dados de Proveniência", e a segunda relacionada a "Análise de Dados de Proveniência".

O Visionary Framework foi detalhado no Capítulo 4. As etapas relacionadas ao uso do Visionary foram especificadas, incluindo a captura dos dados, as análises dos dados utilizando ontologias e redes complexas e a visualização das informações.

O Visionary foi desenvolvimento considerando o modelo padrão de proveniência PROV e por conta disso, herda as características do modelo, sendo genérico e flexível para ser adaptado a diferentes contextos. Desta forma, as análises utilizando ontologias e redes complexas podem ser utilizadas em qualquer extensão realizada no modelo, desde de que suas características básicas sejam mantidas.

A avaliação preliminar do Visionary apresentou indícios de sua capacidade em auxiliar a compreensão e análise dos dados de proveniência, auxiliando usuários na tomada de decisão e aprimoramento dos processos que geraram os dados. Apesar disso, o framework precisa ser mais explorado, ampliado e avaliado.

6.1 CONTRIBUIÇÕES

As principais contribuições desta dissertação são destacadas a seguir:

- Mapeamento e Revisão Sistemáticas considerando as duas principais tecnologias envolvidas na abordagem, i.e., visualização de dados de proveniência e análises de dados de proveniência. Com estas revisões, foi possível identificar os principais trabalhos na área, além de apresentar as principais contribuições destes trabalhos. Além disso, foi possível identificar requisitos específicos ainda não abordados pela literatura revisada e, a partir destes, identificar novas funcionalidades que deram

origem a proposta do framework Visionary.

- Especificação e implementação de funcionalidades relacionadas a análises do grafo de proveniência, que podem ser utilizadas em qualquer contexto e adaptação do modelo PROV, capazes de fornecer conhecimento estratégico sobre os dados.
- Uma abordagem para a visualização dos dados de proveniência, de código aberto e baseada na web, que simplifica a compreensão e exploração dos dados.

6.2 LIMITAÇÕES

A seguir são apresentadas algumas limitações da abordagem como um todo e na implementação dos seus recursos.

- Uma avaliação com mais participantes deve ser conduzida para melhor embasar os resultados obtidos com o estudo experimental.
- A visualização das informações fica sobrecarregada com um conjunto muito grande de dados. Apesar dos recursos de filtragem, destaque e localização, é conveniente a implementação de um sistema de zoom semântico capaz de selecionar os dados apresentados.
- As análises do grafo também apresentam uma demora excessiva com um conjunto muito grande de dados. Neste caso é importante a otimização do algoritmo e/ou estratégias para análise que sejam mais rápidas.

6.3 TRABALHOS FUTUROS

O framework Visionary pode ser ampliado nas diversas etapas. A partir da versão inicial do framework e sua implementação, foram geradas novas demandas a serem trabalhadas, como descrito nas limitações. Algumas atividades que podem ser mencionadas como trabalhos futuros são:

- Especificação de novos algoritmos e regras para as análises do grafo, permitindo a geração de novas informações que podem ampliar as análises atualmente realizadas.

- Novas visualizações podem ser especificadas. Múltiplas visões vão ofertar pontos de vista diferentes dos dados, facilitando a sua exploração e compreensão.

- Um sistema de adaptação e instalação semiautomático pode ser desenvolvido, permitindo usuários do PROV configurar rápido e facilmente o framework para seu sistema.

REFERÊNCIAS

ACAR, U. A.; AHMED, A.; CHENEY, J.; PERERA, R. A core calculus for provenance. **POST**, Springer, v. 7215, p. 410–429, 2012.

ÁLVAREZ, A. C. **Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem**. Tese (Doutorado) — Universidade de São Paulo, 2007.

AMARAL, L. A.; OTTINO, J. M. Complex networks. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 38, n. 2, p. 147–162, 2004.

ANAND, M. K.; BOWERS, S.; LUDÄSCHER, B. Provenance browser: Displaying and querying scientific workflow provenance graphs. In: IEEE. **Data Engineering (ICDE), 2010 IEEE 26th International Conference on**, 2010. p. 1201–1204.

ANTONIOU, G.; HARMELEN, F. V. Web ontology language: Owl. In: **Handbook on ontologies**, 2004. p. 67–92.

ARSHAD, B.; MUNIR, K.; MCCLATCHEY, R.; LIAQUAT, S. Position paper: Provenance data visualisation for neuroimaging analysis. **arXiv preprint arXiv:1502.01556**, 2015.

BASILI, V.; CALDIERA, G.; ROMBACH, D. Gqm paradigm. **Computer Encyclopedia of Software Engineering**, John Wiley and Sons, 1994.

BOAVENTURA, P. O. **Grafos: teoria, modelos, algoritmos**, 2012.

BORKIN, M. A.; YEH, C. S.; BOYD, M.; MACKO, P.; GAJOS, K. Z.; SELTZER, M.; PFISTER, H. Evaluation of filesystem provenance visualization tools. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 19, n. 12, p. 2476–2485, 2013.

BOWERS, S.; MCPHILLIPS, T.; LUDÄSCHER, B. Declarative rules for inferring fine-grained data provenance from scientific workflow execution traces. In: SPRINGER. **International Provenance and Annotation Workshop**, 2012. p. 82–96.

BOWERS, S.; MCPHILLIPS, T.; LUDÄSCHER, B.; COHEN, S.; DAVIDSON, S. B. A model for user-oriented data provenance in pipelined scientific workflows. In: SPRINGER. **International Provenance and Annotation Workshop**, 2006. p. 133–147.

BUNEMAN, P.; CHAPMAN, A.; CHENEY, J.; VANSUMMEREN, S. A provenance model for manually curated data. **IPAW**, Springer, v. 6, p. 162–170, 2006.

BUNEMAN, P.; KHANNA, S.; TAN, W. C. Why and where: A characterization of data provenance. In: SPRINGER. **ICDT**, 2001. v. 1, p. 316–330.

CALLAHAN, S. P.; FREIRE, J.; SANTOS, E.; SCHEIDEGGER, C. E.; SILVA, C. T.; VO, H. T. Vistrails: visualization meets data management. In: ACM. **Proceedings of the 2006 ACM SIGMOD international conference on Management of data**, 2006. p. 745–747.

CAO, B.; PLALE, B.; SUBRAMANIAN, G.; ROBERTSON, E.; SIMMHAN, Y. Provenance information model of karma version 3. In: IEEE. **Services-I, 2009 World Conference on**, 2009. p. 348–351.

CASERTA, P.; ZENDRA, O. Visualization of the static aspects of software: A survey. **IEEE transactions on visualization and computer graphics**, IEEE, v. 17, n. 7, p. 913–933, 2011.

CEOLIN, D.; GROTH, P.; HAGE, W. R. V.; NOTTAMKANDATH, A.; FOKKINK, W. Trust evaluation through user reputation and provenance analysis. In: CEUR-WS. ORG. **Proceedings of the 8th International Conference on Uncertainty Reasoning for the Semantic Web-Volume 900**, 2012. p. 15–26.

CEOLIN, D.; GROTH, P.; MACCATROZZO, V.; FOKKINK, W.; HAGE, W. R. V.; NOTTAMKANDATH, A. Combining user reputation and provenance analysis for trust assessment. **Journal of Data and Information Quality (JDIQ)**, ACM, v. 7, n. 1-2, p. 6, 2016.

CHEAH, Y.-W.; PLALE, B. Provenance analysis: Towards quality provenance. In: IEEE. **E-Science (e-Science), 2012 IEEE 8th International Conference on**, 2012. p. 1–8.

CHEN, P.; PLALE, B.; CHEAH, Y.-W.; GHOSHAL, D.; JENSEN, S.; LUO, Y. Visualization of network data provenance. In: IEEE. **High Performance Computing (HiPC), 2012 19th International Conference on**, 2012. p. 1–9.

CHEN, Y. V.; QIAN, Z. C.; WOODBURY, R.; DILL, J.; SHAW, C. D. Employing a parametric model for analytic provenance. **ACM Transactions on Interactive Intelligent Systems (TiiS)**, ACM, v. 4, n. 1, p. 6, 2014.

CHENEY, J.; CHITICARIU, L.; TAN, W.-C. et al. Provenance in databases: Why, how, and where. **Foundations and Trends® in Databases**, Now Publishers, Inc., v. 1, n. 4, p. 379–474, 2009.

COSTA, G. C.; SCHOTS, M.; OLIVEIRA, W. E.; LO, H.; DALPRA, C. M.; BRAGA, R.; DAVID, J. M. N.; MARCOS, A.; MIGUEL, V. S.; CAMPOS, F. Sppv: Visualizing software process provenance data. **Sociedade Brasileira de Computação–SBC**, p. 49, 2016.

DAI, C.; LIN, D.; BERTINO, E.; KANTARCIOGLU, M. An approach to evaluate data trustworthiness based on data provenance. **Secure Data Management**, Springer, v. 5159, p. 82–98, 2008.

DALPRA, H. L. O. **PROV-Process: proveniência de dados aplicada a processos de desenvolvimento de software**. Dissertação (Mestrado) — Universidade Federal de Juiz de Fora (UFJF), 2016.

DAVIDSON, S. B.; FREIRE, J. Provenance and scientific workflows: challenges and opportunities. In: ACM. **Proceedings of the 2008 ACM SIGMOD international conference on Management of data**, 2008. p. 1345–1350.

DEELMAN, E.; GANNON, D.; SHIELDS, M.; TAYLOR, I. Workflows and e-science: An overview of workflow system features and capabilities. **Future generation computer systems**, Elsevier, v. 25, n. 5, p. 528–540, 2009.

DIEHL, S. **Software visualization: visualizing the structure, behaviour, and evolution of software**, 2007.

EBDEN, M.; HUYNH, T.; MOREAU, L.; RAMCHURN, S.; ROBERTS, S. Network analysis on provenance graphs from a crowdsourcing application. **Provenance and Annotation of Data and Processes**, Springer, p. 168–182, 2012.

GIL, Y.; DEELMAN, E.; ELLISMAN, M.; FAHRINGER, T.; FOX, G.; GANNON, D.; GOBLE, C.; LIVNY, M.; MOREAU, L.; MYERS, J. Examining the challenges of scientific workflows. **Computer**, IEEE, v. 40, n. 12, 2007.

GIL, Y.; MILES, S. Prov model primer. **W3C Working Draft, 11th December**, 2012.

GROTH, P.; DEELMAN, E.; JUVE, G.; MEHTA, G.; BERRIMAN, B. Pipeline-centric provenance model. In: ACM. **Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science**, 2009. p. 4.

GROTH, P.; MOREAU, L. Prov-overview. an overview of the prov family of documents. World Wide Web Consortium, 2013.

GUARINO, N. et al. Formal ontology and information systems. In: **Proceedings of FOIS**, 1998. v. 98, n. 1998, p. 81–97.

HARARY, F. Graph theory. addison. **Reading, MA**, 1969.

HOEKSTRA, R.; GROTH, P. Prov-o-viz-understanding the role of activities in provenance. In: SPRINGER. **International Provenance and Annotation Workshop**, 2014. p. 215–220.

HUNTER, J.; CHEUNG, K. Provenance explorer-a graphical interface for constructing scientific publication packages from provenance trails. **International Journal on Digital Libraries**, Springer, v. 7, n. 1, p. 99–107, 2007.

HUYNH, T. D.; EBDEN, M.; VENANZI, M.; RAMCHURN, S. D.; ROBERTS, S.; MOREAU, L. Interpretation of crowdsourced activities using provenance network analysis. In: **First AAAI Conference on Human Computation and Crowdsourcing**, 2013.

JACOBSON, I.; BOOCH, G.; RUMBAUGH, J.; RUMBAUGH, J.; BOOCH, G. **The unified software development process**, 1999.

KADIVAR, N.; CHEN, V.; DUNSMUIR, D.; LEE, E.; QIAN, C.; DILL, J.; SHAW, C.; WOODBURY, R. Capturing and supporting the analysis process. In: IEEE. **Visual**

Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, 2009. p. 131–138.

KARSAI, L. **Clustering Provenance**. Tese (Doutorado) — Ph. D. thesis, University of Sydney, 2016.

KHAN, S.; KANTURSKA, U.; WATERS, T.; EATON, J.; BAÑARES-ALCÁNTARA, R.; CHEN, M. Ontology-assisted provenance visualization for supporting enterprise search of engineering and business files. **Advanced Engineering Informatics**, Elsevier, v. 30, n. 2, p. 244–257, 2016.

KIENLE, H. M.; MULLER, H. A. Requirements of software visualization tools: A literature survey. In: IEEE. **Visualizing Software for Understanding and Analysis, 2007. VISSOFT 2007. 4th IEEE International Workshop on**, 2007. p. 2–9.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.

KITCHENHAM, B. A.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. In: **Technical report, Ver. 2.3 EBSE Technical Report. EBSE**, 2007.

KOHWALTER, T.; OLIVEIRA, T.; FREIRE, J.; CLUA, E.; MURTA, L. Prov viewer: a graph-based visualization tool for interactive exploration of provenance data. In: SPRINGER. **International Provenance and Annotation Workshop**, 2016. p. 71–82.

LEBO, T.; SAHOO, S.; MCGUINNESS, D.; BELHAJJAME, K.; CHENEY, J.; CORSAR, D.; GARIJO, D.; SOILAND-REYES, S.; ZEDNIK, S.; ZHAO, J. Prov-o: The prov ontology. **W3C recommendation**, v. 30, 2013.

LETHBRIDGE, T. C.; SIM, S. E.; SINGER, J. Studying software engineers: Data collection techniques for software field studies. **Empirical software engineering**, Springer, v. 10, n. 3, p. 311–341, 2005.

LIM, C.; LU, S.; CHEBOTKO, A.; FOTOUHI, F. Prospective and retrospective provenance collection in scientific workflow environments. In: IEEE. **Services Computing (SCC), 2010 IEEE International Conference on**, 2010. p. 449–456.

LOPES, C. T.; FRANZ, M.; KAZI, F.; DONALDSON, S. L.; MORRIS, Q.; BADER, G. D. Cytoscape web: an interactive web-based network browser. **Bioinformatics**, Oxford University Press, v. 26, n. 18, p. 2347–2348, 2010.

LUDÄSCHER, B.; ALTINTAS, I.; BERKLEY, C.; HIGGINS, D.; JAEGER, E.; JONES, M.; LEE, E. A.; TAO, J.; ZHAO, Y. Scientific workflow management and the kepler system. **Concurrency and Computation: Practice and Experience**, Wiley Online Library, v. 18, n. 10, p. 1039–1065, 2006.

MARGO, D. W.; SMOGOR, R. Using provenance to extract semantic file attributes. In: **TaPP**, 2010.

MATTOSO, M.; GLAVIC, B. **Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings**, 2016.

MCGRATH, R. E.; FUTRELLE, J. Reasoning about provenance with owl and swrl rules. In: **AAAI Spring Symposium: AI Meets Business Rules and Process Management**, 2008. p. 87–92.

MISSIER, P.; BELHAJJAME, K. A prov encoding for provenance analysis using deductive rules. In: SPRINGER. **IPAW**, 2012. p. 67–81.

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**, Elsevier, v. 27, n. 6, p. 743–756, 2011.

MOREAU, L.; KWASNIKOWSKA, N.; BUSSCHE, J. Van den. The foundations of the open provenance model. 2009.

MOREAU, L.; MISSIER, P. Prov-dm: The prov data model. **Retrieved July**, v. 30, p. 2013, 2013.

MUNISWAMY-REDDY, K.-K.; HOLLAND, D. A.; BRAUN, U.; SELTZER, M. I. Provenance-aware storage systems. In: **USENIX Annual Technical Conference, General Track**, 2006. p. 43–56.

MUNISWAMY-REDDY, K.-K.; SELTZER, M. Provenance as first class cloud data. **ACM SIGOPS Operating Systems Review**, ACM, v. 43, n. 4, p. 11–16, 2010.

NEWMAN, M. **Networks: an introduction**, 2010.

OINN, T.; LI, P.; KELL, D. B.; GOBLE, C.; GODERIS, A.; GREENWOOD, M.; HULL, D.; STEVENS, R.; TURI, D.; ZHAO, J. Taverna/my grid: aligning a workflow system with the life sciences community. **Workflows for e-Science**, Springer, p. 300–319, 2007.

OLIVEIRA, W.; AMBRÓSIO, L. M.; BRAGA, R.; STRÖELE, V.; DAVID, J. M.; CAMPOS, F. A framework for provenance analysis and visualization. **Procedia Computer Science**, Elsevier, v. 108, p. 1592–1601, 2017.

OMG, B. P. M. Notation (bpmn) version 2.0 (2011). **Available on: http://www. omg. org/spec/BPMN/2.0**, 2011.

PACKER, H. S.; MOREAU, L. Sentence templating for explaining provenance. 2014.

PETERSEN, K. Measuring and predicting software productivity: A systematic map and review. **Information and Software Technology**, Elsevier, v. 53, n. 4, p. 317–343, 2011.

PRAT, N.; MADNICK, S. Measuring data believability: A provenance approach. In: IEEE. **Hawaii International Conference on System Sciences, Proceedings of the 41st Annual**, 2008. p. 393–393.

PRIKLADNICKI, R.; AUDY, J. L. N. Process models in the practice of distributed software development: A systematic review of the literature. **Information and Software Technology**, Elsevier, v. 52, n. 8, p. 779–791, 2010.

RICHARDSON, D. P.; MOREAU, L. Towards the domain agnostic generation of natural language explanations from provenance graphs for casual users. In: SPRINGER. **International Provenance and Annotation Workshop**, 2016. p. 95–106.

RIO, N. D.; SILVA, P. P. D. Probe-it! visualization support for provenance. In: SPRINGER. **International Symposium on Visual Computing**, 2007. p. 732–741.

RUNESON, P.; HOST, M.; RAINER, A.; REGNELL, B. **Case study research in software engineering: Guidelines and examples**, 2012.

SILVA, P. P.; MCGUINNESS, D. L.; FIKES, R. A proof markup language for semantic web services. **Information Systems**, Elsevier, v. 31, n. 4, p. 381–395, 2006.

SIMMHAN, Y. L.; PLALE, B.; GANNON, D. A survey of data provenance in e-science. **ACM Sigmod Record**, ACM, v. 34, n. 3, p. 31–36, 2005.

SIRQUEIRA, T. F. M. et al. E-seco proversion: uma arquitetura para manutenção e evolução de workflows científicos. Master's thesis, Universidade Federal de Juiz de Fora (UFJF), 2016.

SOUZA, V. F. et al. **ECOS PL-Science: Uma Arquitetura para Ecossistemas de Software Cient´ ıfico Apoiada por uma Rede Ponto a Ponto**. Dissertação (Mestrado) — Universidade Federal de Juiz de Fora (UFJF), 2015.

SPENCE, R. **Information Visualization: Design for Interaction second edition by: Robert Spence**, 2007.

STITZ, H.; LUGER, S.; STREIT, M.; GEHLENBORG, N. Avocado: Visualization of workflow–derived data provenance for reproducible biomedical research. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**, 2016. v. 35, n. 3, p. 481–490.

STRUBULIS, C.; TZITZIKAS, Y.; DOERR, M.; FLOURIS, G. Evolution of workflow provenance information in the presence of custom inference rules. In: **3rd Intern. Workshop on the role of Semantic Web in Provenance Management (SWPM'12), co-located with ESWC'12, Heraklion, Crete**, 2012.

TOWNEND, P.; WEBSTER, D.; VENTERS, C. C.; DIMITROVA, V.; DJEMAME, K.; LAU, L.; XU, J.; FORES, S.; VIDUTO, V.; DIBSDALE, C. et al. Personalised provenance reasoning models and risk assessment in business systems: A case study. In: IEEE. **Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on**, 2013. p. 329–334.

WARE, C. **Information visualization: perception for design**, 2012.

WASSERMAN, S.; FAUST, K. **Social network analysis: Methods and applications**, 1994.

WOHLIN, C.; RUNESON, P.; HÖST, M.; OHLSSON, M. C.; REGNELL, B.; WESSLÉN, A. **Experimentation in software engineering**, 2012.

YIN, R. K. **Estudo de Caso-: Planejamento e Métodos**, 2015.

YU, L. **A developer's guide to the semantic Web**, 2011.

Apêndice A - FORMULÁRIO DE 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

3.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

4.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

5.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

6.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

7.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

8.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

9.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

10. 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

11. 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

12. 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

13. 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

14.O que mais gostou na abordagem PROV-Process?

15.O que menos gostou na ferramenta PROV-Process?

16.O que mudaria na ferramenta PROV-Process?

Apêndice B - TERMO DE CONCENTIMENTO E LIVRE ESCLARECIMENTO (TCLE)

Condutor do estudo: Weiner Esmério Batista de Oliveira

Instituição: Universidade Federal de Juiz de Fora

Introdução

Este termo de consentimento é referente ao estudo de avaliação aplicado ao framework Visionary para identificar suas contribuições e limitações.

O framework Visionary foi desenvolvido para auxiliar na compreensão e uso dos dados de proveniência através de técnicas de visualização de software, ontologias e análise de redes complexas. O framework captura os dados de proveniência e gera novas informações usando ontologias e análise do grafo de proveniência. A visualização apresenta e destaca as inferências e os resultados obtidos com a análise. O Visionary é um framework livre de contexto que pode ser adaptado para qualquer sistema que utiliza o modelo PROV de proveniência.

A proveniência é reconhecida hoje como um desafio central para estabelecer confiabilidade e prover segurança em sistemas computacionais. Em workflows científicos, por exemplo, a proveniência é considerada essencial para apoiar a reprodutibilidade dos experimentos, a interpretação dos resultados e o diagnóstico de problemas. Consideramos que estes benefícios podem também ser utilizados em outros contextos, como em processos de software.

Procedimentos

O framework foi implementado e será utilizada para executar todas as etapas de avaliação. O (A) participante deverá seguir os seguintes passos para realização da avaliação: (i) ler e assinar, caso de acordo, o TCLE (este documento); (ii) ler e preencher o Questionário de Caracterização do Participante; (iii) ler o Roteiro de Avaliação e retirar suas dúvidas, se houver, com o pesquisador presente; (iv) preencher o questionário de avaliação.

Para todos os dados coletados, serão retiradas informações pessoais, que não serão utilizadas em nenhum momento durante a análise ou apresentação dos resultados.

Tratamento de Riscos

Todas as providências necessárias serão tomadas durante a coleta de dados, visando

garantir a sua privacidade. Os dados coletados durante o estudo destinam-se apenas às atividades relacionadas com a solução proposta. Benefícios e Custos

Espera-se que este estudo seja capaz de fornecer novos conhecimentos sobre questões relacionadas à proveniência de dados, bem como aspectos de visualização de software. Este estudo visa contribuir com a área de proveniência e visualização, apresentando métodos de análise dos dados de proveniência.

A participação neste estudo não envolve nenhum gasto ou ônus. O (A) participante também não receberá qualquer espécie de reembolso ou gratificação devido à participação na pesquisa.

Confidencialidade da Pesquisa

As informações coletadas neste estudo são confidenciais. Seus dados pessoais, não serão divulgados de nenhuma forma.

Participação

Sua participação neste estudo é muito importante e voluntária. Você tem o direito de não querer participar ou de sair deste estudo a qualquer momento, sem penalidades. Em caso de você decidir se retirar do estudo, favor notificar o pesquisador responsável. O pesquisador responsável por este estudo irá fornecer explicações sobre o estudo. Para isto, entre em contato com um dos pesquisadores a partir dos seguintes endereços de e-mails:

Weiner Esmério Batista de Oliveira – woliveira82@gmail.com

Declaração de Consentimento

Li as informações contidas neste documento antes de assinar este termo de consentimento. Declaro, para os devidos fins, que toda a linguagem técnica utilizada na descrição deste estudo de pesquisa foi explicada satisfatoriamente e que recebi respostas para minhas dúvidas. Confirmo também que sou livre para me retirar do estudo em qualquer momento, sem qualquer penalidade. Declaro ter mais de 18 anos e dou meu consentimento de livre e espontânea vontade para participar deste estudo.

_____

Local e Data

_____

CPF do Participante

_____

Nome do Participante

Assinatura do Participante

Apêndice C - QUESTIONÁRIO DE CARACTERIZAÇÃO

Número do participante: _____

1.Formação:

( ) Doutorado
( ) Doutorado (cursando)
( ) Mestrado
( ) Mestrado (cursando)
( ) Graduação
( ) Graduação (cursando)

3.Como você avalia seus conhecimentos em relação à proveniência de dados?

( ) Muito bom
( ) Bom
( ) Moderado
( ) Baixo
( ) Muito baixo

4.Como você avalia seus conhecimentos em relação à visualização de software?

( ) Muito bom
( ) Bom
( ) Moderado
( ) Baixo
( ) Muito baixo

5.Como você avalia seus conhecimentos em relação à análise de redes complexas?

( ) Muito bom
( ) Bom

( ) Moderado

( ) Baixo

( ) Muito baixo

6.Como você avalia seus conhecimentos em relação Business Process Model Notation (BPMN)?

( ) Muito bom

( ) Bom

( ) Moderado

( ) Baixo

( ) Muito baixo

Apêndice D - ROTEIRO DE AVALIAÇÃO 1

Diante da ferramenta Visionary, execute os passos e responda as perguntas a seguir:
Número do participante: ___Hora de início: ___: ___Hora de término: ___: ___

1. Qual o nome da ontologia está sendo analisada? _____
2. Quantos nós ou vértices estão sendo exibidos pela visualização? _____
3. Existe informações inferidas na visualização? _____
4. Localize o item (nó) com o nome "fRelCargaCancelar" e informe seu tipo e tipo específico:

   Tipo _____Tipo Específico _____
5. Localize o item com o nome "clsCalculoNota" e informe a quantidade de itens que ele está se relacionado:

   _____

   _____
6. Informe também o nome dos itens que o item acima está se relacionando:

   _____
7. Qual o nome do agente mais importante na rede? _____
8. Qual o nome da atividade mais importante na rede? _____
9. Qual o nome da entidade mais importante na rede? _____
10. Qual dos itens apresentados apresenta maior dificuldade de substituição, caso seja necessário? _____
11. Qual o item mais similar à "Implementação_da_Solução_11"? _____
12. Quantas relações estão presentes entre os itens "Pesquisador_1" e "Abertura_da_Requisição_de_Mudança_4"? _____

Apêndice E - ROTEIRO DE AVALIAÇÃO 2

Diante da ferramenta Visionary, execute os passos e responda as perguntas a seguir:
Número do participante: ___Hora de início: ___: ___Hora de término: ___: ___

1. Qual o nome da ontologia está sendo analisada? _____
2. Quantos nós ou vértices estão sendo exibidos pela visualização? _____
3. Existe informações inferidas na visualização? _____
4. Localize o item (nó) com o nome "Bernardo" e informe seu tipo e tipo específico:

   Tipo _____Tipo Específico _____
5. Localize o item com o nome "Resolução_do_Caso_19" e informe a quantidade de itens que ele está se relacionado:

   _____

   _____
6. Informe também o nome dos itens que o item acima está se relacionando:

   _____
7. Qual o nome do agente mais importante na rede? _____
8. Qual o nome da atividade mais importante na rede? _____
9. Qual o nome da entidade mais importante na rede? _____
10. Qual dos itens apresentados apresenta maior dificuldade de substituição, caso seja necessário? _____
11. Qual o item mais similar à "Resolução_do_caso_15"? _____
12. Quantas relações estão presentes entre os itens "Solicitar_novo_recurso_9" e "Grupo_pesquisa_C"? _____

Apêndice F - ROTEIRO DE AVALIAÇÃO 3

Diante da ferramenta Visionary, execute os passos e responda as perguntas a seguir:
Número do participante: ___Hora de início: ___: ___Hora de término: ___: ___

1. Qual o nome da ontologia está sendo analisada? _____
2. Quantos nós ou vértices estão sendo exibidos pela visualização? _____
3. Existe informações inferidas na visualização? _____
4. Localize o item (nó) com o nome "_7339" e informe seu tipo e tipo específico:

   Tipo _____Tipo Específico _____
5. Localize o item com o nome "WebNovoCodigo.vb_9883" e informe a quantidade de itens que ele está se relacionado:

   _____

   _____
6. Informe também o nome dos itens que o item acima está se relacionando:

   _____
7. Qual o nome do agente mais importante na rede? _____
8. Qual o nome da atividade mais importante na rede? _____
9. Qual o nome da entidade mais importante na rede? _____
10. Qual dos itens apresentados apresenta maior dificuldade de substituição, caso seja necessário? _____
11. Qual o item mais similar à "fTiposPesquisador_9894"? _____
12. Quantas relações estão presentes entre os itens "VB6_7338" e "Implementação_da_solução_11021"? _____

Apêndice G - ROTEIRO DE AVALIAÇÃO 4

Diante da ferramenta Visionary, execute os passos e responda as perguntas a seguir:
Número do participante: ___Hora de início: ___: ___Hora de término: ___: ___

1. Qual o nome da ontologia está sendo analisada? _____
2. Quantos nós ou vértices estão sendo exibidos pela visualização? _____
3. Existe informações inferidas na visualização? _____
4. Localize o item (nó) com o nome “Bernardo” e informe seu tipo e tipo específico:

   Tipo _____Tipo Específico _____
5. Localize o item com o nome “_9586” e informe a quantidade de itens que ele está se relacionado:

   _____

   _____
6. Informe também o nome dos itens que o item acima está se relacionando:

   _____
7. Qual o nome do agente mais importante na rede? _____
8. Qual o nome da atividade mais importante na rede? _____
9. Qual o nome da entidade mais importante na rede? _____
10. Qual dos itens apresentados apresenta maior dificuldade de substituição, caso seja necessário? _____
11. Qual o item mais similar à “_9661”? _____
12. Quantas relações estão presentes entre os itens “Equipe_de_teste_7286” e “Resolução_do_caso_10499”? _____

Apêndice H - QUESTIONÁRIO DE AVALIAÇÃO

Responda as questões abaixo em relação à ferramenta Visionary:

Número do participante: _____

1. Você considera o Visionary uma ferramenta de fácil utilização?
   ( ) Concordo totalmente
   ( ) Concordo parcialmente
   ( ) Indiferente
   ( ) Descordo parcialmente
   ( ) Descordo totalmente

2. Os recursos do Visionary são facilmente identificados?
   ( ) Concordo totalmente
   ( ) Concordo parcialmente
   ( ) Indiferente
   ( ) Descordo parcialmente
   ( ) Descordo totalmente

3. Você considera que a ferramenta Framework auxilia na compreensão dos dados de proveniência?
   ( ) Muito bom
   ( ) Bom
   ( ) Moderado
   ( ) Baixo
   ( ) Muito baixo

4. Você considera os recursos de visualização bem aplicados na ferramenta Visionary?
   ( ) Muito bom
   ( ) Bom
   ( ) Moderado
   ( ) Baixo
   ( ) Muito baixo

5.Você considera que a ferramenta Framework auxilia na análise dos dados de proveniência?

( ) Muito bom

( ) Bom

( ) Moderado

( ) Baixo

( ) Muito baixo

6.Como você avalia seus conhecimentos sobre as atividades requisitadas no Roteiro de Avaliação?

( ) Muito bom

( ) Bom

( ) Moderado

( ) Baixo

( ) Muito baixo

7.Você considerou de fácil realização a atividade 2?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

8.Os recursos do Visionary auxiliam na realização da atividade 2?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

9.Você considerou de fácil realização a atividade 3?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

10.Os recursos do Visionary auxiliam na realização da atividade 3?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

11.Você considerou de fácil realização a atividade 4?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

12.Os recursos do Visionary auxiliam na realização da atividade 4?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

13.Você considerou de fácil realização a atividade 5?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

14.Os recursos do Visionary auxiliam na realização da atividade 5?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

15.Você considerou de fácil realização a atividade 6?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

16.Os recursos do Visionary auxiliam na realização da atividade 6?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

17.Você considerou de fácil realização a atividade 7?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

18.Os recursos do Visionary auxiliam na realização da atividade 7?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

19.Você considerou de fácil realização a atividade 8?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

20.Os recursos do Visionary auxiliam na realização da atividade 8?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

21.Você considerou de fácil realização a atividade 9?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

22.Os recursos do Visionary auxiliam na realização da atividade 9?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

23.Você considerou de fácil realização a atividade 10?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

24.Os recursos do Visionary auxiliam na realização da atividade 10?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

25.Você considerou de fácil realização a atividade 11?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

26.Os recursos do Visionary auxiliam na realização da atividade 11?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

27.Você considerou de fácil realização a atividade 12?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

28.Os recursos do Visionary auxiliam na realização da atividade 12?

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Indiferente

( ) Descordo parcialmente

( ) Descordo totalmente

29.Relate problemas encontrados durante a realização da avaliação?

_____

_____

_____

30. Você tem alguma sugestão de melhoria ou modificação do Framework?

31. Sinta-se livre para realizar qualquer comentário: