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

Lenita Martins Ambrósio

Apoiando o Reúso em uma Plataforma de Ecossistema de Software Científico Através do Gerenciamento de Contexto e de Proveniência

Juiz de Fora
2018

Lenita Martins Ambrósio

Apoiando o Reúso em uma Plataforma de Ecossistema de Software Científico Através do Gerenciamento de Contexto e 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 a obtenção do grau de Mestre em Ciência da Computação.

Orientador: José Maria Nazar David.

Juiz de Fora

2018

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

Ambrósio, Lenita Martins.
Apoiando o Reúso em uma Plataforma de Ecossistema de Software Científico Através do Gerenciamento de Contexto e de Proveniência / Lenita Martins Ambrósio. -- 2018.
122 p. : il.

Orientador: José Maria Nazar David
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós Graduação em Ciência da Computação, 2018.

1. Proveniência de Dados. 2. Elementos de Contexto. 3. Reúso. 4. Experimentos Científicos. 5. E-Science. I. David, José Maria Nazar, orient. II. Título.

Lenita Martins Ambrósio

Apoiando o Reúso em uma Plataforma de Ecossistema de Software Científico Através do Gerenciamento de Contexto e 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 a obtenção do grau de Mestre em Ciência da Computação.

Aprovada em 14 de setembro de 2018

BANCA EXAMINADORA

_____

Prof. D.Sc. José Maria Nazar David - Orientador

Universidade Federal de Juiz de Fora

_____

Prof.ª D.Sc. Regina Maria Maciel Braga

Universidade Federal de Juiz de Fora

_____

Prof. D.Sc. Wagner Antônio Arbex

Universidade Federal de Juiz de Fora

_____

Prof.ª D.Sc. Neide dos Santos

Universidade do Estado do Rio de Janeiro

_____

Prof.ª D.Sc. Fernanda Araújo Baião Amorim

Universidade Federal do Estado do Rio de Janeiro

Ao meu marido Yuri.

AGRADECIMENTOS

Agradeço a Deus por iluminar minhas escolhas nos momentos de incerteza, por me oferecer boas oportunidades e por ter colocado boas pessoas ao longo deste caminho.

Ao meu marido Yuri que me ajudou com tudo que estava ao seu alcance, foi paciente nos meus momentos de estresse, foi compreensivo com minha ausência e me encheu de amor e carinho nos momentos mais difíceis.

Aos meus pais Getúlio e Regina por me apoiarem em todas as minhas decisões.

À minha tia Rosana por me incentivar e me mostrar que eu seria capaz de conquistar este e muitos outros sonhos.

Aos meus irmãos Aline e Davi que seguem meu caminho e por isso me motivam a ir cada vez mais longe.

Aos meus avós, pelo incentivo e oração.

Ao meu orientador Professor José Maria pela confiança depositada em mim e no meu trabalho, por todo tempo dedicado a este trabalho e pelo conhecimento compartilhado.

À Prof.ª Regina Braga pelos ensinamentos e por suas inúmeras contribuições com este trabalho.

À pesquisadora Mariana Campos pela atenção e auxílio na condução do estudo de caso na Embrapa Gado de Leite.

Aos professores Fernanda Baião, Neide dos Santos e Wagner Arbex que enriqueceram este trabalho com suas críticas e contribuições.

A todos os demais professores do PPGCC por compartilharem comigo seus conhecimentos.

Aos amigos que fiz durante o curso por estarem sempre presentes compartilhando comigo todas as dificuldades e conquistas, especialmente à amiga Tatiane pela companhia que tornou esta jornada mais leve e divertida.

Aos demais familiares e amigos que contribuíram de alguma forma para esta conquista.

"*We are drowning in information but starved for knowledge*"

John Naisbitt

RESUMO

Considerando o cenário de experimentação científica atual e o crescente uso de aplicações em larga escala, o gerenciamento de dados de experimentação está se tornando cada vez mais complexo. O processo de experimentação científica requer suporte para atividades colaborativas e distribuídas. O gerenciamento de informações contextuais e de proveniência desempenha um papel fundamental no domínio neste domínio. O registro detalhado das etapas para produzir resultados, bem como as informações contextuais do ambiente de experimentação, pode permitir que os cientistas reutilizem esses resultados em experimentos futuros e reutilizem o experimento ou partes dele em outro contexto. O objetivo deste trabalho é apresentar uma abordagem de gerenciamento de informações de proveniência e contexto que apoie pesquisadores no reúso de conhecimento sobre experimentos científicos conduzidos em uma plataforma colaborativa e distribuída. Primeiramente, as fases do ciclo de vida do gerenciamento de contexto e proveniência foram analisadas, considerando os modelos existentes. Em seguida, foi proposto um *framework* conceitual para apoiar a análise de elementos contextuais e dados de proveniência de experimentos científicos. Uma ontologia capaz de extrair conhecimento implícito neste domínio foi especificada. Essa abordagem foi implementada em uma plataforma de ecossistema científico. Uma avaliação realizada por meio de estudos de caso evidenciou que essa arquitetura é capaz de auxiliar os pesquisadores durante a reutilização e reprodução de experimentos científicos. Elementos de contexto e proveniência de dados, associados a mecanismos de inferência, podem ser utilizados para apoiar a reutilização no processo de experimentação científica.

Palavras-chave: Proveniência de Dados. Elementos de Contexto. Reúso. Experimentos Científicos. Workflows Científicos. E-Science.

ABSTRACT

Considering the current experimentation scenario and the increasing use of large-scale applications, the experiment data management is growing complex. The scientific experimentation process requires support for collaborative and distributed activities. Managing contextual and provenance information plays a key role in the scientific domain. Detailed logging of the steps to produce results, as well as the environment context information could allow scientists to reuse these results in future experiments and reuse the experiment or parts of it in another context. The goal of this work is to present a provenance and context metadata management approach that support researchers in the reuse of knowledge about scientific experiments conducted in a collaborative and distributed platform. First, the context and provenance management life cycle phases were analyzed, considering existing models. Then it was proposed a conceptual framework to support the analysis of contextual elements and provenance data of scientific experiments. An ontology capable of extracting implicit knowledge in this domain was specified. This approach was implemented in a scientific ecosystem platform. An evaluation conducted through case studies shown evidences that this architecture is able to help researchers during the reuse and reproduction of scientific experiments. Context elements and data provenance, associated with inference mechanisms, can be used to support the reuse in scientific experimentation process.

Keywords: Data Provenance. Contextual Elements. Reuse. Scientific Experiments. Scientific Workflows. E-Science.

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EA	Eficiência Alimentar
ECOS	Ecossistema de Software
ECOSC	Ecossistema de Software Científico
E-SECO	e-Science *Ecosystem*
LPSC	Linha de Produtos de Software Científico
OPM	*Open Provenance Model*
OWL	*Web Ontology Language*
PRIME	*PRagmatic Interoperability to MEaningful collaboration*
RDF	*Resource Description Framework*
SGWfC	Sistema Gerenciador de *Workflow* Científico
UFJF	Universidade Federal de Juiz de Fora
W3C	*World Wide Web Consortium*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta o contexto relacionado ao problema abordado nesta dissertação, as motivações para o estudo deste problema, os objetivos do presente trabalho, bem como sua organização.

1.1 CONTEXTUALIZAÇÃO

Ao longo das últimas décadas a ciência passou a explorar novas possibilidades de experimentação científica. Fenômenos complexos passaram a ser simulados por supercomputadores através de ferramentas computacionais. A ciência está evoluindo progressivamente para a *e-Science* que é considerada como o terceiro paradigma da ciência (DEELMAN *et al.*, 2009). Com o uso de simulações cada vez mais complexas, a ciência computacional enfrenta atualmente um novo desafio: gerenciar e analisar o grande volume de dados produzido por essas simulações. Neste contexto, um quarto paradigma emerge, relacionado à Ciência Aberta e Intensiva em Dados (*Data-Intensive Science and Open Science*) (HEY, 2009). Nesse novo cenário de experimentação, a informação é o principal produto. E é fundamental que os cientistas possam compartilhar essas informações com outros membros da comunidade, bem como reutilizar os dados de seus pares (TENOPIR *et al.*, 2015).

Um experimento científico caracteriza-se por uma série de operações ligadas entre si, as quais são modeladas e executadas através de *workflows* científicos (GOBLE *et al.*, 2010). Um *workflow* científico é um modelo ou *template* composto por serviços, scripts ou outros *workflows*, que representam uma sequência de atividades científicas implementadas por ferramentas, a fim de alcançar um determinado objetivo (DEELMAN *et al.*, 2009). Para apoiar os pesquisadores durante a modelagem e execução dos *workflows* científicos, surgiram os Sistemas Gerenciadores de *Workflows* Científicos (SGWfCs).

SGWfCs, como o Kepler (ALTINTAS *et al.*, 2004), o Taverna (OINN *et al.*, 2007) e o VisTrails (FREIRE *et al.*, 2006) modelam explicitamente a dependência entre processos dentro de um experimento e coordenam o comportamento dos processos em tempo de execução (BELLOUM *et al.*, 2011). Estas ferramentas envolvem etapas de análises de dados vindos de diversas fontes e computação em larga escala. Além disso, requerem suporte à colaboração entre os cientistas geograficamente distribuídos (DEELMAN *et al.*, 2009).

No cenário de desenvolvimento de software, uma das abordagens usadas para lidar com as necessidades de colaboração e distribuição em um ambiente heterogêneo são os Ecossistemas de Software (ECOSs) (MANIKAS, 2016). No domínio de *e-Science*, denomina-

se Ecossistema de Software Científico (ECOSC) um subconjunto de ECOS, caracterizado pelas relações entre fornecedores de software científico, institutos de pesquisa, órgãos de fomento, instituições financiadoras e as partes interessadas em fornecer e/ou reutilizar os resultados da pesquisa, apoiado por uma infraestrutura tecnológica (BOSCH, 2009; FREITAS *et al.*, 2015).

Tendo em vista a complexidade do gerenciamento de experimentos científicos, e a necessidade dos pesquisadores de uma ferramenta de apoio a essas atividades, FREITAS *et al.* (2015) especificaram a plataforma E-SECO (*E-Science ECOsystem*). Esta plataforma é baseada na abordagem de ECOSC, e tem como objetivo principal apoiar os pesquisadores durante todas as etapas do ciclo de vida de um experimento científico.

1.2 MOTIVAÇÃO

Frequentemente, os dados utilizados e gerados pelos experimentos científicos necessitam ser reutilizados, bem como partes dos experimentos. Existem vários motivos para o compartilhamento de dados científicos, dentre eles podemos citar: (i) do ponto de vista ético, os resultados da pesquisa financiada, publicamente, devem ser disponibilizados; (ii) a ciência aberta deve tornar a pesquisa mais transparente ao permitir a reprodutibilidade dos resultados da pesquisa; e (iii) a ciência aberta deve promover uma ciência mais colaborativa e eficiente, maximizando assim os benefícios sociais e econômicos da pesquisa (MICHEL, 2017).

Considerando o cenário de experimentação científica atual e o crescente uso de aplicações em larga escala, o gerenciamento de dados de experimentação está se tornando cada vez mais complexo. Metadados que descrevem os produtos de dados utilizados e gerados por essas aplicações são essenciais para desambiguar os dados e permitir sua reprodutibilidade e reutilização. Adicionalmente, estes metadados auxiliam na interpretação correta dos dados por diferentes pesquisadores e grupos de pesquisa. Assim, o gerenciamento de informações sobre o contexto e a proveniência dos dados dos experimentos científicos desempenham um papel fundamental na *e-Science*.

Informações de proveniência descrevem a origem, a derivação, a propriedade, e a história dos dados (LIM *et al.*, 2010). As informações sobre proveniência proporcionam a verificação da precisão e atualidade dos dados, auxiliam na compreensão dos dados e aumentam sua confiabilidade. Consequentemente, aumentam as chances de reúso dos dados (MISSIER, 2016). Por essa razão, o gerenciamento de proveniência tem sido considerado um ponto chave na arquitetura de SGWfCs e amplamente reconhecido na comunidade científica (LIM *et al.*, 2010).

Contexto é uma descrição complexa do conhecimento compartilhado sobre circunstâncias físicas, sociais, históricas ou outras circunstâncias em que uma ação ou um evento ocorre (RITTENBRUCH, 2002). O gerenciamento de informações de contexto é uma tarefa fundamental na realização de atividades colaborativas. O resultado do trabalho individual precisa ser conhecido pelos participantes do grupo, caso contrário, não haverá um real trabalho em grupo, mas um conjunto incoerente de atividades isoladas. Assim, para entender completamente muitas atividades ou eventos, é necessário ter acesso a informações contextuais relevantes (BRÉZILLON *et al.*, 2004).

O cenário de experimentação científica, colaborativa e distribuída, também requer que aspectos sociais e organizacionais possam ser considerados, visto que o conhecimento sobre a forma como os experimentos são realizados pode ser tácito e, muitas vezes, permanece com o pesquisador. Assim, armazenar e recuperar informações contextuais durante o processo de experimentação pode ser crítico se suas atividades são realizadas para serem reproduzidas ou reutilizadas (MAYER *et al.*, 2014).

No domínio da experimentação científica, pode-se considerar informações de proveniência como um tipo de elemento contextual que descreve informações no passado. Assim, são primordiais para que os pesquisadores possam compreender, reproduzir, examinar e auditar os resultados obtidos pelo experimento, bem como reutilizar o experimento ou partes dele. Além disso, também são importantes neste contexto o uso de mecanismos para descoberta de informações implícitas, bem como mecanismos para facilitar a visualização destas informações.

Embora várias abordagens tenham sido propostas para capturar e modelar a proveniência (CUEVAS-VICENTTÍN *et al.*, 2014; DE OLIVEIRA *et al.*, 2010; GROTH *et al.*, 2005) dos experimentos científicos, a associação explícita dessas abordagens ao gerenciamento de informações contextuais, e à abordagem da proveniência como um tipo de informação contextual na experimentação científica é um tópico que necessita ser explorado. Além disso, não foram identificadas na literatura soluções que considerem as informações contextuais e de proveniência para promoverem o reúso em um ECOSC.

1.3 OBJETIVO

Este trabalho tem por objetivo apoiar o reúso de conhecimento em uma plataforma de ECOSC, através de uma abordagem para o gerenciamento de informações de contexto e de proveniência dos experimentos científicos. Esta abordagem denominada *ContextProv*, visa auxiliar os pesquisadores no entendimento dos experimentos já realizados, e assim, favorecer o reúso de

conhecimento neste domínio. Considerando conhecimento quaisquer artefatos gerados durante o processo de experimentação, o que abrange desde *workflows*, tarefas, serviços, dados e resultados obtidos, até o experimento com um todo. Além disso, este trabalho tem como objetivo secundário a descoberta de informações de contexto e de proveniência implícitas, através do uso de ontologias e mecanismos de inferência.

Para alcançar estes objetivos, foram considerados os seguintes objetivos específicos:

- Identificar quais são as principais fases do ciclo de vida de informações contextuais e de proveniência em plataformas de ECOSC;
- Identificar as informações contextuais e de proveniência relevantes em um ambiente colaborativo e distribuído de experimentação científica;
- Especificar uma ontologia capaz de extrair conhecimento implícito sobre a proveniência e o contexto de experimentos científicos;
- Implementar a solução proposta em uma plataforma de ECOSC;
- Identificar visualizações apropriadas para a apresentação de informações de contexto e de proveniência de experimentos científicos.

1.4 ORGANIZAÇÃO

Este trabalho está dividido em cinco capítulos. O Capítulo 2 apresenta os conceitos envolvidos no presente trabalho, bem como os trabalhos relacionados. O Capítulo 3 apresenta a solução proposta para apoiar o reúso de experimentos científicos em plataformas de ECOSC, detalhando os aspectos conceituais e a implementação da solução. O Capítulo 4 apresenta a avaliação da solução, destacando seu planejamento, execução e resultados obtidos. O Capítulo 5 apresenta as considerações finais, destacando as contribuições do trabalho, suas limitações e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são descritos os principais conceitos relacionados à proposta deste trabalho. Considerações sobre e-Science, proveniência de dados, ontologias, contexto, integração de dados e ecossistemas de software são apresentadas a fim de embasar a abordagem proposta. Também são apresentados a plataforma E-SECO e os trabalhos relacionados à solução proposta neste trabalho.

2.1 E-SCIENCE

Atualmente, a computação deixou de ser apenas uma ferramenta de apoio para se tornar parte fundamental da ciência e de seus métodos científicos. A sinergia entre ciência da computação e as outras áreas do conhecimento criou um novo modo de se fazer ciência – a *e-Science* (ou e-ciência) – que unifica teoria, experimentos e simulação, ao mesmo tempo em que lida com uma grande quantidade de informação (HEY, 2009).

Uma das principais atividades da *e-Science* é a criação e a utilização de processos para a concepção de experimentos científicos, análise de dados e descoberta de conhecimento. Neste contexto, os maiores desafios da ciência atualmente é conseguir simular fenômenos complexos através de ferramentas computacionais, e também gerenciar e analisar a grande quantidade de dados gerados neste processo. Assim, os *workflows* científicos se tornaram uma ferramenta importante para que os cientistas possam automatizar estes processos. Além disso, auxiliam na publicação, no compartilhamento e no reúso destes processos ou dos resultados obtidos por eles (DEELMAN *et al.*, 2009).

Um experimento científico é uma série de operações de análise ligadas entre si, as quais são modeladas e executadas através de *workflows* científicos (GOBLE *et al.*, 2010). Um *workflow* científico é um modelo ou *template* composto por serviços, *scripts* ou outros *workflows*, que representa uma sequência de atividades científicas implementadas por ferramentas, afim de alcançar um determinado objetivo. Para apoiar estes experimentos, permitindo que os pesquisadores pudessem focar em suas pesquisas e não no gerenciamento computacional, surgiram os Sistemas Gerenciadores de *Workflows* Científicos (SGWfCs). Os SGWfCs auxiliam a orquestração de vários algoritmos e processamentos computacionais, fazendo-se valer de processamento paralelo e distribuído, bancos de dados, inteligência artificial, dentre outros (BELLOUM *et al.*, 2011).

Acreditar que um *workflow* não sofrerá evolução e mudanças no contexto de um experimento pode ser considerado utópico, pois à medida que novos resultados vão surgindo, a

pesquisa vai tomando novos rumos, sendo necessário replanejamento, modificações ou adaptações (SIRQUEIRA *et al.*, 2016).

Para que seja possível o gerenciamento de experimentos científicos, é importante o controle do ciclo de vida do experimento. Neste sentido, diversos modelos já foram propostos para o ciclo de vida dos experimentos científicos (BELLOUM *et al.*, 2011; DEELMAN *et al.*, 2009; SIRQUEIRA *et al.*, 2016). Em geral, eles envolvem quatro etapas principais, conforme o modelo proposto por BELLOUM *et al.* (2011) ilustrado na Figura 1.

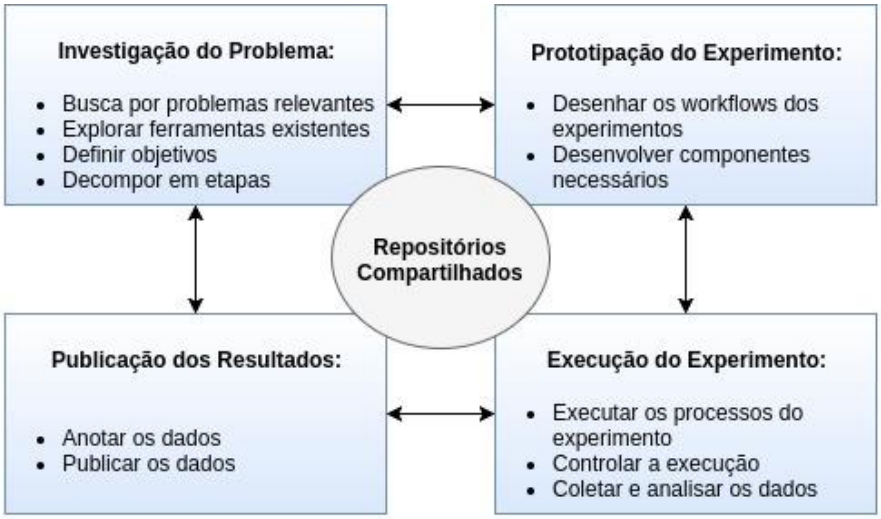

Figura 1. Ciclo de vida de um experimento científico (BELLOUM et al., 2011)

A experimentação se inicia a na fase de investigação do problema, através da qual é definido o escopo da pesquisa. Posteriormente, ocorre a prototipação do experimento, na qual se desenvolve os componentes e *workflows* necessários para o experimento. A etapa principal é a execução do experimento. Nela, o *workflow* é executado de forma controlada e os dados coletados. Por último, ocorre a publicação dos resultados obtidos. Nesse modelo todas as etapas utilizam repositórios compartilhados

2.2 PROVENIÊNCIA DE DADOS

Com o crescente uso de aplicações em larga escala, o gerenciamento de dados está se tornando cada vez mais complexo. Metadados que descrevem os produtos de dados utilizados e gerados por essas aplicações são essenciais para desambiguar os dados e permitir sua reutilização. Além disso, informações de proveniência são cruciais para avaliar se a informação é confiável, como deve ser integrada a outras fontes de informações e como dar crédito a seus criadores ao reutilizá-la. Em um ambiente aberto e inclusivo, como a *web*, onde os usuários encontram

informações muitas vezes contraditórias ou questionáveis, a proveniência pode ajudar esses usuários a avaliar a confiabilidade dessas informações (MISSIER, 2016).

A proveniência de dados, às vezes chamada de linhagem ou *pedigree*, segundo BUNEMAN *et al.* (2001) é a descrição das origens de uma porção de dados e do processo pelo qual chegou em uma base de dados. SIMMHAN *et al.* (2005) definem a proveniência de dados como um tipo de metadados, ou seja, dados sobre dados, que traz o histórico de derivação de um artefato de dados a partir de suas fontes e destacam os principais usos destas informações de proveniência:

- **Qualidade dos dados:** a proveniência pode ser usada para estimar a qualidade e confiabilidade dos dados baseado na fonte dos mesmos e nas transformações sofridas por eles;
- **Rastreabilidade:** a proveniência pode ser usada para traçar uma trilha de auditoria dos dados, determinando os recursos utilizados e os erros ocorridos durante sua derivação;
- **Reprodutibilidade:** informações detalhadas de proveniência permitem a reprodução da derivação dos dados;
- **Atribuição:** a proveniência pode ser usada para estabelecer direitos autorais e propriedade dos dados, permitindo a sua citação e determinando a responsabilidade em caso de dados incorretos;
- **Informação:** uma utilização genérica da proveniência é a consulta com base em metadados de linhagem para a descoberta de dados. A proveniência também pode ser consultada para fornecer um contexto para a interpretação dos dados.

Na experimentação científica os metadados sobre a história de derivação dos dados é essencial para garantir o reúso e a reprodutibilidade dos resultados obtidos através da execução de *workflows* científicos. Adicionalmente, a proveniência proporciona a verificação da precisão e atualidade dos dados. Desta forma, o gerenciamento de proveniência tem sido considerado um ponto chave na arquitetura de SGWfCs e amplamente reconhecido na comunidade científica (LIM *et al.*, 2010). Neste domínio, MISSIER (2016) descreve a proveniência como o resultado da observação da execução de um processo de transformação de dados. Incluindo detalhes de entradas e saídas desse processo e os processos realizados por seres humanos ou apenas parcialmente automatizados.

Especialmente para *workflows* científicos, a proveniência pode ser classificada como prospectiva e retrospectiva. A proveniência prospectiva representa a especificação das

tarefas computacionais que serão executadas. Corresponde aos passos a seguir para alcançar um resultado. A proveniência retrospectiva é dada por atividades executadas e informações sobre o ambiente usado para produzir um resultado, consistindo de um histórico estruturado e detalhado da execução de tarefas computacionais (OLIVEIRA *et al.*, 2018).

Atualmente, os principais modelos para captura de dados de proveniência são o *Open Provenance Model* (OPM) (MOREAU *et al.*, 2011) e o PROV (MOREAU; GROTH, 2013). O modelo OPM é o resultado do esforço da comunidade para alcançar a interoperabilidade dos dados de proveniência. Seus principais objetivos são: (i) permitir que informações de proveniência sejam trocadas entre sistemas, por meio de uma camada de compatibilidade com base em um modelo de proveniência compartilhado; (ii) permitir aos desenvolvedores criar e compartilhar ferramentas que operam em tal modelo proveniência; (iii) definir proveniência de forma precisa; (iv) apoiar uma representação digital de proveniência para qualquer 'coisa', quer produzida por sistemas de computador ou não; (v) permitir a coexistência em vários níveis de descrição; (vi) definir um conjunto de regras que identifiquem as inferências válidas que podem ser feitas na representação de proveniência.

O modelo PROV foi fortemente influenciado pelo OPM, e atualmente é o modelo padrão recomendado pela W3C (*World Wide Web Consortium*). Este modelo permite armazenar dados de proveniência de maneira mais detalhada, focando nas responsabilidades dos agentes nos itens de proveniência. Este fato pôde ser constatado, pois o modelo PROV possui relações específicas para agentes, sem equivalências no OPM, mostrando-se um modelo mais abrangente. Desta forma, o PROV possibilita novas formas de representação do conhecimento inclusive a captura de proveniência centrada em processo, em entidade ou em agente (MOREAU; GROTH, 2013).

O PROV é composto por uma família de doze documentos, mas para utilizá-lo não é preciso estar familiarizado com todos eles. Este modelo foi projetado especificamente para que os usuários e desenvolvedores possam começar com o uso básico, e gradualmente, evoluir para cenários de uso mais avançados. Entre os principais documentos, podem ser citados o PROV-DM, que especifica o modelo de captura de dados, o PROV-CONSTRAINTS, que é um conjunto de restrições aplicáveis ao modelo de dados (PROV-DM) e o PROV-O, uma ontologia para mapeamento do modelo de dados.

Baseado no padrão PROV, MISSIER (2016) propõe o modelo de ciclo de vida de proveniência. Este modelo representa as principais fases do ciclo de vida de dados de proveniência. Conforme ilustrado pela Figura 2, o ciclo se inicia com a captura (*Capture*) dos dados. Em seguida os dados são armazenados (*Store*), permitindo que futuramente sejam

compartilhados (*Sharing*) ou consultados (*Query*). Durante a consulta e o compartilhamento dos dados, é importante que sejam mantidas as associações entre os dados de proveniência e as entidades qual pertencem (*Preserve association to data*). O ciclo termina com a visualização (*Visualize*) e análise (*Analyse*) dos dados. Essas fases tratam os dados brutos que produzirão informações que facilitam a análise desses dados pelo usuário.

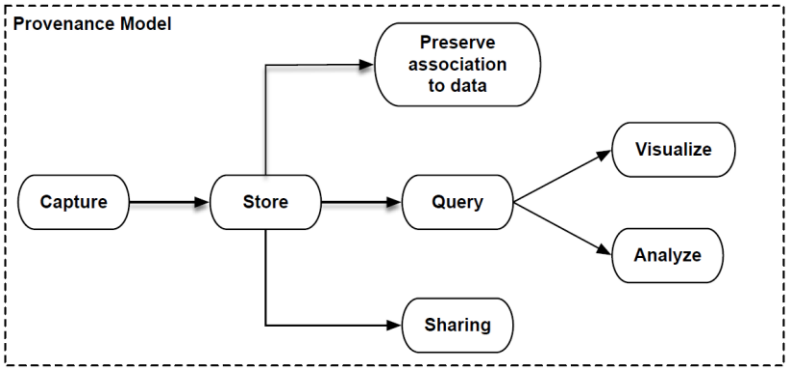

Figura 2. Ciclo de vida de informações de proveniência (MISSIER, 2016)

Considerando este modelo, MISSIER (2016) destaca como principais desafios da área: (i) distinguir em quais eventos a proveniência pode ser observada, e qual o nível de detalhe necessário das informações; (ii) o gerenciamento de grandes volumes de dados; (iii) como gerar conhecimento útil a partir dos dados de proveniência.

SIMMHAN *et al*. (2005) dividem as informações de proveniência em dois grupos: (i) Informações sintáticas, utilizadas em sistemas baseados em anotações, os quais frequentemente adotam a linguagem XML para a representação das informações; (ii) Informações semânticas, utilizadas em sistemas que possuem ontologias de domínio em linguagens como RDF[1] (*Resource Description Framework*) e OWL[2] (*Web Ontology Language).* As ontologias expressam precisamente os conceitos e relacionamentos usados na proveniência e proveem boas informações contextuais. Algumas definições importantes sobre ontologias são descritas a seguir, bem como maiores detalhes sobre a ontologia PROV-O, e outras utilizadas para proveniência em experimentos científicos.

[1] RDF é um modelo padrão para intercâmbio de dados na web.

[2] OWL é uma linguagem da web semântica projetada para representar um conhecimento rico e complexo sobre coisas, grupos de coisas e relações entre as coisas.

2.3 ONTOLOGIA

Para apoiar o compartilhamento e reutilização de conhecimento entre os diferentes sistemas é preciso definir um vocabulário comum de representação deste conhecimento. Neste sentido, GRUBER (1995) tomou o termo ontologia emprestado da filosofia e o definiu para a computação como uma especificação formal e explícita de uma conceituação compartilhada. Essa conceituação é uma visão simplificada e abstrata do mundo que se deseja representar para algum propósito.

Em geral, uma ontologia é uma especificação de um vocabulário de domínio, composto por: definições de classes, relacionamentos e funções. As principais utilizações de uma ontologia são o compartilhamento o entendimento comum sobre a estrutura da informação entre pessoas ou agentes de software; o suporte à reutilização do conhecimento de domínio; a explicitação de suposições sobre o domínio; a separação do conhecimento do domínio do conhecimento operacional e a análise do conhecimento de domínio (GRUBER, 1995).

As ontologias permitem descrever a semântica das classes e propriedades usadas em documentos na *web*. Com isso, se tornou o terceiro componente básico da *Web* Semântica. O tipo mais comum de ontologias para a *web* tem uma taxonomia e um conjunto de regras de inferência. A taxonomia define classes de objetos e as relações de especialização entre eles. As regras de inferência permitem que o conhecimento seja interpretado e inferido logicamente por máquinas (BERNERS-LEE *et al.*, 2001).

No domínio da proveniência de dados as ontologias expressam precisamente os conceitos e relacionamentos usados e proveem boas informações contextuais. Visto isso, o modelo PROV apresentado anteriormente definiu uma ontologia para a modelagem de dados de proveniência chamada PROV-O (MOREAU; GROTH, 2013).

A ontologia PROV-O expressa o modelo de dados do PROV (PROV-DM) usando a OWL. Ela fornece um conjunto de classes, propriedades e restrições que podem ser usados para representar e trocar informações de proveniência gerada em sistemas diferentes e em diferentes contextos. Ela também pode ser especializada para criar novas classes e propriedades para modelar informações de proveniência para diferentes aplicações e domínios.

A Figura 3 representa o modelo inicial de classes e relacionamentos da ontologia PROV-O. Nesse modelo as entidades são representadas por formas ovais em amarelo, as atividades por retângulos azuis, e os agentes são pentágonos em laranja. As setas em preto indicam relacionamentos no passado, e em rosa indicam responsabilidade.

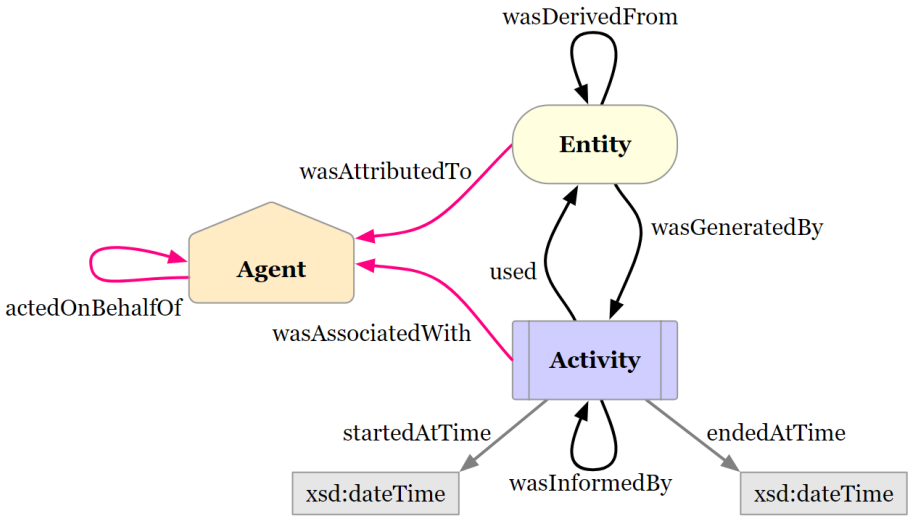

Figura 3. Modelo de classes e relacionamentos do PROV (BELHAJJAME et al., 2013)

Esta ontologia possui em seu modelo inicial três classes, Entidade (*Entity*), Atividade (*Activity*) e Agente (*Agent*). As entidades representam objetos físicos, digitais, conceituais ou outros tipos de objeto com aspectos fixos, por exemplo: uma página da *web*, um gráfico e um documento de texto. As atividades representam algo que atua sobre entidades e ocorre ao longo de um período de tempo, por isso possuem data e hora de início e fim. Representam aspectos dinâmicos do mundo, como ações, processos, etc. Os agentes representam algo que possui algum tipo de responsabilidade por uma atividade, pela existência de uma entidade, ou pela atividade de outro agente. Um agente pode ser uma pessoa, um software, um objeto inanimado, uma organização ou outras entidades que podem ser responsabilizadas. Estas classes se relacionam através de sete propriedades de objeto:

- ***wasGeneratedBy*:** indica que a entidade foi gerada por uma atividade. A entidade não existia antes da execução da atividade e passa a estar disponível para uso após esta geração;
- ***wasDerivedFrom*:** indica que uma entidade foi derivada de outra entidade. Uma derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade baseada em uma entidade pré-existente.
- ***wasAttributedTo*:** indica que a entidade foi atribuída à um agente;
- ***wasInformedBy*:** indica que a atividade é dependente de outra atividade por meio do uso de uma entidade gerada por esta ela;
- ***used*:** indica que a atividade usou uma entidade, e então foi afetada por ela;
- ***wasAssociatedWith*:** indica que a atividade está associada à um agente, ou seja, o agente teve alguma responsabilidade sobre a execução da atividade;

- ***actedOnBehalfOf:*** expressa a responsabilidade de um agente em relação a outro agente. Significa que um agente subordinado atuou em nome de agente responsável em uma atividade.

Apesar de fornecer diversas construções importantes para derivação de conhecimento, no domínio de *e-Science* a ontologia PROV-O não expressa todo o conhecimento necessário para a gerência tanto de experimentos quanto dos *workflows* associados. Uma maneira de trazer esse suporte é através da proposição de regras ontológicas específicas relacionadas a este domínio. Em alguns SGWfCs as informações de proveniência são automaticamente capturadas sob a forma de traços de execução. No entanto, eles muitas vezes utilizam de formatos proprietários dificultando o compartilhamento dessas informações.

A ontologia ProvONE (CUEVAS-VICENTTÍN *et al.*, 2016) é um modelo para a representação de proveniência de *workflow*s científicos criado a fim de preencher estas lacunas. Esta ontologia é uma extensão do padrão PROV, inicialmente denominada D-PROV, com o objetivo de capturar as informações mais relevantes sobre os processos computacionais dos *workflows* científicos e fornecer pontos de extensão para acomodar as especificidades de determinados sistemas de *workflows* científicos (MISSIER *et al.*, 2013).

Criada no contexto da rede DataONE (*Data Observation Network for Earth*), uma rede de dados federados de observações da Terra (CAO, Y *et al.*, 2016), a ontologia ProvONE, adiciona elementos de proveniência (entidades e tipos de relacionamentos) para descrever a estrutura do processo juntamente com as dependências de dados que se originam da execução de um processo. Essa ontologia cobre tanto a proveniência prospectiva quanto a retrospectiva. Além disso, conforme pode-se observar na Figura 4, ela provê informações sobre os aspectos dos *workflows* (classes em azul), dos processos de execução (classes em laranja) e dos artefatos de dados (classes em lilás).

Na representação da proveniência do *workflow*, as várias tarefas são representadas pela classe Programa (*Program)*. Os programas podem ser atômicos ou compostos. Os programas compostos são especificados através da associação *hasSubProgram*. Um determinado programa pode ser distinguido como um *Workflow*. Cada programa pode ter uma série de portas (*Port*) de entrada ou saída. As portas dos vários programas são conectadas por canais (*Channel*). Uma classe controladora (*Controller*) pode ser usada para especificar que a execução de um determinado programa é controlada por outro programa.

Cada instância de execução representa a execução de um programa específico (seu plano), que pode ser um *workflow*, e também pode estar associado a um usuário (*User*) responsável pela execução. Para a execução de um programa, uma série de entidades (*Entity*)

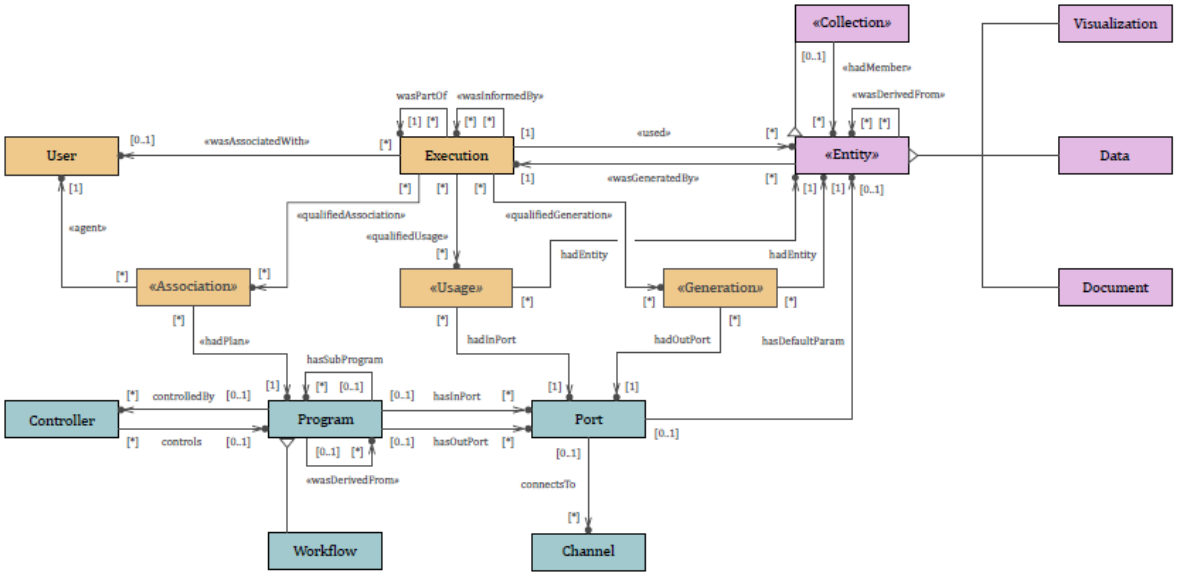

Figura 4. Modelo conceitual da ProvONE (CUEVAS-VICENTTÍN et al., 2016)

de entrada são lidas a partir das portas de entrada, e são usadas para gerar uma série de entidades de saída enviadas por meio das portas de saída. Essas saídas podem ser itens de Dados (*Data*), Visualização (*Visualization*) ou Documento (*Document*), dependendo dos objetivos do *workflow*. Através do uso das classes *Usage* e *Generation*, são registrados os eventos de envio de entidades de uma porta de saída para uma porta de entrada. Essas classes são relacionadas pelas propriedades *hadEntity*, *hadInPort* e *hadOutPort*.

Na representação da estrutura de dados as entidades associadas a instâncias de um *workflow* representadas pelas classes *Data*, *Visualization* e *Document*. A classe *Data* é definida para ser genérica e representa itens de dados de vários tipos, por exemplo arquivos XML, JSON, CSV. As visualizações são uma classe diferenciada destinada a representar vários itens de visualização, por exemplo arquivos JPG, PNG, SVG, MP4, geralmente gerados a partir de *workflows*. A classe *Document* é uma representação genérica de um artigo ou relatório publicado ou não publicado que foi criado como resultado de uma determinada execução de um programa ou *workflow*. Coleções de entidades são representadas por meio da classe *Collection*. Uma coleção pode, por sua vez, representar um conjunto, bolsa, lista ou outra variante de um grupo de itens.

A representação da evolução do *workflow* e das alterações específicas que são executadas durante a especificação do mesmo não são modeladas diretamente no ProvONE, pois espera-se que elas variem entre diferentes SGWfCs. No entanto, as diferentes versões de um *workflow* podem ser representadas usando a associação *wasDerivedFrom* do PROV.

2.4 CONTEXTO

O contexto desempenha um papel fundamental em atividades que envolvem raciocínios como compreensão, interpretação e diagnóstico. Essas atividades dependem de um histórico ou experiência que geralmente não é explícito, mas dá uma dimensão contextual ao conhecimento e à atividade. Assim, o contexto é sempre relativo a algo como, por exemplo, o contexto de uma ação ou de um objeto (BRÉZILLON, 2005).

Contexto é um conceito amplo e aplicável em muitas áreas, por isso existem muitas definições diferentes, relativas à área de conhecimento a qual pertence. BAZIRE e BRÉZILLON (2005) analisaram cerca de 150 definições de contexto, dos mais diversos domínios, e observaram que contexto pode ser descrito de forma geral como o conjunto de circunstâncias que envolvem um evento ou um objeto.

RITTENBRUCH (2002) define contexto como "uma descrição complexa do conhecimento compartilhado sobre circunstâncias físicas, sociais, históricas ou outras circunstâncias em que uma ação ou um evento ocorre". Dessa forma, para entender completamente muitas ações ou eventos, é necessário ter acesso a informações contextuais relevantes.

DEY *et al.* (2001) apresentam uma definição mais específica para o domínio de aplicações sensíveis ao contexto: "qualquer informação que possa ser usada para caracterizar a situação de entidades (ou seja, uma pessoa, local ou objeto) consideradas relevantes para a interação entre um usuário e um aplicativo, incluindo o usuário e o aplicativo".

Contexto em um processo de trabalho tem uma natureza dinâmica, no qual novos eventos surgem e novas decisões são tomadas. Assim, uma organização que não associe informação de contexto às atividades que realiza, e artefatos que gera, possui em sua memória organizacional um imenso conjunto de documentos com pouca ou nenhuma conexão entre eles. Como essa memória não possui um contexto associado, ela é frequentemente ignorada como um recurso de informação (NUNES *et al.*, 2007).

Segundo BRÉZILLON *et al.* (2004), trabalhar em um grupo supõe gerenciar o contexto explicitamente. Não apenas os contextos individuais precisam ser processados, mas também o contexto do grupo, que envolve todo o conhecimento relacionado ao grupo, incluindo composição do grupo, regras, papéis, objetivos, estratégias, procedimentos de coordenação, entre outros. Portanto, o contexto de grupo não é simplesmente a união ou intersecção de contextos individuais. Considerando que a representação explícita do contexto traz benefícios para apoiar a interação entre os membros do grupo, BRÉZILLON *et al.* (2004) propuseram um

framework representando os mecanismos de *groupware* associados a uma representação explícita do contexto, ilustrado na Figura 5. Este *framework* é livre de domínio e fornece diretrizes para o tratamento adequado de informações de contexto e *awareness* ao desenvolver sistemas de *groupware*.

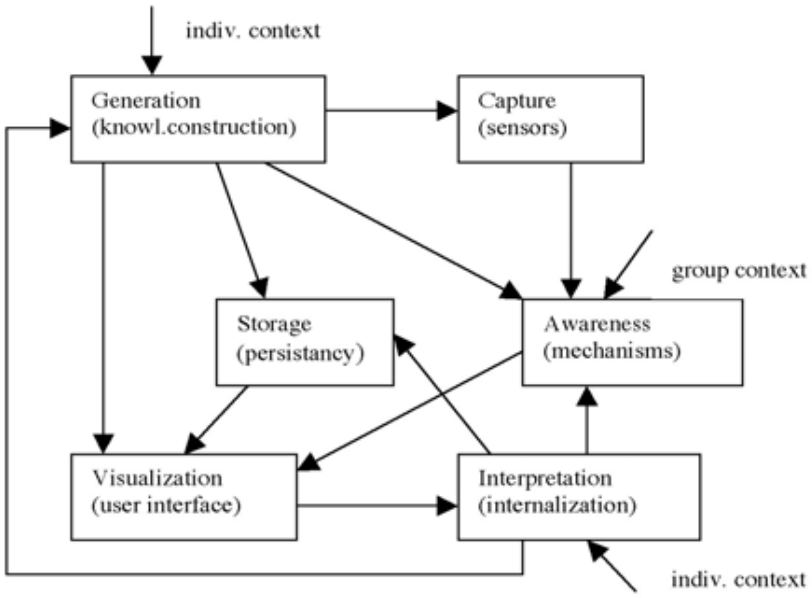

Figura 5. Contexto para processamento de conhecimento em *group work* (BRÉZILLON *et al.*, 2004)

De acordo com o *framework*, as informações individuais são geradas (*Generation*) pelas pessoas. Essas informações são transmitidas ao restante do grupo, apresentadas em uma interface (*Visualization).* A visualização fornece aos usuários a representação física das informações. Além disso, as informações geradas também podem ser armazenadas (*Storage)* de acordo com condições pré-estabelecidas. A etapa de captura (*Capture*) consiste em procedimentos que, através de sensores, reúnem dados físicos da etapa de geração. A etapa de *Awareness* consiste no processamento de informações para fornecê-las aos outros participantes. É importante ressaltar que estas informações são provenientes de várias entradas, e que precisam ser transformadas de alguma forma, talvez resumida ou filtrada, para torná-la disponível para outras pessoas. É nesta fase que o contexto do grupo é adicionado, auxiliando no processamento das demais entradas.

A interpretação (*Interpretation*), assim como a geração, é uma tarefa de usuário. Esta tarefa é realizada considerando as informações visualizadas e o contexto individual do usuário. Desta forma, nesta etapa, as informações são assimiladas em forma de conhecimento. Essas fases não ocorrem necessariamente nessa ordem e, juntas, formam um ciclo de transformação de informações de contexto em conhecimento.

Também com o propósito de fornecer diretrizes para pesquisa e desenvolvimento na área de *groupware* e contexto, ROSA *et al.* (2003) apresentam um *framework* conceitual que identifica e classifica os elementos contextuais mais comuns em *groupware*. Esse *framework* agrupa as informações contextuais em cinco categorias principais: (i) informações sobre pessoas e grupos, (ii) informações sobre tarefas agendadas, (iii) informações sobre a relação entre pessoas e tarefas, (iv) informações sobre o ambiente onde a interação ocorre (v) informações sobre tarefas e atividades já concluídas. Em ambientes síncronos de *groupware*, os membros do grupo precisam trabalhar simultaneamente, mas em ambientes assíncronos, pode haver um intervalo de tempo entre as interações. As necessidades de cada tipo de ambiente são diferentes, por isso este *framework* analisa estas situações de forma diferente.

Este *framework* é um ponto de partida para uma classificação mais específica de elementos contextuais em domínios particulares, onde novos elementos contextuais podem ser considerados relevantes.

2.5 INTEGRAÇÃO DE DADOS

A integração de dados pode ser definida como uma subárea de Banco de Dados que estuda o problema de combinar dados que residem em diferentes fontes, buscando apresentar ao usuário uma visualização unificada desses dados (LENZERINI, 2002). DOAN *et al*. (2012) consideram que a integração de dados busca oferecer acesso uniforme a um conjunto de fontes de dados autônomos e heterogêneos. Desta forma estes sistemas se baseiam em:

- **Consulta:** o foco da maioria dos sistemas de integração de dados é sobre a consulta de dados de diferentes fontes. Embora em alguns casos o interesse pode ser em atualizar dados;
- **Múltiplas fontes:** a integração de dados já é um desafio para um pequeno número de fontes, mas os desafios são exacerbados quando o número de fontes cresce;
- **Heterogeneidade:** um cenário típico de integração de dados envolve fontes de dados que foram desenvolvidas independentemente uma da outra. As fontes terão diferentes esquemas e referências a objetos. Algumas fontes podem estar completamente estruturadas, enquanto outras podem ser desestruturadas ou semiestruturadas;
- **Autonomia:** as fontes não pertencem necessariamente a uma única entidade administrativa, e mesmo quando o fazem, eles podem ser administrados por diferentes sub-organizações. Portanto, não podemos assumir que temos acesso total aos dados em uma fonte ou que podemos acessar os dados sempre que quisermos. Além disso,

as fontes podem alterar seus formatos de dados e padrões de acesso a qualquer momento.

2.6 ECOSSISTEMAS DE SOFTWARE

Os softwares atualmente estão passando por uma forte tendência de customização em massa. Além disso, têm se tornado cada vez mais complexos. A quantidade de funcionalidades necessárias para satisfazer às exigências dos clientes ultrapassa a capacidade de desenvolvimento das empresas. Assim, uma empresa sozinha não consegue desenvolver todas as funcionalidades necessárias em tempo aceitável para os clientes, e com investimento que ainda seja lucrativo (BOSCH, 2009). Desta forma, o software deixou de ser um produto atômico desenvolvido por uma única organização, para ser um conjunto de componentes desenvolvidos colaborativamente por várias empresas. Esta mudança de paradigma deu origem aos Ecossistemas de Software (ECOS).

Os ECOSs estão se tornando cada vez mais comuns no mercado. Através desta forma de desenvolvimento estão surgindo produtos inovadores, os quais reúnem diversos componentes de software que juntos possuem alto valor de mercado. Os maiores exemplos de ecossistemas são as plataformas de software como os sistemas operacionais da Microsoft, da Apple e do Google. Outro exemplo, no ramo dos negócios, é a plataforma de rede social do Facebook (JANSEN; BRINKKEMPER; CUSUMANO, 2013). Recentemente, observa-se ainda um crescimento do número de aplicativos para dispositivos móveis, através dos quais desenvolvedores externos podem contribuir com o desenvolvimento de aplicativos para ECOS como Android, IOS e Windows Phone (FONTÃO *et al.*, 2016).

Segundo BOSCH, (2009) os ecossistemas comerciais e sociais também possuem grande similaridade com os ECOSs. Por isso, contribuíram para a sua definição de ECOS como “um conjunto de soluções que possibilitam, suportam e automatizam as atividades e transações realizados pelos atores em um dado ecossistema social ou de negócio juntamente com as organizações que disponibilizam essas soluções”.

Desde então, a literatura apresenta diversas definições para ECOS. MANIKAS e HANSEN (2013) apresentam diversas definições de ECOS, e identificaram três elementos utilizados frequentemente: software comum, negócios e relacionamentos de conexões. Através da combinação destes termos, apresentam um ECOS em termos da interação e colaboração entre atores que utilizam uma plataforma tecnológica comum, com uma série de soluções de software ou serviços.

A plataforma E-SECO utiliza a abordagem de ECOS voltada para o domínio de *e-Science.* Nela, desenvolvedores internos e externos atuam diretamente no desenvolvimento da plataforma, criando novos recursos para satisfazer às necessidades da comunidade científica (FREITAS *et al.*, 2015). Desta forma, a definição apresentada por BOSCH e BOSCH-SIJTSEMA (2010) é considerada a mais aderente à proposta deste trabalho: "um ecossistema de software consiste em uma plataforma de software, um conjunto de desenvolvedores internos e externos e uma comunidade de especialistas de domínio à serviço de uma comunidade de usuários que compõem elementos de soluções relevantes para satisfazer suas necessidades".

De acordo com essa definição, FREITAS *et al.*, (2015) apresentaram um Ecossistema de Software Científico (ECOSC) como "um subconjunto de ECOS, caracterizado por suas relações com fornecedores de software científico, institutos de pesquisa, órgãos de fomento, instituições financiadoras e partes interessadas nos resultados de pesquisa".

2.7 A PLATAFORMA E-SECO

O processo de experimentação científica envolve interações entre pesquisadores e instituições geograficamente distribuídos. Além disso, demanda a manipulação de grandes volumes de dados, serviços e recursos computacionais distribuídos. Este cenário categoriza um ecossistema de experimentação científica (FREITAS *et al.*, 2015).

Neste contexto, foi proposta uma plataforma de apoio ao processo de experimentação científica denominada E-SECO (*E-Science ECOsystem*) (FREITAS *et al.*, 2015). Essa plataforma é definida por um ambiente colaborativo e distribuído de experimentação científica. Relações entre fornecedores de software científico, institutos de pesquisa, pesquisadores, órgãos de fomento, instituições financiadoras, e as partes interessadas nos resultados de pesquisa caracterizam a plataforma como um ECOSC. A plataforma E-SECO possui código aberto[3], e é acessível através da *web*[4] ou de uma rede de computadores, dependendo da forma como for instalada e configurada.

O processo de experimentação implementado pela plataforma E-SECO foi baseado no ciclo de vida de experimentos científicos definido por BELLOUM *et al.* (2011), apresentado anteriormente. Este ciclo foi estendido por FREITAS *et al.* (2015) e SIRQUEIRA *et al.* (2016) acrescentando atividades de Gerência de Configuração e Gerência de Proveniência aos repositórios compartilhados definidos no modelo original, conforme a Figura 6. Desta forma,

[3] Código disponível no GitHub pelo link: https://github.com/pgcc/e-seco

[4] Uma instância demonstrativa da plataforma pode ser acessada pelo link: http://nenc.ufjf.br:8080/eseco_public

esse novo ciclo de vida enfatiza a necessidade de controle e gerência das diversas versões dos *workflows*, bem como a captura dos processos e dados para posterior análise.

Além disso, propõe: a realização de revisões sistemáticas de literatura durante a etapa de investigação do problema; a utilização do conceito de uma Linha de Produto de Software Científico (LPSC) e a integração com repositórios externos de serviços e *workflows* na etapa de prototipação do experimento; uma plataforma para transformação de *workflows* científicos para permitir a integração entre diversos SGWfCs durante a execução do experimento.

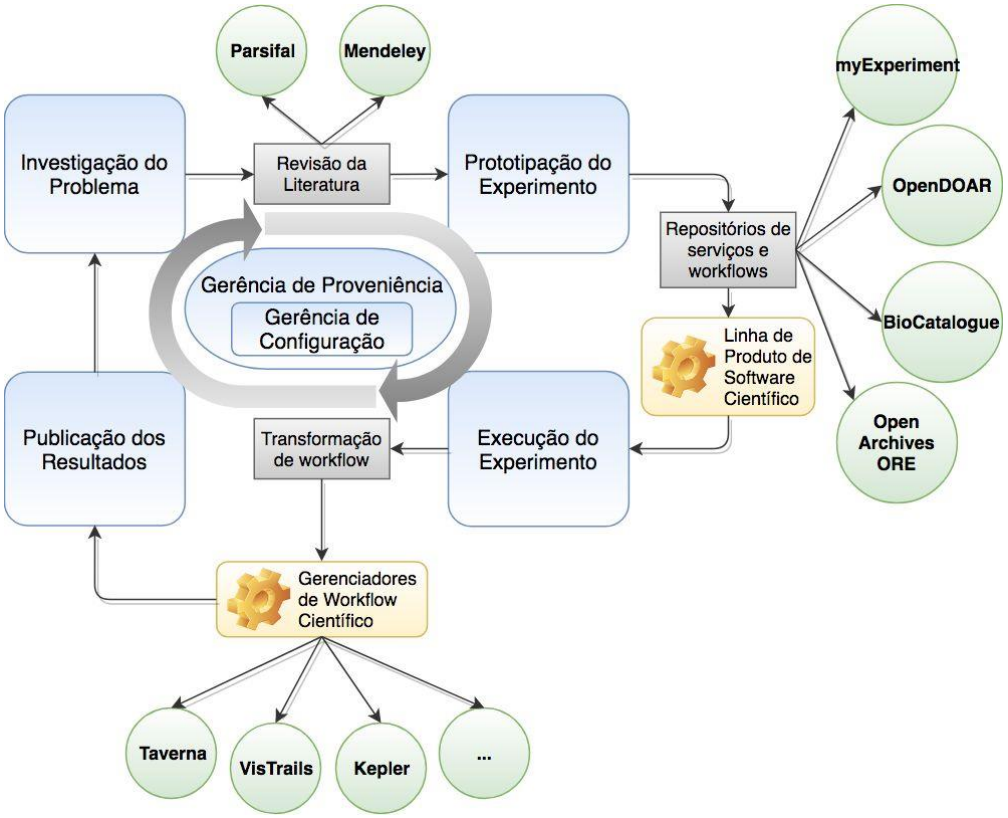

Figura 6. Ciclo de vida de um experimento científico no E-SECO *ProVersion* (SIRQUEIRA et al., 2016)

A Figura 7 apresenta uma visão geral dessa plataforma, onde o núcleo é composto pelo módulo ***Collaborative PL-Science*** (PEREIRA *et al.*, 2016), responsável por todas atividades associadas a uma LPSC. A LPSC é composta por artefatos e serviços de colaboração, e pelos processos de experimentação envolvidos no projeto de *workflows* científicos. O módulo ***PL-Science Product Line*** implementa uma **Rede Ponto a Ponto** (P2P) fazendo com que cada instância da plataforma E-SECO seja um nó nesta rede. Através dessa rede P2P os artefatos são compartilhados com outras aplicações. Esta plataforma inclui também uma **Camada de Interoperabilidade** chamada **PRIME (PRagmatic Interoperability to MEaningful**

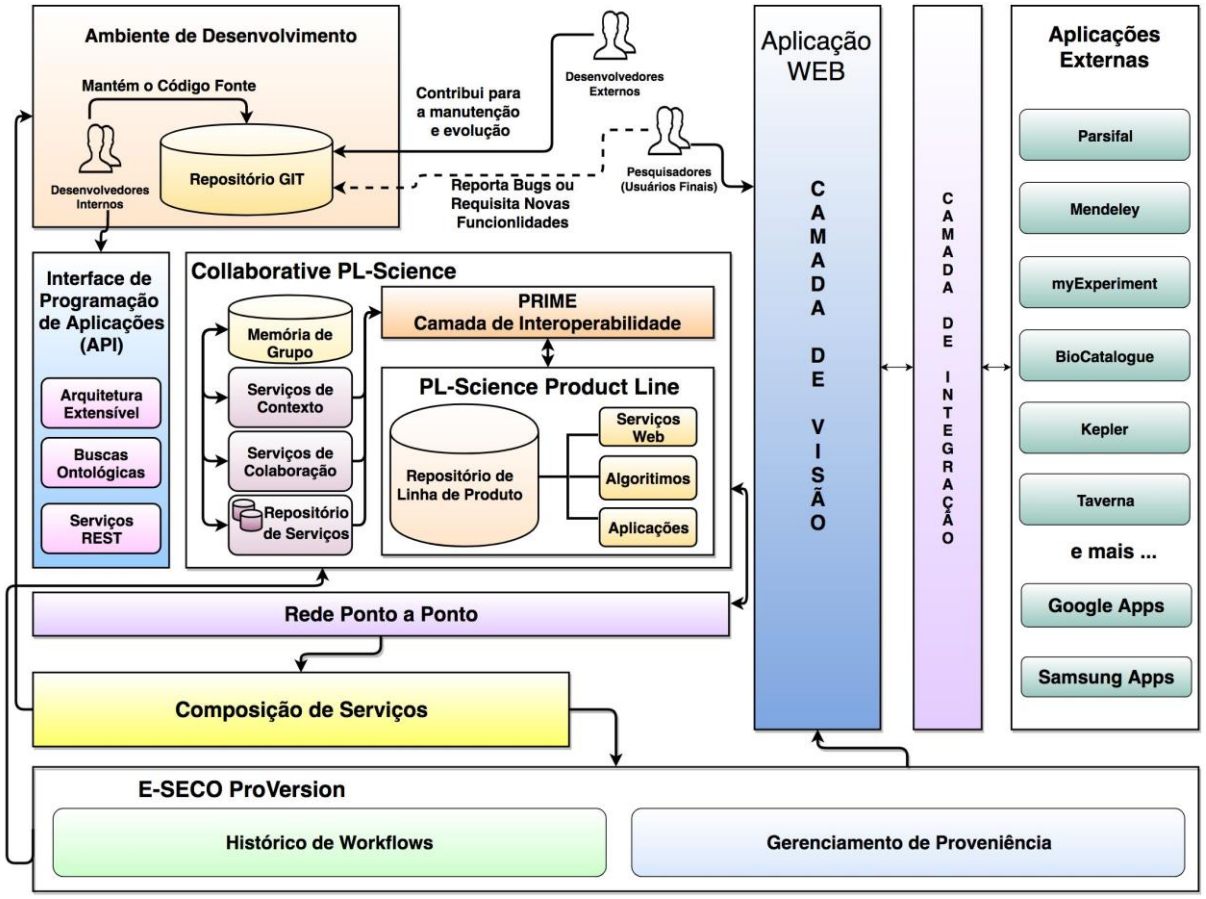

Figura 7. Visão Geral da Plataforma E-SECO

collaboration) (NEIVA *et al.*, 2015). Esta camada oferece suporte às atividades de colaboração nos diferentes níveis: sintático, semântico e pragmático.

Os desenvolvedores externos podem apoiar o desenvolvimento de artefatos para o ecossistema, e propor melhorias através do **Ambiente de Desenvolvimento** gerenciado pela plataforma GitHub. Essas atividades são avaliadas pela equipe de desenvolvimento interna, a qual é responsável pela manutenção e evolução da plataforma. O **Módulo de Composição de Serviços** oferece recursos para a composição de serviços internos e externos na plataforma (MARQUES *et al.*, 2017). Os usuários finais interagem com a plataforma utilizando a **Camada de Visão** através da qual conduzem experimentos e desenvolvem artefatos para a LPSC.

Atualmente, a plataforma E-SECO está integrada com as seguintes **Aplicações Externas**: Parsifal[5], Mendeley[6] (ZAUGG *et al.*, 2011), myExperiment[7], BioCatalogue[8],

[5] https://parsif.al

[6] https://www.mendeley.com

[7] https://www.myexperiment.org

[8] https://www.biocatalogue.org

Kepler[9] (ALTINTAS *et al.*, 2004) e Taverna[10] (OINN *et al.*, 2007). A integração com estas aplicações é feita através da **Camada de Integração**. Além disso, uma **Interface de Programação de Aplicações** (API) permite tratar aspectos relevantes para o sucesso das integrações com a plataforma, tais como: (i) a construção de uma documentação contendo todos os recursos disponibilizados por ela, (ii) o suporte a flexibilidade da plataforma através dos tipos de dados apoiados nos formatos XML e JSON, (iii) o suporte à consistência na implementação de métodos HTTP, (iv) o desenvolvimento de um ambiente para apoiar os testes de integração, e (v) a distribuição de clientes de integração para apoiar desenvolvedores externos.

O gerenciamento de configuração dos experimentos é feito no módulo **E-SECO *ProVersion*** (SIRQUEIRA *et al.*, 2016). No item de **Histórico de *Workflows*** os dados relacionados a tarefas, valores de entradas, saída e informações de experimentos são associados aos respectivos *workflows* e analisados para permitir a identificação de *workflows* semelhantes, características de evolução e/ou necessidades de manutenção. Utiliza como base os dados capturados no item de **Gerenciamento de Proveniência** do **E-SECO *ProVersion***. Embora a plataforma já possua um módulo de proveniência, este módulo não abrange todo o ciclo de vida dos dados de proveniência, não inclui o gerenciamento de contexto e não está focado no apoio ao reúso dos experimentos científicos.

2.8 TRABALHOS RELACIONADOS

O suporte aos experimentos científicos através da proveniência de dados e contexto têm sido explorados em alguns trabalhos na literatura. Diante disso, um mapeamento sistemático da literatura foi conduzido a fim de fundamentar o presente trabalho.

2.8.1. Mapeamento sistemático da literatura

O mapeamento sistemático é uma forma de revisão da literatura, com objetivo fornecer uma visão geral, identificando e categorizando os estudos disponíveis sobre o tema. Estes estudos são projetados para reduzir o viés e proporcionar uma imagem do estado da arte de um campo específico de investigação. Através de um protocolo pré-definido, este tipo de estudo segue

[9] https://kepler-project.org

[10] http://www.taverna.org.uk

passos metodológicos precisos e rigorosos para selecionar e analisar documentos relevantes (KITCHENHAM *et al.*, 2011).

O presente trabalho segue este processo, apoiado pela ferramenta Parsifal[11]. O Parsifal é uma ferramenta online concebida para apoiar pesquisadores na realização de revisões sistemáticas da literatura no contexto da Engenharia de Software.

2.8.1.1. Planejamento

O planejamento consiste basicamente na especificação das questões de pesquisa e no desenvolvimento do protocolo. As questões de pesquisa definem o foco para a identificação dos estudos preliminares, a extração de dados dos estudos e a análise dos dados. Já o protocolo especifica o método a ser utilizado no mapeamento sistemático, de forma que seja reduzido o viés do investigador, e que os resultados sejam reprodutíveis.

A fim de obter uma visão geral das publicações na área, o mapeamento sistemático pretende responder às seguintes questões (QM):

Q1. Quantos estudos foram publicados ao longo dos anos?

Q2. Quem são os autores mais ativos na área?

Q3. Quantos estudos foram encontrados sobre cada assunto (manutenção, evolução, proveniência, contexto)?

Q4. Quais são os principais veículos de publicação de pesquisas na área?

Para facilitar a criação da *string* de busca, e permitir que este processo seja reprodutível, as questões de pesquisa foram modeladas na Tabela 1 seguindo o modelo PICOC sugerido no guia prático para revisões sistemáticas de (PETTICREW e ROBERTS, 2008).

Tabela 1. PICOC

(P)	***Population***	**Experimentos científicos**
(I)	***Intervention***	Soluções para proveniência ou contexto
(C)	***Comparison***	Não se aplica
(O)	***Outcomes***	Solução
(C)	***Context***	Ecossistemas ou plataformas de software

Para criar a *string* de busca, foram definidos os principais termos relacionados aos interesses do mapeamento sistemático. A fim de aumentar o alcance da *string*, foram incluídas

[11] https://parsif.al

variações gráficas dos termos e seus sinônimos. Estes termos foram definidos de acordo com cada uma das partes das questões de pesquisa modeladas no PICOC, e estão listados na Tabela 2.

Tabela 2. Palavras Chave

(P)	***Population***	E-Science, eScience, Scientific Experiment, E-science experiment, In Silico Experiment, Scientific Workflow, E-science Workflow, Workflow in E-science
(I)	***Intervention***	Context-aware, Awareness, Context-awareness, Context Element, Context Information, Context Modeling, Contextual Element, Contextual Information, Provenance, Provenance-aware
(C)	***Comparison***	Não se aplica
(O)	***Outcomes***	Approach, Solution, Framework, Method, Process, Ontology, Model, Technique, Tool, Support
(C)	***Context***	Infrastructure, Architecture, Platform, Scientific Application, Scientific Software, Software Ecosystem, Ecosystem Platform, Information Science

De posse destes termos, a *string* foi gerada da seguinte forma: os termos dentro de um mesmo grupo foram separados pelo operador lógico *OR*, e os grupos foram unidos na *string* pelo operador *AND* conforme a fórmula: *(Population) AND (Intervention) AND (Outcomes) AND (Context).* Desta forma, a Tabela 3 apresenta a *string* de busca utilizada.

Tabela 3. *String* de busca

```
("E-Science" OR "eScience" OR "Scientific Experiment" OR "E-science experiment" OR "In Silico Experiment" OR "Scientific Workflow" OR "E-science Workflow" OR "Workflow in E-science") AND ("Context-aware" OR "Awareness" OR "Context-awareness" OR "Context Element" OR "Context Information" OR "Context Modeling" OR "Contextual Element" OR "Contextual Information" OR "Provenance" OR "Provenance-aware") AND ("Infrastructure" OR "Architecture" OR "Platform" OR "Scientific Application" OR "Scientific Software" OR "Software Ecosystem" OR "Ecosystem Platform" OR "Information Science") AND ("Approach" OR "Solution" OR "Framework" OR "Method" OR "Process" OR "Ontology" OR "Model" OR "Technique" OR "Tool" OR "Support")
```

A validação desta *string* foi feita através da recuperação de alguns artigos previamente conhecidos que tratam de assuntos relevantes na área. Os artigos selecionados para controle foram: (i) *Towards an Adaptive and Distributed Architecture for Managing Workflow Provenance Data* (COSTA *et al*, 2014); (ii) *Managing rapidly-evolving scientific workflows* (FREIRE *et al.*, 2006); (iii) *Implicit provenance gathering through configuration management* (NEVES *et al.*, 2013); e (iv) *Ontologies for Describing the Context of Scientific Experiment Processes* (MAYER *et al.*, 2014).

Visando selecionar os documentos de acordo com as questões apresentadas anteriormente, foram definidos critérios de inclusão e de exclusão dos artigos. O último passo na definição do protocolo é a escolha das bases para a busca. Foram consideradas as principais bases eletrônicas que indexam as pesquisas produzidas na área de Ciência da Computação. Foram excluídas apenas aquelas que não permitem buscas com expressões lógicas; ou não permitem busca em partes específicas do texto como título, resumo e palavras-chave; ou não estão disponíveis na instituição. Os critérios de inclusão e exclusão, bem como as bases utilizadas são apresentados no Apêndice A.

2.8.1.2. Condução

A primeira atividade na condução do mapeamento sistemático é a execução da busca nas fontes selecionadas. Para isso, a *string* de busca foi reescrita seguindo a sintaxe de cada uma das bases. Para evitar que fossem recuperados trabalhos que não estão intimamente ligados ao tema de pesquisa, a busca foi limitada ao título, abstract e palavras-chave. Desta forma, garante-se que os termos da *string* encontrados no artigo são realmente relevantes para o mesmo. As *strings* foram executadas nas bases, e foram recuperados mais de 600 trabalhos. Os trabalhos foram importados para o Parsifal, e foram eliminados os trabalhos que estavam duplicados. Desta forma, ao final restaram 358 trabalhos a serem analisados segundo os critérios de seleção definidos.

A próxima atividade executada foi a análise dos artigos com base na leitura de seus títulos e resumos, segundo os critérios de aceitação e exclusão definidos. Desta forma, foram excluídos aqueles que se encaixaram em algum dos critérios de exclusão, considerados então irrelevantes para esta pesquisa. Após este processo, foram aceitos 73 trabalhos, os quais foram analisados através de uma leitura completa. Após a leitura dos trabalhos aceitos, foram extraídas as informações necessárias para responder às questões de mapeamento (QM) definidas. O relatório final do mapeamento sistemático, com as respostas a estas questões, é apresentado no Apêndice A.

Através deste mapeamento sistemático da literatura foram identificados os principais trabalhos existentes, relacionados ao problema e à solução proposta. Entretanto, não foram identificadas soluções que utilizem o gerenciamento de contexto e proveniência de dados, e relacione estes dois conceitos, para apoiar o reúso de experimentos científicos em plataformas de ECOSC. Assim, foi constatada viabilidade de desenvolvimento da proposta desse trabalho. A seguir, são apresentados os trabalhos relacionados à abordagem proposta.

2.8.2. Karma

Karma (CAO *et al.*, 2009 e SIMMHAN *et al.*, 2006) é um *framework* para captura de proveniência de experimentos científicos focados em *workflows*, utiliza um *web service* para captura dos dados e os armazena em um repositório no formato XML. Captura *metadados* de proveniência uniformes e usáveis, de maneira independente do *workflow*. O modelo Karma captura duas formas de proveniência: proveniência de processo, que são *metadados* que descrevem a execução do *workflow* e invocações associadas; e proveniência de dados, que fornece metadados semelhantes sobre a história da derivação de um produto de dados.

A proveniência neste modelo é dada em dois níveis: o nível de registro, que se relaciona à persistência dos metadados de serviços e dados que podem ser utilizados em uma sequência de execução; e o nível de execução, que modela as instâncias do nível de registro e grava as informações relacionadas a invocações de métodos e aos produtos de dados utilizados ou gerados por cada invocação. Este *framework* fornece diretrizes para o armazenamento e consulta dos dados de proveniência em um banco de dados relacional, em conformidade com o modelo OPM.

Este modelo não utiliza ontologias, desta forma, não é capaz de realizar a inferência de conhecimento implícito e fornecer informações importantes para facilitar a reutilização de dados. Além disso, não fornece visualizações adequadas para usuários com pouco conhecimento de proveniência de dados, nem tampouco considera a execução em plataformas de ECOSC.

2.8.3. PReServ

Provenance Recording for Services (PReServ) (GROTH *et al.*, 2005) é um mecanismo de captura de proveniência para experimentos científicos independente do SGWfC. Ele foi desenvolvido para o contexto de bioinformática e todos os dados coletados são armazenados em metadados no formato XML. É uma implementação baseada em *web services* Java, subjacente à arquitetura Pasoa (GROTH *et al.*, 2006). O PReServ captura as interações entre componentes internos e agrupamento de interações por meio do protocolo PReP (*Provenance Recording Protocol*), que especifica as mensagens que os atores podem trocar com o banco de proveniência.

O PReServ não utiliza um modelo de proveniência padrão, o que dificulta a interoperabilidade dos dados, bem como sua consulta e a análise pelo pesquisador. Além disso, eles não fornecem a inferência do conhecimento implícito por meio de ontologias. Abrange

apenas as fases de captura de dados, armazenamento e consulta. Desta forma, sem uma solução de visualização adequada, a interpretação dos dados é dificultada. Também não considera a execução em plataformas de ECOSC.

2.8.4. SciCumulus

SciCumulus (DE OLIVEIRA, D *et al.*, 2010) é um *middleware* para orquestrar *workflows* científicos por meio do SGWfC em ambientes distribuídos e paralelos. Esta abordagem oferece um serviço de captura de proveniência em tempo real. A proveniência é armazenada com granularidade a nível de atividades e em tempo de execução. Assim, é possível monitorar o estado do *workflow* e avaliar os resultados disponíveis durante a execução. Este serviço é baseado em um modelo de proveniência que considera tanto os descritores dos dados relativos ao ambiente de nuvem quanto aos dados relativos à estrutura e execução dos *workflows* (proveniência prospectiva e retrospectiva).

Nesta abordagem, o repositório de proveniência é mantido utilizando o banco de dados relacional, e o acesso a estas informações é feito através de consultas a este banco de dados. Esta abordagem está focada na proveniência dos *workflows*, sendo assim, não gerencia contexto de forma explícita e completa. Além disso, as visualizações implementadas não estão voltadas para o reúso dos experimentos científicos em plataformas de ECOSC, o que pode dificultar a interpretação do usuário durante o reúso.

2.8.5. ProM

ProM (SILVA *et al.*, 2014) é um *framework* que utiliza algoritmos de comparação para mineração tanto de processos imperativos quanto declarativos. O objetivo desta ferramenta é auxiliar o especialista no planejamento de experimentos científicos, através da descoberta de modelos. Para isso, utiliza os dados de proveniência gerados por *workflows* que alcançaram bons resultados no passado. É uma abordagem genérica, projetada para ser aplicada a qualquer SGWfC compatível com o modelo PROV.

As medidas de qualidade são definidas para obter os melhores resultados de cada cenário de execução. Essas métricas são previamente definidas pelo especialista e coletadas durante a execução das instâncias do *workflow*. O coletor de qualidade é uma atividade incluída na especificação do *workflow*. Através dos modelos declarativos gerados pela ferramenta os especialistas de domínio podem visualizar e compreender melhor as semelhanças entre

instâncias bem-sucedidas do *workflow*. Como consequência, os experimentos podem ser reutilizados, compartilhados ou planejados, de modo a obter melhores resultados.

Apesar de possuir um modelo de dados baseado no padrão PROV, esta abordagem não utiliza a ontologia PROV-O para a extração de conhecimento implícito nos dados de proveniência, e também não inclui na ontologia as informações de contexto dos experimentos. Além disso, não oferece visualizações que apoiem o reúso de experimentos científicos em plataformas de ECOSC, e que facilitem a interpretação dos dados por usuários que desconhecem a linguagem própria dos modelos declarativos.

2.8.6. ProvSearch

ProvSearch (COSTA *et al.*, 2014) é uma arquitetura de gerenciamento de dados de proveniência independente de SGWfC, voltada para experimentos em ambientes distribuídos. Combina técnicas de gerenciamento de *workflows* distribuídos com gerenciamento de dados de proveniência. Permite que os dados de proveniência sejam capturados, armazenados e consultados em tempo de execução, sem interferir na execução do *workflow*.

Nesta abordagem, os dados são fragmentados em múltiplos repositórios de proveniência na nuvem e podem ser acessados por diferentes SGWfCs. Sua arquitetura é composta por quatro componentes: (i) Nós de banco de dados, formam uma rede descentralizada de servidores de bancos de proveniência. Cada nó contém um sistema de gerenciamento de banco de dados distribuído instalado com duas bases de dados diferentes. Uma para armazenar todos os dados de proveniência tradicionais (como a hora inicial e final, etc.) e os resultados do experimento; e a outra apenas com as estatísticas (por exemplo, o tempo médio de execução de um programa específico, porcentagem de erros para uma máquina específica, etc.); (ii) Nó de controle, é responsável por identificar qual o nó de banco de dados irá armazenar os dados de proveniência para uma execução específica; (iii) Depósito integrado e global de proveniência, armazena um resumo de todas a bases locais de dados estatísticas, agindo como um repositório de proveniência, ou seja, as estatísticas de todas as execuções de todas as experiências são armazenadas neste depósito de proveniência integrado e pode ser consultado por qualquer usuário sem acessar resultados de experimentos de terceiros; e (iv) API - interface entre o ProvSearch e os SGWfC existentes.

Os dados de proveniência são tratados em um modelo chamado PROV-Wf, uma extensão do modelo PROV para o domínio dos *workflows* científicos (COSTA *et al.*, 2013). No entanto, essa ontologia não considera as informações contextuais e não é capaz de extrair

informações de proveniência implícitas. Esta abordagem também não possui soluções para a visualização que auxiliem o reúso dos experimentos em plataforma de ECOSC.

2.8.7. PBase

PBase (CUEVAS-VICENTTÍN *et al.*, 2014) é um repositório de proveniência de *workflows* científicos que implementa a ontologia ProvONE, permitindo armazenamento, análise e replicação de experimentos científicos. Este repositório, assim como a ontologia ProvONE, é parte do projeto DataONE: uma rede de dados federados de observações da Terra (MICHENER *et al.*, 2016). A arquitetura do PBase é baseada na arquitetura da plataforma JAVA e possui três níveis: (i) Nível de visualização: um cliente *web* que possui uma interface adaptada para a visualização dos dados de proveniência de *workflows* científicos, tornando a especificação de consultas e a interpretação dos seus resultados mais fácil e eficaz; (ii) Nível de Aplicação: implementação dos serviços *web* para atender às consultas feitas pelo usuário; (iii) Nível de dados: conta com um banco de dados gráfico Neo4j oferecendo assim consultas declarativas e eficientes. Possui uma interface de usuário baseada na *web* que permite aos usuários fazer *upload* de um rastreio de proveniência, visualizar o *workflow* ao lado de seus vários rastreios, emitir consultas e obter visualizações de seus resultados. Entretanto, esta abordagem não oferece apoio ao processo de experimentação em plataformas de ECOSC. O PBase também não usa a ontologia para obter informações implícitas, e não inclui os elementos de contexto do experimento.

2.8.8. E-SECO ProVersion

E-SECO ProVersion (SIRQUEIRA et al., 2016) é uma abordagem de suporte à gerência de configuração na plataforma E-SECO. Esta abordagem utiliza uma extensão do modelo PROV, que abrange tanto a ontologia quanto o modelo de dados, aplicado ao domínio de *workflows* científicos. O módulo de proveniência permite ao pesquisador capturar os dados do *workflow* em diferentes SGWfCs por meio de um *web service* incluído como uma atividade no *workflow*. Estes dados alimentam a ontologia PROV-OEXT que, por meio de regras específicas do domínio, detecta informações sobre a evolução e manutenção em *workflows*. As informações são armazenadas no módulo de histórico dos *workflows*, e disponibilizadas ao pesquisador por meio da interface *web* do E-SECO. Esta abordagem permite a extração de conhecimento implícito através do uso da ontologia. Por outro lado, as informações e proveniência capturadas, se restringem à execução do *workflow*. Com isso, esta abordagem não é capaz de realizar o

gerenciamento de proveniência durante todo o ciclo de experimentação. Além disso, não considera os elementos contextuais do experimento, e não oferece componentes de visualização adequados para apoiar o reúso dos experimentos científicos.

2.8.9. Extended Context-based Framework

BRÉZILLON (2011) e FAN *et al.* (2011) apresentam um *framework* baseado em contexto para melhorar a tomada de decisão sobre *workflows* científicos. Esse *framework* permite tornar o contexto explícito e realizar a contextualização dos *workflows* armazenados no repositório de *workflows* científicos. Esta abordagem visa apoiar os pesquisadores na reutilização de *workflows* científicos presentes nesse repositório. O contexto é explicitado através de Grafos Contextuais (CxGs), um formalismo para representar uniformemente todos os componentes de um processo cooperativo de design de *workflows* científicos.

Segundo o autor, os repositórios científicos contêm *workflows* aplicados com sucesso em contextos específicos. Assim, nenhum *workflow* pode ser reutilizado diretamente, pois um novo experimento envolve um novo contexto. Esta abordagem auxilia o pesquisador a encontrar um *workflow* através de um longo processo de contextualização (identificando o *workflow* publicado que possui um contexto próximo ao desejado). Além disso, apoia a descontextualização, extraindo a parte do *workflow* que pode ser reutilizada de uma forma relativamente genérica, e a recontextualização, desenvolvendo instâncias do *workflow* adaptadas a novos contextos.

Esta abordagem não permite a captura das informações de contexto durante todo o processo de experimentação, e não abrange as informações de proveniência. Além disso não considera as características específicas de plataformas de ECOSC. Desta forma, o contexto relacionado aos *workflows* presentes no repositório pode não ser suficiente para permitir o reúso de elementos deste *workflow*. Além disso, a abordagem permite apenas um tipo de visualização na forma de grafos contextuais, dificultando a interpretação dos dados.

2.8.10. TIMBUS

O projeto TIMBUS é focado no gerenciamento de processos de negócios resilientes. Tem por objetivo tornar o contexto de execução, dentro do qual os dados são processados, analisados, transformados e renderizados, acessíveis por longos períodos. No contexto desse projeto, MAYER *et al.* (2014) propuseram um modelo de contexto para a descrição de experimentos científicos focado especificamente na infraestrutura técnica utilizada como base para o

experimento. Este modelo se baseou na preservação digital dos processos, cujo objetivo permitir a redistribuição ou reconstituição de um processo mesmo quando o ambiente técnico mudou. Esta abordagem visa a preservação dos processos, dos princípios arquiteturais e das ontologias de núcleo e de extensão do experimento, permitindo assim o reúso e a reprodutibilidade dos experimentos.

Define um metamodelo para a captura do contexto no qual o processo está incorporado. Esse contexto pode variar desde aspectos imediatos e locais, como software e hardware que suportam o processo, até aspectos como a organização em que o processo é executado, as pessoas envolvidas, provedores de serviços e até leis e regulamentos. O contexto exato pode diferir significativamente dependendo do domínio do processo. Este metamodelo é descrito através de ontologias, entretanto, a abordagem não faz uso da ontologia para extrair conhecimento implícito. Apesar de considerar as características colaborativas e distribuídas dos experimentos, esta abordagem não está inserida no contexto de uma plataforma de ECOSC. Além disso, é focada no contexto do processo, portanto, não considera todo o ciclo de experimentação científica, limitando seu apoio ao reúso apenas do *workflow*.

2.9 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Conforme apresentado, existem várias soluções para gerenciar a proveniência e o contexto de experimentos científicos. No entanto, essas abordagens apresentam algumas limitações. Algumas abordagens não utilizam um modelo de proveniência padrão e interoperável, outras não apoiam a extração de conhecimento implícito através de inferências. Além disso, muitas abordagens não possuem mecanismos adequados de visualização que apoiem o reúso dos experimentos.

Outra limitação é que essas abordagens analisam as informações de proveniência e de contexto de forma isolada e tratam problemas muito particulares com foco apenas no *workflow* e nos resultados obtidos. Algumas abordagens não consideram as etapas de investigação do problema e prototipação do experimento. Assim, são focadas apenas na execução do *workflow* e publicação dos resultados. Dessa forma, são ignoradas informações acerca da derivação e semelhança entre os experimentos, e sobre os demais artefatos que compõe o experimento. Por não considerarem todas as atividades do ciclo de vida proposto para o experimento que acontece em um ambiente colaborativo, estas abordagens não são capazes de apoiar o reúso durante todo o processo de experimentação científica no contexto de ecossistemas de software.

3 CONTEXTPROV: UMA ABORDAGEM PARA O GERENCIAMENTO DE CONTEXTO E PROVENIÊNCIA

Este capítulo apresenta a solução desenvolvida para alcançar os objetivos da dissertação. Primeiramente são apresentados os aspectos conceituais desenvolvidos para o gerenciamento de contexto e proveniência em plataformas de ECOSC. Esses aspectos envolvem: a definição do ciclo de vida de proveniência e contexto nessas plataformas; a identificação e classificação dos elementos de contexto relevantes em um ambiente colaborativo e distribuído de experimentação científica e a definição de uma ontologia para a modelagem dos dados de proveniência. Posteriormente, são apresentados os aspectos de projeto (requisitos e arquitetura) e implementação da solução na plataforma E-SECO.

3.1 CICLO DE VIDA DE PROVENIÊNCIA E CONTEXTO EM ECOSC

Tendo em vista a complexidade que envolve o processo de experimentação científica em uma plataforma de ECOSC e a dificuldade em reutilizar neste processo, o presente trabalho propõe a abordagem *ContextProv*. Uma solução para o gerenciamento de informações de contexto e de proveniência em uma plataforma de ecossistema de software científico.

Primeiramente buscou-se especificar as fases do ciclo de vida de informações de proveniência e contexto em um ambiente distribuído como uma plataforma de ECOSC. Conforme apresentado anteriormente, BRÉZILLON *et al.* (2004) e MISSIER (2016) definiram modelos de ciclo de vida para elementos de contexto e proveniência respectivamente. Estes ciclos de vida visam modelar o processo de transformação dos dados (de contexto de proveniência) em conhecimento útil para os pesquisadores.

A Figura 8 ilustra as principais fases desses *frameworks*. É notável que eles possuem fases semelhantes. No entanto, MISSIER (2016) não aborda problemas de colaboração durante o ciclo de vida da proveniência. Considerando informações de proveniência como um tipo de elemento de contexto, pode-se estabelecer uma correlação entre as fases desses dois *frameworks*, e assim obter um modelo de ciclo de vida mais adequado para a abordagem proposta neste trabalho. Com este propósito, descrevemos cada uma das fases do *framework* de contexto proposto por BRÉZILLON *et al.* (2004) e sua correspondência com o *framework* de proveniência proposto por MISSIER (2016).

Com relação à obtenção dos dados, o ciclo de vida de proveniência possui apenas uma fase denominada *Capture*. Esta fase consiste na observação da execução de um processo de transformação de dados, incluindo processos automatizados, ou não. Já no ciclo de vida dos

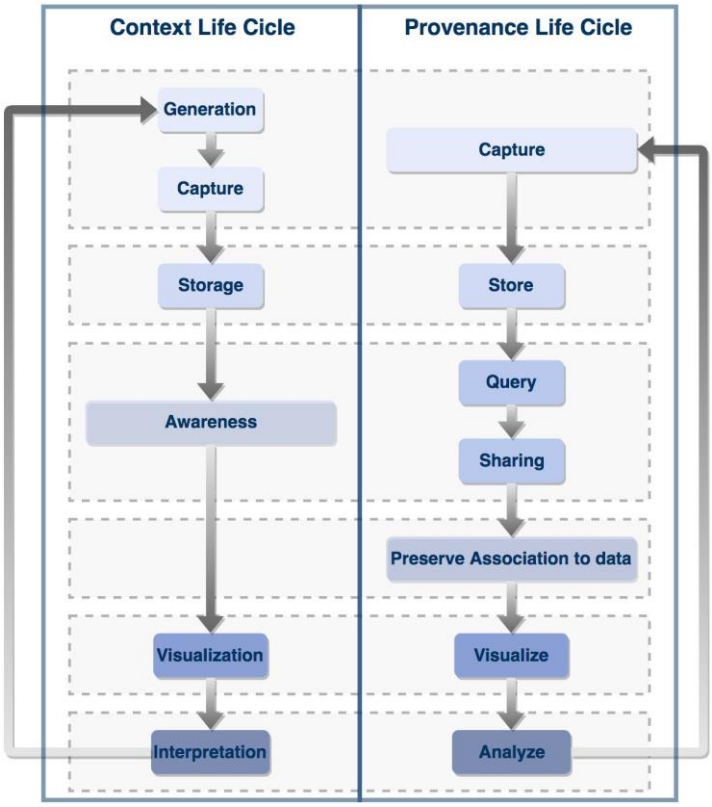

Figura 8. Relacionamento entre as fases do ciclo de vida de contexto e de proveniência.

elementos de contexto, esta fase é dividida em duas. A primeira, denominada *Generation*, ocorre quando um membro contribui com algum conteúdo para o grupo de cientistas. É considerada uma tarefa do usuário, portanto, as informações sobre o contexto individual desse usuário são coletadas neste momento. A segunda fase, denominada *Capture*, consiste em procedimentos para coletar alguns dados físicos relativos à fase anterior. É uma fase de responsabilidade do sistema realizada através de sensores. Na abordagem *ContextProv* essas fases foram unificadas. Sendo assim, denomina-se **Captura** a fase de obtenção das informações de proveniência e contexto, considerando tanto informações geradas pelo usuário quanto capturadas por sensores ou obtidas pela plataforma através da integração com outras plataformas.

As fases denominadas *Store*, no modelo de MISSIER (2016), e *Storage*, no modelo de BRÉZILLON *et al.* (2004), foram incluídas no ciclo de vida de proveniência e contexto utilizado no presente trabalho. A fase de **Armazenamento** consiste em armazenar as informações capturadas de acordo com condições pré-estabelecidas pela plataforma, considerando que se trata de um grande volume de dados.

O ciclo de vida de proveniência possui as fases *Query e Sharing* através das quais as informações obtidas são disponibilizadas para consulta. No modelo proposto por BRÉZILLON *et al.* (2004) na etapa, denominada *Awareness*, os dados são processados para

serem fornecidos a outros cientistas do grupo com o objetivo de compreenderem as atividades realizadas. Neste trabalho, o processamento das informações e a disponibilização das mesmas foi dividido em fases diferentes do ciclo de vida. Portanto, denomina-se **Enriquecimento** a fase onde as informações capturadas são processadas ou tratadas a fim de enriquecer o conhecimento produzido por essas informações. Nessa fase, a abordagem *ContextProv* utiliza ontologias e mecanismos de inferência para extrair informações implícitas. A fase onde as informações de proveniência e contexto são disponibilizadas para consulta denomina-se **Compartilhamento**. Vale ressaltar que as informações ficam disponíveis tanto para os usuários da plataforma como para outros sistemas externos.

Na fase de **Visualização** presente em ambos os modelos (*Visualization* / *Visualize*) as informações são apresentadas na interface do usuário. Nesta fase o conhecimento gerado pela plataforma a partir das informações obtidas deve ser representado de forma adequada visando apoiar o usuário na fase seguinte. A última fase do ciclo é a **Interpretação** (*Interpretation* no modelo de BRÉZILLON *et al.* (2004) e *Analyze* no modelo de MISSIER (2016)). É uma etapa que ocorre quando o usuário interpreta as informações apresentadas como conhecimento. Essa interpretação acontece a partir das informações exibidas e do contexto individual do usuário. Essa fase pode gerar novas contribuições e, assim, fechar o ciclo de processamento das informações. Engloba todas as formas de consumo e exploração do conhecimento disponibilizado pela plataforma por meio de soluções de engenharia de dados.

Além das fases mencionadas anteriormente, o modelo de proveniência de MISSIER (2016) possui uma fase, denominada *Preserve association to data*, que não tem correspondência no modelo de contexto. Esta é uma fase importante para que as informações de origem não sejam perdidas dos dados originais aos quais ela pertence. Entretanto, considerando que esta atividade deve ser realizada durante todo o ciclo, e não em uma fase específica, a abordagem *ContextProv* não inclui esta atividade como uma fase do ciclo de vida de proveniência e contexto.

Visto isso, a Figura 9 apresenta o ciclo de vida de informações de proveniência e contexto definido para plataformas de ECOSC e utilizado na abordagem *ContextProv*. Este ciclo é composto pelas fases de **Captura**, **Armazenamento**, **Enriquecimento**, **Compartilhamento**, **Visualização** e **Interpretação**.

Anotar todos os detalhes críticos de um experimento em um caderno de laboratório é um procedimento científico padrão, especialmente nas ciências experimentais. A principal questão na *e-Science* é que o número e a granularidade dos detalhes críticos são altos. Assim, identificar quais informações são necessárias é um desafio, e por outro lado, armazenar todas

Figura 9. Ciclo de vida de informações de Proveniência e Contexto da abordagem ContextProv

as informações contextuais pode ser um processo demorado e custoso. A subseção seguinte apresenta o *framework* conceitual que visa fornecer diretrizes para o gerenciamento de contexto em sistemas colaborativos desenvolvidos para plataformas de ECOSC.

3.2 *FRAMEWORK* CONTEXT-SE

O *Context-SE* (*Context of Scientific Experiments*) é um *framework* conceitual que identifica e classifica os elementos contextuais e de proveniência relevantes em um ambiente colaborativo e distribuído de experimentação científica. É uma extensão do *framework* proposto por ROSA *et al.* (2003) o qual considera os elementos relevantes para a análise de contexto em aplicações de *groupware*. Para cada categoria, considerada pelos autores, foram identificados outros aspectos contextuais e de proveniência específicos para o domínio da experimentação científica. Esses novos elementos foram baseados em informações encontradas em plataformas científicas como *Mendeley*[12], *Lattes*[13] e *ResearchGate*[14]. Essas plataformas ofereceram informações que envolvem os agentes (pesquisadores, instituições e grupos de pesquisa), e principalmente na ontologia *Prov-SE-O* (descrita na subseção seguinte).

[12] https://www.mendeley.com

[13] http://lattes.cnpq.br

[14] https://www.researchgate.net

A Tabela 4 apresenta as cinco categorias modeladas pelo *Context-SE* e seus objetivos, bem como os elementos contextuais associados que podem influenciar as atividades de experimentação científica do grupo. Os elementos destacados em azul representam os itens incluídos no *framework*.

Tabela 4. *Framework* conceitual *Context-SE* (estendido de (ROSA et al., 2003))

Information Type	**Associated Contexts**	**Goals**	**Examples of contextual elements**	
Agents	Researcher (Synchronous & Asynchronous)	To identify the researchers through the representation of their personal data and profiles.	• Name • Qualifications • Interests • **Degree** • Previous Experience • Location • Working hours • Web page	• **Institution** • **Position (profession)** • **E-mail** • **Awards** • **Skills** • **Languages** • **Publications** • **Research field**
	Research Group or Institutions (Synchronous & Asynchronous)	To identify the research group and the institutions through the representation of its characteristics.	• Name • Members • Roles • Abilities • Previous Experience • Geographical Location	• Organization Structure • Working hours • **Institution** • **Web page** • **E-mail** • **Partners**

Experiment Plan	Experiment (Synchronous & Asynchronous)	To identify the experiments through the representation of its characteristics.	• Name
			• Description
			• Goals
			• Deadlines
			• Estimated effort
			• **Activities**
			• Restrictions
			• Workflow
			  ○ **Title**
			  ○ **Version**
			  ○ **SWMS**
			  ○ **Description**
				○ **Activities**
				• **Similar workflows**
				• **History**
				  ○ **Evolution To**
				  ○ **Evolution Of**
				• **Problem Investigation**
				  ○ **Literature Review**
				  ○ **Related Experiments**
				• **Group in-charge**
				• **Similar experiments**

Relationship between agents and activities during the Experiment Execution	Interaction (Synchronous)	To represent in detail the activities performed during the experiment.	• Group in-charge • Messages exchanged • Presence Awareness • Gesture awareness • **Activities completed** ○ Author	○ Goal ○ Report ○ **Name** • **Input** • **Output** • **Used services**

	Interaction (Asynchronous)	To represent an overview of the activities performed during the experiment.	• Group in-charge
			• Artifacts generated
			  ○ Versions
			  ○ Timestamp
			  ○ **Name**
			• **Activities completed**
				○ Author
				○ Goal
				○ Report
				• **Input**
				• **Output**
				• **Used services**

	Planning (Synchronous & Asynchronous)	To represent the Execution Plan of the activity to be performed.	• Roles in the interaction • Rules • Aim	• Strategies • Coordination Procedures • Working Plan

			• Responsibilities	○ **Activity Name**
Setting	Environment (Synchronous & Asynchronous)	To represent the Environment where the interaction occurs; i.e., characteristics that influence activity execution.	• Quality patterns • Rules • Policies • Institutional deadlines • Organizational structure • **Cultural features**	• Financial constraints • Standard procedures • Standard strategies • **Communication Tool** • **SWMS** • **Geographical Location**
Provenance	Historical (Synchronous & Asynchronous)	To provide understanding about activities completed in the past and their associated contexts.	• **Activities** ○ Task Name ○ **Group in-charge** ○ Goal ○ Justification ○ Date • Versions of the artifacts • Working Plan	• Contextual elements used to carry out the task • Task Goals • **Input** • **Output** • **Used services** • **Sub Activities**

A primeira categoria refere-se a informações sobre os membros do grupo. Elas dizem respeito aos pesquisadores, grupos de pesquisa e instituições. Essas informações são importantes para que outros pesquisadores possam identificar quem está envolvido no experimento, e então entrar em contato com esses pesquisadores caso precisem de sua colaboração no entendimento dos resultados obtidos, ou em experimentos futuros.

A segunda categoria está relacionada a informações sobre tarefas agendadas. No domínio da experimentação científica está relacionada ao planejamento do experimento e caracteriza-se pelas tarefas a serem executadas pelo grupo até a conclusão do experimento. Estas informações se referem ao experimento, aos *workflows* científicos e às tarefas a serem executadas. Contribuem para o reúso do experimento em um novo contexto.

A terceira categoria diz respeito aos relacionamentos entre membros do grupo e tarefas agendadas. Relaciona cada pesquisador, ou grupo de pesquisa, às interações em que estão envolvidos. Essa categoria é dividida em dois tipos de contextos: contexto de interação (informações que representam as ações que ocorreram durante a execução do experimento) e o contexto de planejamento (informações sobre o plano de execução do experimento). Essas informações permitem que os créditos sejam dados aos autores, e que eles sejam responsáveis ou questionados por quaisquer erros que tenham ocorrido durante sua execução.

A quarta categoria reúne informações sobre o ambiente. Abrange tanto questões organizacionais quanto o ambiente tecnológico, ou seja, todas as informações fora do experimento, mas dentro da organização que podem afetar a maneira como as tarefas são executadas. Está relacionada a informações sobre tecnologias usadas, SGWfCs e serviços

externos. São essenciais para a reprodutibilidade do experimento, e também auxiliam no reúso dos experimentos ou parte deles.

Finalmente, a quinta categoria reúne todas as informações sobre as tarefas concluídas. Sua finalidade é fornecer informações básicas sobre as lições aprendidas, seja do mesmo grupo ou de tarefas semelhantes realizadas por outros grupos. Por conseguinte, deve incluir todas as informações contextuais e de proveniência sobre experiências anteriores. As informações sobre os experimentos, *workflows* científicos e tarefas executadas são armazenadas, bem como a proveniência delas. Assim, é possível identificar todos os processos que ocorreram com um artefato, até o final do experimento, bem como identificar o *workflow* e o experimento que deu origem a este artefato, além dos pesquisadores envolvidos.

As informações contextuais presentes no *framework* devem ser capturadas e armazenadas pela plataforma. Para permitir a extração de conhecimento implícito e o enriquecimento das informações armazenadas, as informações de contexto são conectadas a outras informações de proveniência também armazenadas, e são carregadas para uma ontologia. O conhecimento implícito para um experimento pode ser derivado por meio da execução de motores de inferência sob esta ontologia. A subseção seguinte apresenta maiores detalhes da ontologia utilizada.

3.3 ONTOLOGIA PROV-SE-O

Conforme apresentado anteriormente, a ontologia ProvONE (CUEVAS-VICENTTÍN *et al.*, 2016) modela dados de proveniência de experimentos científicos com foco nos *workflows*, suas derivações e sub-*workflows*. Esta ontologia não considera a natureza colaborativa e distribuída do atual processo de experimentação científica. Sendo assim, não oferece suporte para a modelagem do contexto, o que dificulta a representação de experimentos processados de forma distribuída, usando vários *workflows* executados em diferentes SGWfC.

Para suportar estes novos recursos da experimentação científica, foi desenvolvida uma extensão da ontologia ProvONE, chamada *Prov-SE-O* (*Provenance of Scientific Experiments Ontology*). Essa ontologia modela não apenas os workflows, mas experimentos científicos como um todo, considerando informações contextuais relacionadas à sua natureza colaborativa e distribuída.

A ontologia *Prov-SE-O* inclui novas classes, propriedades, cadeias de propriedades[15] e regras em SWRL (Linguagem de Regras da *Web* Semântica) (HORROCKS *et al.*, 2004). A Figura 10 apresenta o modelo conceitual da ontologia, destacando as classes pertencentes a *Prov-SE-O*. As classes em azul representam a proveniência prospectiva dos experimentos, ou seja, a especificação do experimento. As classes em vermelho representam os artefatos de dados produzidos ou consumidos durante a execução dos experimentos. As classes em amarelo representam a proveniência retrospectiva dos experimentos, ou seja, a descrição da execução as atividades e da derivação dos dados. As relações representadas por linhas tracejadas são propriedades adicionadas na nova ontologia, respeitando a nomenclatura das relações do modelo Prov. As relações em vermelho realçam as cadeias de propriedades e em verde as regras de SWRL.

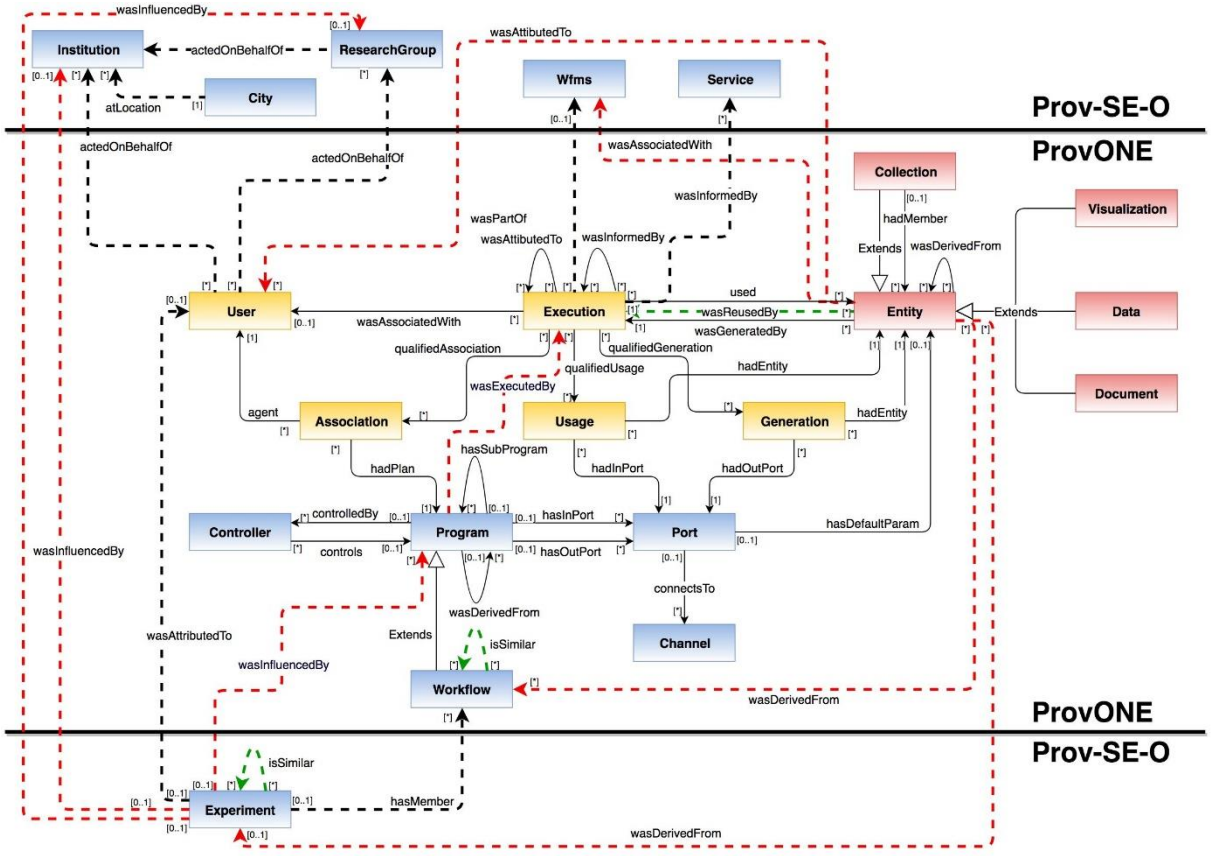

Figura 10. Modelo conceitual da ontologia *Prov-SE-O* (estendido de (CUEVAS-VICENTTÍN et al., 2016))

A Tabela 5 descreve o modelo completo de classes, propriedades, relações e regras SWRL desta ontologia através da Sintaxe Lógica de Descrição (DL - *Description Logic*)

[15] As cadeias de propriedades (*Property Chains*) aparecem na OWL 2 e funcionam classificando os objetos, onde ela permite a transitividade entre várias propriedades.

(BAADER, 2003). As regras SWRL identificam experimentos ou *workflows* similares, e documentos que foram reutilizados. São considerados experimentos similares aqueles que possuem as mesmas palavras chave e utilizam o mesmo *workflow*, ou versões do mesmo *workflow*, e são considerados *workflows* similares aqueles que possuem as mesmas palavras chave e incluem um mesmo programa. Os documentos são considerados como reutilizados, quando são utilizados na execução de um programa que não pertencem ao mesmo *workflow* o qual o documento foi gerado.

Tabela 5. Ontologia *Prov-SE-O* na Sintaxe DL

Classes

Activity
 Activity ⊑ ¬ Entity
Agent
 Agent ⊑ ¬ InstantaneousEvent
Association
 Association ⊑ AgentInfluence
Channel
 Channel ⊑ Entity
City
 City ⊑ Location
 City ⊑ Role
Collection
 Collection ⊑ Entity
Controller
 Controller ⊑ Entity
Data
 Data ⊑ Entity
Developer
 Developer ⊑ User
Document
 Document ⊑ Entity
Entity
 Entity ⊑ ¬ Activity
 Entity ⊑ ¬ InstantaneousEvent
Execution
 Execution ⊑ Activity
Experiment
 Experiment ⊑ Collection
Generation
 Generation ⊑ InstantaneousEvent
 Generation ⊑ ActivityInfluence
Influence
Institution
 Institution ⊑ Organization
Location
 Location ⊑ Role
Organization
 Organization ⊑ Agent
Plan
 Plan ⊑ Entity
Port
 Port ⊑ Entity
Program
 Program ⊑ Entity
 Program ⊑ Plan
ResearchGroup
 ResearchGroup ⊑ Organization
Researcher
 Researcher ⊑ User
Service
 Service ⊑ SoftwareAgent
SoftwareAgent
 SoftwareAgent ⊑ Agent
Usage
 Usage ⊑ InstantaneousEvent
 Usage ⊑ EntityInfluence
User
 User ⊑ Agent
Visualization
 Visualization ⊑ Entity
Wfms
 Wfms ⊑ SoftwareAgent
Workflow
 Workflow ⊑ Program
 Workflow ⊑ Plan
 Workflow ⊑ Entity

Data Properties

academicStatus
 ∃ academicStatus Literal ⊑ User
 ∃ academicStatus Literal ⊑ Researcher
 ⊤ ⊑ ∀ academicStatus string
atTime
 ∃ atTime Literal ⊑ InstantaneousEvent
 ⊤ ⊑ ∀ atTime dateTime
birthday
 ∃ birthday Literal ⊑ Agent
 ⊤ ⊑ ∀ birthday dateTime
description
 ⊤ ⊑ ∀ description string
detail
discipline
 ⊤ ⊑ ∀ discipline string
displayName
 ∃ displayName Literal ⊑ User
 ∃ displayName Literal ⊑ Researcher
 ⊤ ⊑ ∀ displayName string
endedAtTime
 ⊤ ⊑ ∀ endedAtTime dateTime
gender
 ∃ gender Literal ⊑ Agent
 ⊤ ⊑ ∀ gender string
generatedAtTime
 ∃ generatedAtTime Literal ⊑ Entity
 ⊤ ⊑ ∀ generatedAtTime dateTime
id
 ⊤ ⊑ ∀ id integer
interest
 ∃ interest Literal ⊑ Researcher
 ∃ interest Literal ⊑ User
 ⊤ ⊑ ∀ interest string
internalClass
 ∃ internalClass Literal ⊑ Service
 ∃ internalClass Literal ⊑ SoftwareAgent
invalidatedAtTime
 ∃ invalidatedAtTime Literal ⊑ Entity
 ⊤ ⊑ ∀ invalidatedAtTime dateTime
keplerId
 ∃ keplerId Literal ⊑ User
 ∃ keplerId Literal ⊑ Researcher
 ⊤ ⊑ ∀ keplerId string
keyWord
 ⊤ ⊑ ∀ keyWord string
link
 ⊤ ⊑ ∀ link string
mendeleyId
 ∃ mendeleyId Literal ⊑ Researcher
 ∃ mendeleyId Literal ⊑ User
 ⊤ ⊑ ∀ mendeleyId string
name
 ⊤ ⊑ ∀ name string
nature
 ∃ nature Literal ⊑ Service
 ∃ nature Literal ⊑ SoftwareAgent
 ⊤ ⊑ ∀ nature string
phase
 ∃ phase Literal ⊑ Collection
 ∃ phase Literal ⊑ Experiment
 ⊤ ⊑ ∀ phase string
photo
 ∃ photo Literal ⊑ Agent
 ⊤ ⊑ ∀ photo string
shortDescription
 ∃ shortDescription Literal ⊑ SoftwareAgent
 ∃ shortDescription Literal ⊑ Service
 ⊤ ⊑ ∀ shortDescription string
startedAtTime
 ∃ startedAtTime Literal ⊑ Activity
 ⊤ ⊑ ∀ startedAtTime dateTime
status
 ∃ status Literal ⊑ Experiment
 ∃ status Literal ⊑ Collection
 ⊤ ⊑ ∀ status string
title
 ∃ title Literal ⊑ Researcher
 ∃ title Literal ⊑ User
 ⊤ ⊑ ∀ title string
type
updatedAtTime
 ⊤ ⊑ ∀ updatedAtTime dateTime
url
 ∃ url Literal ⊑ SoftwareAgent
 ∃ url Literal ⊑ Service
 ⊤ ⊑ ∀ url string
value
 ∃ value Literal ⊑ Entity
version
 ⊤ ⊑ ∀ version string
webPage
 ⊤ ⊑ ∀ webPage string

Object Properties

actedOnBehalfOf
 $\sqsubseteq$ wasInfluencedBy
$\exists$ actedOnBehalfOf Thing $\sqsubseteq$ Agent
$\top \sqsubseteq \forall$ actedOnBehalfOf Agent
activity
 $\sqsubseteq$ influencer
$\exists$ activity Thing $\sqsubseteq$ ActivityInfluence
$\top \sqsubseteq \forall$ activity Activity
agent
 $\sqsubseteq$ influencer
$\exists$ agent Thing $\sqsubseteq$ AgentInfluence
$\top \sqsubseteq \forall$ agent Agent
alternateOf $\exists$ alternateOf Thing $\sqsubseteq$ Entity
$\top \sqsubseteq \forall$ alternateOf Entity
atLocation
$\exists$ atLocation Thing $\sqsubseteq$ Activity $\sqcup$ Agent $\sqcup$ Entity $\sqcup$ InstantaneousEvent
$\top \sqsubseteq \forall$ atLocation Location
connectsTo
$\exists$ connectsTo Thing $\sqsubseteq$ Entity
$\exists$ connectsTo Thing $\sqsubseteq$ Port
$\top \sqsubseteq \forall$ connectsTo Entity
$\top \sqsubseteq \forall$ connectsTo Channel
controlledBy
$\exists$ controlledBy Thing $\sqsubseteq$ Plan
$\exists$ controlledBy Thing $\sqsubseteq$ Entity
$\exists$ controlledBy Thing $\sqsubseteq$ Program
$\top \sqsubseteq \forall$ controlledBy Controller
$\top \sqsubseteq \forall$ controlledBy Entity
controls
$\exists$ controls Thing $\sqsubseteq$ Entity
$\exists$ controls Thing $\sqsubseteq$ Controller
$\top \sqsubseteq \forall$ controls Plan
$\top \sqsubseteq \forall$ controls Program
$\top \sqsubseteq \forall$ controls Entity
entity
 $\sqsubseteq$ influencer
$\exists$ entity Thing $\sqsubseteq$ EntityInfluence
$\top \sqsubseteq \forall$ entity Entity
executed
 $\sqsubseteq$ influenced
executed $\equiv$ wasExecutedBy$^-$
$\exists$ executed Thing $\sqsubseteq$ Activity
$\top \sqsubseteq \forall$ executed Entity
generated
 $\sqsubseteq$ influenced
generated $\equiv$ wasGeneratedBy$^-$
$\exists$ generated Thing $\sqsubseteq$ Activity
$\top \sqsubseteq \forall$ generated Entity
hadActivity
$\exists$ hadActivity Thing $\sqsubseteq$ Delegation $\sqcup$ Derivation $\sqcup$ End $\sqcup$ Start
$\exists$ hadActivity Thing $\sqsubseteq$ Influence
$\top \sqsubseteq \forall$ hadActivity Activity
hadDerivation
TransitivePropertyhadDerivation
hadEntity
$\exists$ hadEntity Thing $\sqsubseteq$ Activity
$\top \sqsubseteq \forall$ hadEntity Entity
hadGeneration
$\exists$ hadGeneration Thing $\sqsubseteq$ Derivation
$\top \sqsubseteq \forall$ hadGeneration Generation
hadInPort
$\exists$ hadInPort Thing $\sqsubseteq$ Usage
$\top \sqsubseteq \forall$ hadInPort Port
hadMember
 $\sqsubseteq$ wasInfluencedBy
$\exists$ hadMember Thing $\sqsubseteq$ Collection
$\top \sqsubseteq \forall$ hadMember Entity
hadOutPort
$\exists$ hadOutPort Thing $\sqsubseteq$ Generation
$\top \sqsubseteq \forall$ hadOutPort Port
hadPlan
$\exists$ hadPlan Thing $\sqsubseteq$ Association
$\top \sqsubseteq \forall$ hadPlan Plan
hadUsage
$\exists$ hadUsage Thing $\sqsubseteq$ Derivation
$\top \sqsubseteq \forall$ hadUsage Usage
hasDefaultParam
$\exists$ hasDefaultParam Thing $\sqsubseteq$ Port
$\top \sqsubseteq \forall$ hasDefaultParam Entity
hasInPort
$\exists$ hasInPort Thing $\sqsubseteq$ Program
$\exists$ hasInPort Thing $\sqsubseteq$ Entity
$\exists$ hasInPort Thing $\sqsubseteq$ Plan
$\top \sqsubseteq \forall$ hasInPort Port
hasMember
 $\sqsubseteq$ wasInfluencedBy
hasOutPort
$\exists$ hasOutPort Thing $\sqsubseteq$ Entity
$\exists$ hasOutPort Thing $\sqsubseteq$ Program
$\exists$ hasOutPort Thing $\sqsubseteq$ Plan
$\top \sqsubseteq \forall$ hasOutPort Port
hasSubProgram
$\exists$ hasSubProgram Thing $\sqsubseteq$ Entity
$\exists$ hasSubProgram Thing $\sqsubseteq$ Program
$\exists$ hasSubProgram Thing $\sqsubseteq$ Plan
$\top \sqsubseteq \forall$ hasSubProgram Entity
$\top \sqsubseteq \forall$ hasSubProgram Program
$\top \sqsubseteq \forall$ hasSubProgram Plan
influenced
influenced $\equiv$ wasInfluencedBy$^-$
influencer
$\exists$ influencer Thing $\sqsubseteq$ Influence
$\top \sqsubseteq \forall$ influencer Thing
isSimilar
isSimilar $\equiv$ isSimilar$^-$
qualifiedAssociation
 $\sqsubseteq$ qualifiedInfluence
$\exists$ qualifiedAssociation Thing $\sqsubseteq$ Activity
$\exists$ qualifiedAssociation Thing $\sqsubseteq$ Execution
$\top \sqsubseteq \forall$ qualifiedAssociation Association
qualifiedDerivation
 $\sqsubseteq$ qualifiedInfluence
$\exists$ qualifiedDerivation Thing $\sqsubseteq$ Entity
$\top \sqsubseteq \forall$ qualifiedDerivation Derivation
qualifiedGeneration
 $\sqsubseteq$ qualifiedInfluence
$\exists$ qualifiedGeneration Thing $\sqsubseteq$ Activity
$\top \sqsubseteq \forall$ qualifiedGeneration Generation
qualifiedUsage
 $\sqsubseteq$ qualifiedInfluence
$\exists$ qualifiedUsage Thing $\sqsubseteq$ Activity
$\exists$ qualifiedUsage Thing $\sqsubseteq$ Execution
$\top \sqsubseteq \forall$ qualifiedUsage Usage
specializationOf
 $\sqsubseteq$ alternateOf
$\exists$ specializationOf Thing $\sqsubseteq$ Entity
used
 $\sqsubseteq$ wasInfluencedBy
wasReusedBy
 $\equiv$ used$^-$
$\exists$ used Thing $\sqsubseteq$ Activity
$\exists$ used Thing $\sqsubseteq$ Execution
$\top \sqsubseteq \forall$ used Entity
wasAssociatedWith
 $\sqsubseteq$ wasInfluencedBy
$\exists$ wasAssociatedWith Thing $\sqsubseteq$ Activity
$\exists$ wasAssociatedWith Thing $\sqsubseteq$ Execution
$\top \sqsubseteq \forall$ wasAssociatedWith User
$\top \sqsubseteq \forall$ wasAssociatedWith Agent
wasAttributedTo
 $\sqsubseteq$ wasInfluencedBy
$\exists$ wasAttributedTo Thing $\sqsubseteq$ Entity
$\top \sqsubseteq \forall$ wasAttributedTo Agent
wasDerivedFrom
 $\sqsubseteq$ wasInfluencedBy
TransitivePropertywasDerivedFrom
$\exists$ wasDerivedFrom Thing $\sqsubseteq$ Plan
$\exists$ wasDerivedFrom Thing $\sqsubseteq$ Entity
$\exists$ wasDerivedFrom Thing $\sqsubseteq$ Program
$\top \sqsubseteq \forall$ wasDerivedFrom Entity
$\top \sqsubseteq \forall$ wasDerivedFrom Program
$\top \sqsubseteq \forall$ wasDerivedFrom Plan
wasExecutedBy
 $\sqsubseteq$ wasInfluencedBy
executed $\equiv$ wasExecutedBy$^-$
wasGeneratedBy
 $\sqsubseteq$ wasInfluencedBy
generated $\equiv$ wasGeneratedBy$^-$
$\exists$ wasGeneratedBy Thing $\sqsubseteq$ Entity
$\top \sqsubseteq \forall$ wasGeneratedBy Execution
$\top \sqsubseteq \forall$ wasGeneratedBy Activity
wasInfluencedBy
influenced $\equiv$ wasInfluencedBy$^-$
$\exists$ wasInfluencedBy Thing $\sqsubseteq$ Activity $\sqcup$ Agent $\sqcup$ Entity
$\top \sqsubseteq \forall$ wasInfluencedBy (Activity $\sqcup$ Agent $\sqcup$ Entity)
wasInformedBy
 $\sqsubseteq$ wasInfluencedBy
wasPartOf
 $\equiv$ wasInformedBy$^-$
$\exists$ wasInformedBy Thing $\sqsubseteq$ Execution
$\exists$ wasInformedBy Thing $\sqsubseteq$ Activity
$\top \sqsubseteq \forall$ wasInformedBy Activity
$\top \sqsubseteq \forall$ wasInformedBy Execution
wasMemberOf
wasPartOf
wasPartOf $\equiv$ wasInformedBy$^-$
$\exists$ wasPartOf Thing $\sqsubseteq$ Activity
$\exists$ wasPartOf Thing $\sqsubseteq$ Execution
$\top \sqsubseteq \forall$ wasPartOf Execution
wasReusedBy
wasReusedBy $\equiv$ used$^-$
wasRevisionOf $\sqsubseteq$ wasDerivedFrom
$\exists$ wasRevisionOf Thing $\sqsubseteq$ Entity
wasStartedBy
 $\sqsubseteq$ wasInfluencedBy
$\exists$ wasStartedBy Thing $\sqsubseteq$ Activity
$\top \sqsubseteq \forall$ wasStartedBy Entity

SWRL Rules

$wasMemberOf(?w,?x) \wedge wasMemberOf(?w,?y) \wedge keyWord(?x,?k) \wedge keyWord(?y,?k) \wedge id(?x,?idx) \wedge id(?y,?idy) \wedge swrlb:notEqual(?idx,?idy) \Rightarrow isSimilar(?x,?y)$
$wasMemberOf(?w,?x) \wedge wasMemberOf(?z,?y) \wedge wasDerivedFrom(?z,?w) \wedge keyWord(?x,?k) \wedge keyWord(?y,?k) \wedge id(?x,?idx) \wedge id(?y,?idy) \wedge swrlb:notEqual(?idx,?idy) \Rightarrow isSimilar(?x,?y)$
$hasSubProgram(?x,?p) \wedge hasSubProgram(?y,?p) \wedge keyWord(?x,?k) \wedge keyWord(?y,?k) \wedge id(?x,?idx) \wedge id(?y,?idy) \wedge swrlb:notEqual(?idx,?idy) \Rightarrow isSimilar(?x,?y)$
$hasMember(?x,?w) \wedge hasMember(?y,?z) \wedge isSimilar(?w,?z) \Rightarrow isSimilar(?x,?y)$
$Document(?d) \wedge Document(?e) \wedge Execution(?x) \wedge used(?x,?d) \wedge wasInfluencedBy(?e,?x) \wedge id(?d,?idd) \wedge id(?e,?ide) \wedge swrlb:notEqual(?idd,?ide) \Rightarrow wasReusedBy(?d,?x)$

Por meio de algoritmos de inferência, esta ontologia permite a derivação de conhecimento implícito como, por exemplo: experimentos e *workflows* semelhantes entre si ou derivados um do outro; pesquisadores, instituições e grupos de pesquisa envolvidos ou que

influenciaram o experimento; *workflows* ou serviços utilizados no processo de experimentação; dados consumidos e gerados pelo experimento, atividades ou entidades produzidas em um experimento e reutilizadas em experimentos futuros, entre outros. A Figura 11 apresenta exemplos de inferências exibidas na ferramenta Protégé[16]. Acredita-se que estas informações possam facilitar o reúso do experimento e de partes do experimento.

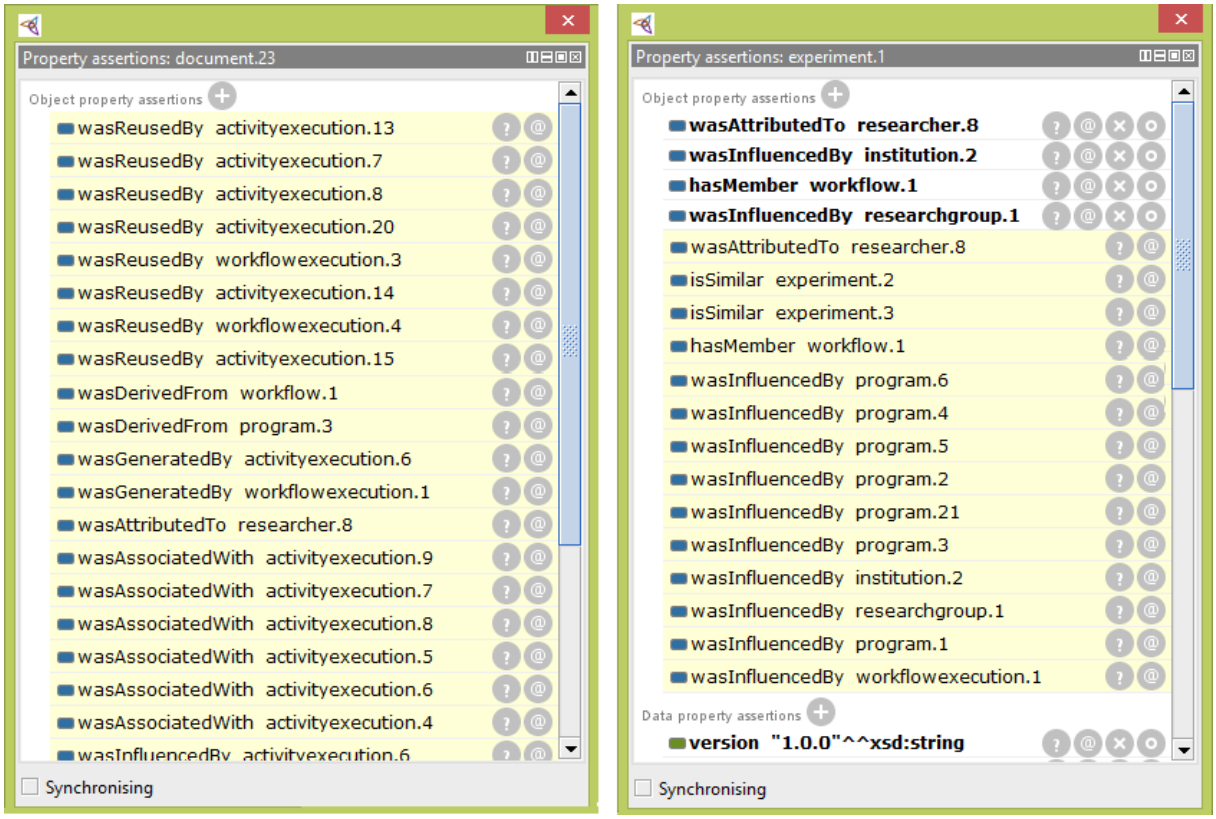

Figura 11. Exemplo de inferências na ferramenta Protégé

Após identificar as fases do ciclo de vida das informações de contexto e de proveniência, e os elementos contextuais relevantes para este domínio, a solução proposta foi implementada na plataforma E-SECO.

3.4 PROJETO E IMPLEMENTAÇÃO

A implementação da abordagem *ContextProv* na plataforma E-SECO deve considerar o ciclo de vida de proveniência e contexto em ECOSC apresentado anteriormente. A Figura 12 ilustra a relação entre o ciclo de experimentação científica na plataforma E-SECO e o ciclo de gerenciamento de proveniência e contexto da abordagem *ContextProv*. Conforme pode ser

[16] https://protege.stanford.edu

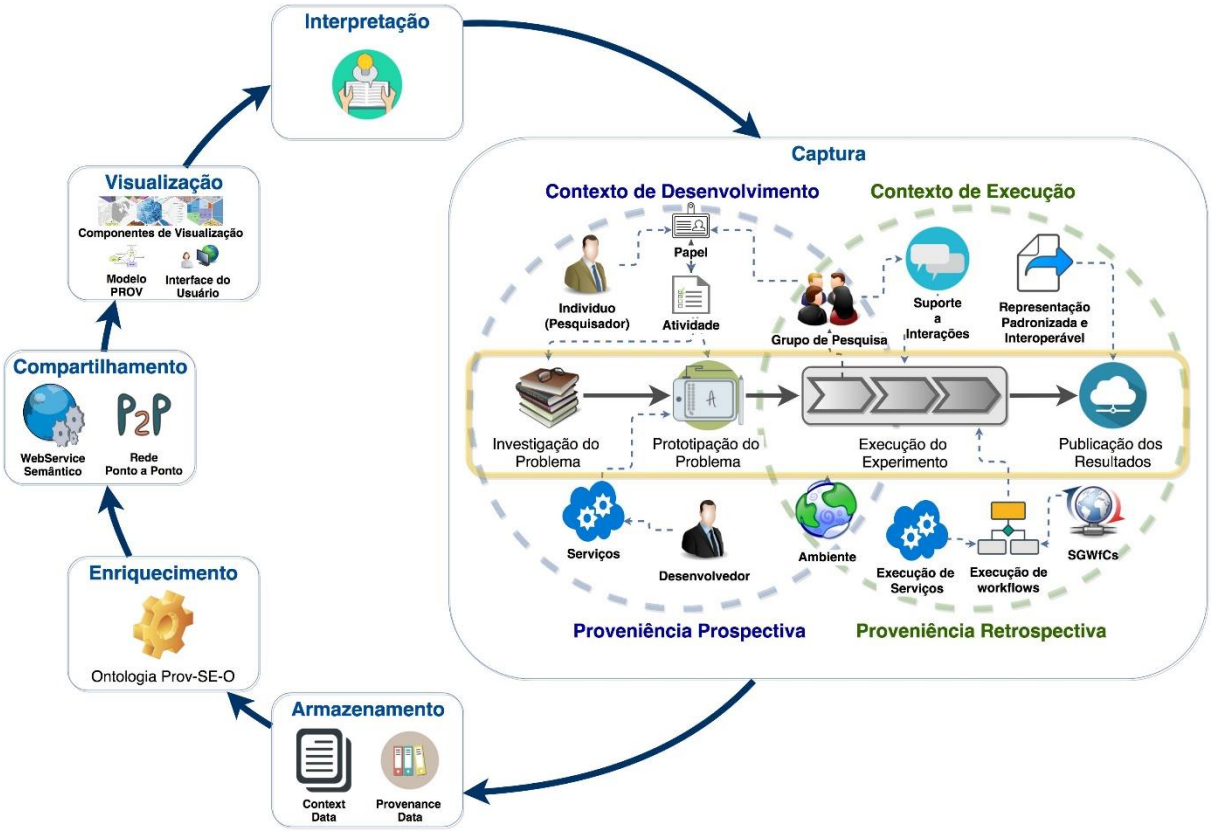

Figura 12. Gerenciamento de Contexto e Proveniência na Plataforma E-SECO

observado, o gerenciamento de contexto e proveniência deve ocorrer durante todas as fases do ciclo de experimentação. Tendo em vista os ciclos apresentados na figura, foram derivados os requisitos funcionais e não funcionais, apresentados a seguir.

3.4.1. Requisitos funcionais

RF 001. A **Captura** das informações de proveniência e contexto deve ocorrer durante todo o ciclo de experimentação científica.

RF 002. A plataforma deve ser capaz de obter todas as informações presentes no *framework Context-SE*.

RF 003. Nas fases de Investigação do Problema e Prototipagem devem ser capturados os elementos do **Contexto de Desenvolvimento** e as informações de **Proveniência Prospectiva**. Essas informações representam uma especificação abstrata do experimento e fornecem diretrizes para a derivação de experimentos futuros.

RF 004. Durante as fases de Execução do Experimento e Publicação dos Resultados devem ser capturados os elementos relativos ao **Contexto de Execução,** e as informações de **Proveniência Retrospectiva**. Essas informações representam, por exemplo, quais tarefas

foram executadas e como os artefatos de dados foram derivados. Auxiliam na verificação dos resultados obtidos pelo experimento.

RF 005. A plataforma deve permitir que o usuário cadastre as informações contextuais e de proveniência através de formulários, mas deve também permitir que estas informações sejam obtidas por meio de sensores ou da integração com outras plataformas de experimentação utilizadas durante o processo.

RF 006. A plataforma deve conter um repositório para o **Armazenamento** das informações contextuais e de proveniência capturadas ao longo do processo de experimentação.

RF 007. Os dados do repositório devem passar pela fase de **Enriquecimento**, para isso devem ser carregados para a ontologia *Prov-SE-O* e processados por uma máquina de inferência que permite extrair conhecimento implícito.

RF 008. A plataforma deve oferecer mecanismos para o **Compartilhamento** das informações de proveniência e contexto obtidas.

RF 009. A plataforma deve oferecer mecanismos de **Visualização** adequados para todas as informações capturadas e inferidas pela ontologia, e assim, apoiar os pesquisadores na **Interpretação** das informações.

3.4.2. Requisitos não funcionais

RNF 001. Na fase de compartilhamento, as informações devem ser fornecidas de forma padronizada e interoperável para que plataformas externas possam fazer uso deste conhecimento. Para isso, os dados de proveniência devem ser compartilhados em um formato padrão para compartilhamento de dados na *web* (XML ou JSON) e respeitar o padrão PROV.

RNF 002. A solução deve respeitar os princípios de extensibilidade, flexibilidade, interoperabilidade e ausência de estado da plataforma E-SECO. Para isso, deve seguir o modelo MVC (*Model-View-Controller*) e SOA (*Service-Oriented Architecture*) já utilizado pela plataforma.

RNF 003. A solução deve considerar a linguagem principal da plataforma, Java para Web para respeitar os princípios de portabilidade a plataforma.

3.4.3. Arquitetura geral

Para atender aos requisitos funcionais apresentados, devem ser implementadas ferramentas de apoio para cada etapa do ciclo de vida de proveniência e contexto em ECOSC. A Figura 13

apresenta uma visão geral da arquitetura da solução proposta. Os componentes destacados com bordas tracejadas são aqueles já existentes na plataforma E-SECO, e que não precisam ser alterados.

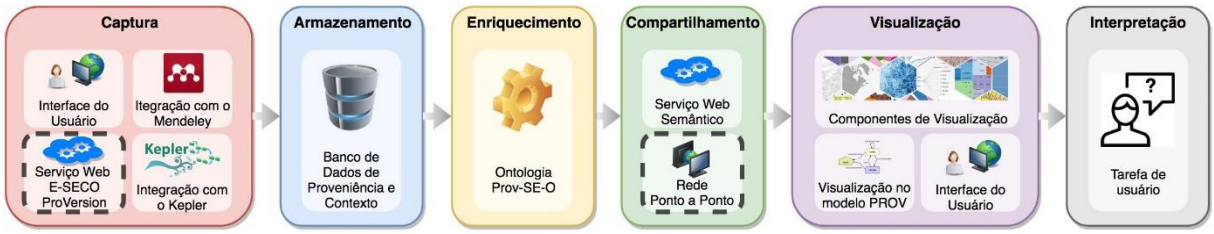

Figura 13. Visão geral da abordagem *ContextProv*

A **captura** das informações é feita através da interface do usuário, através da qual o próprio pesquisador pode cadastrar dados de proveniência e as informações de contexto durante todo o processo de experimentação. Para obter informações sobre o perfil científico dos pesquisadores que utilizam a plataforma, sem que o próprio usuário precise alimentar este cadastro manualmente, a plataforma deve estar integrada à plataforma Mendeley durante o cadastro do perfil do usuário. A escolha desta plataforma se deve ao fato de ser a única plataforma científica aberta que oferece um serviço para busca de informações. Através desta integração são obtidas as informações sobre o pesquisador, seus interesses de pesquisa, instituições de ensino e grupos de pesquisa.

A plataforma E-SECO apoia a execução de experimentos científicos, mas não tem como objetivo executar um *workflow* científico. Desta forma, a captura das informações de proveniência e contexto relacionadas à execução dos *workflows* ocorre por meio da integração com os SGWfCs. O E-SECO *ProVersion* implementa um serviço *web* que se comunica diretamente com os SGWfCs. Através deste serviço, são obtidas informações sobre as diversas execuções do experimento como: o momento de início e término da execução, os dados de entrada e saída de cada tarefa executada pelos *workflows* e o resultado final obtido. Porém, este serviço não é capaz de obter informações de proveniência de forma padronizada e detalhada. Informações como o agente responsável pela execução de cada atividade e o relacionamento entre os *workflows*, os agentes e as atividades não são contempladas. Para obter essas informações, foram analisados diversos SGWfCs. O Kepler (ALTINTAS *et al.*, 2004) se destacou, pois, além de fornecer os resultados da execução do experimento, ele fornece informações de proveniência de todas as etapas da execução do *workflow*, e ainda utiliza o modelo de proveniência PROV (MOREAU e GROTH, 2013).

O **armazenamento** dos dados capturados é realizado a partir do uso de repositórios distribuídos, modelados com base no modelo conceitual de dados do ProvONE (CUEVAS-VICENTTÍN *et al.*, 2016). Com isso, os dados presentes nesta base seguem o formato padronizado facilitando sua interpretação, e sua interoperabilidade com outros sistemas. Apesar de se tratar de um grande volume de dados, a plataforma E-SECO gerencia esses dados com o apoio de uma rede ponto a ponto (CLASSE *et al.*, 2017). Desta forma, cada nó da rede possui um repositório de dados do E-SECO, permitindo que os dados sejam armazenados de forma descentralizada.

O **enriquecimento** dos dados é feito através da ontologia *Prov-SE-O*. Nesta fase, os dados capturados, e os já armazenados na base de dados são inseridos na ontologia, e através de algoritmos de inferência pode ser derivado conhecimento implícito, tais como: experimentos e *workflows* similares entre si, ou derivados um do outro; SGWfC utilizados para a execução dos *workflows* (quando estes não são informados explicitamente); pesquisadores, instituições e grupos de pesquisa envolvidos ou que influenciaram o experimento; *workflows* ou serviços externos utilizados no processo de experimentação; dados consumidos e gerados ou reutilizados pelo experimento, entre outros.

O **compartilhamento** de informações é realizado na plataforma E-SECO através da uma rede ponto a ponto. Desta forma, usuários desta plataforma já possuem acesso direto aos dados publicados por outros pesquisadores através desta rede. No entanto, considerando a necessidade de um suporte ao reúso de dados de proveniência e contexto, foi desenvolvido um serviço específico para o compartilhamento dessas informações de forma padronizada e interoperável, através do uso do modelo PROV. Este serviço possibilita que informações sobre a proveniência e contexto dos experimentos científicos, bem como de seus *workflows*, execuções e dados sejam consultados pelos próprios responsáveis e também compartilhados com plataformas externas.

A **visualização** das informações é feita através da interface de usuário. Um dos objetivos dessa abordagem é identificar visualizações apropriadas para a apresentação de informações de contexto e de proveniência de experimentos científicos. Desta forma, para que as informações sejam exibidas de forma padronizada e através de abstrações que facilitem sua compreensão, foram desenvolvidos componentes de visualização para representar: o *workflow* utilizado e o reúso de atividades do *workflow*; o relacionamento entre pesquisadores; o relacionamento entre os pesquisadores e os experimentos; a proveniência das entidades produzidas e reutilizadas pelo experimento. Além disso, a abordagem conta com um grafo de proveniência e contexto modelado de acordo com as convenções de representação de

atividades, entidades e agentes estabelecidas no padrão PROV, o qual permite a visualização e navegação por todas as informações capturadas pela plataforma e inferidas pela ontologia.

3.4.4. Implementação

A abordagem proposta foi implementada na plataforma E-SECO através de serviços. Estes serviços interagem com os módulos da plataforma e com aplicações externas. A Figura 14 apresenta um diagrama de componentes da solução. Os componentes em branco são aqueles já existentes na plataforma, e as caixas em preto representam as aplicações externas que foram integradas à plataforma. Todos os novos componentes foram implementados em Java, seguindo o padrão MVC e a arquitetura orientada a serviços, garantindo os padrões utilizados pela plataforma.

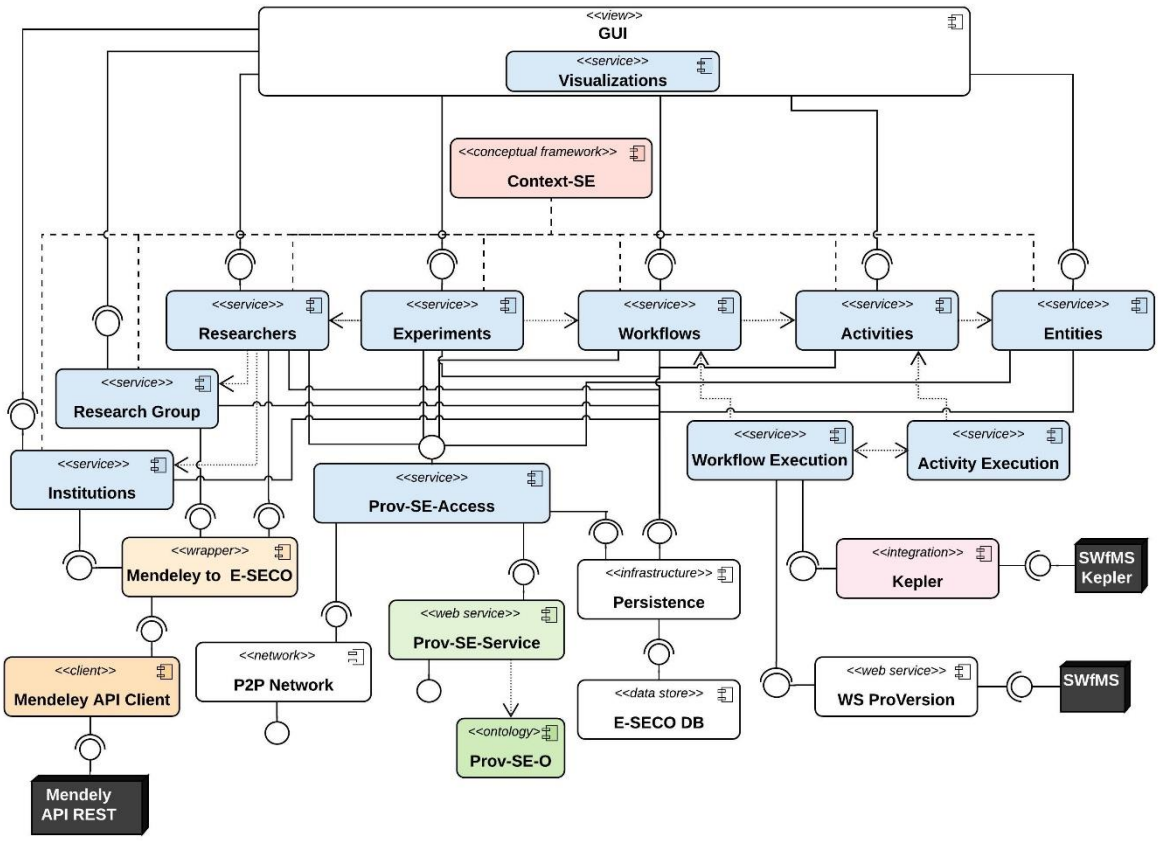

Figura 14. Diagrama de componentes da abordagem *ContextProv*

A interface do usuário (GUI) permite que os pesquisadores interajam com a plataforma para cadastrar, visualizar, listar e excluir informações de proveniência e contexto dos experimentos científicos. A Figura 15 apresenta algumas telas da interface do usuário. A interface existente na plataforma foi estendida para comportar o cadastro e a visualização das

informações de contexto e proveniência definidas no *framework Context-SE*. Assim, esta interface interage com os serviços de controle das informações presentes na plataforma.

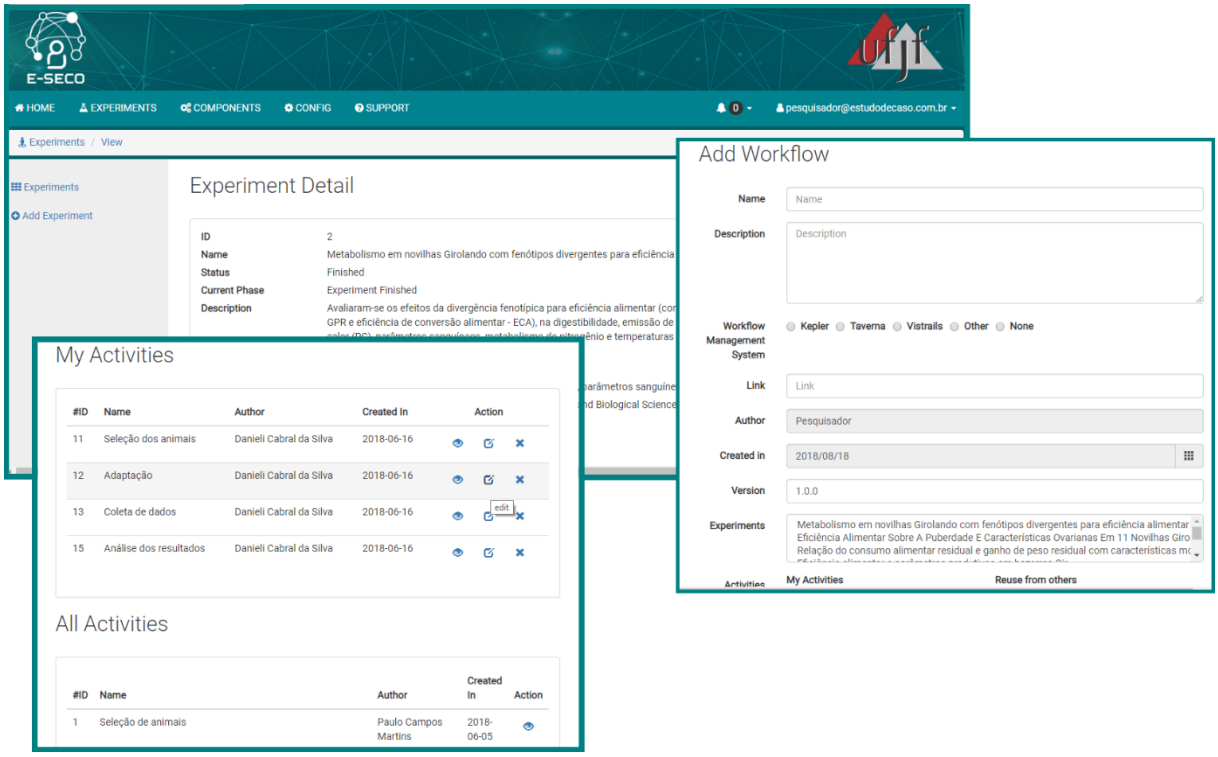

Figura 15. E-SECO GUI

Estes serviços controlam informações de contexto e proveniência dos seguintes objetos: Pesquisadores; Grupos de Pesquisa; Instituições; Experimentos; *Workflows;* Atividades; Entidades; SGWfCs; Mudança de fase do experimento; Execução de *workflows*; Execução de atividades e Detalhes do experimento e atividade. Estes serviços foram modelados de acordo com as informações definidas pelo *framework* conceitual *Context-SE*, e também baseados no modelo de proveniência PROV. A Figura 16 ilustra o modelo de classes desses objetos na plataforma.

Alguns desses serviços são integrados a plataformas externas de onde são obtidas informações de proveniência e contexto. Mais detalhes dobre estas integrações são apresentadas a seguir.

3.4.4.1. Integração com a plataforma Mendeley

Mendeley (ZAUGG *et al.*, 2011) é uma plataforma aberta para organizar citações de pesquisa e anotar artigos e os arquivos em formato PDF que os acompanham. A plataforma disponibiliza um ambiente colaborativo através do qual pesquisadores podem interagir e criar anotações em artigos compartilhados com um determinado grupo. Desta forma, essa plataforma integra o

gerenciamento de referências com recursos de uma rede social acadêmica que permite a colaboração entre pesquisadores geograficamente distribuídos.

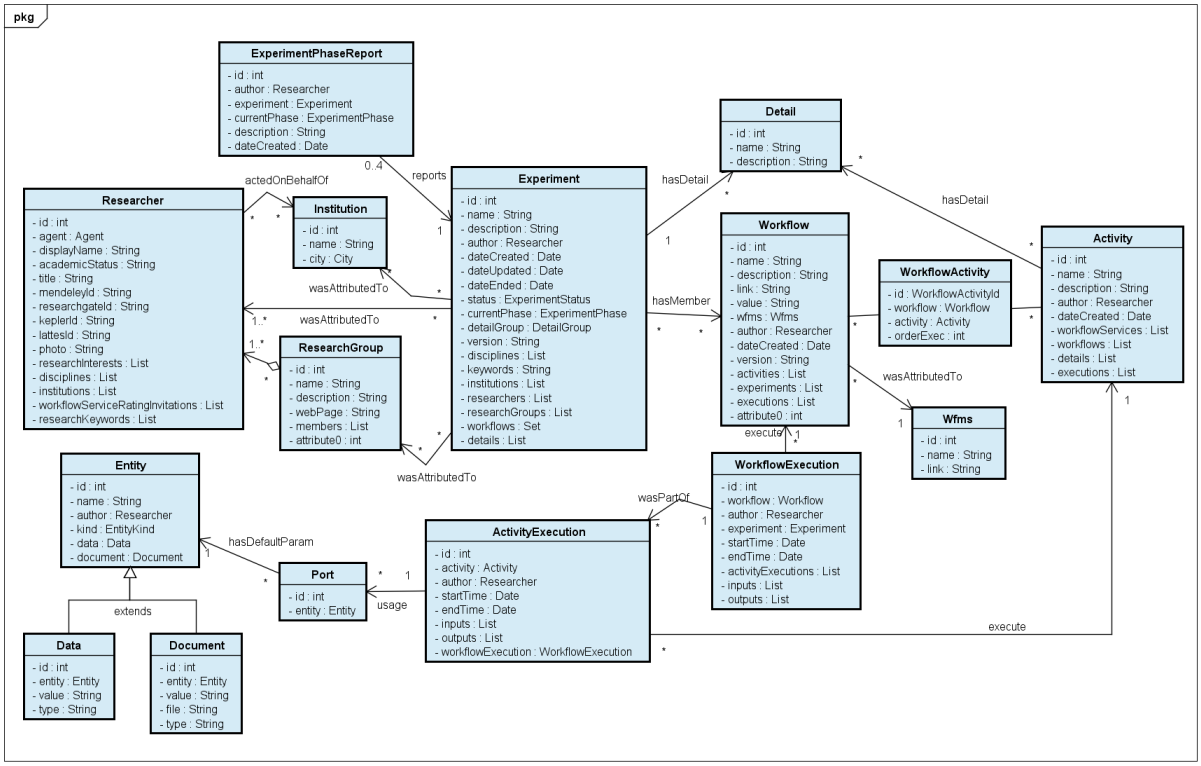

Figura 16. Modelo de classes na plataforma E-SECO

Dentre os recursos disponibilizados pelo Mendeley está uma aplicação *desktop* para os principais sistemas operacionais como Windows, Mac OSX, Linux, além de versões para iOS e Android. Além disso, oferece as seguintes funcionalidades: (i) apoia o gerenciamento e anotação de artigos, e (ii) oferece um *plugin* de integração com Word, LibreOffice e BibTeX para gerenciamento de referências e uma API REST para desenvolvedores externos. Adicionalmente, possui uma aplicação *web* para explorar os aspectos sociais da plataforma, tais como: seguir atualizações de um pesquisador ou criar grupos de pesquisa.

A API REST disponibilizada pela plataforma Mendeley permite a busca por documentos, pastas, anotações, entre outros, além de fornecer acesso aos dados do perfil dos pesquisadores, das instituições e dos grupos de pesquisa cadastrados na plataforma. Este serviço fornece os dados no formato XML e JSON, além de garantir conexões HTTPS e autorização de acesso com OAuth 2.0. Para realizar a integração entre as plataformas E-SECO e Mendeley, foi necessária a implementação de uma aplicação cliente responsável por autenticar a plataforma E-SECO no serviço REST da plataforma Mendeley, a fim de obter o *token* de autorização de acesso, e realizar as requisições ao serviço.

Além disso, a aplicação implementa um *wrapper* que realiza a transformação dos dados obtidos no formato JSON em objetos do modelo de dados do E-SECO. Os *wrappers* são os componentes de um sistema de integração de dados que se comunicam com as fontes de dados. Este componente é responsável pelo envio de consultas dos níveis mais altos do sistema de integração de dados para as fontes, e depois a conversão das respostas para um formato que pode ser manipulado pelo processador de consulta (DOAN *et al.*, 2012). A complexidade do *wrapper* depende da natureza da fonte de dados. No caso mais simples, pode envolver apenas a interação com um *driver* JDBC. No caso da integração com a plataforma Mendeley, o *wrapper* analisa os dados semiestruturados no formato JSON e os transforma para o modelo de dados implementado na plataforma E-SECO.

Quando o pesquisador se cadastra na plataforma, ele tem a opção de preencher seu perfil utilizando esta integração. Assim, conforme ilustra a Figura 17, são obtidas informações sobre: seu status acadêmico; seus interesses de pesquisa; as disciplinas envolvidas em seu contexto de pesquisa; a instituição a qual ele está vinculado; seus grupos de pesquisa e ainda as palavras chave de seus artigos publicados organizadas por ano de publicação. Estas informações são importantes para caracterizar o pesquisador e aumentar a confiabilidade de seus experimentos para que outros pesquisadores possam reutilizá-los.

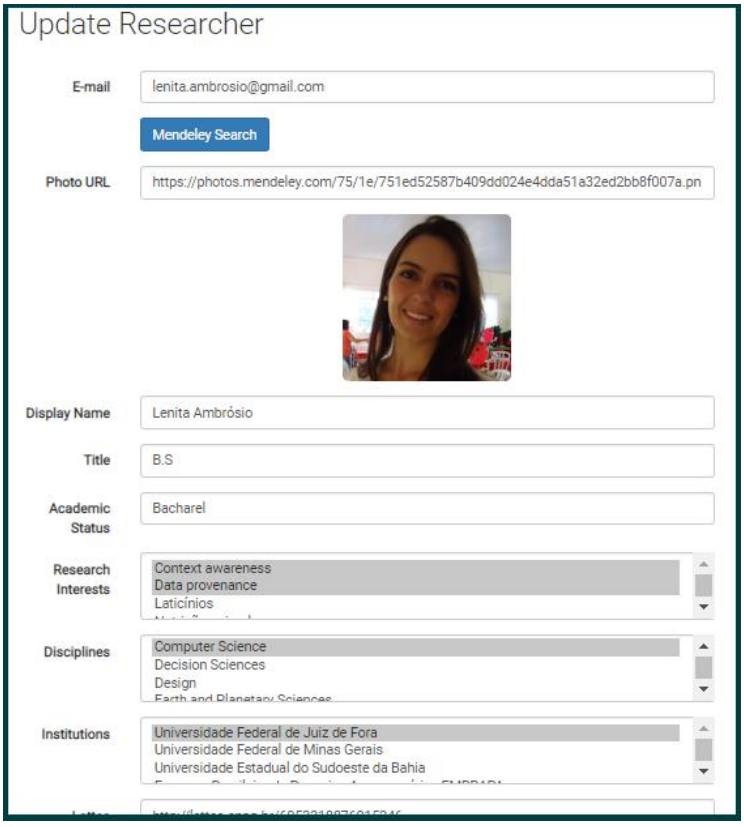

Figura 17. Cadastro do Pesquisador - Integração com o *Mendeley*

3.4.4.2. Integração com o Kepler

O Kepler (ALTINTAS *et al.*, 2004) é um SGWfC projetado para ajudar cientistas, analistas e programadores a criar, executar e compartilhar modelos e análises em uma gama de disciplinas científicas e de engenharia. O Kepler pode operar com dados armazenados em uma variedade de formatos, local e *online*. è um ambiente eficaz para integrar componentes de software diferentes e facilitar a execução remota e distribuída de modelos.

Uma das formas de integração entre sistemas é através de um formato de dados comum. Este método garante que diversas aplicações possam compartilhar informações sem a necessidade de converter dados para todos os formatos de outras aplicações. O uso de um formato de dados comum também garante que não haverá perda semântica durante a conversão.

A integração com o Kepler acontece por meio de um formato de dados comum. O modelo de dados utilizado é o PROV. A família de documentos PROV define um modelo e suas correspondentes serializações, bem como outras definições de apoio para permitir o intercâmbio de informações de proveniência em ambientes heterogêneos na *web*.

O Kepler realiza a captura de informações de proveniência durante a execução do *workflow* e, ao final, permite que estas informações sejam exportadas de acordo com o modelo PROV. Isso fez com que a integração entre o Kepler e a plataforma E-SECO fosse possível. Para isso, o pesquisador precisa configurar o módulo de proveniência do Kepler para gerar os dados de proveniência no modelo PROV e em arquivos no formato JSON. Após configurar o módulo de proveniência do Kepler, o pesquisador inicia a execução do *workflow*. A Figura 18 apresenta um exemplo de *workflow* no Kepler, e a configuração para exportação de proveniência no modelo PROV e em arquivo do tipo JSON.

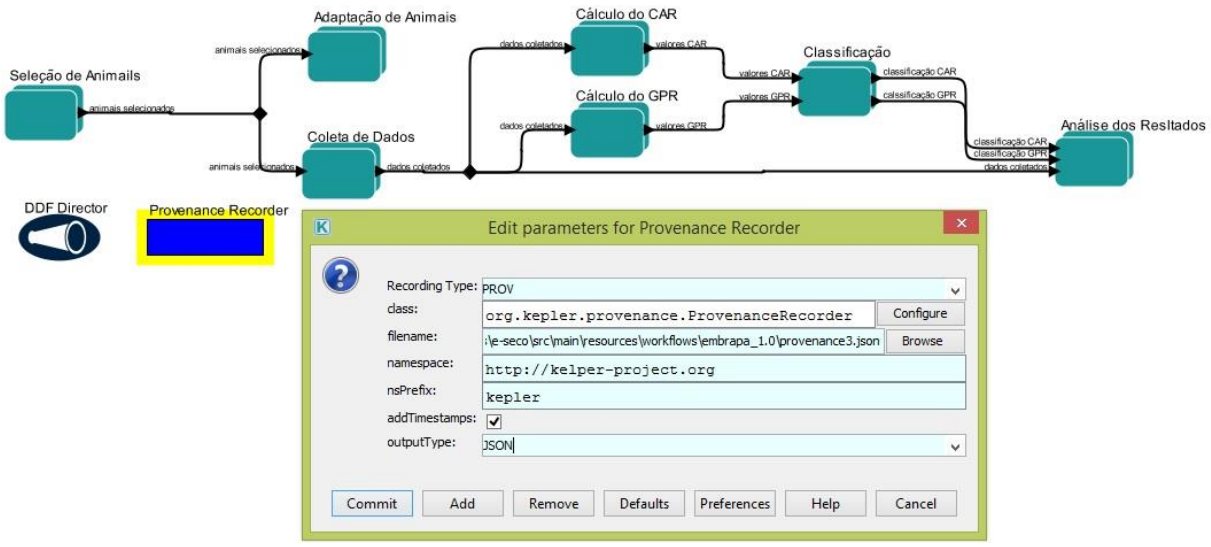

Figura 18. Exemplo de *workflow* no Kepler e da configuração para exportação dos dados de proveniência.

Ao término da execução, o arquivo contendo os dados de proveniência já terá sido gerado pelo SGWfC. O pesquisador precisa então importar este arquivo para a plataforma E-SECO. Para isso, ele precisa selecionar o *workflow* que foi executado, e fazer o *upload* do arquivo. A Figura 19 ilustra a importação do arquivo de proveniência. Como o modelo de dados da plataforma E-SECO também segue o padrão de proveniência baseado no modelo PROV (MOREAU e GROTH, 2013), estes dados são importados sem perda semântica.

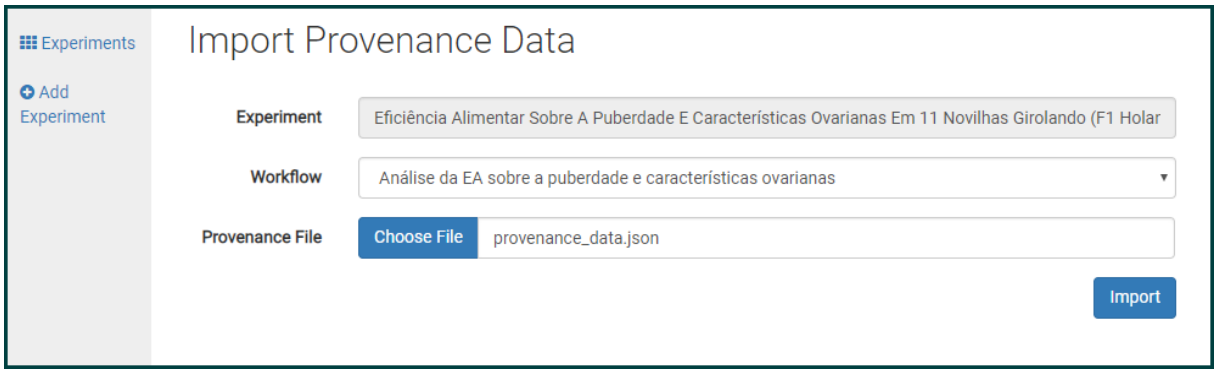

Figura 19. Importação dos dados de proveniência

Todas as informações de proveniência e contexto informadas pelo usuário e obtidas por meio das integrações são enviadas para a ontologia, onde é extraído conhecimento implícito nos dados. Para isso, a solução conta com dois componentes. O primeiro é um serviço *web* RESTFul, independente da plataforma, denominado Prov-SE-Service. Este serviço tem acesso direto a ontologia e disponibiliza métodos para a carga e recuperação de dados da ontologia. O segundo, é um serviço da plataforma E-SECO que busca as informações no banco de dados da plataforma, envia as informações para o Prov-SE-Service por meio de uma requisição POST e busca as inferências utilizando métodos GET. A seguir, são apresentados detalhes da implementação do componente Prov-SE-Service.

3.4.4.3. Prov-SE-Service

É um serviço *web* semântico RESTFul que provê acesso à ontologia. Este serviço permite que plataformas externas enviem dados para a ontologia, e recuperem o conhecimento obtido por inferência. Possui dois métodos de acesso: (i) o método POST utilizado para a entrada de dados recebe arquivos no formato JSON contendo as informações dos experimentos científicos modelados de acordo com o modelo PROV; (ii) o método GET utilizado para a busca de dados na ontologia. Recebe o nome do objeto como parâmetro e retorna um arquivo JSON contendo todas as informações relativas a este objeto, inclusive as inferidas pela ontologia. O processo de carga e recuperação dos dados na ontologia é feito por meio de três classes.

A classe *OntologyController* é responsável pela interface entre o sistema e a ontologia. Para isso, utilizou-se a API do Apache Jena[17], um *framework* JAVA *open source* que possui suporte a ontologias na linguagem OWL. Em conjunto com esta API, foi utilizado o motor de inferência *Pellet*, o qual permitiu a execução da ontologia, e a extração de conhecimento a partir das relações e regras implementadas, e das consultas implementadas na linguagem SPARQL. A classe *DataHandler* insere na ontologia as informações recebidas pelo serviço no formato JSON. A classe *InferenceLayer* é responsável por executar o motor de inferência e extrair conhecimento implícito nos dados. Além disso, implementa a consulta dos dados da ontologia utilizada pelo serviço para fornecer estas informações a plataformas externas.

A ontologia Prov-SE-O faz parte deste serviço, e foi desenvolvida com o auxílio da ferramenta Protégé[18], utilizando a linguagem OWL. Esta ontologia foi criada a partir da importação da ontologia ProvONE, e em seguida foram adicionadas novas classes relações, *property chains* e regras em SWRL (*Semantic Web Rule Language*) para permitir que a ontologia englobasse os experimentos científicos como um todo, e fosse capaz de explicitar conhecimentos implícitos sobre a proveniência dos mesmos. Além disso, para que esta ontologia pudesse modelar também os elementos de contexto dos experimentos, os dados de contexto definidos no *framework Context-SE* foram adicionados à ontologia como *object properties*.

Para permitir que usuários e agentes de software possam descobrir, invocar e compor serviços utilizando o *Prov-SE-Service*, este serviço foi descrito semanticamente através da ontologia OWL-S (FILHO e FERREIRA, 2009). OWL-S é uma ontologia padrão da W3C para a descrição serviços *web*, a qual descreve semanticamente o perfil do serviço, como o serviço pode ser utilizado, e como é possível interagir com o serviço (MARTIN *et al.*, 2004). Utilizou-se o *framework* Jersey para a criação automatizada do arquivo WADL (*Web Application Description Language*) no Netbeans. As ontologias de descrição dos serviços foram feitas manualmente. As ontologias *Service*, *Profile*, *Process* seguem o padrão OWL-S (MARTIN *et al.*, 2004). A ontologia *Grounding*, a qual descreve como interagir com o serviço, utiliza uma abordagem alternativa para mapeamento de arquivos WADL à OWL-S proposta por (FILHO e FERREIRA, 2009), uma vez que a OWL-S suporta apenas serviços em WSDL (*Web Services Description Language*). A anotação semântica deste serviço, utilizando OWL-

[17] https://jena.apache.org/

[18] https://protege.stanford.edu

S, e o modelo de proveniência utilizado baseado no modelo PROV melhoram sua interoperabilidade com plataformas externas.

A correta interpretação das informações obtidas por este serviço, e sua transformação em conhecimento pelo pesquisador depende diretamente da forma como essas informações são apresentadas. Com o propósito de apoiar a interpretação dos dados pelos pesquisadores, e apoiar o reúso dos experimentos, foram implementadas na interface do usuário visualizações específicas para a apresentação de informações de contexto e proveniência dos experimentos. Essas visualizações são apresentadas a seguir.

3.4.4.4. Visualizações

Os componentes de visualização foram desenvolvidos utilizando a biblioteca D3[19], uma biblioteca JavaScript para manipulação de documentos baseados em dados. Esses componentes possuem propriedades dinâmicas que permitem a interação do usuário para uma melhor interpretação dos dados.

Na tela de visualização dos experimentos, foi utilizado um componente que oferece a visualização em forma de grafo para apresentar as informações de proveniência do experimento. O experimento que está sendo visualizado é representado pelo nó ao centro do grafo. Os demais nós representam os objetos que se relacionam com este experimento, entre eles podemos citar: *workflows*, atividades, documentos, pesquisadores, outros experimentos, etc. As arestas do grafo representam as relações entre o experimento e os outros objetos presentes na plataforma. As arestas tracejadas representam relações inferidas pela ontologia.

Esta visualização se baseia no modelo PROV. Cada nó possui um ícone que o relaciona com as classes definidas no modelo Prov, e as relações entre os nós seguem a nomenclatura deste modelo. Conforme ilustra a Figura 20, o Experimento que é o nó principal por exemplo, recebe o ícone de uma oval amarela, usada no PROV para representar as entidades. Enquanto o nó que representa a Universidade Federal de Minas Gerais, por exemplo, recebe como ícone um pentágono laranja, usado para representar os agentes. A relação entre o experimento e a Universidade, *was Influenced By,* é uma das relações utilizadas pelo PROV para representar o relacionamento entre uma entidade e um agente. Através de duplo clique no nó principal, o pesquisador pode visualizar as informações de contexto deste objeto. E com um duplo clique nos demais nós o pesquisador pode navegar para a visualização de proveniência

[19] https://d3js.org

deste nó. Assim, o pesquisador pode por exemplo navegar do experimento para o workflow, e do *workflow* para uma de suas atividades.

Acredita-se que este componente auxilia os pesquisadores na reutilização dos experimentos, ou partes dos experimentos, uma vez que permite uma visualização completa tanto das informações de proveniência como dos elementos de contexto de todos os objetos cadastrados na plataforma. E a navegação nesse grafo permite que o pesquisador possa rastrear a proveniência dos objetos, aumentando ou diminuindo o nível de profundidade das informações visualizadas.

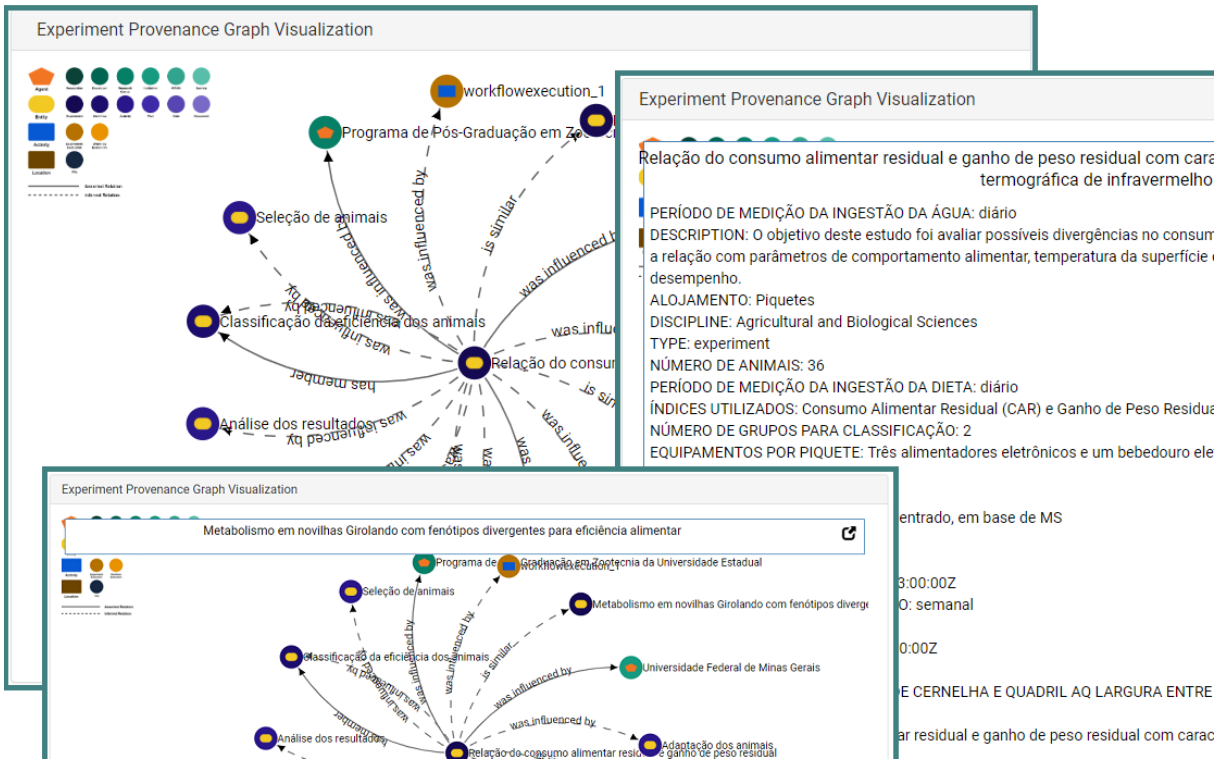

Figura 20. Grafo de visualização de proveniência

Com relação aos *workflows*, foi implementado um componente de visualização que apresenta o conjunto de atividades do *workflow* e a ordem de execução dessas atividades. Conforme pode ser observado na Figura 21, as atividades em laranja são atividades reutilizadas de outros experimentos. Enquanto as atividades em verde representam atividades que pertencem ao *workflow* que está sendo visualizado, e que já foram reutilizadas por outros experimentos. Ao passar o mouse sobre a atividade, é possível visualizar maiores informações sobre a atividade. Desta forma, esta visualização traz informações de proveniência e contexto do *workflow* e das atividades, em um nível alto de abstração, fornecendo ao pesquisador uma visão ampla do processo de experimentação e de reúso.

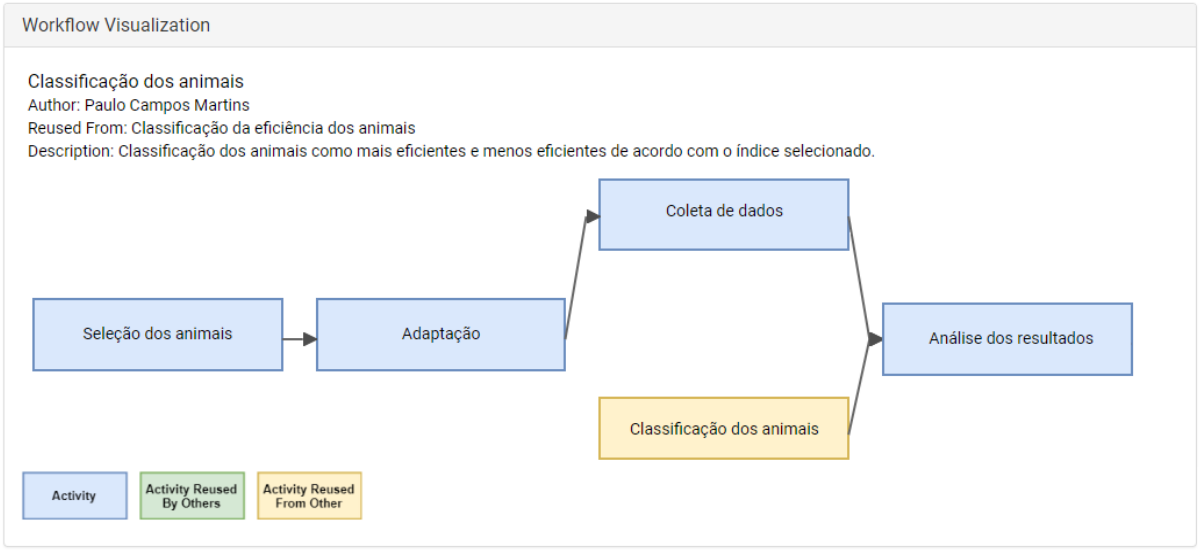

Figura 21. Visualização do *workflow*

Na tela de visualização de entidades, que são documentos ou dados produzidos ou consumidos pelo experimento, um componente interativo permite a visualização de sua derivação e reutilização. Conforme ilustra a Figura 22, o círculo em azul representa o *workflow* onde o documento, que neste caso é uma planilha, foi gerado. Por outro lado, o círculo em laranja representa o *workflow* onde a planilha foi reutilizada. Este componente também traz, em cinza, a atividade onde a entidade foi gerada ou reutilizada. Ao passar o mouse sobre o *workflow* ou sobre a atividade, é possível visualizar a descrição, o autor e a dada e hora na qual a entidade em questão foi gerada ou reutilizada. Esta visualização é importante para que os pesquisadores possam rastrear e analisar todo o processo de produção dos documentos, ou dados, cadastrados na plataforma. Como resultado, podem decidir se este processo é seguro e confiável o bastante para que este documento, ou dados, possa ser reutilizado em seu experimento.

Um componente de visualização no formato de gráfico de acordes ilustra o relacionamento entre os pesquisadores e os experimentos cadastrados na plataforma. Clicando em um dos experimentos o componente apresenta todos os pesquisadores que participaram do experimento. Clicando em um pesquisador (coluna esquerda) o componente apresenta todos os experimentos (coluna direita) dos quais o pesquisador participou. O componente auxilia na visualização da proveniência dos experimentos, e permite que ao identificar os responsáveis pelo experimento, o pesquisador possa entrar em contato com esses pesquisadores caso tenha algum problema ou dificuldade durante o reúso deste experimento. Esta visualização permite ainda identificar qual pesquisador é mais atuante na área, de acordo com o número de

Figura 22. Visualização de Entidades (Dados ou Documentos)

experimentos dos quais ele participou. Isso pode ser importante caso o pesquisador esteja procurando por outros pesquisadores para colaborarem com seu experimento. E também mensurar a experiência do pesquisador, o que influencia diretamente na confiabilidade de seus experimentos para o caso de reúso. A Figura 23 ilustra este componente de visualização.

Outro componente de visualização dos pesquisadores cadastrados na plataforma foi desenvolvido utilizando o formato de grafo, para indicar a colaboração entre os pesquisadores. Este grafo é direcionado, de forma que as arestas que chagam ao nó do pesquisador indicam os pesquisadores que colaboraram com ele na realização de seu experimento. As arestas que saem do nó do pesquisador apontam para os pesquisadores com os quais ele colaborou. Este componente é interativo, e permite, com um duplo clique no nó do pesquisador, visualizar os experimentos de sua autoria. A Figura 24 ilustra este componente.

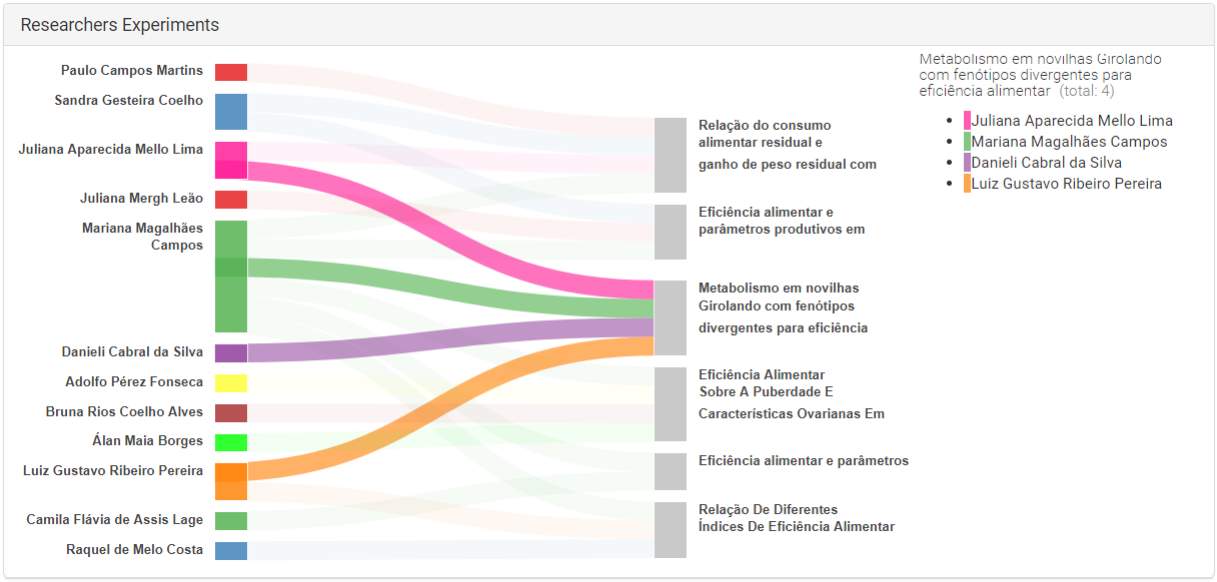

Figura 23. Visualização do relacionamento entre Pesquisadores e Experimentos

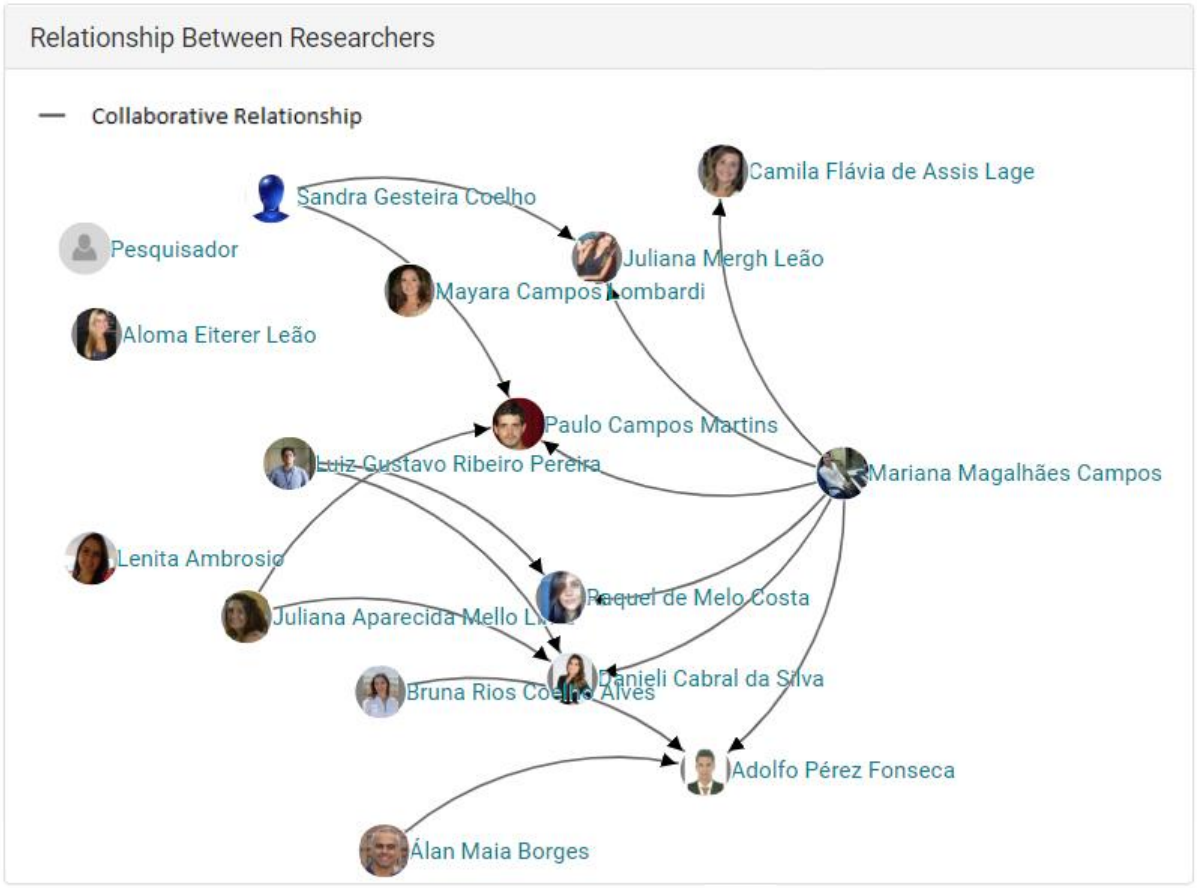

Figura 24. Visualização do relacionamento entre Pesquisadores

3.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou uma abordagem de gerenciamento de proveniência e contexto em plataformas de ECOSC, desenvolvida com o objetivo de apoiar o reúso de conhecimento gerado a partir do processo de experimentação conduzido na plataforma. Inicialmente, foram apresentados os elementos teóricos desenvolvidos para a solução, e em seguida o projeto e a implementação da solução na plataforma E-SECO. O capítulo seguinte apresenta a avaliação da abordagem proposta realizada através de estudos de caso.

4 AVALIAÇÃO DA SOLUÇÃO

Este capítulo descreve a avaliação da solução proposta e implementada na plataforma E-SECO. Primeiramente apresenta as hipóteses de pesquisa e a metodologia utilizada na avaliação das mesmas, em seguida descreve detalhes sobre a condução da avaliação, e por fim os resultados obtidos.

4.1 DEFINIÇÃO DO ESCOPO

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 definido com base na estrutura do método GQM (BASILI e WEISS, 1984) da seguinte forma:

'**Analisar** a influência do gerenciamento de contexto e de proveniência de dados **com o propósito de** apoiar a reutilização de experimentos científicos **sob o ponto de vista de** pesquisadores **no contexto de** uma plataforma de ecossistema de software científico'.

Atendendo ao escopo especificado, a seguinte questão de pesquisa (QP) foi elaborada: 'Como o gerenciamento de contexto e de proveniência de dados potencializa a reutilização de experimentos em uma Plataforma de Ecossistema de Software Científico?'. A partir desta questão, foram definidas a hipótese nula (H0) e a hipótese alternativa (H1):

H0. 'O gerenciamento de contexto e de proveniência de dados não potencializa a reutilização de experimentos em uma Plataforma de Ecossistema de Software Científico.'

H1. 'O gerenciamento de contexto e de proveniência de dados potencializa a reutilização de experimentos em uma Plataforma de Ecossistema de Software Científico.'

A escolha da melhor estratégia de pesquisa depende de basicamente de três condições: a) o tipo de questão da pesquisa; b) o controle que o pesquisador possui sobre os eventos comportamentais efetivos; c) o foco em fenômenos históricos, em oposição a fenômenos contemporâneos (YIN, 2015). Visto que a questão apresentada é do tipo 'COMO', o que remete a questões explanatórias, o pesquisador tem pouco controle sobre os eventos, e o foco se encontra em fenômenos contemporâneos inseridos em algum contexto da vida real, a estratégia mais adequada para responder à questão previamente apresentada é o estudo de caso.

Segundo YIN (2015), "um estudo de caso é uma investigação empírica que investiga um fenômeno contemporâneo dentro de seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos". Especificamente na área de Engenharia de Software, RUNESON *et al.* (2012) apresentam um

estudo de caso como sendo “uma investigação empírica que se baseia em múltiplas fontes de evidência para investigar uma instância (ou um pequeno número de instâncias) de um fenômeno de engenharia de software contemporâneo em seu contexto da vida real, especialmente quando os limites entre o fenômeno e contexto não estão claramente definidos”. Desta forma, estudos de caso foram conduzidos baseados nas definições apresentadas.

4.2 CENÁRIO DA AVALIAÇÃO

Em 2014 foi iniciado na Empresa Brasileira de Pesquisa Agropecuária (Embrapa), um projeto de pesquisa focado em eficiência alimentar (EA) para bovinos leiteiros. Segundo CAMPOS *et al.* (2015), os gastos com alimentação representam o principal custo da atividade pecuária, diferenças entre os animais na conversão da dieta consumida em leite e carne são de grande relevância. Animais que utilizam os alimentos de forma mais eficiente necessitam consumir menos para atingir o mesmo nível de produção e, dessa forma, são mais lucrativos e produzem mais alimento por unidade de área. Além disso, o aumento da eficiência alimentar proporciona menor desperdício e excreção de nutrientes, com implicações ambientais positivas.

Pesquisadores da Embrapa trabalham nesse projeto em conjunto com estudantes de mestrado em Zootecnia e em Ciência Animal da Universidade Federal de Minas Gerais (UFMG) e da Universidade Estadual do Sudoeste da Bahia (UESB). A fim de calcular diferentes índices de EA, e analisar a relação destes índices com diferentes características desses animais, os pesquisadores precisam realizar experimentos em campo. Estes experimentos são realizados na Fazenda Experimental e no Complexo Multiusuário de Bioeficiência e Sustentabilidade da Pecuária da Embrapa, localizada em Coronel Pacheco, Minas Gerais, Brasil.

A realização de experimentos em campo demanda muito tempo, uma vez que os animais precisam ser observados durante meses, para que se consiga coletar todos os dados necessários. Além disso, esses experimentos têm um alto custo financeiro, pois é preciso manter os animais e toda a infraestrutura durante o período de experimentação. Desta forma, o reúso de experimentos é uma estratégia importante para maximizar os resultados obtidos a partir de cada experimento.

Considerando o cenário apresentado, a avaliação da solução proposta no presente trabalho foi realizada através de um estudo de caso de experimentos de Eficiência Alimentar conduzidos pela Embrapa.

4.3 PLANEJAMENTO

Visto que os experimentos no cenário apresentado demandam meses para serem concluídos, seria difícil acompanhar um experimento durante todo o período de sua execução em tempo hábil para a avaliação do presente trabalho. Além disso, seria necessário acompanhar casos nos quais ocorrem o reúso de experimentos. Desta forma, optou-se por realizar o estudo de caso utilizando as informações de experimentos já concluídos.

As evidências para um estudo de caso podem vir de seis fontes distintas: documentos, registros em arquivo, entrevistas, observação direta, observação participante e artefatos físicos. O uso dessas seis fontes requer habilidades e procedimentos metodológicos sutilmente diferentes. A utilização de múltiplas fontes permitiu o desenvolvimento de linhas convergentes de investigação, chamado de triangulação de dados (YIN, 2015).

No presente estudo, foram utilizados como fonte de coleta de dados: (i) documentos, (ii) registros em arquivo, (iii) entrevistas informais, (iv) questionário de avaliação e (v) observação direta. Vale ressaltar que, os experimentos foram conduzidos anteriormente à realização deste estudo. Portanto, para que fosse possível a observação direta, foi solicitado aos participantes que fizessem uma nova análise dos dados destes experimentos através da plataforma.

Conforme descrito no escopo do projeto, este estudo pretende avaliar a abordagem proposta sob o ponto de vista dos pesquisadores. Desta forma, participaram deste estudo pesquisadores da Embrapa e da UFJF, com diferentes níveis de conhecimento em relação a experimentação científica, proveniência de dados e análise de contexto. Todos os participantes assinaram um termo de consentimento, conforme apresentado no Apêndice A, e responderam ao questionário de caracterização do participante, presente no Apêndice C.

O objetivo do estudo de caso foi avaliar se a abordagem proposta apresentou benefícios de apoio a reutilização dos experimentos. Para isso, a abordagem *ContextProv* foi apresentada aos pesquisadores. Em seguida, foi solicitado que eles analisem os experimentos cadastrados na plataforma, bem como suas informações de proveniência e contexto, com o propósito de reutiliza-los em experimentos futuros. Durante essa análise, foi feita a observação direta da interação dos pesquisadores com a plataforma. Por fim, foram feitas as entrevistas e aplicado o questionário. O questionário de avaliação utilizado é apresentado no Apêndice D. A fim de obter mais informações sobre o processo de experimentação e sobre o reúso sem o apoio da plataforma, este questionário também contempla questões sobre este processo.

Segundo YIN (2015), conduzir um estudo de caso piloto auxilia os pesquisadores na hora de aprimorar os planos para a coleta de dados tanto em relação ao conteúdo dos dados quanto aos procedimentos que devem ser seguidos. Dessa forma, a fim de identificar possíveis incoerências neste planejamento, antes de realizar o estudo de caso (regular), foi realizado um estudo de caso piloto.

4.4 PREPARAÇÃO

Antes da condução do estudo de caso, foi preciso analisar o processo de experimentação científica conduzido sem o uso da abordagem proposta, focando principalmente nos experimentos que fizeram reúso de partes de outros experimentos anteriores, e selecionar os casos a serem estudados. Além disso, foi preciso coletar os dados de proveniência e contexto dos experimentos selecionados. Para isso, foram utilizadas as seguintes fontes de coleta: documentos, registros em arquivo e entrevistas.

Os documentos e registros em arquivo analisados foram as dissertações de mestrado dos pesquisadores que conduziram os experimentos, e planilhas com os dados coletados durante a realização de cada um. As entrevistas foram conduzidas de maneira informal, durante reuniões com os pesquisadores, e uma visita a Fazenda Experimental em Coronel Pacheco. A partir dessa análise, foi possível entender melhor os experimentos e identificar a ocorrência do reúso no processo utilizado.

A partir da análise dos documentos, foi possível observar que os pesquisadores utilizam um sistema automatizado para a coleta de informações sobre os animais durante a execução do experimento. Entretanto, não era utilizada nenhuma ferramenta de apoio ao processo de experimentação científica capaz de gerenciar todas as etapas desse processo. Assim, as informações sobre o planejamento, prototipação, e execução dos experimentos ficavam exclusivamente sob a responsabilidade dos pesquisadores. Ao final do experimento, os resultados obtidos eram publicados através de artigos, dissertações e teses desses pesquisadores. O pesquisador que quisesse reutilizar algo, precisava ler todas as publicações sobre o experimento, e buscar pelas informações de contexto. Além de ser um processo demorado, a confiabilidade desses dados era prejudicada, pois muitas informações contextuais importantes para o reúso podiam ser omitidas nas publicações.

Após analisar todo o material disponibilizado, identificou-se casos nos quais ocorreu o reúso de partes de outros experimentos, o que é bastante interessante para o presente estudo de caso. Então, foram selecionados três experimentos realizados no período de abril de 2015 a outubro de 2016 para compor este estudo.

No ‘Experimento 1’, o pesquisador teve como objetivo estudar divergências no consumo alimentar residual (CAR) e ganho de peso residual (GPR) em novilhas F1 Girolando e as relações com consumo, desempenho, comportamento alimentar e de ingestão de água, temperatura da superfície do corpo e características morfológicas. Este experimento ocorreu entre abril e agosto de 2015 na Fazenda Experimental da Embrapa através da execução de um *workflow* composto pelas seguintes atividades: (i) seleção de animais; (ii) adaptação dos animais; (iii) coleta dos dados; (iv) cálculo do CAR; (v) cálculo do GPR; (vi) classificação e (vii) análise dos resultados.

No ‘Experimento 2’, o objetivo foi avaliar os efeitos da divergência fenotípica para eficiência alimentar (CAR, GPR e eficiência de conversão alimentar - ECA), na digestibilidade, emissão de metano entérico (CH4), partição energética, produção de calor (PC), parâmetros sanguíneos, metabolismo de nitrogênio e temperaturas de diferentes regiões corporais em novilhas Girolando em condições tropicais. Este experimento ocorreu entre agosto e novembro de 2015 no Complexo Multiusuário de Bioeficiência e Sustentabilidade da Pecuária da Embrapa.

Tendo em vista o alto custo financeiro e o tempo para realizar um novo experimento de EA, e sabendo que já havia ocorrido um experimento de EA em novilhas Girolando F1, o pesquisador optou por reutilizar as atividades já realizadas no experimento anterior e continuar seu experimento com os mesmos animais. Entretanto, uma novilha precisou ser removida do segundo experimento devido à uma fratura de membro posterior direito, o que impediu o pesquisador de reutilizar a etapa de classificação dos animais. Assim, foram reutilizadas apenas as três primeiras etapas do experimento anterior. O cálculo dos índices de EA foram refeitos, incluindo também o índice ECA, e foi feita uma nova classificação dos animais. Por fim, o pesquisador comparou a classificação com os parâmetros de interesse de sua pesquisa.

No Experimento 3, o objetivo foi abordar os parâmetros reprodutivos em novilhas F1 HG (Girolando) desde a puberdade até a primeira concepção e correlacionar a eficiência nutricional com alguns desses parâmetros. Este experimento ocorreu entre agosto de 2015 e outubro de 2016 na Embrapa, em duas etapas. Na primeira etapa, o objetivo foi estudar os parâmetros reprodutivos até a primeira concepção, incluindo-se idade e peso à puberdade, comportamento de estro, dinâmica folicular, resposta a protocolos hormonais e fertilidade. Esta etapa foi realizada através das seguintes atividades: (i) coleta dos dados e (iii) análise dos resultados.

O objetivo da segunda etapa do Experimento 3 foi relacionar a EA com parâmetros reprodutivos em novilhas F1 HG. Nesta fase, o pesquisador observou que já existiam

experimentos que mediram a eficiência alimentar desse mesmo grupo de novilhas. Então reutilizou atividades desses experimentos anteriores que coletaram os dados de consumo desses animais em duas idades diferentes, pós-desaleitamento com idade de 149,1 ± 30,49 dias a 337,6 ± 30,2 dias (Experimento 1) e pós-puberdade com idade de 530,7 ± 27,8 dias a 610,8 ± 27,8 Dias (Experimento 4 - não apresentado). Assim, o pesquisador precisou realizar apenas as atividades de classificação e análise dos resultados.

Através da análise dos documentos e das entrevistas, foram coletadas as informações de proveniência e contexto dos experimentos selecionados, de acordo com os dados modelados pelo *framework Context-SE* e com a ontologia *Prov-SE-O*, para o estudo de caso, por exemplo: índices de EA e o número de grupos utilizados para a classificação; a duração do período de adaptação dos animais; o número de animais que participaram no experimento; a idade e o peso inicial dos animais; e informações sobre a composição da dieta dos animais. Essas informações, entre outras coletadas, foram cadastradas na plataforma, e utilizadas na condução dos estudos de caso apresentados a seguir.

4.5 ESTUDO DE CASO PILOTO

Para a realização do estudo de caso piloto, buscou-se participantes com algum conhecimento sobre experimentação científica, e com diferentes níveis de conhecimento sobre proveniência de dados e análise de contexto de experimentos científicos. Assim, este estudo contou com a participação de três pesquisadores da UFJF.

Conforme pode-se observar no Gráfico 1, o questionário de caracterização do participante pode constatar que os participantes têm conhecimento moderado ou bom em experimentação científica, proveniência de dados e análise de contexto.

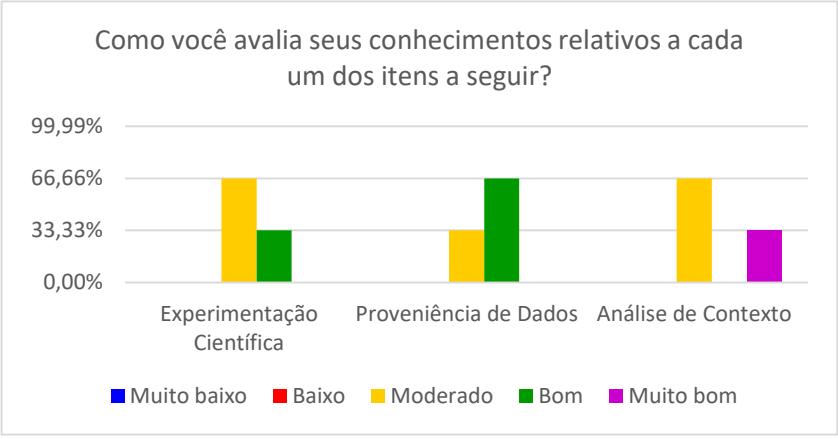

Gráfico 1. Caracterização dos participantes (Estudo de Caso Piloto)

A plataforma E-SECO e os recursos de gerenciamento de proveniência e contexto da abordagem *ContextProv* foram apresentados aos participantes. Em seguida, os pesquisadores tiveram acesso à plataforma, através da qual puderam interagir com as visualizações e analisar as informações de proveniência e contexto do processo de experimentação realizado no 'Experimento 1'. O problema de pesquisa tratado no 'Experimento 2' foi apresentado aos participantes, e foi solicitado que eles avaliassem as informações apresentadas pela plataforma considerando a possibilidade de reúso de partes dos experimentos. Após esta análise, os participantes responderam ao questionário de avaliação da solução. A fim de sanar quaisquer dúvidas dos participantes tenham tido ao longo da avaliação, após responderem ao questionário os participantes foram entrevistados.

4.5.1. Resultados obtidos

A seguir são apresentados os resultados obtidos pelo estudo de caso piloto. Estes resultados foram divididos de acordo com as fontes de coleta utilizadas.

4.5.1.1. Observação direta

Através da observação direta, foi possível perceber que os pesquisadores conseguiram encontrar as informações de proveniência e contexto para avaliar a possibilidade de reúso de partes dos experimentos. Todos os pesquisadores optaram por utilizar os atalhos apresentados na página inicial para acessar as informações, isso indica que o componente de atalho pode facilitar o uso da plataforma por usuários iniciantes.

Entretanto, foram observadas algumas dificuldades no uso da plataforma. No grafo de visualização de proveniência por exemplo, nenhum dos participantes conseguiu acessar os detalhes do grafo e navegar para outros itens. Esses recursos são acessíveis com um duplo clique no item. O fato dos pesquisadores não conseguirem acessar estes recursos pode indicar que o treinamento oferecido não foi completo o suficiente, ou um problema de usabilidade no componente. Um dos pesquisadores também relatou dificuldade com as cores da legenda deste grafo, pois são utilizadas cores muito parecidas, o que dificultou diferenciar os itens. Esse mesmo pesquisador também teve dificuldade na interpretação da visualização de proveniência das entidades. Ele ficou em dúvida sobre o que os círculos em cinza representam na visualização. Outro relatou dificuldade em encontrar os experimentos similares, visto que esta informação só está presente no grafo de proveniência.

4.5.1.2. Questionário

A análise dos resultados obtidos através do questionário foi dividida de acordo com os aspectos da solução que estão sendo avaliados.

A primeira parte, composta pelas questões 1, 2 e 3, buscou compreender melhor o processo de experimentação sem o uso da plataforma E-SECO. As respostas obtidas indicaram que os experimentos eram executados sem nenhuma plataforma de apoio, e por isso, segundo um dos participantes "*o processo era complexo, pouco eficiente e com muitas chances de erro*". Também foi possível observar que os pesquisadores consideram muito importante o reúso nesse domínio, foram citados motivos como: a possibilidade de obter comparativos entre cenários variados sem a necessidade de construí-los; a economia de recursos, principalmente de tempo; e permitir observar aspectos até então não observados. Apesar disso, os resultados mostram que o reúso nesse domínio ainda enfrenta muitos desafios. Conforme pode ser observado na Tabela 6.

Tabela 6. Avaliação do reúso sem o apoio da plataforma E-SECO (Estudo de Caso Piloto)

Como você avalia o processo de reúso de experimentos científicos adotado atualmente?
3 respostas
O reúso atualmente é baixo devido à dificuldade no acesso aos dados e a insegurança em utiliza-los, uma vez que, por muitas vezes, são omitidos os critérios adotados no experimento para a confecção dos dados.
O processo atual, que é centralizado em uma pessoa, não é nenhum pouco confiável. O ser humano não é eficiente ao lidar com dados e está sujeito à várias adversidades, dentre elas, a morte. Fora isso, há o problema da dificuldade em se utilizar o que foi feito anteriormente e apenas documentado em dissertações. Os textos de dissertações podem não abordar em sua plenitude toda a experimentação, deixando de fora pontos importantes.
Por não utilizar uma plataforma, analisar dados de experimentos científicos pode se tornar uma tarefa mais complexa. Acredito que o processo atual de reúso não é o ideal, por não permitir analisar dados contextuais.

As questões 4, 5, 6 e 7 buscaram avaliar se as informações contidas no *framework* de contexto são relevantes e suficientes para apoiar o reúso de experimentos científicos. Conforme ilustra o Gráfico 2, as respostas obtidas para a questão 4 indicaram que para a maioria dos itens, as informações presentes no *framework* são suficientes para o reúso desses experimentos. Exceto para o item 'Instituição', para o qual um dos pesquisadores respondeu que as informações apresentadas são indiferentes para o reúso dos experimentos. As respostas para a questão 5 confirmam este resultado visto que a maioria dos pesquisadores disseram não ser necessário acrescentar nenhuma informação aos itens apresentados. Um dos pesquisadores sugeriu a inserção de um campo de descrição textual sobre as instituições. Segundo o

pesquisador, este campo poderia auxiliar na identificação de instituições que realizam experimentos semelhantes, contribuindo assim com o reúso dos experimentos.

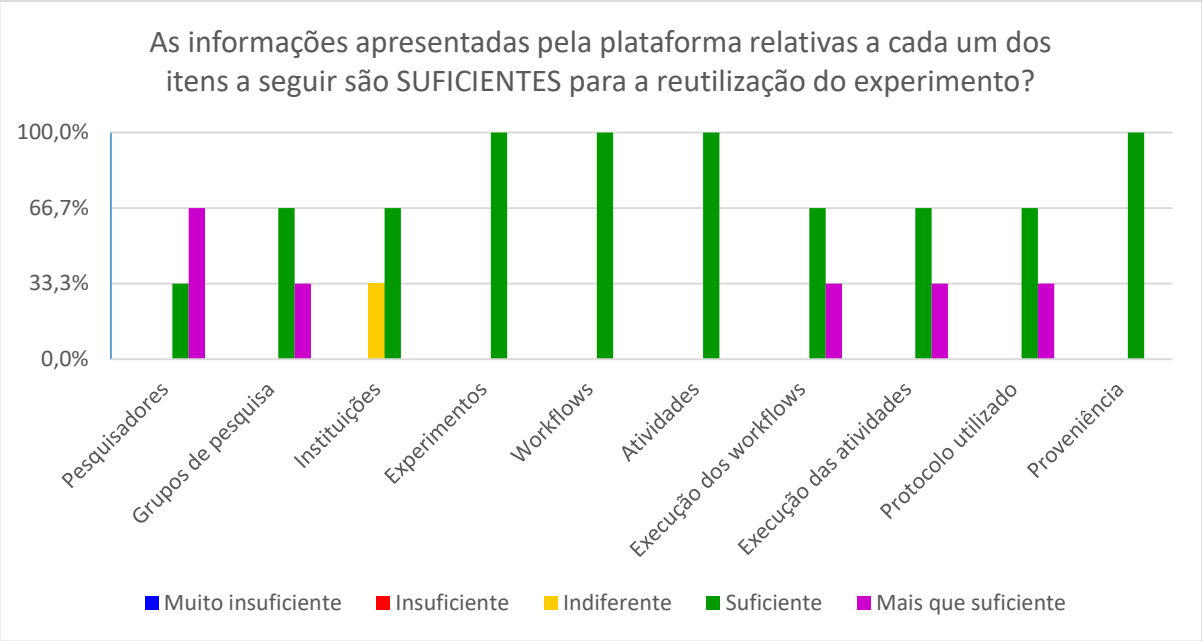

Gráfico 2. Avaliação do *framework* de contexto (Estudo de Caso Piloto)

As questões 6 e 7 visaram identificar informações presentes no *framework* que não são relevantes para o reúso de experimentos científicos. O excesso de informações pode sobrecarregar o espaço de trabalho dificultando a análise do usuário, e prejudicando sua compreensão. De acordo com as respostas obtidas, 70% das informações apresentadas foram consideradas relevantes, ou muito relevantes, por todos os pesquisadores. As informações sobre os pesquisadores, grupos de pesquisa e instituições foram consideradas irrelevantes por um dos pesquisadores. A resposta desse pesquisador para a questão 7, conforme apresenta a Tabela 7, justifica a sua posição. Os demais pesquisadores afirmaram não ser necessário remover nenhuma informação da plataforma.

Tabela 7. Avaliação da relevância das informações do *framework* de contexto (Estudo de Caso Piloto)

Você removeria alguma informação? Quais?
3 respostas
O meu ponto de vista é que dados pessoais e institucionais são irrelevantes ao reúso para esse contexto, pois todos os pesquisadores e pesquisas são relevantes ao processo, mesmo que não haja tanta proximidade entre os participantes. Dessa forma, eu removeria esses três itens iniciais.
Não removeria
Não.

As questões 8 a 13 objetivaram avaliar se as informações inferidas pela ontologia podem apoiar o reúso de experimentos científicos. Com relação à apresentação de experimentos similares 66,7% dos pesquisadores afirmaram ser muito relevante, e 33,3% afirmaram ser relevante para o reúso dos experimentos. Sobre a apresentação das atividades responsáveis pela criação de um documento, 100% dos pesquisadores consideram ser muito relevante. A apresentação (i) das atividades que já reutilizaram um documento, (ii) das atividades do experimento que foram reutilizadas DE outro experimento e (iii) das atividades do experimento que foram reutilizadas POR outros experimentos foram consideradas relevantes por 66,7% dos pesquisadores, e muito relevante por 33,3%. Por outro lado, a apresentação das instituições envolvidas na realização dos experimentos foi considerada relevante para 66,7%, enquanto 33,3% consideraram indiferente para o reúso do experimento.

A questão 14 teve como objetivo avaliar se as informações provenientes da integração com o Mendeley são relevantes para apoiar o reúso de experimentos científicos. O Gráfico 3 apresenta o resultado obtido para esta questão. É possível observar que a maior parte dos pesquisadores consideraram essas informações relevantes ou muito relevantes.

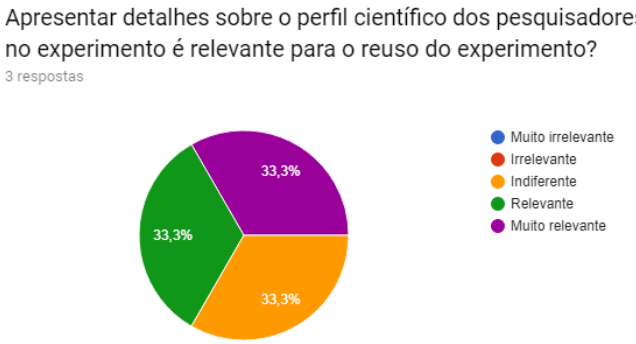

Gráfico 3. Avaliação da integração com o Mendeley (Estudo de Caso Piloto)

Sobre a integração da plataforma com o SGWfC Kepler, as questões 15 e 16 visaram avaliar se as informações obtidas através dessa integração podem apoiar o reúso de experimentos científicos. Com relação à data e hora de início e fim de cada atividade executada, 66,7% dos pesquisadores consideram essas informações muito relevantes e 33,3% consideram relevantes para o reúso dos experimentos. Sobre a apresentação das entradas e saídas (documentos utilizados e documentos produzidos) de cada atividade executada, 67,7% afirmam ser relevante e 33,3% muito relevante para o reúso dos experimentos.

As questões 17 a 21 avaliaram como as visualizações presentes na plataforma podem contribuir para o reúso de experimentos científicos. 33,3% dos pesquisadores

concordam fortemente que a visualização que relaciona os pesquisadores envolvidos com os experimentos e a visualização que apresenta a colaboração entre os pesquisadores contribuem para o reúso dos experimentos científicos. Enquanto 33,3% concordam com essa afirmação, e 33,3% a consideram indiferente. Sobre a visualização que apresenta a atividade que gerou um documento e as atividades que já reutilizaram este documento, 66,7% dos participantes concordam que essa visualização contribui para o reúso dos experimentos, e 33,7% concordam fortemente com essa afirmação. Com relação a visualização que apresenta o fluxo de atividades do experimento, todos os pesquisadores concordam que contribui para o reúso dos experimentos. E por último, sobre a visualização que apresenta as informações de proveniência do experimento, 66,7% dos participantes concordam fortemente que essa visualização contribui para o reúso dos experimentos, e 33,7% concordam com essa afirmação.

As questões 22 e 23 buscaram evidências sobre como os pesquisadores avaliam o processo de experimentação científica apoiado pela plataforma. As respostas indicam que o processo com o apoio da plataforma é mais ágil, fácil, seguro e confiável. Alguns motivos citados foram: (i) a plataforma procura levar dados e informações importantes para o pesquisador no processo da experimentação, de forma automática; (ii) as visualizações colaboram para o entendimento dos dados; (iii) a conexão com os demais pesquisadores permite entrar em contato para diversos fins; (iv) a organização e simplicidade da interface; (v) a plataforma fornece informações para apoiar as decisões sobre reúso, a partir dos dados de proveniência. Com relação ao apoio oferecido pela plataforma ao reúso de experimentos, a Tabela 8 apresenta as respostas dos pesquisadores, que indicaram que a abordagem proposta é capaz de facilitar as atividades relacionadas ao reúso dos experimentos na plataforma E-SECO.

Tabela 8. Avaliação do apoio da plataforma ao reúso dos experimentos científicos (Estudo de Caso Piloto)

Quais são os pontos observados na plataforma E-SECO que podem contribuir para a reutilização do experimento?
3 respostas
1º O acesso e a interpretação das informações dos experimentos auxiliam o pesquisador e proporcionam segurança neste reúso; 2º As visualizações proporcionam o acesso simplificado aos dados; 3º Os protocolos dos experimentos garantem que o pesquisador não tenha dúvidas no processo de reúso.
Todas as informações contextuais de cada experimento são muito úteis para o reúso. Essas informações, fora da plataforma, teriam que ser extraídas e processadas manualmente e de diversas fonte, uma atividade bastante demorada e desgastante. Na plataforma, o pesquisador tem ao seu dispor muitas dessas informações já extraídas e processadas, facilitando e gerando confiança. As informações de proveniência entram também nesse contexto, pois são igualmente importantes para que o pesquisador possa analisar o comportamento do experimento, realizado em diversos contextos. Com as informações de proveniência, o pesquisador consegue filtrar e encontrar um experimento que se adeque às suas necessidades.
Visualização de dados, análise de proveniência de dados, análise da colaboração entre pesquisadores.

A última questão buscou sugestões dos pesquisadores para a melhoria da plataforma. Um dos pesquisadores sugeriu explorar melhor os filtros nas visualizações, para permitir ao pesquisador encontrar de forma mais eficiente dados e recursos para serem reutilizados. Os demais pesquisadores consideram que a plataforma já possui recursos suficientes para apoiar o processo de experimentação e reúso.

4.5.1.3. Entrevista

Os três pesquisadores relataram ter gostado das funcionalidades e informações oferecidas pela plataforma. Um dos pesquisadores ressaltou que o grafo de proveniência permite ter uma visão geral do experimento em relação a todos os outros dados presentes na plataforma. Outro destacou a importância dos dados sobre a execução dos *workflows* e das atividades, principalmente a data e hora de início e término. Segundo ele, essas simples informações de contexto podem esclarecer diversas coisas sobre o experimento, por exemplo se um ensaio de EA ocorre no verão, é natural que os animais tenham ingerido maior quantidade de água que em outro ensaio feito no inverno.

Um dos participantes disse que considera as informações sobre pesquisadores, instituições e grupos de pesquisa irrelevantes para decisões sobre o reúso dos experimentos. Mas, quando questionado sobre a confiabilidade dos experimentos desenvolvidos por terceiros, ele afirmou ter analisado a possibilidade de reúso apenas dentro da mesma instituição. Neste caso, os dados seriam confiáveis, independente do pesquisador, ou do grupo de pesquisa ao qual ele está vinculado. Por outro lado, considerando um cenário de reúso de experimentos desenvolvidos fora da instituição, o pesquisador admitiu que estas informações podem ser relevantes.

4.5.1.4. Análise dos resultados

Os resultados obtidos indicam que o *framework* de contexto é suficiente, isto é, possui as informações necessárias para apoiar o reúso dos experimentos científicos cadastrados na plataforma. Isso pode ser observado através do questionário onde que todos os itens foram avaliados como suficientes. Com relação à relevância das informações contidas no *framework*, três itens foram avaliados como irrelevantes por um dos pesquisadores. Isso foi confirmado por ele na entrevista, na qual declarou não considerar importante para o reúso informações pessoais

sobre os pesquisadores, instituições e grupos de pesquisa. Ainda assim, a maioria dos pesquisadores consideram relevantes ou muito relevantes os demais itens analisados.

Apesar das dificuldades observadas durante a análise do grafo de proveniência pelos pesquisadores, todas as informações inferidas pela ontologia foram avaliadas como relevantes, ou muito relevantes, para o reúso pela maioria dos pesquisadores. Isso evidencia que estas dificuldades não comprometeram a análise do grafo, e que a ontologia *Prov-SE-O* é capaz de potencializar o reúso de experimentos na plataforma E-SECO.

As informações sobre o perfil dos pesquisadores também foram consideradas relevantes, ou muito relevantes, pela maioria dos participantes, dando indícios de que esta integração pode oferecer benefícios para o reúso dos experimentos. A importância dos dados sobre a execução dos *workflows* e das atividades, obtidos pela integração com o Kepler, foi destacada por um dos participantes durante a entrevista. O resultado obtido no questionário também evidencia esse fato, visto que todos os pesquisadores avaliaram esses itens como relevantes, ou muito relevantes, para o reúso dos experimentos na plataforma. Isso ressalta que a integração com esta plataforma externa pode trazer benéficos para o reúso de experimentos científicos na plataforma E-SECO.

Todas as visualizações tiveram um resultado positivo pela maioria dos pesquisadores tanto no questionário tanto na entrevista. Isso reafirma que as dificuldades observadas não comprometeram o resultado das análises, e que as visualizações contribuem para o reúso dos experimentos.

A análise apresentada revela que todos os pontos da abordagem ContextProv receberam uma avaliação positiva com relação ao reúso de experimentos científicos na plataforma E-SECO. Desta forma, este estudo piloto apresentou indícios de que as informações de proveniência e contexto apresentadas pela plataforma podem apoiar o reúso de experimentos científicos em uma plataforma de ECOSC. Também identificou melhorias para a condução do estudo de caso regular.

4.5.2. Ajustes para o estudo de caso regular

Algumas dificuldades relatadas no uso da plataforma evidenciam a necessidade de melhorar o treinamento dado aos participantes antes da avaliação. Para o estudo de caso regular, o treinamento precisa ser mais detalhado principalmente com relação ao uso das visualizações.

A caracterização do participante também ficou incompleta, visto que o estudo de caso foi feito no domínio de eficiência alimentar / nutrição animal e o formulário de caracterização do participante não avaliava os conhecimentos do pesquisador nesse domínio.

As primeiras questões do formulário, que avaliam o processo de experimentação sem o uso da plataforma E-SECO, também precisam ser reformuladas. Isso se deve ao fato de que foi utilizado o termo ‘atualmente’ para se referir a este processo (sem a plataforma), o que não ficou claro para todos os participantes e gerou dúvida durante a avaliação.

Além disso, considerando a característica distribuída do processo de experimentação, foi observado que não seria possível reunir todos os participantes envolvidos nos experimentos selecionados para a realização do estudo de caso regular. Isso ocorre porque muitos desses pesquisadores não residem nas cidades onde o estudo de caso foi conduzido. Assim, foi observada a necessidade de permitir que os participantes fizessem o treinamento e participasse da avaliação remotamente. Para isso, a plataforma foi publicada em um servidor da UFJF, permitindo o acesso remoto dos pesquisadores. Além disso, no estudo de caso regular, o treinamento foi feito através de um vídeo. Os formulários em ambas as versões do estudo de caso foram respondidos online.

4.6 ESTUDO DE CASO REGULAR

Os experimentos selecionados para o estudo de caso regular são experimentos de análise da EA de bovinos leiteiros. Desta forma, para o estudo de caso regular, buscou-se pela participação de pesquisadores com bom conhecimento no domínio de agropecuária e análise da eficiência alimentar de bovinos. Assim, o estudo foi conduzido exclusivamente com pesquisadores da Embrapa. Devido à impossibilidade de reunir todos os participantes para a condução deste estudo de forma presencial, 7 pesquisadores participaram do estudo remotamente. Considerando a importância da observação direta para o presente estudo, os demais pesquisadores participaram do estudo de forma presencial. O estudo presencial foi conduzido em dois locais: (i) na sede da Embrapa Gado de Leite em Juiz de Fora, onde contou com a participação de 2 pesquisadores; e (ii) na fazenda experimental da Embrapa em Coronel Pacheco, onde participaram 7 pesquisadores. Assim, um total de 16 pesquisadores, em sua maioria mestres e doutores em Zootecnia ou Ciência Animal, fizeram parte deste estudo de caso.

Através do questionário de caracterização do participante, conforme apresenta o Gráfico 4, foi possível constatar que: (i) os participantes têm bom conhecimento em experimentação científica; (ii) a maior parte dos pesquisadores possuem conhecimento entre moderado e bom sobre proveniência de dados e análise de contexto; e (iii) todos os pesquisadores possuem conhecimento no mínimo moderado com relação à análise da EA de bovinos.

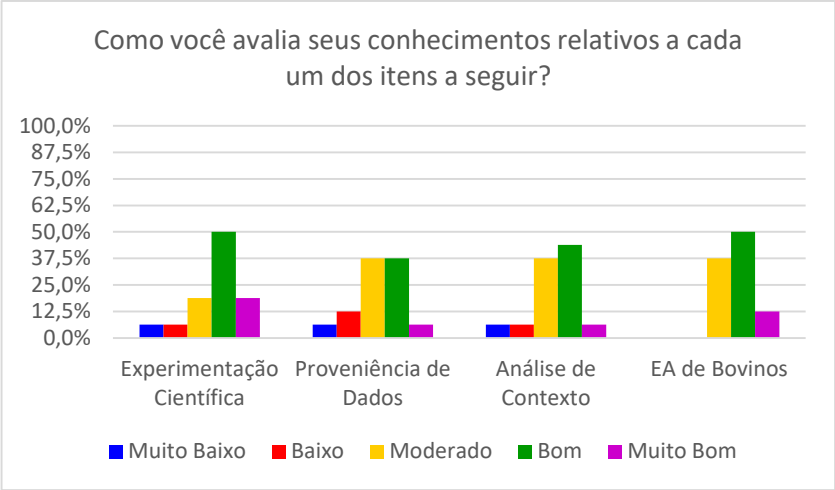

Gráfico 4. Caracterização dos participantes (Estudo de Caso Regular)

A plataforma E-SECO e os recursos de gerenciamento de proveniência e contexto da abordagem *ContextProv* foram apresentados aos participantes. Para os pesquisadores que participaram remotamente, esta apresentação foi feita por meio de um vídeo. Os pesquisadores receberam o acesso à plataforma, e puderam então interagir com sua interface de usuário para cadastrar novos experimentos, e analisar as informações de proveniência e contexto do processo de experimentação realizado no 'Experimento 1'. O problema de pesquisa tratado no 'Experimento 3' foi apresentado aos participantes. Em seguida, foi solicitado que eles avaliassem a possibilidade de reúso de partes do experimento cadastrado. Após esta análise, os participantes responderam ao questionário de avaliação da solução. A fim de obter mais evidências acerca da avaliação da solução e sanar as dúvidas dos participantes, após responderem ao questionário os pesquisadores que participaram do estudo presencial foram entrevistados (entrevista informal).

4.6.1. Resultados obtidos

A seguir são apresentados os resultados obtidos a partir da observação direta, dos formulários preenchidos e das entrevistas do estudo de caso regular.

4.6.1.1. Observação direta

Foi possível observar que os pesquisadores conseguiram encontrar rapidamente as informações relativas ao experimento que deveriam analisar. Alguns utilizaram o atalho da página inicial, enquanto outros preferiram utilizar o menu principal. Isso demonstra que ambas as formas de acesso são importantes para o sistema, e que as informações principais, que são relativas ao experimento e ao protocolo do experimento, são facilmente acessíveis.

Por outro lado, alguns pesquisadores tiveram dificuldade em encontrar informações de granularidade menor. Um dos pesquisadores relatou não ter encontrado os documentos utilizados como entrada e como saída na execução das atividades. Isso pode ter ocorrido pois os documentos, assim como os dados, são acessíveis através do item '*Entities'*, visto que a solução é baseada no modelo PROV. E, conforme foi observado na caracterização dos participantes, os pesquisadores possuem pouco conhecimento de proveniência de dados, e portando podem não conhecer o modelo PROV.

Outro pesquisador teve dificuldade no entendimento das fases do ciclo de vida dos experimentos cadastrados na plataforma. Isso pode ter ocorrido pelo desconhecimento do pesquisador em relação ao ciclo de vida de experimentos científicos proposto por BELLOUM *et al.* (2011), que é utilizado pela plataforma E-SECO. Ou ainda, é possível que no domínio de eficiência alimentar, exista outro ciclo de vida específico.

Com relação às visualizações, apesar de neste estudo o treinamento ter sido mais detalhado com relação às opções de interação das visualizações, ainda tiveram participantes que não conseguiram acessar os detalhes do grafo de proveniência. As outras visualizações foram acessadas corretamente pelos participantes.

4.6.1.2. Questionário

Os resultados obtidos a partir do questionário são apresentados de acordo com os aspectos da solução que cada questão pretende avaliar.

As questões 1, 2 e 3 tiveram o objetivo de avaliar como é o processo de experimentação científica sem o uso da plataforma E-SECO. Como resultado, a maioria das respostas indicam problemas e dificuldades nesse processo. Por exemplo: (i) "*Eficiente porém com grande dificuldade de interação e comparação dos dados obtidos e também difícil difusão dos resultados.*"; (ii) "*O processo de experimentação atualmente é complicado, falta comunicação entre os pesquisadores ou a mesma se torna complicada devido a distância ou falta de conhecimento.*"; (iii) "*Os experimentos são delineados de maneira independente, tornando o processo lento, mais susceptível a erros.*".

Com relação à prática do reúso neste domínio, os resultados indicaram que os pesquisadores consideram muito importante pois "*gerar dados, principalmente em experimentação animal é muito desgastante, tanto para o pesquisador quando para instituição, o reúso 'evita' este desgaste uma vez que permite gerar novas perspectivas dos dados gerados sem obrigatoriamente realizar outro experimento*". Além disso, o reúso "*diminui o custo, o tempo e o número de animais utilizados para o experimento, e acelera o retorno da informação*

para à população científica". Apesar disso, o reúso ainda é pouco praticado, visto que "*é trabalhoso e realizado de maneira manual*". Um dos pesquisadores relatou que "*os dados não ficam agrupados. Assim, a possibilidade de se perder informações é grande*".

As questões 4, 5, 6 e 7 buscaram avaliar se as informações contidas no *framework* de contexto são relevantes e suficientes para apoiar o reúso de experimentos científicos. De acordo com as respostas obtidas para a questão 4, apresentadas no Gráfico 5, as informações visualizadas através da plataforma E-SECO, para todos os itens listados, foram consideradas suficientes pela maioria dos pesquisadores. Dois pesquisadores marcaram vários itens como insuficientes. Entretanto, na questão 5, onde deveriam listar as informações faltantes, eles não justificaram suas respostas. As respostas destes pesquisadores podem estar relacionadas a alguma dificuldade com o uso da plataforma.

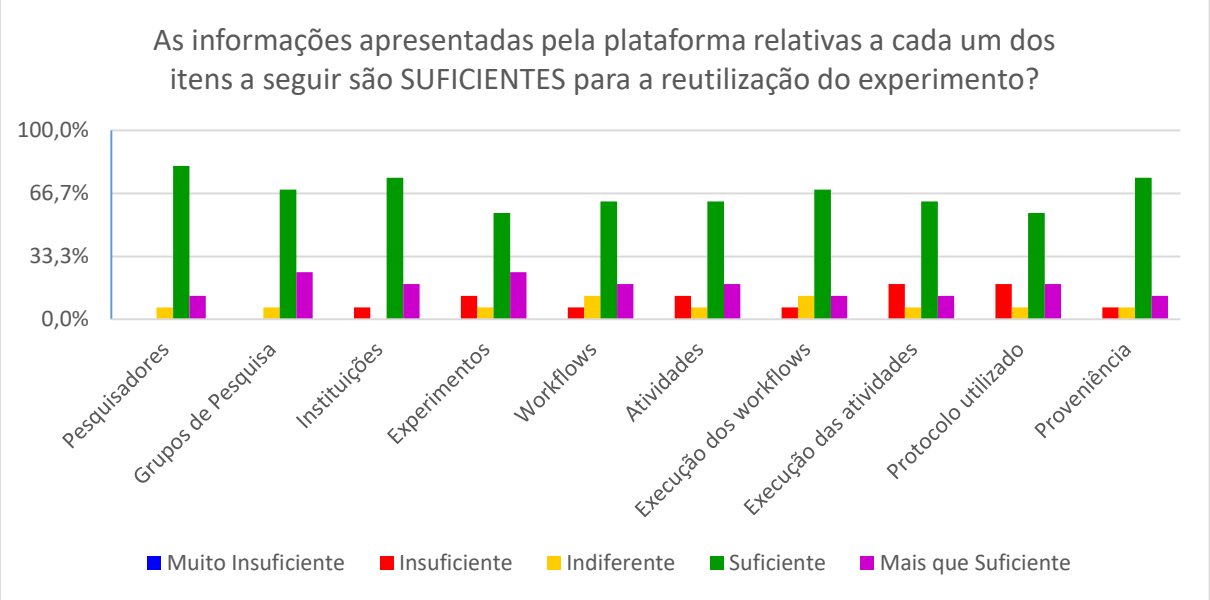

Gráfico 5. Avaliação do framework de contexto (Estudo de Caso Regular)

O excesso de informações presentes na plataforma, dificulta o dia a dia do pesquisador que precisa inserir estas informações, e ainda pode sobrecarregar o espaço de trabalho. Isso prejudica a visualização e interpretação das informações. Desta forma, as questões 6 e 7 avaliaram a relevância das informações presentes no *framework*. Na questão 6, todos os itens foram considerados relevantes por mais de 80% dos pesquisadores. 6,3% dos participantes consideraram alguns itens (Pesquisadores, Atividades, Execução dos *workflows*, Execução das atividades e Protocolo) como irrelevantes para o reúso dos experimentos. Na questão 7, estes pesquisadores tiveram oportunidade de listar as informações que consideraram irrelevantes e que poderiam ser removidas da plataforma, caso considerassem que estas informações estavam sobrecarregando o ambiente de trabalho. Entretanto, os pesquisadores não

sugeriram a remoção de nenhum item. Isso indica que eles preferem manter estes itens na plataforma, pois podem ser úteis para outras análises, ou consideram que estes itens não sobrecarrega a visualização dos dados.

As questões 8 a 13 objetivaram avaliar se as informações inferidas pela ontologia podem apoiar o reúso de experimentos científicos. A apresentação de experimentos similares foi considerada muito relevante por 50% dos participantes, relevante por 43,8% e indiferente por 6,3%. Com relação à apresentação das atividades responsáveis pela criação de um documento 62,5% dos pesquisadores consideraram relevantes, e 31,3% consideraram muito relevantes para o reúso dos experimentos. A apresentação das atividades que já reutilizaram um documento foi considerada relevante por 95% dos pesquisadores.

O Gráfico 6 apresenta a avaliação dos pesquisadores sobre a relevância da apresentação das atividades do experimento que foram reutilizadas DE outro experimento. É possível observar que a maior parte dos pesquisadores consideraram estas informações relevantes para o reúso dos experimentos. Sobre a apresentação das atividades do experimento que foram reutilizadas POR outros experimentos, um dos pesquisadores considerou indiferente. Os demais pesquisadores consideraram muito relevante (25%) ou relevante (68,8%). A apresentação das instituições envolvidas na realização dos experimentos também foi considerada relevante pela maioria dos pesquisadores.

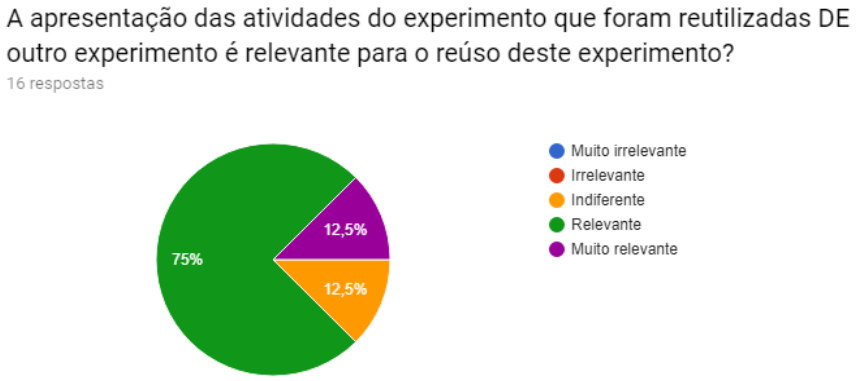

Gráfico 6. Relevância da apresentação das atividades reutilizadas (Estudo de Caso Regular)

A questão 14 teve como objetivo avaliar se as informações provenientes da integração com o Mendeley são relevantes para apoiar o reúso de experimentos científicos. A maior parte dos pesquisadores consideram que as informações relativas ao perfil científico do pesquisador são relevantes para o reúso dos experimentos.

Em relação à integração da plataforma com o SGWfC Kepler, as questões 15 e 16 buscaram avaliar se as informações obtidas através dessa integração apoiam o reúso de experimentos científicos na plataforma E-SECO. 75% dos pesquisadores consideraram que informar a data e hora de início e fim de cada atividade executada é relevante para o reúso dos experimentos. Entretanto, 18% considerou indiferente e 6,3% irrelevante. Sobre a apresentação das entradas e saídas (documentos utilizados e documentos produzidos) de cada atividade executada, conforme ilustra o Gráfico 7, a maior parte dos pesquisadores consideraram estas informações relevantes para o reúso dos experimentos.

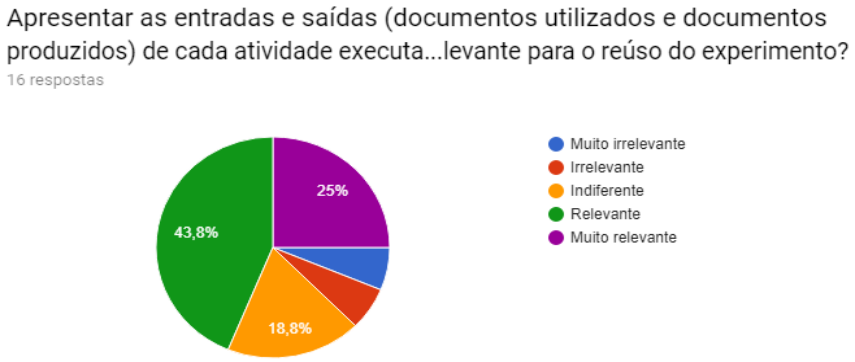

Gráfico 7. Relevância das entradas e saídas das atividades executadas (Estudo de Caso Regular)

As questões 17 a 21 avaliaram como as visualizações presentes na plataforma podem contribuir para o reúso de experimentos científicos. 31,5% dos participantes concordam fortemente e 37,5% concordam que a visualização que relaciona os pesquisadores envolvidos com os experimentos contribui para o reúso do experimento. 25% consideraram esta visualização indiferente para o reúso dos experimentos. Com relação à visualização que apresenta a atividade que gerou um documento e as atividades que já reutilizaram este documento, 68,8% concordam que esta visualização contribui para o reúso dos experimentos, e 12,5% concordam fortemente com esta afirmação. 18,8% considerou esta visualização indiferente. Sobre a visualização que apresenta o fluxo de atividades do experimento, 87,5% dos pesquisadores concordam que contribui para o reúso dos experimentos. Os demais consideraram esta visualização indiferente, ou seja, não influencia no reúso dos experimentos na plataforma. Por fim, sobre a visualização que apresenta as informações de proveniência do experimento, 12,5% dos participantes concordam fortemente que essa visualização contribui para o reúso dos experimentos, e 43,8% concordam com essa afirmação. 6,3% discordam desta afirmação e os demais pesquisadores a consideraram indiferente.

As questões 22 e 23 buscaram evidências sobre como os pesquisadores avaliam o processo de experimentação científica com o apoio da plataforma E-SECO e da abordagem *ContextProv*. As respostas para essas questões evidenciam que os pesquisadores consideram o que o processo de experimentação com o apoio da plataforma é mais organizado, permite maior exploração dos dados de um experimento e então o aprofundamento da pesquisa. De acordo com um dos pesquisadores, a plataforma oferece "*uma forma eficiente de armazenar e reutilizar dados de experimentação para dar suporte a futuros estudos relacionados às áreas*". As respostas obtidas para a questão 23 apresentam as características da abordagem proposta que oferecem apoio ao reúso dos experimentos científicos, dentre elas se destacam: (i) *a concentração de informações relacionadas a vários experimentos*; (ii) *a possibilidade de inserir os dados do protocolo do experimento e também das atividades individualmente*; (iii) *facilita o acompanhamento remoto do processo de experimentação por todos os pesquisadores envolvidos no experimento*; (iv) *facilita a reutilização de protocolos e de animais*; (v) *apresenta toda a estrutura de documentos, atividades realizadas e até mesmo as informações acerca dos grupos envolvidos*.

A última questão buscou obter sugestões dos pesquisadores para a melhoria da plataforma. A Tabela 9 apresenta as sugestões recebidas para a implementação de melhorias.

Tabela 9. Sugestões para a melhoria da plataforma (Estudo de Caso Regular)

Gostaria de deixar alguma sugestão que possa melhorar o processo de experimentação científica através da plataforma E-SECO?
6 respostas
Definição de protocolos mínimos e estruturação para receber o banco de dados dos experimentos de maneira organizada e padronizada
Abrir espaço para uma discussão ativa sobre a relevância de futuros experimentos
Identificação de todas as pessoas que utilizam a plataforma e proteção dos dados contidos nela.
Criação de um aplicativo para que os alunos de graduação interessados em desenvolver pesquisas e cursar pós-graduação possam visualizar as características do pesquisar e a área de atuação do mesmo como forma de facilitar a aproximação e uma possível orientação na pós-graduação.
Considerar em cada fase, os envolvidos. Os resultados são importantes, mas sem a mão-de-obra por trás deles, não chegariam informações a plataforma.
Sempre garantir a informação do responsável (nome) por gerar determinados dados (planilhas) disponibilizados.

4.6.1.3. Entrevista

Ao ser entrevistado um dos participantes relatou que "*o processo de experimentação utilizado na empresa está em crescente evolução, de estrutura e recursos. Apesar disso o planejamento dos experimentos e a comunicação entre os pesquisadores ainda têm muito a ser melhorado*". Outro pesquisador apontou que "*os pesquisadores enfrentam muitas dificuldades em reunir os dados para o experimento, e que durante o período de experimentação alguns deles não tomam os devidos cuidados para que os dados possam ser reutilizados por outros pesquisadores*". Sobre a importância do reúso neste domínio, um pesquisador ressaltou que "*existe uma gama de dados já gerados com grande potencial de ser avaliado sob uma nova visão gerando novos dados tão importantes quanto os gerados no primeiro estudo*". Essas declarações demonstram a necessidade de uma ferramenta de apoio ao processo de experimentação, através da qual o planejamento e execução dos experimentos sigam uma metodologia bem definida, possa ser melhorada a comunicação entre os pesquisadores, e o registro dos dados seja feito de forma segura e organizada. Além disso, demonstra a importância do reúso de experimentos dentro desta instituição.

Diante das dificuldades apresentadas, durante a entrevista os pesquisadores foram bastante receptivos em relação à plataforma. Sobre as informações armazenadas e apresentadas pela plataforma, um dos pesquisadores relatou que "*a organização e apresentação de todas as etapas do experimento facilitam o aproveitamento dos dados uma vez que diminui a perda de informações*". Segundo este participante, "*muitas informações importantes para o reúso dos dados são perdidas pois o pesquisador responsável não anotou, e acabou se esquecendo*". Além disso, de acordo com outro pesquisador, "*existe o risco de perda da própria planilha de dados, pois elas são armazenadas apenas pelo pesquisador, que muitas vezes não tem cuidados com relação a realização de backups destes dados*". Assim, este pesquisador afirmou que *a possibilidade de fazer upload dos documentos produzidos pelo experimento, e armazená-los de forma segura e acessível a todos os pesquisadores do grupo. É o principal ponto positivo da abordagem proposta em relação ao reúso dos experimentos"*. Essas respostas evidenciam que o gerenciamento das informações de proveniência e contexto dos experimentos, bem como dos próprios dados produzidos pelo experimento, contribui para a melhoria do processo de experimentação e para o reúso dos experimentos.

Outro pesquisador ressaltou que "*o reúso dos dados exige que as informações relativas ao protocolo utilizado no experimento sejam armazenadas em um nível de detalhe muito grande. Porém, informações como, por exemplo, a calibragem dos aparelhos utilizados*

na coleta dos dados, geralmente não são disponibilizadas pelo pesquisador nos relatórios ou nas publicações. Assim, neste ponto, o apoio da plataforma é muito importante". Ainda com relação ao protocolo do experimento, um pesquisador afirmou ser "*indiferente para o reúso dos experimentos, porque a plataforma dá liberdade para o pesquisador cadastrar ou não as informações*". Apesar disso, o pesquisador concordou que caso o cadastro do protocolo seja feito de forma correta, essas informações podem melhorar o processo de reúso dos experimentos.

Com relação à apresentação dos experimentos similares, um pesquisador afirmou ser "*muito importante para o reúso dos experimentos, pois ao visualizar experimentos similares o pesquisador pode verificar as atividades executadas neste tipo de experimento. Assim, se uma determinada atividade aparece em experimentos similares, então ela pode ser uma atividade importante para experimentos futuros*". Outro pesquisador afirmou que "*apresentar os experimentos similares é importante pois permite a comparação dos resultados obtidos por experimentos semelhantes realizados em contextos diferentes*". Também sobre a apresentação de experimentos similares, outro pesquisador ressaltou que "*muitas vezes os experimentos realizados dentro da instituição ocorrem com os mesmos animais. Assim, analisar os experimentos similares permite analisar os mesmos animais em diferentes contextos, e em diferentes fases de vida*. *Isso permite, por exemplo, avaliar se uma bezerra classificada como eficiente na fase de cria (0 a 3 meses) e recria (3 a 9 meses) se manteve eficiente nos ensaios de EA a pasto (9 a 15 meses) e em confinamento (15 a 18 meses).* Os relatos desses pesquisadores evidenciam a contribuição da ontologia e das informações inferidas para o processo de experimentação.

Sobre a visualização que apresenta as atividades que reutilizaram um documento, um participante afirmou que "*é muito importante para o reúso dos experimentos porque permite ao pesquisador conhecer todos os trabalhos que já foram conduzidos a partir daquele conjunto de dados, e assim evita que ele refaça a mesma análise sobre o mesmo conjunto de dados*". Outro pesquisador ressaltou a importância dessa visualização, mas com o ponto de vista de autor do documento. Neste caso ele considera "*muito importante apresentar as atividades ou experimentos que reutilizaram o documento, para que o autor possa acompanhar o impacto e contribuição de sua pesquisa para comunidade científica*". Por outro lado, um pesquisador demonstrou preocupação com relação ao reúso dos dados, e questionou sobre a disponibilidade dos dados para pesquisadores fora da instituição, e sobre a possibilidade de limitar o acesso mesmo dentro da instituição antes da publicação dos resultados. Isso demonstra que a cultura do reúso ainda não está totalmente difundida dentro da empresa.

No que se refere ao cadastro das instituições envolvidas no experimento, um participante ressaltou que "*permite ao pesquisador encontrar outras instituições interessadas em seu domínio de pesquisa, e então buscar novas conexões e novos colaboradores para o experimento*". Sobre a visualização que relaciona os pesquisadores aos experimentos, um pesquisador afirmou que "*é útil para o reúso do experimento, pois em caso de dúvida sobre o experimento, é possível entrar em contato não só com o autor principal, mas com todos os envolvidos*". Também em relação às informações referentes aos pesquisadores, mas agora sobre o seu perfil científico apresentado pela plataforma, um participante afirmou que considera essas informações irrelevantes para o reúso dos experimentos. Quando questionado sobre a confiabilidade dos experimentos, o pesquisador pontuou que "*pesquisadores com pouca experiência também podem realizar bons trabalhos, e por isso seus experimentos também deveriam ser considerados confiáveis*". Neste caso, foi esclarecido ao pesquisador, que o objetivo da plataforma é apresentar as informações de proveniência e contexto, oferecendo suporte à tomada de decisão. Mas a interpretação dos dados é uma tarefa individual dos pesquisadores.

Em relação às informações sobre a execução das atividades e dos *workflows*, um pesquisador ressaltou que "*apresentar o período de execução das atividades é muito importante, não só por ajudar identificar o contexto em que a atividade aconteceu, mas também por permitir identificar os momentos em que os animais estavam fora dos experimentos*". Segundo ele, "*é muito importante conhecer os períodos em que os animais não estão participando dos experimentos, pois nestes momentos eles podem por exemplo ficar soltos a pasto. E então sofrer influências que vão impactar os futuros experimentos realizados com estes animais*". Estes depoimentos reforçam a importância da integração com o SGWfC, e do acompanhamento da execução dos *workflows*.

4.6.1.4. Análise dos resultados

As informações de proveniência e contexto modeladas pelo *framework Context-SE* foram consideradas suficientes, para apoiar o reúso dos experimentos, pela maior parte dos pesquisadores. Nas entrevistas, diversos pesquisadores ressaltaram a importância das informações sobre os documentos produzidos durante o experimento, e principalmente sobre o protocolo do experimento. Entretanto, o protocolo do experimento ainda foi considerado insuficiente por 18,75% dos participantes. O motivo para esta avaliação foi esclarecido pelos pesquisadores durante a entrevista. Segundo eles, a liberdade dada aos pesquisadores ao preencher o protocolo pode comprometer a qualidade e a relevância dessas informações. Além

disso, foi observado que alguns pesquisadores tiveram dificuldade em encontrar informações de granularidade menor, como os documentos gerados pelas atividades. Isso se refletiu nas respostas do formulário, através das quais as informações sobre a execução das atividades foram consideradas insuficientes por 18,78% dos participantes. Contudo, as avaliações para este aspecto da solução foram em sua maioria positivas, evidenciando que as informações de proveniência e contexto modeladas pelo *framework Context-SE* são relevantes e suficientes para apoiar o reúso dos experimentos científicos na plataforma E-SECO.

Com relação à ontologia *Prov-SE-O*, apesar das dificuldades observadas no uso do grafo de proveniência, e no entendimento dos termos do modelo PROV, os resultados apresentam indícios de que a ontologia contribui para o reúso dos experimentos. Isso foi identificado no formulário, onde todas as informações geradas pela ontologia foram consideradas relevantes, ou muito relevantes, pela maioria dos pesquisadores. Além disso, durante as entrevistas, os pesquisadores ressaltaram a importância de informações como os experimentos similares, a proveniência dos documentos e os casos de reúso de documentos e de atividades.

Com relação à integração da plataforma E-SECO com a plataforma Mendeley, as respostas obtidas pelo formulário mostram que 62,5% dos pesquisadores consideram que acesso ao perfil científico do pesquisador é relevante para o reúso do experimento. As entrevistas também evidenciaram a relevância das informações sobre os pesquisadores, visto que elas facilitam o contato com os responsáveis pelo experimento para quaisquer esclarecimentos necessários, e ainda permitem que sejam dados créditos aos responsáveis pelos experimentos. Por outro lado, alguns participantes pontuaram que conhecer o perfil científico dos pesquisadores não deve ser determinante para a confiabilidade dos experimentos no caso de reúso.

A integração com a plataforma Kepler também se mostrou relevante para o reúso dos experimentos na plataforma E-SECO. De acordo com o formulário, mais de 60% dos participantes consideram que as informações coletadas durante a execução do *workflow* são relevantes para o reúso dos experimentos. Isso foi reafirmado pelos pesquisadores nas entrevistas. Nelas, eles falaram sobre a importância de conhecer os períodos em que os animais estavam em experimentação ou ociosos, e de poder armazenar os documentos produzidos durante a execução dos experimentos.

Sobre as visualizações implementadas pela solução, no formulário, todas foram consideradas relevantes para o reúso dos experimentos científicos pela maioria dos pesquisadores. Vale destacar (i) a visualização do *workflow* e (ii) a visualização da proveniência

dos documentos, que foram consideradas relevantes por 87,5% e 81,3% dos participantes, respectivamente. Nas entrevistas, os pesquisadores também ressaltaram a importância da visualização da proveniência dos documentos, através da qual são apresentadas as atividades que reutilizaram um documento específico. Outros pesquisadores destacaram a relevância da visualização que relaciona os experimentos com os pesquisadores envolvidos. Os resultados obtidos mostram que as dificuldades observadas, durante a interação dos pesquisadores com o grafo de proveniência, não comprometeram a relevância desta visualização para o reúso dos experimentos. Estes resultados apresentam indícios de que as visualizações implementadas também contribuem para o reúso dos experimentos científicos na plataforma E-SECO.

De acordo com a análise apresentada, todos os aspectos da solução foram considerados relevantes para o reúso de experimentos científicos pela maior parte dos pesquisadores. É importante ressaltar que esses resultados foram obtidos a partir de três fontes diferentes (observação direta, questionário e entrevista), e que através do processo de triangulação de dados apresentado, foi verificada a convergência entre os resultados obtidos por estas três fontes. Visto isso, podemos concluir que este estudo de caso apresentou indícios de que o gerenciamento de proveniência e contexto, implementados pela abordagem *ContextProv*, na plataforma E-SECO contribui para o reúso dos experimentos científicos cadastrados na plataforma. Desta forma, foram reveladas evidências de que a abordagem *ContextProv* pode apoiar o reúso dos experimentos científicos em plataformas de ECOSC. A aceitação da hipótese H1 se dá por não terem sido encontrados indícios que pudessem rejeitá-la, portanto, '*o gerenciamento de contexto e de proveniência de dados potencializa a reutilização de experimentos em uma Plataforma de Ecossistema de Software Científico*'. Entretanto, estudos de casos adicionais devem ser conduzidos no sentido de transferir conhecimentos para outros contextos nos quais os experimentos serão conduzidos.

4.7 AMEAÇAS À VALIDADE

A validade de um estudo denota a confiabilidade dos resultados e até que ponto os resultados não são tendenciosos pelo ponto de vista subjetivo dos pesquisadores (RUNESON *et al.*, 2012). Para avaliar as ameaças à validade dos estudos de caso apresentados, foram considerados os aspectos de qualidade definidos por YIN (2015): validade de constructo, validade interna, validade externa e validade de conclusão.

4.7.1. Validade do constructo

A validade do constructo está relacionada ao entendimento correto de todas as questões envolvidas na avaliação, por parte dos participantes. Nesta avaliação, utilizou-se várias fontes de coleta para a avaliação da solução, visto que várias fontes de evidências fornecem essencialmente várias avaliações do mesmo fenômeno. Entretanto, caso o participante não tenha entendido o exato significado e propósito de cada pergunta do questionário ou da entrevista, então existe uma ameaça à validade do projeto.

4.7.2. Validade interna

A validade interna diz respeito ao conhecimento sobre todos os fatores que possam influenciar o objeto de estudo. Quando o autor da avaliação estuda se algum fator influencia o objeto de estudo, e existem outros fatores que possam também estar influenciando este mesmo objeto de estudo. Nesta avaliação, foi dado um treinamento aos participantes para que conhecessem as ferramentas e recursos disponíveis na plataforma E-SECO. Porém, alguns participantes tiveram que fazer o treinamento remoto, através de um vídeo. Assim, é possível que algum participante não tenha entendido alguma parte do treinamento e que, durante a interação com a plataforma, desconhecesse sobre a existência ou utilidade de algum recurso. Neste caso existe uma ameaça à validade interna.

4.7.3. Validade externa

A validade externa define condições que limitam a generalização dos resultados. O estudo de caso piloto contou com a participação de alunos de pesquisadores da área de engenharia de software. Isso reflete em uma ameaça quando os pesquisadores não estão relacionados ao contexto, tendo em vista que os pesquisadores de outros contextos podem possuir uma percepção diferente da arquitetura, se comparados à população utilizada. Para o estudo de caso regular foram escolhidos apenas pesquisadores relacionados ao contexto dos experimentos utilizados. A abordagem proposta está inserida em um domínio específico. Portanto, os resultados obtidos pelos estudos de caso apresentados, se limitam ao domínio de plataformas de ECOSC e às condições do ambiente no qual o estudo foi conduzido, e não podem ser generalizados, mas transferidos para outros contextos.

4.7.4. Validade de conclusão

A validade de conclusão está relacionada à habilidade de chegar a uma conclusão correta entre o tratamento e o resultado. Assim, o número de participantes no estudo de caso é uma ameaça, pois a presença de mais participantes poderia influenciar nos resultados aqui apresentados. Além disso, os estudos de caso apresentados, devido a uma limitação de tempo, foram conduzidos após o término dos experimentos. Assim, estudos conduzidos durante todo o período de experimentação poderiam obter resultados diferentes. Com relação à integração com as plataformas Mendeley e Kepler, a empresa que participou do estudo não faz uso destas plataformas. Portanto, a relevância dos dados obtidos por meio dessas integrações foi avaliada de forma indireta. Para isso, os dados que deveriam ser obtidos através dessas integrações foram adicionados à plataforma manualmente. Visto isso, outros estudos precisam ser conduzidos para explorar outras situações que envolvem o reúso de experimentos científicos em plataformas de ECOSC.

4.8 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi avaliado como o gerenciamento de contexto e de proveniência de dados potencializa a reutilização de experimentos em uma plataforma de Ecossistema de Software Científico. Após a condução de um estudo de caso piloto e um estudo de caso regular, dados coletados de três fontes distintas (questionários, observação direta e entrevista) foram triangulados para obter os resultados finais. Estes resultados apresentam indícios que o gerenciamento de contexto e proveniência proposto pela abordagem *ContextProv* pode apoiar o reúso de experimentos científicos. Apesar dos indícios observados, devem ser realizados novos estudos para considerar outros contextos.

5 CONCLUSÕES

Este trabalho apresentou uma fundamentação teórica com os principais conceitos que envolvem o gerenciamento de contexto e de proveniência de dados em plataformas de ecossistemas de software científico. Primeiramente, foi apresentado o domínio de *e-Science* e os desafios para o reúso durante o processo de experimentação científica em ECOSC. Em seguida, foram apresentados conceitos sobre proveniência de dados, ontologia, contexto e integração de dados, utilizados na solução proposta. Por fim, os conceitos relacionados aos ecossistemas de software que fazem parte do enfoque da solução.

Considerando as dificuldades da reutilização de conhecimento em plataformas de ECOSC, este trabalho buscou apoiar o reúso neste domínio, através de uma abordagem de gerenciamento de proveniência e contexto. Para tanto, a abordagem *ContextProv* foi proposta. Ela realiza a captura, o armazenamento, o enriquecimento, o compartilhamento e a visualização de informações de proveniência e contexto, durante todo o ciclo de vida de experimentação científica. Posteriormente, estudos de caso foram conduzidos para avaliar a solução proposta. Eles apresentaram indícios sobre a viabilidade de utilização dos recursos apresentados para apoiar o reúso de conhecimento no contexto de um ECOSC.

O presente trabalho apresentou também contribuições que vão além do apoio ao reúso e fornecem suporte para grupos de pesquisa que trabalham com plataformas de ECOSC, são elas:

- O mapeamento sistemático da literatura, que identificou e categorizou os principais trabalhos existentes no domínio de proveniência de dados e elementos de contexto em *e-Science*. Este mapeamento contribui para a comunidade científica, fornecendo uma visão geral do problema de gerenciamento de proveniência e contexto de experimentos científicos, bem como as principais soluções propostas até o momento;
- A modelagem de um ciclo de vida de informações contextuais e de proveniência em plataformas de ECOSC, o qual fornece diretrizes para que outras plataformas científicas implementem o gerenciamento de proveniência e contexto;
- A especificação de um *framework* conceitual que modela as informações contextuais e de proveniência em um ambiente colaborativo e distribuído de experimentação científica. Este *framework* também contribui para a comunidade científica fornecendo suporte à seleção de elementos de contexto e de proveniência relevantes para este domínio;

- O desenvolvimento de uma ontologia capaz de modelar e extrair conhecimento implícito sobre a proveniência e o contexto de experimentos científicos. Esta ontologia está acessível através de um web service semântico e, por utilizar o modelo PROV, permite a interoperabilidade com plataformas externas;
- A integração entre a plataforma E-SECO e as plataformas Mendeley e Kepler, que aumentou o suporte da plataforma E-SECO ao processo de experimentação. Além disso, gerou conhecimento sobre a integração entre plataformas através de APIs e através de modelos de dados comuns, o que abre oportunidades para a implementação de outras integrações;
- A implementação do gerenciamento de proveniência e contexto na plataforma E-SECO que é capaz de aumentar o reúso do conhecimento produzido durante o processo de experimentação. Além disso, pode oferecer suporte ao desenvolvimento e composição de serviços, uma vez que, durante o reúso de um experimento, é possível que alguns dos serviços utilizados não interoperem mais com os mesmos componentes, então as informações de proveniência e contexto vão auxiliar no desenvolvimento de novos serviços, ou na realização da composição com outros serviços;
- A identificação e implementação de visualizações para a apresentação de informações de contexto e de proveniência de experimentos científicos;
- Auxílio na promoção do conhecimento organizacional, sistematizando a organização do conhecimento gerado através dos experimentos e promovendo a cultura do reúso.

Este trabalho foi desenvolvido com o objetivo de potencializar o reúso de conhecimento em plataformas de ECOSC. Desta forma, o gerenciamento de proveniência e contexto implementado está limitado a este propósito, não podendo ser generalizado. A utilização da abordagem proposta em outros contextos, por exemplo para promover a reprodutibilidade dos experimentos científicos, necessita de novas análises e avaliações. No entanto, os resultados obtidos podem ser transferidos para outros contextos. Além disso, para o volume de dados utilizados neste trabalho, pode-se dizer que a ontologia atendeu às expectativas. Porém, sabe-se que a ontologia possui restrições para o processamento de grandes volumes de dados, podendo ocasionar lentidão na obtenção dos resultados.

Como trabalhos futuros, outras integrações com sistemas gerenciadores de *workflows* científicos, com sensores e com plataformas externas precisam ser implementadas, a fim de permitir outras formas de captura das informações de proveniência e contexto. Em

relação ao enriquecimento dos dados, novas regras podem ser implementadas na ontologia, ou ainda, a ontologia pode ser integrada com ontologias específicas do domínio dos experimentos, para ampliar sua capacidade de extração de conhecimento. A condução de novos estudos de caso, a fim de avaliar o apoio oferecido pela abordagem *ContextProv* ao reúso de experimentos em outros domínios é outra possibilidade de trabalhos futuros.

REFERÊNCIAS

ALTINTAS, I *et al.* Kepler: an extensible system for design and execution of scientific workflows. 2004, Santorini Island, Greece: 16th International Conference on Scientific and Statistical Database Management, 2004. p.423–424.

BAADER, F. **The description logic handbook: Theory, implementation and applications**. New York, NY, USA: Cambridge university press, 2003. .

BASILI, V R; WEISS, D M. A Methodology for Collecting Valid Software Engineering Data. **IEEE Transactions on Software Engineering** n. 6, p. 728–738 , 1984.

BAZIRE, M; BRÉZILLON, P. Understanding context before using it. 2005, Paris, France: International and Interdisciplinary Conference on Modeling and Using Context, 2005. p.29–40.

BELLOUM, A *et al.* Collaborative e-science experiments and scientific workflows. **IEEE Internet Computing** p. 39–47 , 2011.

BERNERS-LEE, T *et al.* The semantic web. **Scientific american** v. 284, n. 5, p. 28–37 , 2001.

BOSCH, J. From Software Product Lines to Software Ecosystems. SPLC, 2009, Pittsburgh, PA, USA: Proceedings of the 13th International Software Product Line Conference, 2009. p.111–119.

BOSCH, J; BOSCH-SIJTSEMA, P M. Softwares product lines, global development and ecosystems: collaboration in software engineering. **Collaborative Software Engineering**. Berlin, Heidelberg: Springer, 2010. p. 77–92.

BRÉZILLON, P *et al.* Context-Based Awareness in Group Work. 2004, Miami Beach, Florida, USA: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, 2004. p.575–580.

BRÉZILLON, P. Contextualization of scientific workflows. 2011, Karlsruhe, Germany:

International and Interdisciplinary Conference on Modeling and Using Context, 2011. p.40–53.

BRÉZILLON, P. Task-Realization Models in Contextual Graphs. 2005, Paris, France: International and Interdisciplinary Conference on Modeling and Using Context, 2005. p.55–68.

BUNEMAN, P; KHANNA, S; WANG-CHIEW, T. Why and Where: A characterization of data provenance. 2001, London, United Kingdom: International conference on database theory, 2001. p.316–330.

CAMPOS, M M *et al.* Tecnologias de precisão na avaliação da eficiência alimentar. **Embrapa Gado de Leite-Artigo em periódico indexado (ALICE)** v. 79, p. 73–85 , 2015.

CAO, B *et al.* Provenance Information Model of Karma Version 3. 2009, Los Angeles, CA, USA: Proceedings of the 2009 Congress on Services-I, 2009. p.348–351.

CAO, Y *et al.* DataONE: a data federation with provenance support. 2016, McLean, VA, USA: Proceedings of the 6th International Provenance and Annotation Workshop on Provenance and Annotation of Data and Processes, 2016. p.230–234.

CLASSE, T *et al.* A Distributed Infrastructure to Support Scientific Experiments. **Journal of Grid Computing** v. 15, n. 4, p. 475–500 , 2017.

COSTA, F *et al.* Capturing and Querying Workflow Runtime Provenance with PROV: A Practical Approach. 2013, New York, NY, USA: Proceedings of the Joint EDBT/ICDT 2013 Workshops, 2013. p.282–289.

COSTA, F; DE OLIVEIRA, D; MATTOSO, M. Towards an Adaptive and Distributed Architecture for Managing Workflow Provenance Data. 2014, Sao Paulo, Brazil: Proceedings of the 2014 IEEE 10th International Conference on e-Science, 2014. p.79–82.

CUEVAS-VICENTTÍN, V *et al. ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance* Acessado: 14/09/2016. Disponível em: <http://jenkins-

1.dataone.org/jenkins/view/Documentation Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html>. Acesso em: 14 set. 2016.

CUEVAS-VICENTTÍN, V *et al.* The PBase scientific workflow provenance repository. **International Journal of Digital Curation** p. 28–38 , 2014.

DE OLIVEIRA, D *et al.* Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. 2010, Miami, FL, USA: Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on, 2010. p.378–385.

DEELMAN, E *et al.* Workflows and e-Science: An overview of workflow system features and capabilities. **Future Generation Computer Systems** p. 528–540 , 2009.

DEY, A K; ABOWD, G D; SALBER, D. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. **Human-computer interaction** p. 97–166, 2001.

DOAN, AH; HALEVY, A; IVES, Z. **Principles of data integration**. 1st. ed. San Francisco, CA, USA: Elsevier, 2012.

FAN, X *et al.* A context-based framework for improving decision making in scientific workflow. 2011, Shanghai, China: Computer Research and Development (ICCRD), 2011 3rd International Conference on, 2011. p.15–19.

FILHO, O F F; FERREIRA, M A G V. Semantic web services: a restful approach. 2009, Rome, Italy: Proceedings of the IADIS International Conference WWW/Internet, 2009. p.169–180.

FONTÃO, A D L *et al.* MSECO-DEV: Application development process in mobile software ecosystems. 2016, San Francisco, USA: Proceedings of the International Conference on Software Engineering and Knowledge Engineering, 2016. p.317–322.

FREIRE, J *et al.* Managing Rapidly-evolving Scientific Workflows. IPAW'06, 2006, McLean,

USA: Proceedings of the 2006 International Conference on Provenance and Annotation of Data, 2006. p.10–18.

FREITAS, V *et al.* Uma Arquitetura para Ecossistema de Software Científico. 2015, Belo Horizonte, Brasil: Workshop em Desenvolvimento Distribuído de Software, Ecossistemas de Software e Sistemas-de-Sistemas (WDES), 2015. p.41–48.

GOBLE, C A *et al.* myExperiment: a repository and social network for the sharing of bioinformatics workflows. **Nucleic acids research** p. 677–682 , 2010.

GROTH. P et al. **An architecture for provenance systems**. [S.l.]: PROVENANCE Enabling and Supporting Provenance in Grids for Complex Problems, 2006.

GROTH, P; MILES, S; MOREAU, L. Preserv: Provenance recording for services. **Engineering and Physical Sciences Research Council - EPSRC** , 2005.

GRUBER, T R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **International journal of human-computer studies** p. 907–928 , 1995.

HEY, T. **The fourth paradigm: data-intensive scientific discovery**. Washington, DC, USA: Microsoft Research, 2009. .

HORROCKS, I *et al. SWRL: A semantic web rule language combining OWL and RuleML*. Disponível em: <http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/>. Acesso em: 14 set. 2016.

JANSEN, S; BRINKKEMPER, S; CUSUMANO, M A. **Software Ecosystems: Analyzing and Managing Business Networks in the Software Industry**. [S.l.]: Edward Elgar Publishing, Incorporated, 2013. 350 p. .

KITCHENHAM, B A; BUDGEN, D; BRERETON, O P. Using Mapping Studies As the Basis for Further Research - A Participant-observer Case Study. **Information and Software Technology**

p. 638–651 , 2011.

LENZERINI, M. Data integration: A theoretical perspective. 2002, Madison, Wisconsin: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2002. p.233–246.

LETHBRIDGE, T C; SIM, Susan El; SINGER, J. Studying software engineers: Data collection techniques for software field studies. **Empirical software engineering** p. 311–341 , 2005.

LIM, C *et al.* Prospective and Retrospective Provenance Collection in Scientific Workflow Environments. 2010, Miami, Florida: Services Computing (SCC), 2010 IEEE International Conference on, 2010. p.449–456.

MANIKAS, K. Revisiting software ecosystems research: A longitudinal literature study. **Journal of Systems and Software** p. 84–103 , 2016.

MANIKAS, K; HANSEN, K M. Software ecosystems--A systematic literature review. **Journal of Systems and Software** p. 1294–1306 , 2013.

MARQUES, P *et al.* Apoiando a Composição de Serviços em Ecossistemas de Software Científico. 2017, São Paulo, Brasil: 14º SBSC - Simpósio Brasileiro de Sistemas Colaborativos, 2017.

MARTIN, D *et al. OWL-S: Semantic Markup for Web Services*. Disponível em: <https://www.w3.org/Submission/OWL-S/>. Acesso em: 14 set. 2016.

MAYER, R; MIKSA, T; RAUBER, A. Ontologies for describing the context of scientific experiment processes. 2014, Sao Paulo, Brasil: 10th International Conference on e-Science, 2014. p.153–160.

MICHEL, F. **Integrating heterogeneous data sources in the Web of data**. Université Côte d'Azur, 2017.

MICHENER, W *et al. Data Observation Network for Earth (DataONE)*. Disponível em: <https://www.dataone.org/>. Acesso em: 14 set. 2016.

MISSIER, P *et al.* D-PROV: extending the PROV provenance model with workflow structure. 2013, Lombard, IL: Proceedings of the 5th USENIX conference on Theory and Practice of Provenance, 2013.

MISSIER, P. The Lifecycle of Provenance Metadata and Its Associated Challenges and Opportunities. **Building Trust in Information: Perspectives on the Frontiers of Provenance** p. 127–137 , 2016.

MOREAU, L *et al.* The Open Provenance Model core specification (v1.1). **Future Generation Computer Systems** p. 743–756 , 2011.

MOREAU, L; GROTH, P. *An Overview of the PROV Family of Documents*. Disponível em: <https://www.w3.org/TR/prov-overview>. Acesso em: 14 jan. 2018.

NEIVA, F W *et al.* PRIME: Pragmatic Interoperability Architecture to Support Collaborative Development of Scientific Workflows. 2015, Belo Horizonte, Brasil: Proceedings of the 2015 IX Brazilian Symposium on Components, Architectures and Reuse Software, 2015. p.50–59.

NEVES, V C; BRAGANHOLO, V; MURTA, L. Implicit Provenance Gathering Through Configuration Management. 2013, Piscataway, NJ, USA: Proceedings of the 5th International Workshop on Software Engineering for Computational Science and Engineering, 2013. p.92–95.

NUNES, V T; SANTORO, F M; BORGES, Ma R S. Um modelo para gestão de conhecimento baseado em contexto. 2007, Rio de Janeiro, Brasil: XXVII Simpósio Brasileiro de Sistemas Colaborativos (SBSC), 2007. p.69–82.

OINN, T *et al.* Taverna/myGrid: Aligning a Workflow System with the Life Sciences Community. **Workflows for e-Science: Scientific Workflows for Grids** p. 300–319 , 2007.

OLIVEIRA, W; OLIVEIRA, D De; BRAGANHOLO, V. Provenance Analytics for Workflow-Based Computational Experiments: A Survey. **ACM Computing Surveys (CSUR)** v. 51, n. 3, p. 53 , 2018.

PEREIRA, A F *et al.* An Architecture to Enhance Collaboration in Scientific Software Product Line. 2016, Koloa, HI, USA: 49th Hawaii International Conference on System Sciences (HICSS), 2016. p.338–347.

PETTICREW, M; ROBERTS, H. **Systematic Reviews in the Social Sciences: A Practical Guide**. [S.l.]: John Wiley & Sons, 2008. .

RITTENBRUCH, M. Atmosphere: a framework for contextual awareness. **International Journal of Human-Computer Interaction** p. 159–180 , 2002.

ROSA, M G P; BORGES, M R S; SANTORO, F M. A conceptual framework for analyzing the use of context in groupware. 2003, Autrans, France: International Conference on Collaboration and Technology, 2003. p.300–313.

RUNESON, P *et al.* **Case Study Research in Software Engineering: Guidelines and Examples**. 1st. ed. [S.l.]: Wiley Publishing, 2012. .

SILVA, M F; BAIÃO, F A; REVOREDO, K. Towards Planning Scientific Experiments through Declarative Model Discovery in Provenance Data. 2014, São Paulo, Brasil: 10th International Conference on e-Science, 2014. p.95–98.

SIMMHAN, Y L *et al.* Performance Evaluation of the Karma Provenance Framework for Scientific Workflows. 2006, Chicago, IL, USA: Provenance and Annotation of Data: International Provenance and Annotation Workshop, IPAW, 2006. p.222–236.

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

SIRQUEIRA, T F M *et al.* E-SECO ProVersion: An Approach for Scientific Workflows Maintenance and Evolution. **Procedia Computer Science** p. 447–556 , 2016.

TENOPIR, C *et al.* Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. **PLOS ONE** p. 1–24 , 2015.

YIN, R K. **Estudo de Caso: Planejamento e Métodos**. 5. ed. [S.l.]: Bookman editora, 2015. .

ZAUGG, H *et al.* Mendeley: Creating communities of scholarly inquiry through research collaboration. **TechTrends: Linking Research and Practice to Improve Learning** p. 32–36, 2011.

APÊNDICE A. MAPEAMENTO SISTEMÁTICO DA LITERATURA

A.1 CRITÉRIOS DE INCLUSÃO

CI 1. O artigo relata uma solução (método, técnica, modelo, ferramenta, framework ou ontologia) de contexto de experimentos científicos;

CI 2. O artigo relata um modelo (método, técnica, solução, ferramenta, framework ou ontologia) de proveniência de experimentos científicos.

A.2 CRITÉRIOS DE EXCLUSÃO

CE 1. Não é um artigo, e sim apenas resumo de uma conferência ou seminário;

CE 2. O artigo não foi publicado em uma conferência ou revista com revisão por pares;

CE 3. O artigo não está escrito em Inglês;

CE 4. O artigo não tem o texto completo disponível na instituição do pesquisador;

CE 5. A solução proposta não é aplicada em e-Science, *workflows* científicos ou experimentos científicos;

CE 6. O artigo não discute sobre os elementos do contexto, ou sobre proveniência de experimentos científicos.

A.3 BASES DE BUSCA UTILIZADAS

- ACM (www.portal.acm.org);
- Compendex (www.engineeringvillage.com);
- IEEE Xplore (www.ieeexplore.com.br);
- ScienceDirect (www.sciencedirect.com);
- Scopus (www.scopus.com);
- Web of Science (www.isiknowledge.com).

A.4 RELATÓRIO

Foram recuperados mais de 600 trabalhos. Eliminados os duplicados, restaram 358 trabalhos a serem analisados segundo os critérios de seleção definidos anteriormente. A Tabela 10 apresenta o total de trabalhos encontrados em cada base, após a eliminação das duplicidades.

Tabela 10. Total de trabalhos encontrados por base

BASE	TOTAL DE TRABALHOS
ACM Digital Library	8
EI Compendex	72
IEEE Digital Library	70
ISI Web of Science	46
Science@Direct	17
Scopus	145
TOTAL	**358**

Após a análise dos artigos com base na leitura de seus títulos e resumos, segundo os critérios de aceitação e exclusão definidos, 285 (79,7%) trabalhos foram eliminados. O restante foi considerado aceito, e foram melhor avaliados através da leitura completa. O Gráfico 8 ilustra a relação entre o total de artigos aceitos em cada base.

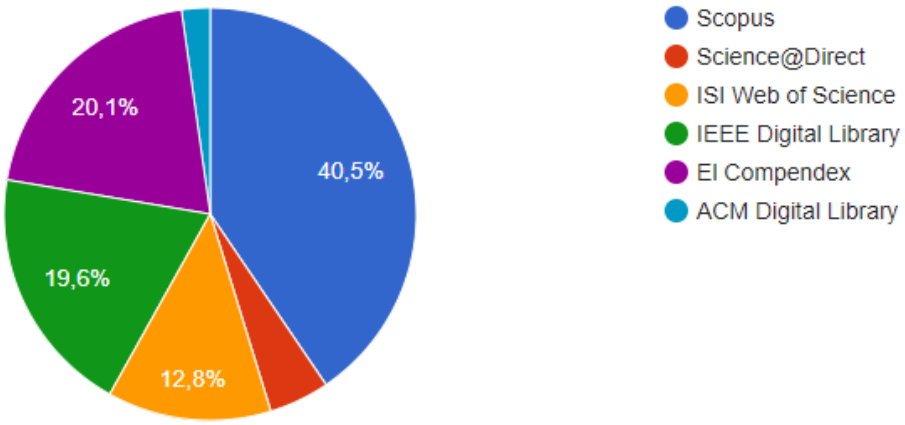

Gráfico 8. Trabalhos aceitos por base

Os 73 artigos escolhidos durante a condução do estudo foram analisados, a fim de extrair as informações necessárias para responder às questões de mapeamento (QM) definidas. O Gráfico 9 representa a resposta para a QM 1 (Quantos estudos foram publicados ao longo dos anos?). Pode-se observar que apesar da pesquisa não impor nenhuma restrição com relação ao ano de publicação, os resultados obtidos foram publicados a partir de 2003. Vale ressaltar, que este mapeamento foi feito em 2017, sendo assim, apresenta apenas os trabalhos publicados até esta data.

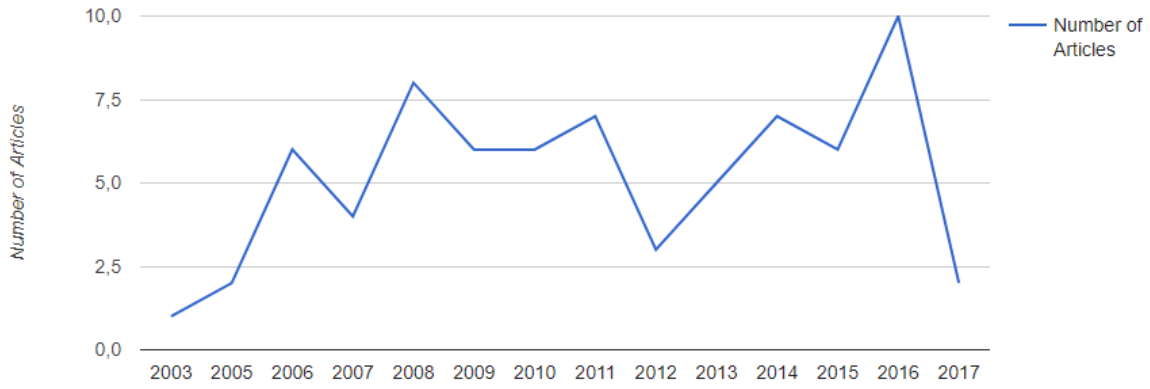

Gráfico 9. Trabalhos aceitos ao longo dos anos

Com relação à QM 2 (Quem são os autores mais ativos na área?), pode-se observar na Tabela 11 que Marta Mattoso com sete artigos, Regina Braga, Fernanda Campos, Shiyong Lu e Bertram Ludäscher com cinco artigos são os principais autores na área pesquisada. Além destes, a tabela apresenta um ranking de publicações entre os autores que publicaram três artigos ou mais.

Tabela 11. Publicações por autor

AUTOR	PUBLICAÇÕES
MATTOSO, M.	7
BRAGA, R.	5
CAMPOS, F.	5
LU, S.	5
LUDÄSCHER, B.	5
DA CRUZ, S.M.S	4
DAVID, J.M.N.	4
ALTINTAS, I.	3
BOWERS, S.	3
CAMPOS, M.L.M	3
COHEN-BOULAKIA, S.	3
DE OLIVEIRA, D.	3
FREIRE, J.	3
MISSIER, P.	3
MOREAU, L.	3
SILVA, C.T.	3

Para responder à QM 3 (Quantos estudos foram encontrados sobre cada assunto (proveniência, contexto)?) os artigos foram classificados por assunto. O Gráfico 10 demonstra que a análise da proveniência em experimentos científicos já é bastante explorada na literatura,

tratada em 65 dos 73 artigos encontrados nesta pesquisa. Já o assunto contexto, que recuperou apenas 8 artigos, ainda não é muito abordado. Contudo, o contexto é também representado parcialmente pelos artigos que falam sobre proveniência, uma vez que as informações sobre proveniência podem ser consideradas informações contextuais.

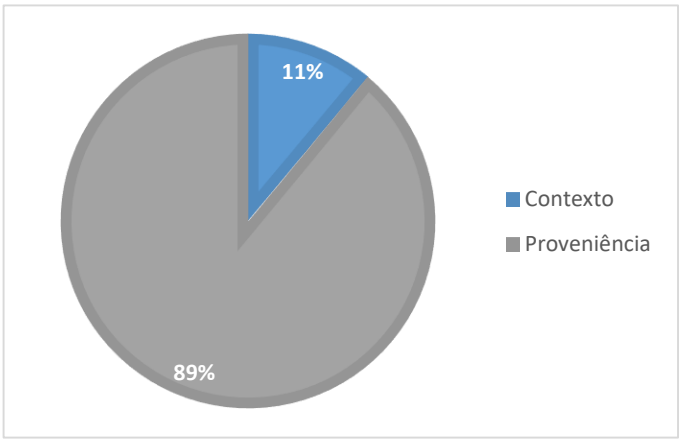

Gráfico 10. Artigos por assunto pesquisado

De acordo com a QM 4 (Quais são os principais veículos de publicação de pesquisas na área?), pode-se observar através do Gráfico 11 que as conferências são o principal meio de publicação de trabalhos na área de proveniência e contexto em experimentos científicos. Com 38 artigos selecionados, as conferências representam 52% do total, em segundo lugar vem os workshops (ou simpósios, ou congressos) onde foram publicados 23% dos trabalhos. As revistas possuem menor representatividade na área, com apenas 22% do total. E por último os livros com 3%.

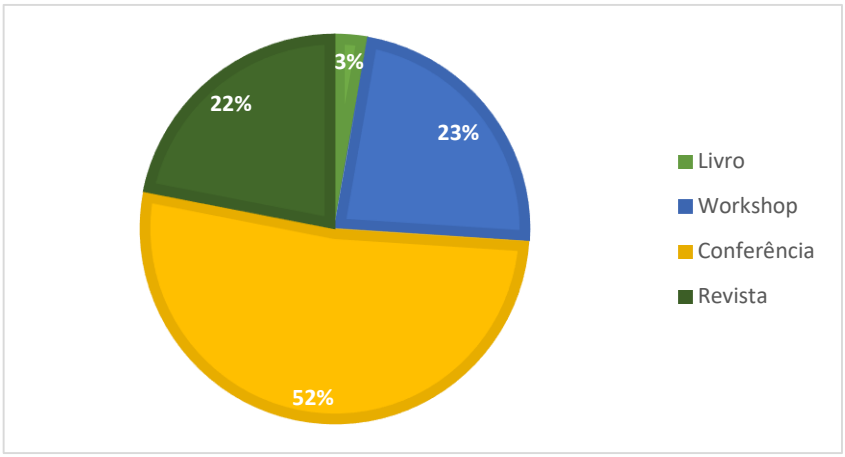

Gráfico 11. Artigos aceitos por veículo

Dentre os veículos citados anteriormente alguns tiveram um maior número de trabalhos aceitos. Desta forma pode-se destacar os seguintes:

- *IEEE International Conference on e-Science*, com 9 artigos aceitos;
- *International Provenance and Annotation Workshop*, com 6 artigos;
- *Journal of Future Generation Computer Systems*, com 5 artigos.

A.5 AMEAÇAS À VALIDADE

Este mapeamento da literatura teve como objetivo identificar, categorizar e analisar soluções para o gerenciamento de informações de contexto e/ou proveniência em experimentos científicos. Contudo, como qualquer método, possui limitações e ameaças à sua validade. Os resultados deste estudo podem ter sido influenciados por algumas limitações como o número de pesquisadores que trabalharam na seleção dos artigos, os termos utilizados na *string* de busca, e as fontes escolhidas para a busca dos trabalhos.

A fim de mitigar a possibilidade de viés com relação ao número de pesquisadores selecionando os artigos, todo o protocolo e planejamento do processo foi revisado por mais de um pesquisador, de forma a garantir que ele possa ser reproduzido por terceiros.

Com relação aos termos utilizados na *string* de busca, por conta de uma limitação de tempo, não incluímos na *string* o termo ‘contexto’ de forma isolada, uma vez que este termo é muito utilizado em artigos, mas não no sentido de ‘elementos de contexto’ o qual estamos buscando. Desta forma, este termo traria muitos falsos positivos, fazendo com que fosse necessário um tempo muito maior para a seleção dos artigos. Por outro lado, muitos trabalhos importantes para a área poderiam não ser recuperados pela *string* devido à ausência deste termo.

Para minimizar este problema, com relação ao termo ‘contexto’ e também aos demais termos de interesse, utilizamos os artigos de controle para a extração de palavras-chave e sinônimos importantes para o estudo. Assim, a *string* final foi capaz de recuperar estes estudos, indicando que é capaz de recuperar também outros artigos importantes para a área.

As fontes escolhidas para a busca também podem ser uma ameaça à validade deste estudo uma vez que, de acordo com os critérios de escolha estabelecidos, não foram consideradas todas as bases existentes. Entretanto, acreditamos que as bases selecionadas foram suficientes para se obter uma grande representação da área pesquisada, uma vez que a única base importante excluída foi a Springer[20], e muitos trabalhos indexados nesta base são também indexados pelas demais.

[20] http://link.springer.com

APÊNDICE B. TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Condutor do Estudo: Lenita Martins Ambrósio - lenita.ambrosio@gmail.com
Instituição: Universidade Federal de Juiz de Fora

Prezado(a) Senhor(a), a plataforma E-SECO tem como objetivo apoiar pesquisadores durante as diversas etapas que compõem o ciclo de vida de um experimento científico. A partir disso, este trabalho busca apoiar a reutilização de experimentos científicos com base em informações de proveniência e contexto dos mesmos. Este estudo de caso busca avaliar como o gerenciamento de contexto e de proveniência de dados potencializam a reutilização de experimentos em uma Plataforma de Ecossistema de Software Científico.

Procedimentos: Inicialmente, a plataforma E-SECO será apresentada e o participante receberá um treinamento sobre o uso da mesma. O participante terá um tempo para explorar a plataforma, e visualizar as informações de contexto e proveniência de experimentos realizados por seu grupo de pesquisa. Finalmente, o participante responderá questões referentes ao processo de experimentação apoiado pela plataforma E-SECO. Para participar deste estudo, solicitamos uma colaboração especial para: (i) permitir um estudo dos dados resultantes para a avaliação da proposta apresentada; (ii) participar da entrevista e/ou responder aos questionários. Para todos os 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 lhe fornecer novos conhecimentos sobre questões relacionadas à proveniência e ao contexto de experimentos científicos, bem como aspectos de visualização. A participação neste estudo não envolve nenhum gasto ou ônus. O participante também não receberá qualquer espécie de reembolso ou gratificação devido à participação na pesquisa.

Confidencialidade da Pesquisa: A informação coletada neste estudo é confidencial. As respostas informadas, bem como seus dados pessoais, tais como número de documentos e e-mail, não serão identificados de modo algum.

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. Para isso, entre em contato com o pesquisador responsável a partir do endereço de e-mail informado anteriormente.

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.

Juiz de Fora ___ de agosto de 2018.

CPF do Participante: _____
Nome do Participante: _____
E-mail do Participante: _____

_____
Assinatura do Participante

APÊNDICE C. FORMULÁRIO DE CARACTERIZAÇÃO DO PARTICIPANTE

1- Nome: _____

2- Formação acadêmica (Graduação/Mestrado/Doutorado + Nome do Curso): _____

3- Selecione sua área de pesquisa:

( ) Nutrição animal / Eficiência alimentar

( ) Qualidade do Leite

( ) Engenharia de Software

( ) Inteligência Computacional

( ) Outro: _____

4- Como você avalia seus conhecimentos em Experimentação Científica?

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

5- Como você avalia seus conhecimentos em Proveniência de Dados?

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

6- Como você avalia seus conhecimentos em Análise de Contexto?

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

7- Como você avalia seus conhecimentos em Análise da Eficiência Alimentar de Bovinos?

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

APÊNDICE D. QUESTIONÁRIO DO ESTUDO DE CASO

1- Como você avalia o processo de experimentação científica adotado atualmente (sem o apoio da plataforma E-SECO)?

_____

_____

2- Você considera importante a prática do reúso na experimentação científica? Porque?

_____

_____

3- Como você avalia o processo de reúso de experimentos científicos adotado atualmente (sem o apoio da plataforma E-SECO)?

_____

_____

4- As informações apresentadas pela plataforma relativas a cada um dos itens a seguir são suficientes para a reutilização do experimento?

	muito insuficiente	insuficiente	indiferente	suficiente	mais que suficiente
Pesquisadores					
grupos de pesquisa					
Instituições					
Experimentos					
Workflows					
Atividades					
execução dos *workflows*					
execução das atividades					
protocolo utilizado					
Proveniência					

5- Você acrescentaria alguma informação? Quais?

_____

_____

6- As informações relativas a cada item a seguir são relevantes para a reutilização do experimento?

	muito irrelevante	irrelevante	indiferente	relevante	muito relevante
Pesquisadores					
grupos de pesquisa					
Instituições					
Experimentos					
Workflows					
Atividades					
execução dos *workflows*					
execução das atividades					
Protocolo utilizado					
Proveniência					

7- Você removeria alguma informação? Quais?

_____

_____

8- A apresentação de experimentos similares é relevante para o reúso do experimento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

9- A apresentação das atividades responsáveis pela criação de um documento (por exemplo uma planilha) é relevante para o reúso deste documento em outros experimentos?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

10- A apresentação das atividades que já reutilizaram um documento é relevante para uma nova reutilização do documento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

11- A apresentação das atividades do experimento que foram reutilizadas DE outro experimento é relevante para o reúso deste experimento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

12- A apresentação das atividades do experimento que foram reutilizadas POR outros experimentos é relevante para o reúso deste experimento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

13- A apresentação das instituições envolvidas na realização dos experimentos é relevante para o reúso do experimento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

14- Apresentar detalhes sobre o perfil científico dos pesquisadores envolvidos no experimento é relevante para o reúso do experimento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

15- Apresentar a data e hora de início e fim de cada atividade executada é relevante para o reúso do experimento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

16- Apresentar as entradas e saídas (documentos utilizados e documentos produzidos) de cada atividade executada é relevante para o reúso do experimento?

( ) muito irrelevante ( ) irrelevante ( ) indiferente ( ) relevante ( ) muito relevante

17- A visualização que relaciona os pesquisadores envolvidos com os experimentos, que está apresentada abaixo, contribui para a reutilização do experimento?

( ) discordo fortemente ( ) discordo ( ) indiferente ( ) concordo ( ) concordo fortemente

18- A visualização que apresenta a colaboração entre os pesquisadores, que está apresentada abaixo, contribui para a reutilização do experimento?

( ) discordo fortemente ( ) discordo ( ) indiferente ( ) concordo ( ) concordo fortemente

19- A visualização que apresenta a atividade que gerou um documento e as atividades que já reutilizaram este documento, que está apresentada abaixo, contribui para a reutilização do experimento?

( ) discordo fortemente ( ) discordo ( ) indiferente ( ) concordo ( ) concordo fortemente

20- A visualização que apresenta o fluxo de atividades do experimento contribui para a reutilização do experimento?

( ) discordo fortemente ( ) discordo ( ) indiferente ( ) concordo ( ) concordo fortemente

21- A visualização que apresenta as informações de proveniência do experimento, que está apresentada abaixo, contribui para a reutilização do experimento?

( ) discordo fortemente ( ) discordo ( ) indiferente ( ) concordo ( ) concordo fortemente

22- Como você avalia o processo de experimentação científica apoiado pela plataforma E-SECO?

_____

_____

23- Quais são os pontos observados na plataforma E-SECO que podem contribuir para a reutilização do experimento?

_____

_____

24- Gostaria de deixar alguma sugestão que possa melhorar o processo de experimentação científica através da plataforma E-SECO?

_____

_____