Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

UnBGOLDProv: Arquitetura de proveniência de dados para um workflow de publicação de dados abertos governamentais

Cleyton Peixoto dos Reis Júnior

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientadora
Prof.a Dr.a Maristela Terto de Holanda

Brasília
2020

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

PP99u

Peixoto dos Reis Júnior, Cleyton
UnBGOLDProv: Arquitetura de proveniência de dados para um workflow de publicação de dados abertos governamentais / Cleyton Peixoto dos Reis Júnior; orientador Maristela Terto de Holanda. -- Brasília, 2020.
80 p.

Dissertação (Mestrado - Mestrado em Informática) -- Universidade de Brasília, 2020.

1. Proveniência de Dados. 2. Dados Abertos Governamentais. 3. PROV-DM. 4. Dados Conectados. 5. Linked Open Government Data. I. Terto de Holanda, Maristela, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

UnBGOLDProv: Arquitetura de proveniência de dados para um workflow de publicação de dados abertos governamentais

Cleyton Peixoto dos Reis Júnior

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof.a Dr.a Maristela Terto de Holanda (Orientadora)
CIC/UnB

Prof. Dr.a Aletéia Patrícia Favacho de Araújo
CIC/UnB

Prof. Dr. Sérgio Lifschitz
DI/PUC - Rio de Janeiro

Prof.a Dr.a Genaína Nunes Rodrigues
Coordenadora do Programa de Pós-graduação em Informática

Brasília, 30 de janeiro de 2020

Dedicatória

Dedico este trabalho primeiramente a Deus, o qual sempre recorro, nas horas difíceis e que nunca me abandonou, dando forças para seguir nessa jornada. À minha esposa Eva e minhas filhas, Maria Eduarda, Heloísa e Cecília, que vivenciaram comigo as dificuldades, me apoiaram e sempre tiveram paciência para vencermos juntos cada etapa. Aos meus pais, Cleyton e Eliete, e meus irmãos, Thiago, Érica e Cínthia, que sempre me apoiam em todos os momentos e estão presentes, sempre dispostos a ajudar. À minha avó Dolores, que sempre me apoiou com sabedoria e aos demais familiares que sempre demonstram carinho e apoio. Por fim, dedico este trabalho também, ao meu avô Erasto (*in memoriam*), que torcia pelo meu sucesso e estava sempre preocupado com o meu desempenho no mestrado.

Agradecimentos

Agradeço primeiramente a Deus, à minha orientadora, professora Maristela Holanda, pelas sábias orientações, pela paciência, pelo compartilhamento de conhecimento, ajuda inestimável e por ter acredito em mim, me dando a oportunidade de fazer o mestrado. Ao professor Márcio Victorino, sempre solicito e disposto a ajudar. Aos amigos que fiz na UnB, com os quais estudamos juntos, passamos dificuldades e que me apoiaram em diversos momentos. Aos professores do programa de pós-graduação que foram extremamente importantes nesta caminhada. E por fim, aos meus amigos, com os quais escrevi os artigos científicos e que foram muito importantes para que concluísse este mestrado.

Resumo

O governo brasileiro aderiu à política de publicação de dados abertos governamentais que possibilita uma administração mais transparente e aberta, permitindo maior participação da sociedade, fortalecimento da democracia e combate à corrupção. No entanto, a forma como os dados abertos são publicados levanta questões como a origem e a autoria dos dados. A realização da proveniência sobre esse dados disponibiliza informações de como, quando e por que os dados foram criados e publicados. Diante desse cenário, considerou-se que a combinação de dados e sua proveniência enriquecem a rastreabilidade dos dados, expondo os métodos e os agentes envolvidos em sua criação, além de promover a possibilidade de reprodutibilidade desses dados. Este trabalho apresenta uma solução tecnológica de proveniência de dados para aprimorar a publicação de dados públicos abertos governamentais, empregando uma arquitetura de informações que pode fornecer a proveniência de dados abertos governamentais públicos, usando o PROV-DM e um banco de dados de grafos. É apresentado como estudo de caso, a implementação de uma arquitetura de informações proposta para coleta, gerenciamento, armazenamento e publicação da proveniência do *workflow* de publicação de dados abertos governamentais conectados. Essa publicação da proveniência, coletada da execução da publicação de um conjunto de dados coletados dos sistemas de informação da UnB, é realizada na plataforma CKAN em conjunto com os dados conectados publicados na plataforma.

Palavras-chave: Proveniência de Dados, Dados Abertos Governamentais, PROV-DM, Dados Conectados

Abstract

The Brazilian Government has adhered to the Linked Open Government Data publication policy that allows for a more transparent and open administration, allowing greater participation of society, strengthening democracy and fighting corruption. However, the way open data is published raises questions such as the origin and authorship of the data. Making the provenance of this data provides information on how, when, and why the data was created and published. Given this scenario, we consider that the combination of data and its origin enriches the traceability of the data, exposing the methods and agents involved in its creation, and promoting the possibility of reproducibility of this data. This paper presents a data provenance technology solution to enhance public open government data publishing by employing an information architecture that can provide the provenance of open government public data using PROV-DM and a graph database. The implementation of a proposed information architecture for collecting, managing, storing and publication of the provenance of the textit workflow for publication of linked open government data is presented as a case study. This provenance publication, collected from the execution of the publication of a set of data collected from UnB's information systems, is carried out on the CKAN platform together with the linked data published on the platform.

Keywords: Provenance, Open Government Data, PROV-DM, Linked Data

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os dados abertos, que podem ser livremente utilizados, foram estimulados por movimentos globais, especialmente após um memorando do presidente Barack Obama sobre transparência e dados governamentais [5] em 2009, e a criação do portal de dados abertos do governo dos EUA. Nesse cenário, a *Open Government Partnership*[1] foi formalmente lançada em 2011, quando os oito governos fundadores (Brasil, Indonésia, México, Noruega, Filipinas, África do Sul, Reino Unido e Estados Unidos) assinaram a Declaração do Governo Aberto e apresentaram seus Planos de Ação. No Brasil, essa política foi consolidada por meio da Lei de Acesso à Informação (LAI) [6], que garante o acesso a qualquer informação de interesse público, exceto aquelas relacionadas à segurança da sociedade e do Estado [6].

Com o desenvolvimento da Web semântica, surgiram padrões e formatos para integrar dados e informações de diferentes fontes. Dessa maneira, o padrão de dados conectados, do inglês *Linked Data*, permite que qualquer instituição publique dados de modo que possam ser lidos por pessoas e processados por máquinas. O termo *Linked Data* surgiu em 2006 com a publicação de *Design Issues* [7], com uma subseção da Web Semântica exclusivamente para *Linked Data*. A chamada *"Web of Data"* surgiu de um conjunto de práticas recomendadas para publicar e conectar dados estruturados na Web a partir de dados conectados [8].

No entanto, a adoção da Web de dados pode trazer algumas preocupações, tais como:

i) o mesmo conjunto de Dados Conectados pode ser replicado e hospedado em locais distintos na Web por meio de diferentes URIs (Uniform Resource Identifier);

ii) os conjuntos de dados podem ser criados individualmente, conectados por diferentes triplas RDF (*Resource Description Framework*) e mantidos por diferentes editores.

Diante disso, surgem algumas perguntas, como é possível confiar nos dados e *links* publicados? Quais dos objetos *Linked Data* fornecem as informações mais confiáveis

[1] `https://www.opengovpartnership.org/`

ou atualizadas sobre a entidade? Essas perguntas levam a avançar além dos dados da entidade, também obtendo informações sobre como os dados se tornaram disponíveis e, nesse contexto, são necessárias informações de proveniência de dados na Web de dados conectados [9].

A proveniência de dados fornece informações sobre uma origem de recurso, como quem o criou, como foi criado e quando foi modificado. É amplamente aceita que esse tipo de informação possa ser adotada para diversos fins relacionados aos dados, como controle de versão, estabelecimento de propriedades, avaliação de qualidade e confiabilidade, descoberta de novos dados, além de processos de auditoria e reprodutibilidade de experimentos [10], [9]. Assim sendo, espera-se, por exemplo, que os dados publicados por agências e organizações oficiais do governo sejam mais confiáveis do que os provenientes de redes sociais ou colaboradores independentes. O registro de metadados de proveniência durante os vários estágios do ciclo de vida dos dados é fundamental para alcançar seus benefícios. A implementação de mecanismos que adicionam informações de proveniência de dados na execução de processos de integração pode aumentar o valor e promover o uso disseminado dos dados conectados, pois torna os processos transparentes e verificáveis.

Portanto, no contexto de publicação de dados abertos governamentais conectados, este trabalho apresenta uma arquitetura capaz de realizar a proveniência de dados de maneira automatizada, que utiliza o modelo de dados genérico, PROV-DM[2], para representação da proveniência do *workflow* de publicação e realiza armazenamento dos metadados de proveniência em um banco de dados baseado em grafo. Essa arquitetura expande um trabalho anterior, adicionando o recurso de gerenciamento de proveniência de dados ao *workflow* de publicação de dados abertos governamentais que utiliza o UnBGOLD (UnB *Government Linked Open Data*) [4].

1.1 Objetivos

O objetivo deste trabalho é propor uma arquitetura que visa definir uma solução tecnológica para capturar, gerenciar e publicar os metadados de proveniência gerados durante o processo de publicação de dados governamentais abertos conectados. Esta arquitetura prevê a criação de um modelo de dados de proveniência baseado no PROV-DM, que captura informações sobre o processo de publicação de dados abertos governamentais, com o objetivo de organizar e estruturar os metadados de proveniência.

Objetivos Específicos

Para alcançar o objetivo geral, foram estabelecidos os seguintes objetivos específicos:

[2] `https://www.w3.org/TR/prov-dm/`

- Estabelecer um conjunto mínimo de metadados de proveniência a serem capturados e utilizar ontologias para representá-los;
- Capturar a proveniência do *workflow* de publicação de dados abertos governamentais conectados através de uma requisição HTTP.
- Validar a proposta por meio da captura da proveniência da execução do *workflow* de publicação de dados abertos governamentais conectados dos sistemas de informação acadêmicos da Universidade de Brasília (UnB);
- Publicar esses metadados de proveniência, em conjunto com os dados abertos conectados, na instância de dados abertos da UnB na plataforma CKAN *(Comprehensive Knowledge Archive Network).*

1.2 Estrutura do Trabalho

O restante do documento está estruturado da seguinte maneira:

- O Capítulo 2 apresenta o referencial teórico necessário para o desenvolvimento desta pesquisa, tais como a *web* semântica, metadados, RDF, ontologia, dados conectados, dados abertos governamentais, proveniência de dados, o modelo PROV-DM, bancos de dados baseado em grafos;
- O Capítulo 3 apresenta alguns trabalhos relacionados à pesquisa;
- O Capítulo 4 detalha a proposta deste trabalho;
- O Capítulo 5 apresenta o estudo de caso;
- O Capítulo 6 apresenta a conclusão e os trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta uma revisão dos principais conceitos relacionados ao tema deste trabalho, envolvendo captura, gerenciamento e publicação de metadados de proveniência de um *workflow* de publicação de dados abertos governamentais. Na Seção 2.1 são apresentados os conceitos relacionados a *Web* Semântica. Na Seção 2.2 é descrito o conceito de metadados. A Seção 2.3 apresenta os conceitos relacionados com o *Resource Description Framework*. Na Seção 2.4 está descrito o conceito de ontologia. A Seção 2.5 apresenta os conceitos relativos a dados conectados. A Seção 2.6 apresenta os conceitos de dados abertos e dados abertos governamentais. A Seção 2.7 apresenta os conceitos relacionados a proveniência e ao modelo de dados PROV-DM. Por fim, a Seção 2.9 apresenta os conceitos relacionados a banco de dados de grafo.

2.1 *Web* Semântica

Em 2001, Tim Berners-Lee [11] divulgou um estudo no qual foi apresentado o conceito de *web* semântica, vislumbrando uma evolução da *web*, na qual os computadores poderiam entender o contexto da fala e do pensamento humanos, para poder "entender" o significado da informação e dos dados. Segundo [11], "a *web* semântica é uma extensão da *web* atual, na qual é dada à informação um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação".

A semântica acessível por máquina é potencializada por meio da especificação de documentos *web* em uma linguagem que permita que os *links* sejam criados com valor em seu relacionamento. Isso faz com que os recursos tenham uma semântica associada, permitindo a execução automática de atividades como compra de produtos personalizados, negociação por pacotes turísticos, agendamento de consultas, entre outras [1].

As tecnologias da *web* semântica podem atuar em *"background"*, resultando em uma melhor experiência do usuário, em vez de influenciar diretamente a "aparência" no nave-

gador. A *web* semântica fornece uma estrutura comum que permite que os dados sejam compartilhados e reutilizados nos limites de aplicativos, empresas e comunidades.

O termo *web* semântica se refere a um conjunto de tecnologias que viabilizam a *web* de dados, possibilitando o relacionamento entre os dados disponíveis na *Internet*. Para demonstrar o relacionamento entre essas tecnologias, apresentamos, na Figura 2.1 o modelo proposto por Tim-Berners Lee conhecido como "Bolo de noiva" ou também *"Semantic Web Stack"* [12], no qual cada camada explora e utiliza capacidades das camadas abaixo. Essa Figura 2.1 apresenta, uma visão mais atualizada, do modelo que descreve os recursos e as linguagens que envolvem a *web* semântica. As tecnologias estão disposta de modo a tornar a *web* semântica possível, mostrando que a *web* semântica não é uma substituição da *web* atual, e sim uma extensão.

A integração das linguagens ou das tecnologias *eXtensible Markup Language* (XML), *Resource Description Framework* (RDF), arquiteturas de metadados, ontologias, agentes computacionais, entre outras, promove, cada vez mais, o aparecimento de serviços *web* que garantem a interoperabilidade e a cooperação entre computadores e pessoas. Essa cooperação entre computadores e pessoas se dá por meio da combinação entre páginas *web* e os recursos utilizados para identificar as páginas *web*, conhecidos como metadados.

The Semantic Web Activity no W3C agrupa todos os Grupos de Trabalho e Interesses cujos objetivos são melhorar as atuais tecnologias da *web* semântica ou contribuir para sua disseminação. Logo, trata-se de um esforço colaborativo liderado pelo W3C com a participação de um grande número de pesquisadores e parceiros industriais.

Ultimamente, tem-se associado a *web* semântica à chamada *Web* 3.0, como um próximo movimento da Internet que sucede a *Web* 2.0 (também chamada de *web* social), que tem como conceito *web* enquanto plataforma, envolvendo *wikis*, aplicativos baseados em redes sociais, *blogs* e Tecnologia da Informação.

A *web* semântica é um recurso natural para representar a proveniência, pois contém suporte explícito para representar e inferir conexões entre dados e processos, bem como para adicionar anotações a dados [13].

2.2 Metadados

Metadados, literalmente "dados sobre dados", especificamente, metadados descritivos, são dados estruturados sobre qualquer coisa que possa ser nomeada, como páginas da *web*, livros, artigos de periódicos, imagens, músicas, produtos, processos, pessoas (e suas atividades), dados, conceitos e serviços de pesquisa [14]. Os metadados possibilitam a indexação do conteúdo e a organização do conhecimento.

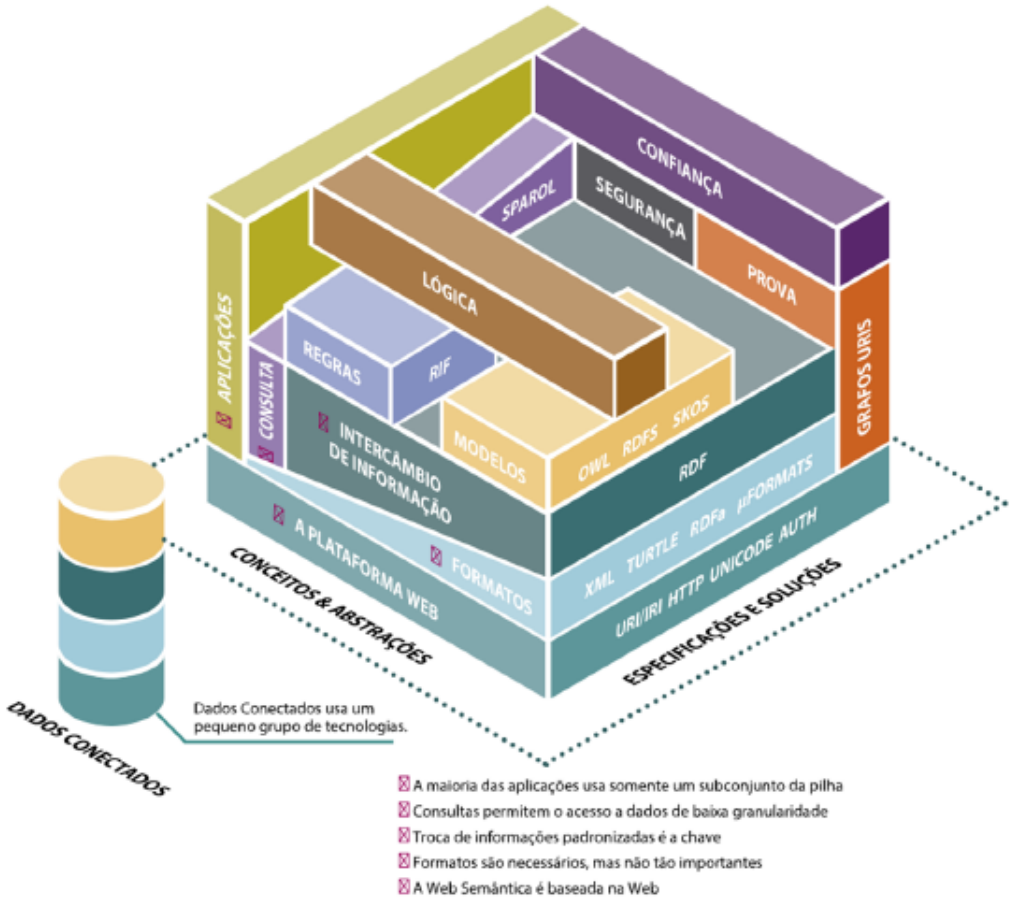

Figura 2.1: Camadas da *Web* Semântica [1].

A principal finalidade dos metadados é organizar e documentar, de forma estruturada, os dados das organizações, com o objetivo de minimizar a duplicação de esforços e facilitar a manutenção dos dados. As corporações necessitam de um maior controle de seus dados, precisam conhecer melhor o conteúdo e a qualidade dos mesmos de forma rápida, automatizada e eficiente. Um outro motivo importante para se estabelecer padrões é a necessidade de disseminação da informação e o acesso à informação de propriedade de outras organizações. Esses metadados são frequentemente governados por padrões e melhores práticas desenvolvidos e promovidos pela comunidade, a fim de garantir qualidade, consistência e interoperabilidade. Linguagens de marcação como HTML e XML fornecem uma maneira padronizada de estruturar e expressar esses padrões para processamento, publicação e implementação da máquina.

Os metadados são cruciais no gerenciamento de informações pessoais e no arquivamento digital, e para garantir recuperação e prestação de contas efetivas na manutenção de registros - algo que está se tornando cada vez mais importante com o aumento do comércio eletrônico e o uso de ferramentas e conteúdos digitais pelos governos. Em todas

essas diversas interpretações, os metadados não apenas identificam e descrevem um objeto de informação, como também documenta como esse objeto se comporta, sua função e uso, seu relacionamento com outros objetos de informação e como deve ser, e foi gerenciado ao longo do tempo.

A definição de metadados na literatura apresenta conceitos divergentes e também semelhantes nas inúmeras pesquisas e estudos sobre o tema. Nesse contexto, o que se pode perceber é a existência de grande variedade de "pontos de vista" que variam de acordo com a área de atuação do pesquisador.

De acordo com Gilliland [15], os metadados podem ser categorizados como:

- Administrativo: metadados usados no gerenciamento e na administração de coleções e recursos de informação. Por exemplo: informações de aquisição e avaliação, informações de localização, critérios de seleção para digitalização, documentação de repatriamento digital;
- Descritivo: metadados usados para identificar, autenticar e descrever coleções e recursos de informações confiáveis relacionados. Por exemplo: catalogando registros, localizando auxílios, controle de versão, índices especializados, relacionamentos conectados entre recursos;
- Preservação: metadados relacionados ao gerenciamento de preservação de coleções e de recursos de informação. Por exemplo: documentação da condição física dos recursos, documentação das ações tomadas para preservar versões físicas e digitais dos recursos (por exemplo, atualização e migração de dados), documentação de quaisquer alterações que ocorram durante a digitalização ou preservação;
- Técnico: metadados relacionados ao comportamento de um sistema ou de outros metadados. Por exemplo: documentação de hardware e software, informações procedurais geradas pelo sistema (por exemplo, metadados de roteamento e eventos), informações técnicas de digitalização (por exemplo, formatos, taxas de compressão, rotinas de dimensionamento), rastreamento de tempos de resposta do sistema, autenticação e dados de segurança (por exemplo, chaves de criptografia, senhas);
- Uso: metadados relacionados ao nível e ao tipo de uso de coleções e recursos de informação. Por exemplo: registros de circulação, registros de exposições físicas e digitais, rastreamento de uso e usuário, informações de reutilização e multiversão de conteúdo, registros de pesquisa, metadados de direitos

2.3 *Resource Description Framework* - RDF

O RDF é um padrão desenvolvido pelo *World Wide Web Consortium (W3C)* para codificar descrições de recursos em formato legível por máquina, por meio do acréscimo de metainformação a esses recursos, para que os computadores possam "entender", compartilhar e processar as informações de maneira significativa e útil. Os metadados do RDF são normalmente codificados usando sintaxe padrão, como *Extensible Markup Language* (XML) e *Turtle*[1]. Assim sendo, como o nome indica, o RDF fornece uma estrutura para a descrição dos recursos, isto é, fornece a sintaxe formal ou a estrutura da descrição do recurso, mas não fornece os valores reais dos dados a serem expressos. A semântica ou o significado devem ser especificados para um domínio ou comunidade em particular, para que os computadores possam entender os metadados codificados. A semântica é especificada por um vocabulário RDF, que é uma representação ou modelo de conhecimento dos metadados que identifica, sem ambiguidade, o que cada elemento de metadados individual significa e como ele se relaciona com os outros elementos no domínio. Os vocabulários de RDF podem ser expressos como esquemas de RDF[2] ou, quando transmitem relacionamentos mais complexos entre elementos de dados, como ontologias da *Web Ontology Language(OWL)*[3] [16].

Dessa maneira, usando a estrutura altamente extensível e robusta de RDF, esquemas RDF e ontologias OWL, podem ser criadas descrições ricas de metadados de recursos digitais que se baseiam em um conjunto teoricamente ilimitado de vocabulários semânticos. A interoperabilidade para processamento automatizado é mantida porque a sintaxe subjacente estrita exige que cada vocabulário seja especificado explicitamente [17].

A descrição sintática pode ser feita por uma série de propriedades, como tipo (*type*) e valor (*value*). Além da descrição, o RDF permite representar a relação e o significado com outros recursos (descrição semântica), por meio das chamadas triplas RDF. Esta representação é realizada por meio de uma tripla de informação que é composta inicialmente pelo “Sujeito”, que é uma entidade indexada que tem relação com uma outra entidade que representa uma informação, que é o “Objeto”. A relação entre eles é chamada de “Predicado”.

Na Figura 2.2 é apresentada a representação gráfica de um grafo da tripla RDF. O IRI (*Internationalized Resource Identifier*) é utilizado para identificar um recurso, que representa o sujeito. As propriedades que descrevem o recurso representam o predicado. O objeto é representado pela informação. A tabela descreve o conteúdo da figura.

[1] `https://www.w3.org/TR/turtle/`
[2] `https://www.w3.org/TR/rdf-schema/`
[3] `https://www.w3.org/OWL/`

Tabela 2.1: Descrição do Grafo de Tripla RDF.

Sujeito (Recurso)	**Predicado (propriedade)**	**Objeto (valor)**
http://www.cic.unb.br/ ~Ana/index.html	http://purl.org/dc/ elements/1.1/creator	"Ana Maria"
http://www.cic.unb.br/ ~Ana/index.html	http://purl.org/dc/ elements/1.1/title	"Página pessoal da professora Ana Maria"
http://www.cic.unb.br/ ~Ana/index.html	http://purl.org/dc/ elements/1.1/date	"4 de outubro de 2019"

Como o RDF fornece um número limitado de elementos predefinidos, foi criado então o RDFS - RDF Schema[4], que é uma extensão do RDF para permitir que comunidades independentes possam desenvolver seus próprios vocabulários, ou seja, novas classes e propriedades, particulares ao seu domínio de aplicação.

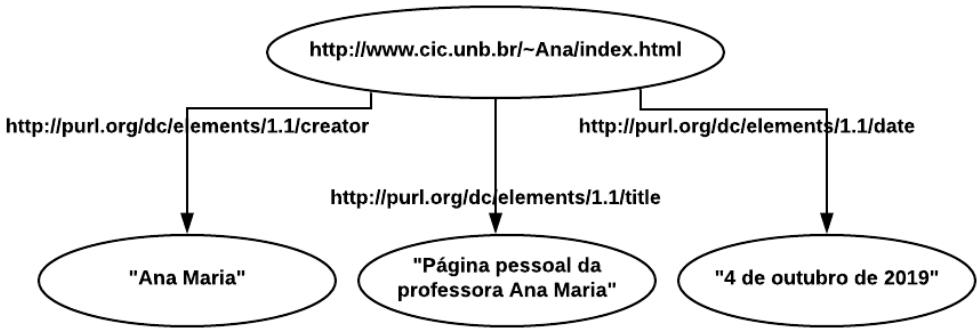

Figura 2.2: Grafo de conceito de Tripla RDF.

2.4 Ontologia

Ontologia é um termo originário da filosofia, que se tornou uma das palavras incorporadas pela Ciência da Computação, no qual é dado um sentido técnico específico, bem diferente do original. No contexto de Ciência da Computação pode-se adotar a definição de ontologia dada por [18], que define uma ontologia como "uma especificação formal explícita de uma conceitualização compartilhada". O autor Fensel [19] analisa esta definição identificando quatro principais conceitos envolvidos:

1. Conceitualização - um modelo abstrato de algum fenômeno que identifica conceitos relevantes para o mesmo;

[4] `https://www.w3.org/2001/sw/wiki/RDFS`

2. Explícita - indica elementos e restrições claramente definidos;
3. Formal - no qual a ontologia deve ser passível de processamento automático;
4. Compartilhada - indica um conhecimento consensual.

Para resumir, segundo a W3C, ontologia é a definição dos termos utilizados na descrição e na representação de uma área do conhecimento [20].

Dessa maneira, a ontologia é um tópico de pesquisa popular em várias áreas, como engenharia do conhecimento, integração de sistemas inteligentes, processamento de linguagem natural, sistemas de informação cooperativos, gerenciamento de conhecimento, software baseado em agentes, e comércio eletrônico. Na Ciência da Computação, ontologias foram desenvolvidas em inteligência artificial de modo a facilitar o compartilhamento e a reutilização da informação [19].

Em outras palavras, uma ontologia fornece uma conceitualização explícita (isto é, meta-informação) que descreve a semântica dos dados, além de um entendimento compartilhado e comum de um domínio que pode ser comunicado entre pessoas e sistemas de aplicativos heterogêneos, e amplamente difundidos. Segundo [19], as ontologias possuem uma função semelhante a de um esquema de banco de dados, no entanto com as seguintes diferenças:

- Uma linguagem para definir ontologias é sintática e semanticamente mais rica que as abordagens comuns para bancos de dados;
- As informações descritas por uma ontologia consistem em textos semiestruturados em linguagem natural e não em informações tabulares;
- Uma ontologia deve ser uma terminologia consensual e compartilhada, pois é usada para compartilhamento e troca de informações;
- Uma ontologia fornece uma teoria de domínio e não a estrutura de um contêiner de dados.

Utilizando a generalidade da ontologia como o critério principal, segundo [21] as ontologias são classificadas como:

- Ontologias de nível superior - descrevem conceitos muito genéricos, tais como espaço, tempo, e eventos. Estes seriam, em princípio, independentes de domínio e poderiam ser reutilizados na confecção de novas ontologias;
- Ontologias de domínio - descrevem o vocabulário relativo a um domínio específico por meio da especialização de conceitos presentes na ontologia de alto nível;

- Ontologias de tarefas - descrevem o vocabulário relativo a uma tarefa genérica ou atividade por meio da especialização de conceitos presentes na ontologia de alto nível;
- Ontologias de aplicação - são as ontologias mais específicas. Conceitos em ontologias de aplicação correspondem, de maneira geral, a papéis desempenhados por entidades do domínio no desenrolar de alguma tarefa.

As ontologias baseiam-se em triplas sujeito-predicado-objeto expressas em RDF/RDF-Schema, que formam a base da *Web Ontology Language* (OWL)[5]. As triplas são declarações que descrevem entidades no domínio de modelagem, suas propriedades e seus relacionamentos com outras entidades. O OWL estende a expressividade do RDF Schema para caracterizar classes e propriedades. Ele foi projetado para ser usado quando as informações precisam ser processadas por máquinas, em vez de apenas serem apresentadas. O OWL estende o RDFS definindo novas primitivas de idioma, e fornece um conjunto muito mais rico de vocabulários do que o RDFS.

2.5 Dados Conectados

As melhores práticas que permitem a criação da *Web of Data* são basicamente quatro princípios que ficaram conhecidos como princípios *Linked Data* [2]. Esses princípios exigem a identificação de entidades com URIs HTTP que possam ser resolvidos pela *web* em dados que descrevam a entidade identificada. Esses dados são representados usando o *Resource Description Framework* (RDF).

O termo Dados Conectados surgiu em 2006, com a publicação por Tim Berners-Lee, do documento *Design Issues* [22], com uma subseção de *web* semântica exclusiva para Dados Conectados. Dados Conectados (do inglês, *Linked Data*) pode ser definido como um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na *Web*, com o intuito de criar uma "*Web* de Dados" [23].

Para um melhor entendimento sobre a *Web* de Dados, pode-se estabelecer um paralelo com a *Web* de Documentos (a *Web* atual). A *Web* de Documentos faz uso do padrão HTML (*HyperText Markup Language*) para acessar dados, enquanto na *web* de dados, os dados são acessados a partir do padrão RDF. Na *web* de documentos, *hiperlinks* são usados para navegar entre as páginas, enquanto na *web* de dados os *links* RDF são usados para acessar dados de diversas fontes [1].

Berners-Lee [22] delineou um conjunto de "regras"para publicar dados na *web* de forma que todos os dados publicados se tornem parte de um único espaço global de dados:

[5] `https://www.w3.org/OWL/`

1. Use URIs como nomes para coisas;
2. Use HTTP URIs para que as pessoas possam procurar esses nomes;
3. Quando alguém procura um URI, forneça informações úteis, usando os padrões (RDF, SPARQL);
4. Inclua *links* para outros URIs, para que eles possam descobrir mais coisas.

Em 2007 surgiu o projeto de Dados Abertos Conectados, (do inglês *Linked Open Data*), da comunidade de *web* semântica, iniciado por um grupo de interesse do W3C, com o objetivo de fazer com que os dados fossem publicados e conectados de forma aberta. A tecnologia de Dados Conectados, diferente de qualquer outra tecnologia, permite que qualquer comunicação de dados possa ser composta de diferentes vocabulários.

Para que esse objetivo fosse atingido, Tim Berners-Lee propôs um conjunto de princípios conhecido como "Sistema de 5 estrelas", que classifica por estrelas o grau de abertura do dado, conforme descrito na Figura 2.3 [1].

Figura 2.3: Sistema de 5 estrelas [1].

O Sistema de 5 estrelas dos Dados Abertos Conectados apresenta a seguinte classificação:

- Se o dado for disponibilizado na *Internet* em qualquer formato, desde que seja com licença aberta, ele é classificado como uma estrela;
- Se o dado for disponibilizado na *Internet* de maneira estruturado, ele é classificado como duas estrelas;

- Se o dado for disponibilizado na *Internet*, de maneira estruturada, em formato não proprietário, ele é classificado como três estrelas;
- Se o dado for disponibilizado na *Internet*, seguindo todas as regras anteriores, mas dentro dos padrões estabelecidos pelo W3C, ele é classificado como quatro estrelas;
- Se o dado for disponibilizado na *Internet*, seguindo todas as regras anteriores, mas conectar seus dados a outros dados, de forma a fornecer um contexto, então ele é cinco estrelas.

Para auxiliar na publicação de Dados Conectados o *Word Wide Web Consortium* (W3C) criou um grupo de trabalho sobre dados abertos governamentais que definiu um conjunto de de boas práticas para facilitar o desenvolvimento e a entrega de dados abertos governamentais como Dados Abertos Conectados:

1. **Preparar os *Stakeholders*:** capacitar os usuários para criar e manter os dados conectados;
2. **Selecionar o conjunto de dados:** definir quais dados serão publicados e conectálos para reuso;
3. **Modelar dados:** os *Stakeholders* definem como irão representar os dados e como eles se relacionam com os demais dados;
4. **Especificar a licença:** especificar uma licença de dados abertos apropriada;
5. **Nomear bons URIs:** considerações sobre nomeação de objetos, suporte multilíngue, alteração de dados ao longo do tempo e estratégia de persistência são os elementos básicos para os dados conectados úteis;
6. **Usar vocabulários-padrão:** definir se irá construir um vocabulário para publicação do dado ou, preferencialmente, utilizar um vocabulário já existente, facilitando conexão com outros dados;
7. **Converter os dados:** converter os dados oriundos de uma fonte original para dados conectados;
8. **Prover acesso aos dados:** definir como humanos e máquina terão acesso aos dados;
9. **Anunciar novo conjunto de dados:** lembre-se de anunciar novos conjuntos de dados em um domínio autoritativo;
10. **Reconhecer o contrato social:** para que o publicador dos dados se comprometa em fomentar a publicação ao longo do tempo.

2.6 Dados Abertos

Dados abertos são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa, sujeitos à exigência de atribuição à fonte original e ao compartilhamento pelas mesmas licenças em que as informações foram apresentadas [24]. Portanto, esses dados são publicados e distribuídos na Internet, compartilhados em formato aberto para que possam ser lidos por pessoas e por máquinas, permitindo o cruzamento com outros dados de diferentes fontes [20]. Por esta perspectiva, a definição do termo dados abertos carrega três normas fundamentais [25]:

- Disponibilidade e acesso: os dados devem estar disponíveis como um todo e sob custo não maior que um custo razoável de reprodução, e preferencialmente devem ser possíveis de ser baixados via *Internet*. Os dados devem também estar disponíveis de uma forma conveniente e modificável;
- Reúso e redistribuição: os dados devem ser fornecidos sob termos que permitam a reutilização e a redistribuição, inclusive a combinação com outros conjuntos de dados;
- Participação universal: todos devem ser capazes de usar, reutilizar e redistribuir – não deve haver discriminação contra áreas de atuação ou contra pessoas ou grupos. Por exemplo, restrições de uso "não comercial" que impediriam o uso "comercial", ou restrições de uso para certos fins (ex.: somente educativos) excluem determinados dados do conceito de "abertos".

Os dados abertos têm sido estimulados por movimentos globais. Nesta direção, o movimento *Open Data* ganhou maior popularidade com o lançamento de iniciativas governamentais de dados abertos no Reino Unido e nos EUA [26] [27].

2.6.1 Dados Abertos Governamentais

Dados abertos governamentais são dados produzidos pelos governos, que devem ser colocados à disposição de qualquer cidadão e para qualquer fim [20]. O conceito de dados governamentais abertos foi inspirada na filosofia de código aberto (do inglês, *open source*), fundamentada por três pilares conceituais [28]:

1. **Abertura:** contribuir com informações valiosas para o mundo;
2. **Participação:** aumentar a conscientização do cidadão sobre as funções do governo para permitir maior responsabilidade;

3. **Colaboração:** permitir que o governo, o país e o mundo funcionem com mais eficiência.

Abrir e publicar dados brutos, como mapas, estatísticas de emprego, pesquisas meteorológicas, estatísticas agrícolas e registros educacionais, juntamente com suas APIs associadas, enquanto reforça e respeita as políticas de privacidade, possibilita a capacitação do governo como um todo, como provedor de infraestrutura de dados e plataforma, que promove um novo nível de transparência. Isso Promove o aumento da fiscalização dos cidadãos sobre seus governos, bem como a uma cooperação mais estreita com eles.

A política de dados abertos no Governo Federal está presente em toda administração pública, seguindo as diretivas apresentadas no Decreto 8.777 de 16 de maio de 2016 [29], que estabelece, para todos os órgãos da administração pública, a publicação do Plano de Dados Abertos - PDA. No PDA consta quais dados o órgão considera ser pertinente à publicação para a sociedade, o mapeamento sobre o domínio das informações, origem da fonte dos dados, quem são os responsáveis pelos dados e quais serão os atores que monitoram e dão sustentação à política de dados abertos. Assim sendo, alguns setores do Governo disponibilizam dados na *web* em formatos não padronizados (Excel, PDF, etc), criando barreiras para o consumo e a análise desses dados por terceiros, reduzindo a transparência efetiva do governo.

Embora muitos conjuntos de dados governamentais abertos estejam disponíveis para acesso público, a maioria foi publicada usando formatos que não permitem *links* distribuídos e não ajudam os consumidores a entenderem seu conteúdo. Como é necessário um esforço humano substancial para tornar os conjuntos de dados brutos compreensíveis e reutilizáveis, foram definidos em [30] os seguintes princípios que norteiam a publicação de Dados Abertos Governamentais (DAG):

1. **Completo:** os dados publicados não estão sujeitos a limitações de privacidade, segurança ou privilégios válidas;
2. **Primário:** Os dados são extraídos diretamente da fonte e devem ter o maior nível de granularidade possível sem tratamentos, como agregação ou modificação;
3. **Atual:** os dados devem ser atualizados o mais rápido possível para garantir maior valor;
4. **Acessível:** os dados devem ser disponibilizados para o maior número de usuários possíveis para atender o maior número de finalidades;
5. **Processável por Máquina:** os dados devem estar o minimamente estruturado para possibilitar o processamento automatizado por máquina;

6. **Não Discriminatório:** não deve conter nenhum tipo de exigência de registro para obter acesso aos dados;
7. **Não Proprietário:** os dados devem ser disponibilizados em um formato não proprietário que possibilite a livre leitura;
8. **Livre de licença:** os dados não devem estar sujeitos a nenhum tipo de regulamentação de direitos autorais, patentes, marcas registradas ou segredos comerciais.

Os dados do governo estão sendo colocados *online* para aumentar a responsabilidade, contribuir com informações valiosas sobre o mundo e permitir que o governo, o país e o mundo funcionem com mais eficiência. Dados governamentais conectados (do inglês *Linked Government Data* (LGD)) é uma técnica promissora que tem sido usada para permitir esse acesso mais eficiente aos dados do governo. Os dados governamentais conectados fazem com que os dados façam parte da *web*, onde podem ser interligados a outros dados que fornecem documentação, contexto adicional ou informações básicas necessárias [28].

2.6.2 Dados Abertos Governamentais Conectados

Dados abertos governamentais conectados [28] é uma técnica promissora para permitir acesso mais eficiente aos dados do governo. Essa abordagem faz com que os dados façam parte da web, onde podem ser interligados a outros dados que fornecem documentação, contexto adicional ou informações básicas necessárias. No entanto, perceber esse potencial é caro. Os esforços de criação de dados abertos governamentais conectados, pioneiros nos EUA e Reino Unido, mostraram que a criação de dados conectados de alta qualidade, a partir de arquivos de dados brutos, exige investimentos consideráveis em engenharia reversa, documentando elementos de dados, limpeza de dados, mapeamento de esquemas e correspondência de instâncias [31] [32] [33].

2.7 Proveniência de Dados

Os sistemas que consomem dados conectados devem avaliar a qualidade e a confiabilidade dos dados. Uma abordagem comum para a avaliação da qualidade de dados é a análise da informação de proveniência. As informações de proveniência sobre um item de dados são informações sobre o histórico do dado, desde a sua criação. A proveniência de dados ajuda a descobrir os erros ao acompanhar o histórico de criação, o processo de criação de dados, o tempo de criação, quem esteve envolvido no processo de criação e quais materiais foram usados durante a criação, sendo a gestão do dado feita desde o processo de criação até a derivação do produto de dados [34].

Em outras palavras, a proveniência pode ser definida como um registro que descreve as pessoas, as instituições, as entidades e as atividades envolvidas na produção, influência ou entrega de um dado ou coisa. Em um ambiente aberto e inclusivo, como a *web*, em que os usuários encontram informações muitas vezes contraditórias ou questionáveis, a proveniência pode ajudar esses usuários a fazerem julgamentos de confiança [20].

Assim sendo, a proveniência pode ser utilizada para diversos fins, conforme descritos por [35]:

- Qualidade dos dados: pode ser usada para estimar a qualidade e a confiabilidade dos dados com base nos dados e transformações da origem, podendo, também, fornecer instruções de prova na derivação de dados;
- Auditoria: pode ser usada para auditar os dados, determinar o uso de recursos e detectar erros na geração dos dados;
- Replicação: podem permitir a repetição de derivação de dados, e ser uma fórmula para a replicação;
- Autoria: pode estabelecer os direitos autorais e propriedade dos dados, permitir a sua reprodutibilidade, e determinar a responsabilidade em caso de dados errados;
- Informativo: pode realizar consultas, com base nos metadados de linhagem, com o objetivo de descoberta de informação. A proveniência também pode ser utilizada para fornecer um contexto para a interpretação dos dados.

2.7.1 Representação da Informação de Proveniência

Uma questão fundamental da proveniência de dados é como representar a informação gerada pela mesma. Segundo Davidson & Freire [34], existem três formas distintas de proveniência:

- Prospectiva, que captura a especificação abstrata do *workflow* como uma receita para derivação de dados futuros;
- Retrospectiva, que captura os passos que foram executados, bem como informações sobre os ambientes de execução utilizados para obter um produto de dado específico. Compreende desde o tempo de duração de cada atividade executada até a origem dos dados de entrada. Além disso, não depende do tratamento da proveniência prospectiva para ser utilizado;
- Dados Definidos pelo Usuário, podendo ser qualquer informação que o usuário julgar necessária para futura utilização. Isto inclui a documentação que não pode ser

automaticamente capturada, mas registram decisões e notas importantes. Esses dados são frequentemente capturados na forma de anotações.

A obtenção da proveniência pode seguir duas abordagens, as quais são: a abordagem preguiçosa (*lazy*) na qual a obtenção da proveniência é executada somente no momento que é solicitada; e a abordagem ansiosa (*eager*) na qual a proveniência é obtida durante a geração da informação e é armazenada para permitir futuras consultas [36].

De acordo com Freire et al. [37], uma solução de gerenciamento de proveniência consiste em três componentes principais: um mecanismo de captura, um modelo de representação, e uma infraestrutura para o acesso, o armazenamento e as consultas, conforme detalhados a seguir.

Captura de Dados de Proveniência

Um mecanismo de captura de proveniência pode trabalhar em três níveis principais [37]:

- *workflow* - o sistema gerenciador de *workflow* deve ser adaptado para capturar os dados dos diferentes processos executados;
- Sistema Operacional - usam as funcionalidades do sistema operacional para capturar informações de proveniência. Estes mecanismos não são acoplados com o *workflow* em todos os processos e, portanto, necessitam de pós-processamento para extrair as relações entre as chamadas do sistema e as tarefas;
- Atividade - tenta mesclar as melhores características dos outros dois níveis. Neste nível cada atividade do *workflow* é responsável por coletar as suas próprias informações de proveniência. A vantagem deste nível é a independência dos *workflows*, assim como no nível de Sistema Operacional. As informações mais precisas são recolhidas, assim como no nível de *workflow*. O problema deste nível é a necessidade de adaptação de atividades pré-existentes para incorporar funcionalidades de coleta de proveniência.

Modelo de Representação - *Provenance Data Model*

O *Provenance Data Model* (PROV-DM) é o modelo de dados conceitual, genérico, que forma uma base para a família de especificações de proveniência do W3C (PROV). O PROV-DM distingue as estruturas centrais, formando a essência das informações de proveniência, de estruturas estendidas que atendem a usos mais específicos de proveniência. O PROV-DM está organizado em seis componentes, respectivamente: (1) entidades e atividades, e a época em que foram criadas, usadas ou finalizadas; (2) derivações de entidades; (3) agentes, responsáveis pelas entidades geradas e atividades que aconteceram;

(4) uma noção de pacote, um mecanismo para apoiar a proveniência da proveniência; (5) propriedades para ligar entidades que se referem à mesma coisa; e, (6) coleções formando uma estrutura lógica para os seus membros [2].

O modelo de dados PROV distingue as estruturas principais das estruturas estendidas, ou seja, as estruturas centrais formam a essência da informação de proveniência, e são comumente encontradas em vários vocabulários específicos de domínio que lidam com proveniência ou tipos similares de informação. Estruturas estendidas aprimoram e refinam estruturas centrais com recursos mais expressivos para atender a usos mais avançados de proveniência. O modelo de dados PROV, compreendendo estruturas centrais e estendidas, é um modelo agnóstico de domínio, mas com pontos de extensibilidade claros, permitindo que outras extensões específicas de domínio e específicas à aplicação sejam definidas [2].

A proveniência descreve o uso e a produção de entidades por atividades, que podem ser influenciadas de várias maneiras pelos agentes. Esses tipos principais e seus relacionamentos são ilustrados pelo diagrama UML da Figura 2.4. Esses conceitos, disponíveis no núcleo do PROV-DM, são descritos com mais detalhes, a seguir [2]:

- Entidade: é algo que se deseja descrever. Eles podem ser físicos, digitais ou conceituais. Pode ser real ou imaginária por exemplo registros, conjuntos de dados, papéis e modelos;
- Atividade: é algo que ocorre durante um período de tempo e age sobre ou com entidades, ou seja, as atividades utilizam as entidades e também, as atividades podem produzir as entidades. Por fim as atividades podem incluir o consumo, e processamento, e transformação, e modificação, e realocação, e uso ou a geração de entidades;
- Agente: é algo que tem alguma forma de responsabilidade por uma atividade que ocorre, pela existência de uma entidade ou pela atividade de outro agente. Um agente pode ser um tipo específico de entidade ou atividade. Isso significa que o modelo pode ser usado para expressar a proveniência dos próprios agentes.

Assim, os relacionamentos podem ser descritos como [2]:

- Geração (*WasGeneratedBy*): é a produção de uma nova entidade por uma atividade;
- Uso (used): é a utilização de uma entidade por uma atividade;
- Comunicação (*wasInformedBy*): a geração de uma entidade por uma atividade e seu uso subsequente por outra atividade, ou seja, é a troca de alguma entidade não especificada por duas atividades, uma atividade usando alguma entidade gerada pela outra;

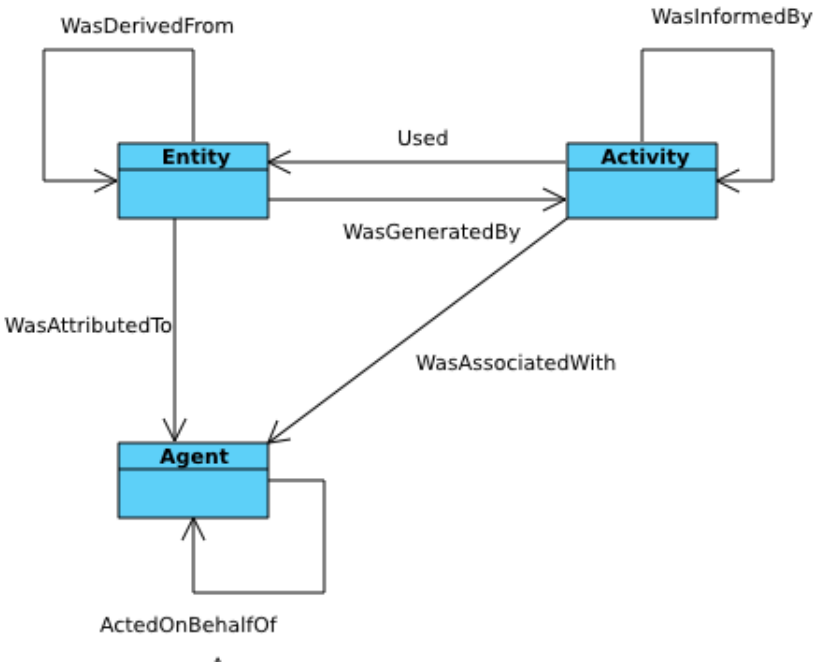

Figura 2.4: Estruturas Principais do PROV [2].

- Derivação (*wasDerivedFrom*): 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. O foco da derivação é conectar uma entidade gerada a uma entidade usada;
- Atribuição (*wasAttributedTo*): é a atribuição de uma entidade a um agente;
- Associação (*wasAssociatedWith*): é uma atribuição de responsabilidade a um agente de uma atividade, indicando que o agente tinha uma função na atividade;
- Delegação (*actedOnBehalfOf*): é a atribuição de autoridade e responsabilidade de um agente (por si ou por outro agente) para realizar uma atividade específica como um delegado ou representante.

Descrições de proveniência podem ser representadas graficamente. A ilustração não pretende representar todos os detalhes do modelo, mas pretende mostrar a essência de um conjunto de descrições de proveniência. A representação é feita por meio de um grafo. Entidades, Atividades e Agentes são representados como nós, com formas oval, retangular e pentagonal, respectivamente. Uso, Geração, Derivação e Associação são representados como arestas direcionadas. A Figura 2.5 mostra um exemplo de grafo baseado no modelo PROV-DM.

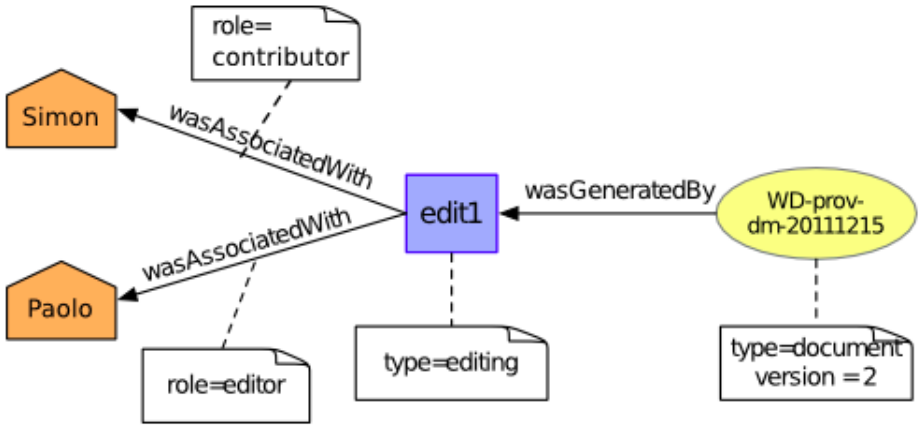

Figura 2.5: Exemplo de grafo baseado no modelo PROV-DM [2].

A escolha do modelo PROV-DM deu-se baseada em [38] que apresentou uma comparação entre os principais modelos de proveniência de dados disponíveis na literatura. Os principais requisitos considerados foram:

- Capacidade de representação da proveniência de uma peça de dado, descrevendo os processos e os insumos utilizados em sua geração;
- Representação gráfica adequada, com diferentes símbolos para cada elemento, e relações suficientes para demonstrar a proveniência de forma objetiva;
- Símbolo para representar grandes conjuntos de dados;
- Extenso material disponível cobrindo diferentes aspectos da proveniência de dados;
- O fato de ser recomendado pela W3C desde 2013;
- Capacidade do modelo de proporcionar o intercâmbio de informações entre diferentes sistemas.

Infraestrutura para acesso, armazenamento e consulta

Segundo [37], a infraestrutura para a consulta efetiva e eficiente dos dados de proveniência é um componente necessário de um sistema de gerenciamento de proveniência, especialmente quando grandes volumes de informações são capturados.

A capacidade de consultar a proveniência permite a reutilização do conhecimento. As informações de proveniência geralmente são associadas a produtos de dados (como imagens ou grafos), portanto, esses dados também ajudam os usuários a fazer consultas estruturadas sobre dados não estruturados. Um recurso comum em muitas abordagens para consultar a proveniência é que suas soluções estão intimamente ligadas aos modelos de armazenamento usados. No entanto, mesmo as consultas que usam uma linguagem

projetada para proveniência provavelmente são muito complicadas para muitos usuários, porque a proveniência contém informações estruturais representadas como um grafo [37].

Alguns modelos de proveniência usam a tecnologia da Web semântica para representar e consultar informações de proveniência. Linguagens da Web semântica, como RDF e OWL, fornecem uma maneira natural de modelar grafos de proveniência e a capacidade de representar conhecimentos complexos, como anotações e metadados [37].

2.8 GRAPHED - *Graph Description Diagram for Graph Databases*

GRAPHED foi definido como um diagrama para resolver a falta de uma notação genérica e abrangente para modelagem de bancos de dados orientado a grafos. O GRAPHED formaliza representações de objetos contidos em grafos, para fornecer um diagrama representando o modelo de dados para bancos de dados orientado a grafos.

Essa notação foi projetada para ser independente, e para ser utilizada na modelagem de grafos conceituais. GRAPHED formaliza representações de objetos contidos em grafos, para fornecer um diagrama representando o modelo de dados para bancos de dados orientado a grafos. Dentro de seus elementos básicos, ele suporta a ideia de um rótulo para o domínio do relacionamento no esquema e valores em instâncias, além de notação para peso e tipos. Ele também abrange *hyperedges* e *hypernodes*, generalizando a borda simples e usando um conceito reestruturado do *hypernode*.

Na notação GRAPHED, tem-se o vértice é representado por um retângulo arredondado, com um rótulo obrigatório dentro. O rótulo dentro da caixa é usado principalmente para identificar o tipo de entidade de vértice, mas pode ser usado para apresentar o nome da entidade. A estrutura do grafo suporta atributos, eles são indicados abaixo do Rótulo, após o símbolo de dois pontos ":", tem-se o tipo à direita. Como o modelo precisa mostrar quais atributos identificam exclusivamente uma entidade, o GRAPHED usa um asterisco "*" ao lado do nome do atributo quando ele faz parte da chave. O próximo passo é conectar os vértices por arestas, que representam um tipo de relacionamento entre as entidades. O relacionamento que uma aresta representa pode ser mútuo ou apenas em uma direção. Portanto, existem duas notações para arestas: a borda não direcionada e a borda direcionada. Outras informações opcionais que complementam a aresta estão incluídas como parte do rótulo. Essas informações incluem a cardinalidade entre os vértices para esse relacionamento, representado por números entre parênteses.

2.9 Banco de Dados Orientado a Grafo

Um banco de dados orientado a grafos representa os dados de acordo com a Teoria de Grafos. Formalmente, um grafo é apenas uma coleção de vértices e arestas, que representam um conjunto de nós e os relacionamentos que os conectam. Os grafos representam entidades como nós e as maneiras pelas quais essas entidades se relacionam com o mundo como relacionamentos. Os nós e os relacionamentos podem ter propriedades. Os relacionamentos têm significado direcional, permitindo encontrar padrões entre os nós. Um sistema de gerenciador de banco de dados de grafos é um sistema de gerenciamento de banco de dados *on-line* com os métodos Criar, Ler, Atualizar e Excluir (CRUD) que expõem um modelo de dados de grafos. Os bancos de dados orientado a grafos, geralmente, são criados para uso com sistemas transacionais (OLTP) [3].

No exemplo apresentado na Figura 2.6, foi representada uma pequena rede de usuários do *Twitter* como um grafo. Cada nó é rotulado como Usuário, indicando seu papel na rede. Esses nós são então conectados com relacionamentos, que ajudam a estabelecer ainda mais o contexto semântico.

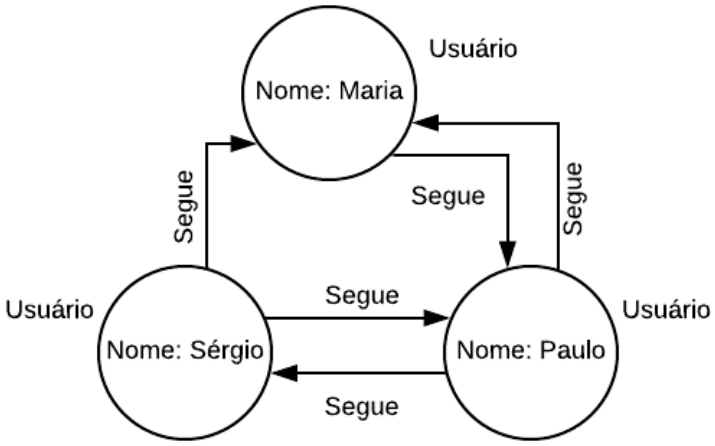

Figura 2.6: Exemplo de grafo [3].

Conforme [39] e [40] os bancos de dados orientado a grafo podem ser classificados em:

- Grafo de Atributos: permite que tanto os vértices como as arestas armazenem informações de atributos como um par de chave e valor. Essa característica permite que certas informações sejam armazenadas próximas aos nós, facilitando a recuperação do dado, além de simplificar a diagramação do modelo, pois são necessários menos vértices, caso alguns atributos não sejam utilizados para expressar relacionamentos

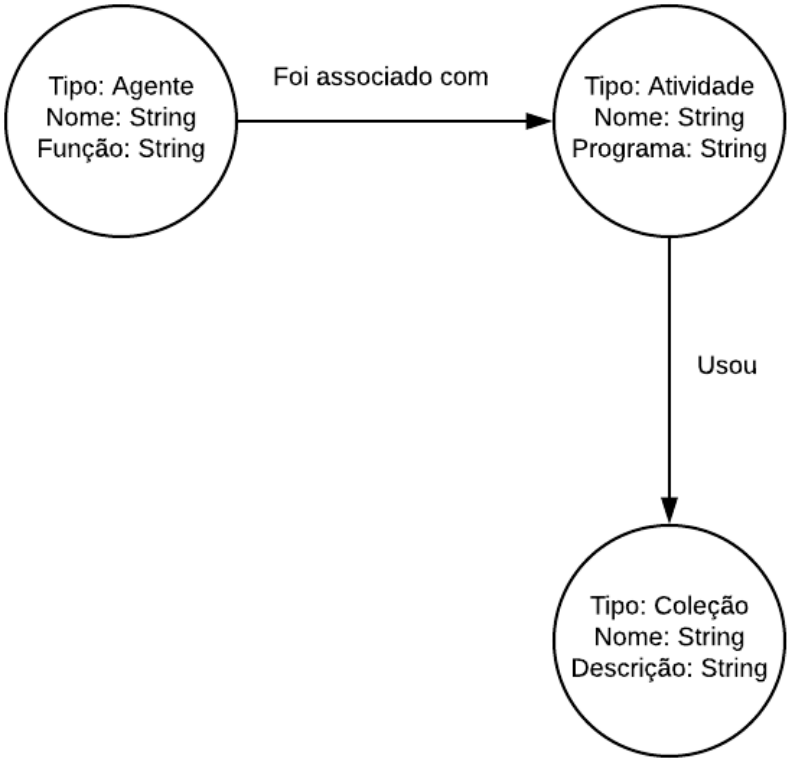

Figura 2.7: Exemplo de grafo de Atributo.

entre as instâncias. Uma vantagem é beneficiar certos tipos de consultas que realizam buscas sobre atributos. Um exemplo de grafo de atributos é apresentado na Figura 2.7;

- Grafo Simples: o modelo de dados tem apenas vértices e arestas cabendo aos rótulos armazenar os valores identificadores das instâncias. Dessa forma, os atributos do grafo transformam-se também em nós, como se fossem entidades. O que cria um modelo mais complexo com uma quantidade maior de relacionamento, porém mais normalizado, pois qualquer atributo que possa ser duplicado é compartilhado, mantendo-se apenas uma instância. A desvantagem dessa modelagem é o aumento da quantidade de elementos no modelo. Além disso, algumas consultas também podem ficar mais complexas pois deve-se buscar os vértices ligados a uma instância se for necessário verificar algum atributo. A Figura 2.8 apresenta um modelo de um grafo simples;
- Multigrafos: diferem dos grafos em que dois nós podem ser conectados por várias arestas. Por exemplo, considerando as cidades de um mapa como nós e as estradas entre elas como arestas, dois nós poderão ter múltiplas conexões, conforme mostrado

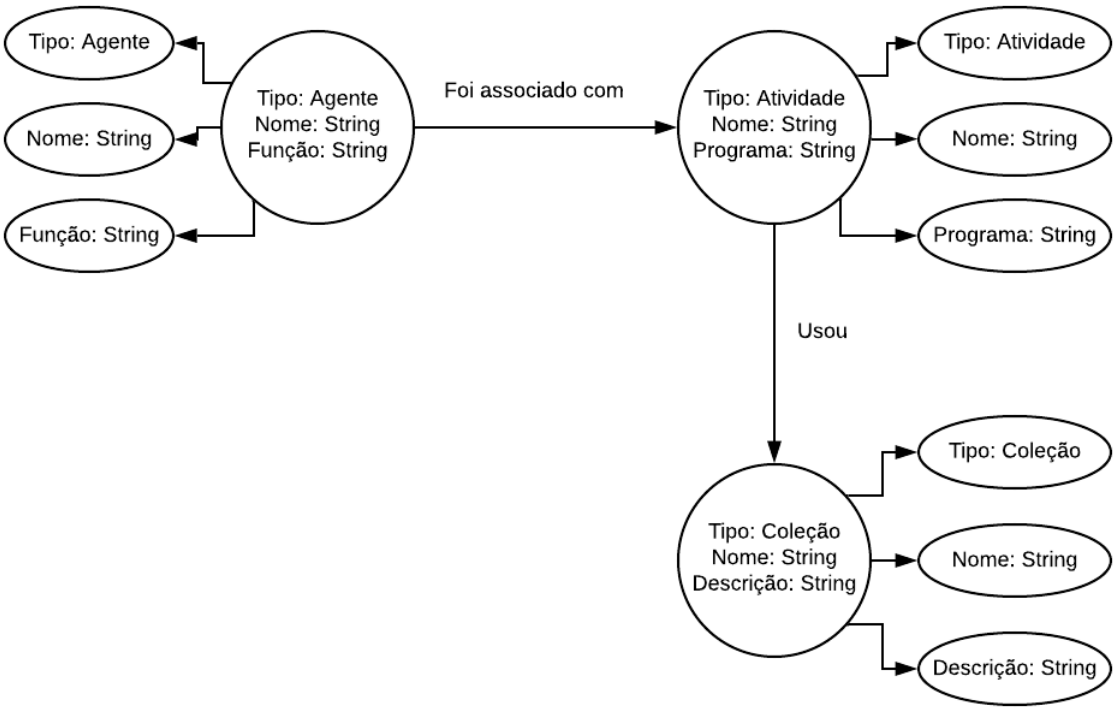

Figura 2.8: Exemplo de grafo Simples.

na Figura 2.9, onde há duas arestas (estradas) entre os nós Fortaleza e Brasília;

- Hipergrafos: um hipergrafo permite que suas arestas liguem mais de dois vértices. Ou seja, a relação entre seus vértices não precisa ser binária. Isso permite que em certas situações onde o mesmo relacionamento ocorre entre várias instâncias seja utilizando apenas um arco. Pode ser usado para modelagem de problemas relacionado ao fluxo de atividades mostrando como representar um fluxo de tarefas.
- Grafo aninhado: são grafos que suportam outros grafos como nós. Um grafo pode ser vinculado a outro, como se fosse um vértice. Um grafo agrupado torna-se um hipernó, sendo tratado como um vértice comum, e o grafo passa a ser chamado de aninhado.

2.10 Arquitetura de Publicação de Dados Abertos Governamentais Conectados

A Universidade de Brasília (UnB) é uma instituição pública criada pela Lei 3.998 em 15 de dezembro de 1961 [41]. A UnB está sujeita à Lei de Acesso à Informação, bem como todas as entidades da administração pública brasileira.

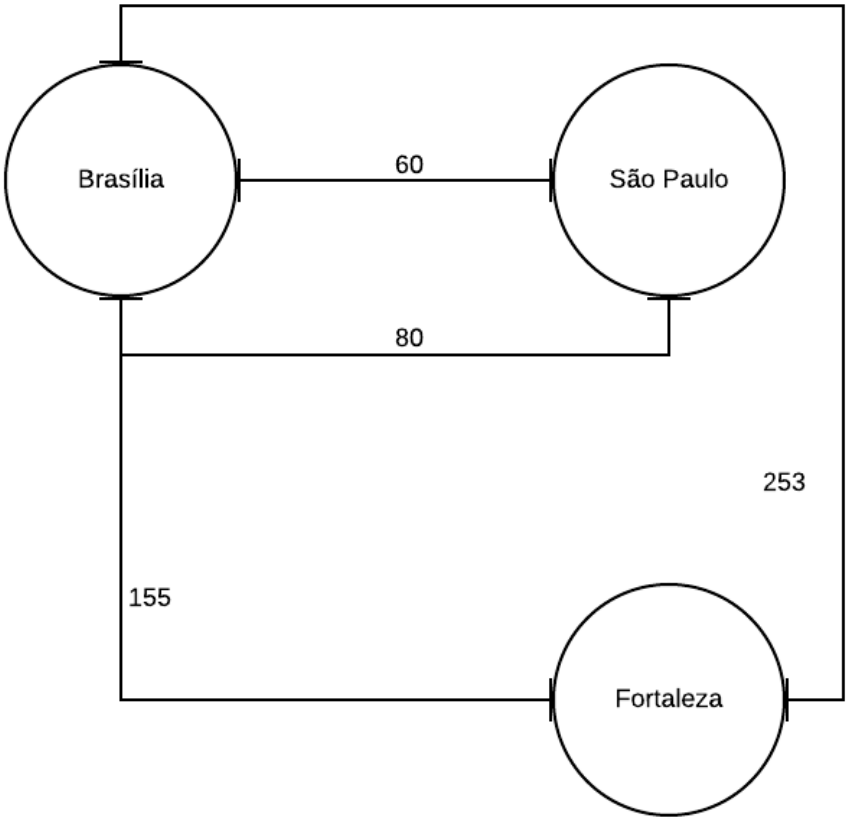

Figura 2.9: Exemplo de Multigrafo.

O UnBGOLD (UnB - *Governamental Linked Open Data*) faz parte de um *workflow* desenvolvido pela UnB para melhorar a qualidade da publicação dos dados da UnB, utilizando como parâmetro de qualidade o índice 5 estrelas, proposto por Berners-Lee [11], e a definição de [30], em que os dados quanto mais atuais possuem maior valor. O UnBGOLD oferece a opção de enriquecer semanticamente os dados por meio de vocabulário controlado, utilizando metadados e ontologias, transformando os dados abertos em dados abertos conectados.

A arquitetura do *workflow* de Publicação de Dados Abertos Governamentais da UnB está dividida nas seguintes camadas:

- Primeira camada: é realizada a extração dos Dados. Nesta camada são definidos os procedimentos de seleção, curadoria e extração dos dados, regras para acesso e disponibilização dos dados;
- Segunda camada: é a camada na qual ocorre a indexação semântica. Esta camada utiliza o UnBGOLD, que oferece uma interface *web* em que os agentes publicadores realizam a indexação semântica dos conjuntos de dados utilizando metadados e

ontologias, e determinam os parâmetros para publicação e atualizam o Catálogo de Dados Abertos Conectados;

- Terceira Camada: é a camada de publicação dos dados. Os dados são publicados pela ferramenta CKAN[6] na qual ficarão disponíveis para os usuários finais utilizarem, além de uma interface de busca onde a pesquisa é realizada no banco de dados do Catálogo de Dados Abertos Conectados com o resultado enriquecido semanticamente.

Uma das características da arquitetura de publicação é o baixo acoplamento entre as camadas, ou seja, a camada de extração pode ser substituída por qualquer outro serviço que ofereça os dados, e a camada de publicação não é vinculada a uma instância CKAN específica, sendo possível que a ferramenta possa publicar em diferentes instâncias. Esta característica é importante porque permite que o UnBGOLD tenha a flexibilidade necessária para ser utilizado por outros órgãos que estejam interessados nesta funcionalidade, desde que este órgão tenha definido as regras de acesso ao serviço de dados, e mantenha uma instância CKAN habilitada para acesso via API do próprio CKAN. A Figura 2.10 apresenta a arquitetura de publicação de dados abertos.

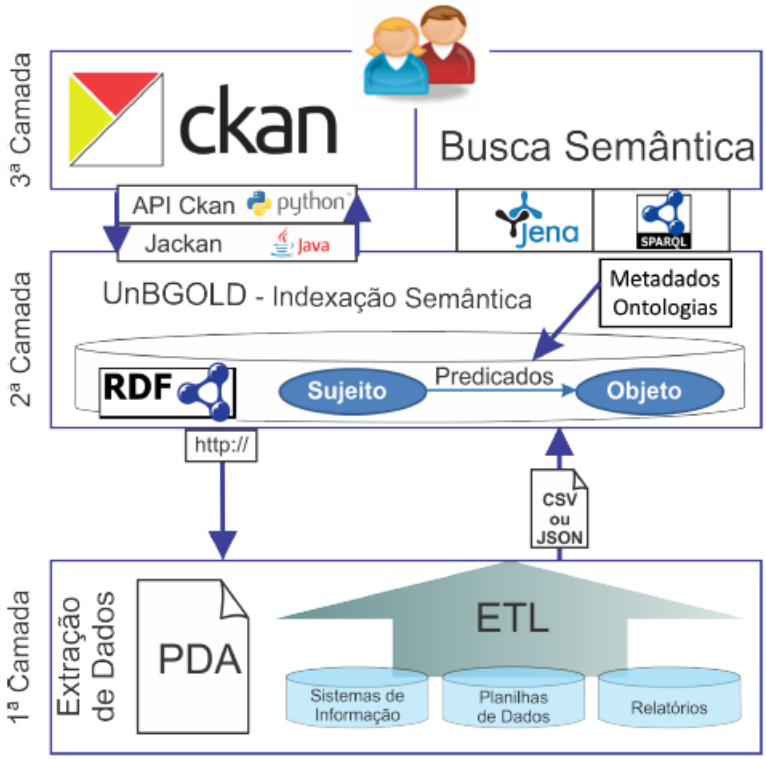

Figura 2.10: Arquitetura de Publicação de Dados Abertos [4].

[6]https://ckan.org/

A camada de extração de dados diz respeito à origem dos dados que passa por um processo de ETL, sendo armazenados em um *Data Warehouse* (DW), que já possui dados limpos originados dos bancos de dados do sistema de informação da UnB.

A camada de indexação semântica consiste em duas partes. A primeira parte, diz respeito a dar significado semântico aos dados que se deseja publicar, tornando-os assim dados abertos conectados. Já a segunda parte, consiste em indexar a fonte de dados, por meio da criação de um catálogo de conjuntos de dados conectados, que irá conter as informações sobre os conjuntos publicados no qual será possível realizar pesquisas sobre as características dos dados e os resultados serão enriquecidos semanticamente.

A camada de publicação de dados diz respeito à interface do consumo final dos usuários. Uma ferramenta de busca semântica está disponível com dados já indexados. A UnB utiliza a solução CKAN, que é uma plataforma *web* projetada pela *Open Knowledge Foundation (OKF)* para a publicação e o compartilhamento de dados abertos.

O CKAN é um sistema de gerenciamento de dados que fornece as ferramentas para publicação, compartilhamento, localização e uso de dados. É utilizado por muitas organizações de vários governos (por exemplo, Estados Unidos, Brasil, União Europeia), para tornar suas enormes fontes de dados abertas e disponíveis. Geralmente, essas organizações implantam suas próprias instâncias do CKAN, personalizando sua interface de usuário padrão e fornecendo seu próprio armazenamento de dados. Os metadados CKAN são expostos por meio de uma *API RESTful* e os *uploaders* de dados precisam preencher os metadados com a solicitação da API. O CKAN é um *"open data hub"*, ou seja, um registro no qual as pessoas podem compartilhar conjuntos de dados publicamente, em conjunto com seus metadados e informações de acesso. O CKAN pode ser visto também como uma plataforma para o *crowdsourcing* (contribuição coletiva) com uma lista abrangente de conjuntos de dados disponíveis. Ele desfruta de uma comunidade ativa que está constantemente melhorando e mantendo as descrições de conjuntos de dados.

Capítulo 3

Trabalhos Relacionados

Proveniência de Dados é um importante tema que vem sendo amplamente pesquisado ao longo do tempo por meio de trabalhos que representam diferentes perspectivas, como é possível observar em *surveys* e revisões de literatura:

- Simmhan et al. em [42] desenvolveram uma taxonomia das características de proveniência de dados com foco em abordagens de *workflows* científicos. Nesta pesquisa os autores categorizam os sistemas de proveniência com base na motivação do registro da proveniência, na descrição, na representação e no armazenamento da proveniência, e nas formas de disseminá-la;
- Cruz et al. apresentou em [43] uma nova taxonomia sobre as características da proveniência, baseada na taxonomia de Simmhan et al. [42], com enfoque no ciclo de vida dos *workflows* científicos, permitindo aos cientistas da computação distinguirem entre diferentes perspectivas de proveniência, e orientar para uma melhor compreensão dos dados de proveniência em geral;
- Buneman & Davidson em [44] discutiram, brevemente, sobre a importância da proveniência, os modelos de proveniência e sua utilização em diferentes áreas, e a granularidade da proveniência. Eles apresentaram também alguns tipos de captura de proveniência, por fim descreveram vários tópicos referentes ao uso e a qualidade dos dados relacionados a proveniência;
- Herschel & Hlawatsch em [45] forneceram uma comparação e uma visão geral dos diferentes tipos de proveniência, além de apresentar alguns conceitos fundamentais de visualização a fim de permitir que os usuários aproveitem a proveniência por meio de visualizações adaptadas;
- Herschel et al. apresentaram em [46] uma visão geral do campo de pesquisa de proveniência, fornecendo uma classificação e a revisão do estado da arte sobre as

seguintes questões: em que a proveniência é usada; quais tipos de proveniência foram definidos e capturados para as diferentes aplicações; e quais recursos e requisitos de sistema que impactam na escolha de uma solução de proveniência específica. .

A proveniência é utilizada para vários propósitos, em diferentes áreas da ciência da computação, como: *workflows* científicos, processos de negócios, bancos de dados, ambientes distribuídos, segurança e também a Web Semântica. Na Bioinformática, conforme pode ser evidenciado em diferentes estudos, [47] [48] [49] [50], a proveniência de dados é utilizada para a reprodutibilidade do experimento, na segurança [51] [52] [53] para rastrear o caminho do dado. No contexto de dados abertos conectados tem-se os artigos apresentados a seguir.

Cordeiro et al. [54] descreveram uma abordagem para apoiar a exposição, o compartilhamento e a associação de recursos de dados na forma de dados conectados, oferecendo um ambiente amigável para estimular a publicação de dados governamentais, que inclui uma ferramenta *Extract Transform Load* (ETL) para gerenciar o processo de publicação; *plugins* projetados para facilitar a extração, transformação e mapeamento de dados brutos em RDF e indexação semântica, além de um repositório para gerenciar e armazenar dados no formato RDF. Para promover uma integração e a reutilização das ferramentas existentes em uma arquitetura extensível, as ferramentas são orquestradas e gerenciadas por meio de uma infraestrutura de *workflow*, que pode ser configurada para permitir a implementação da arquitetura em diferentes domínios e ambientes computacionais. A arquitetura proposta se preocupa com o foco na qualidade dos dados, captação de dados de proveniência e suporte ao tratamento semântico. Esta arquitetura não apresenta mecanismos para auxiliar no processo de curadoria de dados e não realiza a gestão da proveniência. Além de não abordar a gestão da proveniência, a arquitetura não utiliza um modelo de dados para representação da proveniência, impossibilitando que informações de proveniência possam ser trocadas, processadas ou discutidas.

É possível citar também, uma abordagem descrita em [32], que permite aos consumidores de dados fazerem a conversão de dados governamentais conectados, sem esperarem que o governo o faça. Essa transferência da responsabilidade, da conversão dos dados brutos em dados conectados, para o consumidor de dados, chamada de abordagem de auto-atendimento, pode introduzir no processo de publicação, um usuário, não qualificado o bastante, para realizar a tarefa de publicação de dados abertos governamentais, com a qualidade desejada, em virtude da possibilidade deste usuário gerar inconsistências no processo de indexação semântica, por exemplo. Este estudo definiu um *pipeline* de publicação, centrado em torno do *Google Refine*[1], que suporta uma conversão de ponta a ponta de dados governamentais brutos para a dados governamentais conectados. Para

[1] `http://code.google.com/p/google-refine/`

transformação dos dados brutos para o formato RDF e posteriormente, realizar a indexação semântica desses dados, foi desenvolvida uma extensão para o *Google Refine*.O *pipeline* permite a publicação dos dados na plataforma CKAN[2] de registro de dados abertos. O *pipeline* de Publicação de dados governamentais conectados é capaz de capturar as informações de proveniência, representadas de acordo com o OPMV[3] e compartilhando-as junto com os dados no CKAN. Para representação dos dados do governo, foi desenvolvido o vocabulário DCAT [55]. Além disso, foram utilizados também, na representação de dados, diversos vocabulários como o VoID e o *Data Cube Vocabulary*[4]. Também foi descrita a aplicação do *pipeline* de publicação a um catálogo do governo local na Irlanda, resultando em uma quantidade significativa de dados indexados publicados. Apesar de realizar, na plataforma CKAN, a publicação dos dados conectados com a representação da proveniência através do vocabulário OPMV, um vocabulário baseado no vocabulário OPM [56], a abordagem não realiza a modelagem dos dados para representação da proveniência.

Um esforço importante foi apresentado em [57] para coleta de proveniência em dados abertos conectados. A proveniência foi capturada por mecanismo de publicação de dados conectados apoiado por um *workflow* de ETL. A coleta da proveniência, do tipo prospectiva e retrospectiva, foi realizada em três níveis, no qual a camada inferior utiliza o *Open Provenance Model Vocabulary* (OPMV) para descrever a proveniência básica para o *workflow*, a segunda camada usa o vocabulário Cogs[5] para descrever a proveniência relacionada às operações de ETL e a terceira camada usa um vocabulário específico para descrever a proveniência relacionada a objetos específicos do domínio do *workflow*. Os dados de proveniência são publicados assim como os dados conectados. A arquitetura foi validada em um cenário real, através da publicação de dados governamentais abertos, de forma conectada, com a proveniência relacionada. Essa abordagem também não realiza a modelagem dos dados para representação da proveniência.

Mendonça et al. [58] amplia trabalhos anteriores, do próprio autor em [57] e dos autores Campos & Guizzardi [59], que introduziram *workflows* de ETL de dados para publicar dados conectados. Este estudo realiza a coleta e a publicação de metadados de proveniência, do tipo prospectiva e retrospectiva, em diferentes níveis de granularidades, e seu impacto no processo de publicação de dados conectados realizado dentro dos limites de uma organização. A solução utilizou *workflows* ETL para publicação de dados conectados de fontes de dados heterogêneas, com a captura e publicação dos metadados de proveni-

[2] `https://ckan.org/`
[3] `http://open-biomed.sourceforge.net/opmv/ns.html`
[4] `https://www.w3.org/TR/vocab-data-cube/`
[5] `http://vocab.deri.ie/cogs`

ência por meio de triplas RDF anotadas em ontologias preexistentes, PROV-O[6], OPMW[7] e COGS[8], sendo que as duas últimas foram utilizadas como extensão da primeira. Essa abordagem, assim como as outras, não realiza a modelagem dos dados para representação da proveniência.

Foi apresentado em [60] uma abordagem para capturar, gerenciar e publicar os metadados de proveniência gerados nos processos de monitoramento ambiental, cuja arquitetura foi subdividida em um modelo de dados baseado em PROV-DM e *Dublin Core*[9], um repositório de grafos RDF e uma API da *Web* que fornece serviços para coletar, armazenar e consultar metadados de proveniência. Neste estudo foi demonstrada a eficácia na coleta e armazenamento de metadados de proveniência permitindo a consulta de toda a proveniência de conjuntos de dados e produtos de dados, além da reutilização, descoberta e visualização de dados brutos, processos e cientistas envolvidos em sua geração e evolução.

Para Trinh et al. [61], o desafio foi apresentar um modelo baseado na definição de proveniência centrada no *workflow*, que facilita a geração automática de informações de proveniência semântica para processos genéricos de integração de dados conectados. Este modelo foi validado por meio da implementação de um modelo genérico em um ambiente de *mashup* colaborativo e posterior avaliação por meio de um aplicativo de exemplo. Ao desenvolver o modelo de proveniência, utilizando o PROV-DM, a maior parte do vocabulário foi reutilizado da ontologia PROV-O, além disso, vários conceitos foram fornecidos pelos vocabulários VOID[10] e FOAF[11].

Semelhante as abordagens [54], [32], [57] e [58], neste trabalho foi definida uma arquitetura que captura a proveniência de um *workflow* de publicação de dados abertos governamentais. No entanto, este estudo traz como vantagem, o fato de todo o processo de publicação ser realizado por um usuário interno à organização, e não pelo consumidor de dados, como ocorre em [32], proporcionando uma maior confiabilidade no processo publicação de dados abertos governamentais.

Para representação da proveniência, este estudo realizou a modelagem da proveniência, por meio do modelo de dados W3C PROV-DM [2], assim como, as abordagens descritas em [60] e [61]. O PROV-DM, além de se tratar de uma recomendação da W3C, oferece uma representação mais rica (incluindo restrições) e mais formal, quando comparado com outros modelos de proveniência como o OPM.

Do mesmo jeito que ocorre com a abordagem de [32], a arquitetura desenvolvida neste trabalho, realiza a publicação dos metadados de proveniência, em conjunto com os dados

[6] `https://www.w3.org/TR/prov-o/`
[7] `https://www.opmw.org/model/OPMW/`
[8] `http://vocab.deri.ie/cogs`
[9] `http://dublincore.org/`
[10] `http://www.w3.org/TR/void/`
[11] `http://xmlns.com/foaf/spec/`

abertos conectados, na plataforma CKAN de dados abertos.

Na Tabela 3.1 foi apresentado um resumo dos trabalhos relacionados, no contexto de dados abertos conectados, em termos do mecanismo no qual os dados de proveniência foram capturados, da modelagem da proveniência, dos vocabulários utilizados e da representação da proveniência. Como pode ser observado, todos os trabalhos relacionados apresentam um mecanismo de captura da proveniência baseado em um *Workflow*, assim como, a arquitetura proposta neste estudo. Para modelagem da proveniência, apenas as abordagens descritas em [60] e [61] utilizam o PROV-DM. Com relação aos vocabulários utilizados, a abordagem descrita Silva (2016) [60] utiliza os mesmos vocabulários que este estudo utiliza. Por fim, com relação à representação da proveniência, apenas os trabalhos realizados por [57] e [58] trazem essa informação, que difere da abordagem proposta neste estudo, pois além da proveniência prospectiva e retrospectiva, este estudo representa a proveniência de dados definidos pelo usuário.

Este estudo, comparado com os trabalhos relacionados, no contexto de dados conectados, é o que traz uma abordagem mais completa, com todos os componentes necessários para criar uma solução de gerenciamento da proveniência, segundo Freire et al. [37], e que possui agregado à solução, as melhores práticas utilizadas nesses trabalhos relacionados.

Tabela 3.1: Resumo dos trabalhos relacionados.

Trabalhos Correlatos	**Mecanismo de captura**	**Modelagem da Proveniência**	**Vocabulários utilizados**	**Representação da Proveniência**
Cordeiro et al. [54] (2011)	Baseado em um *Workflow* ETL	-	-	Não Informado
Maali et al. [32] (2012)	Baseado em um *Workflow* de publicação de dados conectados	-	OPMV, DCAT, VOID e Data Cube	Não Informado
Mendonça et al. [57] (2013)	Baseado em um *Workflow* ETL	-	OPMV, COGS	Prospectiva e retrospectiva
Mendonça et al. [58] (2016)	Baseado em um *Workflow* ETL	-	PROV-O, OPMW e COGS	Prospectiva e retrospectiva
Silva [60] (2016)	Baseado em um *Workflow* de monitoramento ambiental	PROV-DM	PROV-O, Dublin Core, FOAF	Não Informado
Trinh et al. [61] (2017)	Baseado em um *Workflow* genérico	PROV-DM	PROV-O, VOID e FOAF	Não Informado
Arquitetura Proposta	Baseado em um *Workflow* de publicação de dados conectados	PROV-DM	PROV-O, VOID e FOAF	Prospectiva, retrospectiva e de dados definidos pelo usuário

Capítulo 4

UnBGOLDProv

A Seção 4.1 apresenta o UnBGOLDProv, e seus benefícios. A Seção 4.2 apresenta um conjunto mínimo de metadados de proveniência a ser capturado e o correspondente vocabulário para representação semântica. A Seção 4.3 apresenta o mapeamento de um modelo de dados de proveniência para um modelo de grafos. A Seção 4.4 apresenta a arquitetura do UnBGOLDProv e as tecnologias utilizadas. A Seção 4.5 descreve o processo de captura e publicação dos metadados de proveniência do *workflow* de publicação. Por fim, a Seção 4.6 apresenta as principais características do Neo4J.

4.1 UnBGOLDProv - UnB *Government Linked Open Data Provenance*

Um *workflow* de publicação de dados abertos governamentais conectados oferece muitos benefícios, como a recuperação, a conversão, o aprimoramento e a republicação dos dados de outra organização, no entanto, levantam-se questões importantes sobre a integridade dos produtos de dados resultantes do *workflow*. Com a integração de dados de diferentes fontes, informações de como, de quem, de onde e quando as informações vieram, correm risco de se perder, ou seja, uma exibição de conteúdo integrada pode ocultar respostas importantes sobre como a informação surgiu. A preservação dessas informações tornam-se cada vez mais importantes quando as fontes de integração variam significativamente em graus de autoridade, reputação, políticas e documentação. Assim sendo, a fonte original dos dados deve ser armazenada, juntamente com a descrição completa de todas as operações que foram aplicadas aos dados, permitindo assim, que o processo de publicação possa ser examinado e reproduzido novamente por meio da proveniência de dados.

O UnBGOLDProv visa capturar, gerenciar e publicar os metadados de proveniência gerados durante o processo de publicação de dados governamentais abertos conectados.

A arquitetura proposta expande um trabalho anterior, adicionando o recurso de gerenciamento de proveniência de dados ao *workflow* de publicação de dados abertos governamentais conectados da UnB que utiliza a ferramenta UnBGOLD (UnB *Government Linked Open Data*) [4]. O UnBGOLDProv é uma aplicação *web* desenvolvida em linguagem Java, que utiliza o *framework Spring Boot* [62] e o sistema gerenciador de banco de dados Neo4J. O *Spring Boot* é um *framework* para criação de aplicações *web* em Java, e se destaca por possibilitar o desenvolvimento de aplicações *"stand-alone"*, ou seja, que não necessitam de um servidor para serem executadas, pois a própria aplicação contém um servidor embutido.

Para suportar a captura, a gravação, a consulta e o gerenciamento dos metadados de proveniência, especificou-se uma arquitetura computacional de proveniência composta de quatro componentes principais:

- Um modelo de dados de proveniência, que utiliza o PROV-DM, para organizar e estruturar os metadados de proveniência;
- Um serviço que realiza a captura da proveniência automática, por meio de uma requisição HTTP. Esse serviço captura a proveniência do tipo prospectiva (já que captura os passos que devem ser seguidos para a geração de um dado produto), retrospectiva (pois captura informações obtidas durante a execução dos processos de geração do dado, por exemplo a duração de cada atividade executada) e de dados definidos pelo usuário (captura informações como funções e versões dos programas utilizados). Quanto ao nível de captura realizada é do tipo *workflow*, pois realiza a captura das tarefas que fazem parte do *workflow*;
- Um serviço que realiza a publicação dos metadados de proveniência, por meio da plataforma CKAN. A comunicação entre UnBGOLDProv e o CKAN é realizada por meio de uma API que o CKAN disponibiliza para gerenciamento dos conjuntos de dados [63]. Para fazer a comunicação com a API, agregou-se ao UnBGOLDProv a biblioteca Jackan[1].
- Um repositório de proveniência para armazenar metadados de proveniência e fornecer recursos para consultar e analisar esses dados.

Assim sendo, a proposta oferece os seguintes benefícios:

- Melhora o gerenciamento da implementação e dos resultados da publicação de dados;
- Fornece a capacidade de armazenar e reconstruir cada fase realizada de publicação de dados;

[1] `https://github.com/opendatatrentino/jackan`

- Possui confiabilidade mais proeminente em uma publicação de dados subsequente;
- Permite aos usuários revisar conclusões e fazer descobertas.

4.2 Metadados de Proveniência e Vocabulários

Para realização da captura dos metadados de proveniência foi definido um conjunto mínimo de metadados, baseado no modelo de dados PROV-DM, para representar o *workflow* de publicação de dados abertos governamentais conectados. Uma análise dos principais requisitos e dos vocabulários de domínio foi realizada para garantir que o modelo estendido possa responder às principais questões de proveniência do *workflow* de publicação. Para representação semântica dos metadados de proveniência foram utilizados os seguintes vocabulários:

- *Dublin Core Metadata Element Set* (dc);
- *Friend of a Friend* (foaf);
- *The PROV Ontology* (prov-o).

A Tabela 4.1 apresenta os metadados definidos para captura proveniência, o vocabulário utilizado para representar esse metadado, o termo existente no vocabulário que representa o metadado, e o tipo de dado desse metadado. Foram utilizados neste trabalho os elementos atividade, agente, entidade e coleção, que representam os possíveis nós do grafo de proveniência, e alguns relacionamentos, conforme descrito no modelo PROV-DM. Além disso, foram definidos outros metadados, relacionados com o processo de publicação, para que a proveniência capturada, garanta os benefícios propostos na Seção 4.1 e promova a captura da proveniência prospectiva, retrospectiva e de dados definidos pelo usuário.

4.3 Modelo de Dados de Proveniência em Bancos de Dados Orientado a Grafo

Os sistemas de banco de dados orientado a grafos se tornaram muito populares e foram implantados, principalmente, em situações em que a relação entre dados é significativa, como por exemplo, nas redes sociais. Embora não exija um *design* de esquema específico, um modelo de dados contribui para sua consistência. A modelagem de dados é uma parte essencial do desenvolvimento de sistemas de informação de qualidade. O modelo de dados de grafo é uma estrutura na qual o esquema e/ou instâncias são modeladas como um grafo

Tabela 4.1: Metadados de Proveniência e Vocabulários Utilizados.

Metadado	**Vocabulário**	**Termo**	**Tipo**
Agente	prov	Agent	Literal
Tipo	dc	Type	Literal
Nome	foaf	Name	Literal
Organização	foaf	Organization	Literal
Descrição	dc	Description	Literal
Agente Publicador	dc	publisher	Literal
Texto	dc	Text	Literal
Atividade	prov	Activity	Literal
Software	dcterms	Software	Literal
Versão	dcterms	hasVersion	Literal
Tempo Inicial	prov	startedAtTime	Data
Tempo Final	prov	endedAtTime	Data
Texto	dc	Text	Array
Associado com	prov	wasAssociatedWith	-
Gerado por	prov	wasGeneratedBy	-
Derivado de	prov	wasDerivedFrom	-
Usado	prov	used	-
Requisição	dcterms	URI	Array
Tamanho	dc	extent	inteiro
Localização	dc	Location	Literal
Coleção	prov	Collection	Literal
Entidade	prov	Entity	Literal

Tabela 4.2: Catálogo de Dados de Agente.

Campos	**Descrição**	**Tipo de Campo**
id	Código identificador de Agente	Numérico Inteiro
Nome	Nome de Agente	Literal
Organização	Nome da organização envolvida	Literal
Descrição	Descrição do Agente	Literal
Agente Publicador	Nome do Agente	Literal
Notas	Informações adicionais	Literal

dirigido, podendo ser rotulados ou não. O modelo de grafos usado na arquitetura foi o de atributos, pois permite que tanto vértices como as arestas armazenem informações de atributos como um par de chave e valor. Para auxiliar na modelagem utilizou-se a notação GRAPHED [64].

Os três tipos básicos de modelo PROV-DM, Atividade, Agente e Entidade são representados como nós no banco de dados de grafo, e são diferenciados pela propriedade Tipo. As arestas representam as relações no banco de dados de grafo, sendo diferenciadas também por uma propriedade Tipo. O mapeamento do modelo PROV-DM é mostrado no diagrama da Figura 4.1. Para a construção de modelos de dados em grafo foi escolhido o modelo baseado em atributos, porque se beneficia de certos tipos de consulta que realizam buscas sobre atributos. O diagrama foi projetado usando GRAPHED expondo os nós, seus relacionamentos e cardinalidades. Neste diagrama, é possível identificar diferenças semânticas práticas entre as bordas `Used`, `WasGeneratedBy`, `WasDerivedFrom` e `WasAssociatedWith` e seus relacionamentos com os nós `Activity`, `Collection` e `Agent`.

Nas tabelas 4.2, 4.3 e 4.4 são apresentados, respectivamente, o Catálogo do conjunto de dados de Agente, Atividade e Coleção. Essas tabelas apresentam os Atributos, uma breve descrição sobre cada atributo e o Tipo de Dado desse atributo.

4.4 Arquitetura do UnBGOLDprov

A arquitetura do UnBGOLDProv realiza a coleta de proveniência de dados do *workflow* de publicação de dados abertos governamentais de forma automática, e utiliza o modelo de dados baseado em grafo definido na Seção 4.3.

O projeto da arquitetura foi baseado no princípio *"The Separation of Concerns (SoC)"* [65], sendo que na Figura 4.2 são apresentados os principais componentes da arquitetura:

Tabela 4.3: Catálogo de Dados de Atividade.

Campos	**Descrição**	**Tipo de Campo**
id	Código identificador de Atividade	Numérico Inteiro
Nome	Nome da Atividade	Literal
Programa	Programa utilizado na atividade	Literal
Versão	Versão do programa utilizado na atividade	Literal
Descrição	Descrição da atividade	Literal
Tempo Inicial	Hora de início da atividade	Data
Tempo Final	Hora de finalização da atividade	Data
Software	Software utilizado na atividade	Literal
Notas	Informações adicionais	Array
Used	Indica que a atividade usa uma entidade	-
WasAssociatedWith	Indica que a atividade está associado com um Agente	-

Tabela 4.4: Catálogo de Dados de Coleção.

Campos	**Descrição**	**Tipo de Campo**
id	Código identificador de coleção	Numérico Inteiro
Nome	Nome da coleção	Literal
Tamanho	Tamanho do arquivo da coleção	Literal
Descrição	Descrição do conjunto de dados da coleção	Literal
Localização	Localização do conjunto de dados da coleção	Literal
Notas	Informações adicionais	Literal
WasGeneratedBy	Indica que a coleção foi gerada por uma atividade	-
wasDerivedFrom	Indica que a coleção foi derivada de outra coleção	-

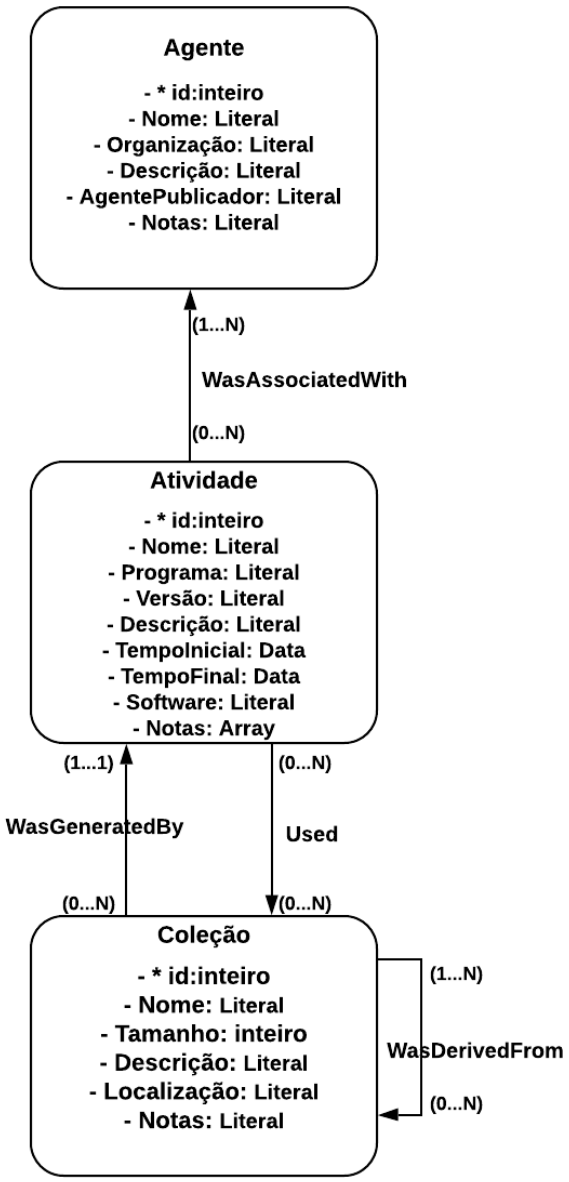

Figura 4.1: Mapeamento das entidades do PROV-DM para o modelo de grafos usando GRAPHED.

- Camada de Controle - contém o fluxo da aplicação e passa os dados de entrada para o serviço;
- Camada de Serviço - é o *middleware* entre a camada de controle e a camada de repositório. Ela reúne os dados do controlador, executa a validação e lógica de negócios, e chama a Camada de Repositório para manipulação de dados;
- Camada de Repositório - camada para interação com a camada de modelo e execução de operações de banco de dados;

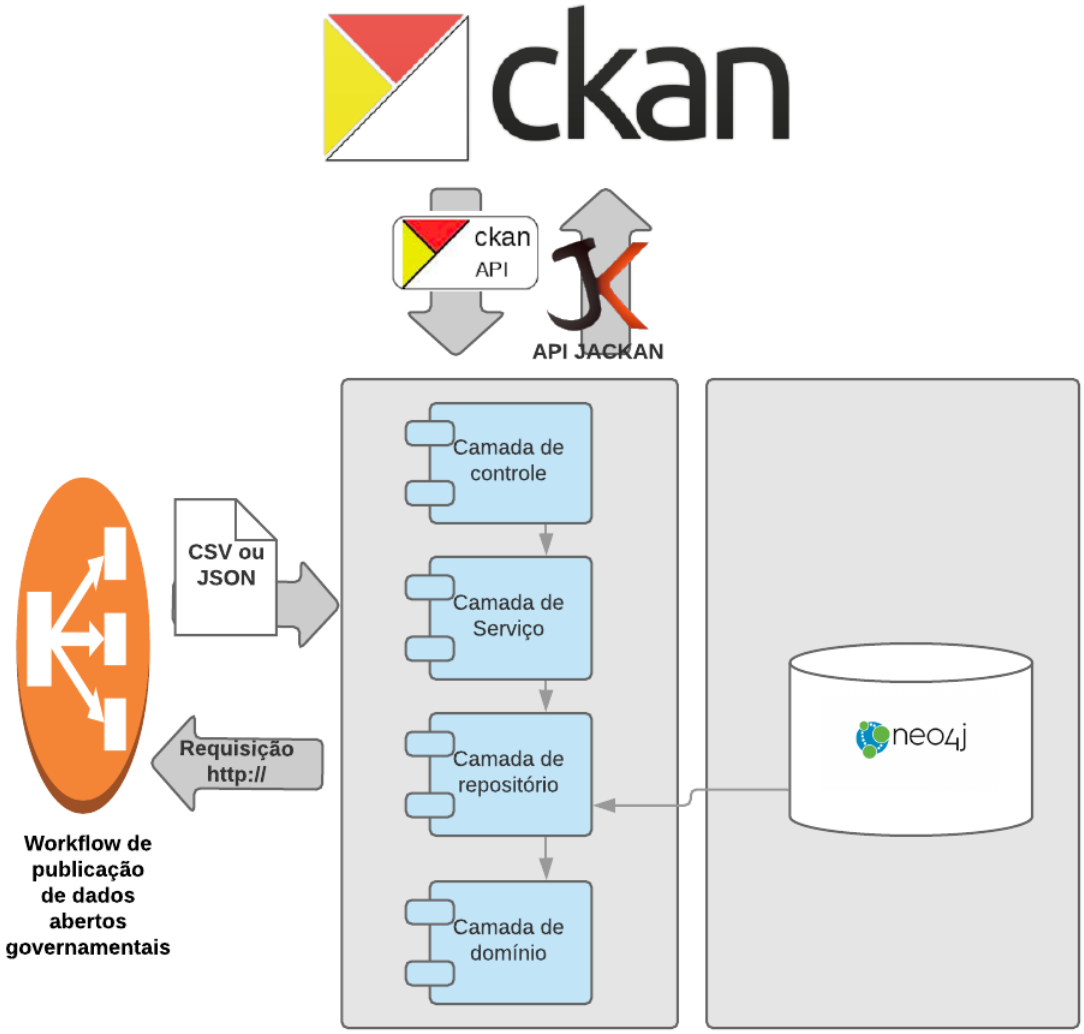

Figura 4.2: Arquitetura do UnBGOLDProv.

- Camada de Domínio - contém as classes de entidade que possui atributos e métodos definidos;
- Banco de Dados de Grafo - é responsável pelo armazenamento dos dados na forma de grafo.

A captura dos metadados de proveniência pelo UnBGOLDProv se dá após a execução do processo de publicação dos dados abertos. Após realizar uma solicitação HTTP e recebimento dos metadados no formato JSON, a Camada de Controle realiza algumas validações de obrigatoriedade e passa para a Camada de Serviço uma ordem de criação do grafo de proveniência. Nesse momento, a Camada de Serviço solicita à Camada de Repositório a criação do grafo de proveniência com os parâmetros informados no banco de dados de grafo. Por fim, a publicação dos conjuntos de dados é realizada por meio

da ligação entre o UnBGOLDProv e a plataforma CKAN, na qual a biblioteca em Java Jackan se conecta à API do CKAN e gerencia os conjuntos de dados abertos.

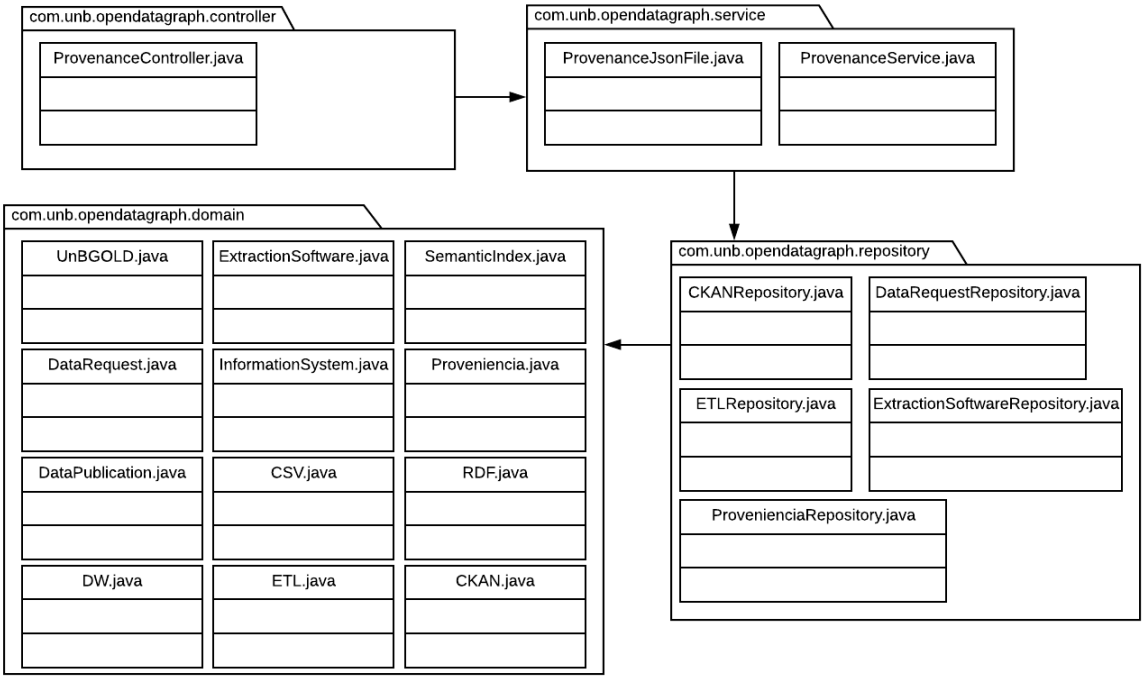

Figura 4.3: Diagrama de pacotes da arquitetura do UnBGOLDProv.

4.4.1 Camada de Controle

Com o objetivo de facilitar a manutenção do sistema e futuras atualizações foi usado o padrão MVC (*Model-View-Controller*), já bem difundido no mercado e na literatura. Esse padrão facilita a separação dos dados ou lógica de negócio (*Model*) da interface com a *Web* (*View*) e do fluxo da aplicação (*Control*). A sua idéia é permitir que a mesma lógica de negócios possa ser acessada e visualizada por meio de diversas interfaces. Por isso, a lógica de negócio (*Model*) não sabe de quantas nem quais interfaces com o usuário estão exibindo seu estado.

Na Camada de Controle da aplicação, foi utilizada a biblioteca HTTP de Java para realizar a comunicação com o barramento de serviços e a biblioteca Jackan para comunicar com a plataforma Ckan. É nesta camada que está presente o fluxo da aplicação. A classe de controle com o nome de *ProvenanceController* pode ser visualizada na Figura 4.3.

4.4.2 Camada de Serviço

A Camada de Serviço define a fronteira de uma aplicação e o seu conjunto de operações disponíveis. Ele encapsula a lógica de negócio, controlando as transações e coordenando as respostas na implementação das operações. O UnBGOLDProv possui os seguintes serviços (demonstrados na Figura 4.3):

- ProvenanceJsonFile: gerencia toda regra de negócio responsável por manipular o arquivo de proveniência no formato JSON, e armazená-lo no formato de uma lista;
- ProvenienciaService: gerencia toda regra de negócio responsável por salvar a lista de proveniência e gerar o grafo.

4.4.3 Camada de Repositório

Esta camada é implementada utilizando-se o padrão *repository*, que representa todos os objetos de um certo tipo como um conjunto conceitual [66]. Este padrão de desenvolvimento atua como uma coleção, exceto que com uma capacidade de consulta mais elaborada. Objetos são adicionados ou removidos, e o mecanismo por trás do padrão *repository* irá inseri-los ou removê-los do banco de dados. Esta definição reúne um conjunto coeso de responsabilidades para garantir o acesso aos dados. As classes que compõem a Camada de Repositório podem ser observadas no pacote *com.unb.opendatagraph.service* da Figura 4.3.

4.4.4 Camada de Domínio

A Camada de Domínio contém as classes de entidade que possui atributos e métodos definidos. Essas classes são baseadas no modelo de dados definido na Figura 4.1. Nelas são representados os agentes, as coleções/entidades, as atividades, e seus respectivos atributos e relacionamentos. A Figura 4.3 apresenta as classes da Camada de Domínio no pacote *com.unb.opendatagraph.domain*.

4.5 Captura e Publicação com Proveniência de Dados

A arquitetura propõe que o gerenciamento da solicitação e entrega dos metadados de proveniência, capturados do *workflow* de publicação de dados abertos governamentais, seja feito pelo barramento de serviços da UnB, ErlangMS [67]. O ErlangMS recebe uma

solicitação por meio de uma requisição via protocolo HTTP e retorna os dados em formato estruturado, ou seja, em formato aberto (CSV ou JSON).

O UnBGOLDProv apresenta uma estratégia de recuperação de proveniência de dados em diferentes estágios da publicação dos dados abertos governamentais conectados: nas etapas de preparação e transformação dos dados, e no processo de indexação semântica, quando as decisões sobre correspondência de esquema e mesclagem de dados devem ser registrado para uso futuro. O mecanismo para coletar a proveniência dos dados é executado juntamente com essas atividades, registrando-os em diferentes níveis de granularidade. Em seguida, a proveniência coletada é publicada, para que os dados abertos conectados e sua respectiva proveniência possam ser explorados juntos, usando a ferramenta UnBGOLD e a plataforma CKAN. A estratégia descrita acima, está representada na Figura 4.4.

A proveniência capturada, gravada e, posteriormente, publicada de forma associada aos seus dados indexados semanticamente, denominada de proveniência aberta conectada, ajuda os usuários ou sistemas de computação a determinar o histórico de derivação desses dados, podendo contribuir significativamente para o uso eficaz desses dados, desempenhando um papel fundamental no enriquecimento do contexto em torno do dados abertos governamentais conectados, apoiando a avaliação de atributos como fidedignidade dos dados.

A arquitetura do UnBGOLDProv, após receber os dados de proveniência no formato JSON, a ontologia PROV-O e os vocabulários *Dublin Core* (DC) e FOAF são utilizados para a representação dos dados de proveniência publicados. Os dados de proveniência capturados são descritos semanticamente por meio de triplas RDF, e publicados na plataforma CKAN, juntamente com os dados de proveniência conectados.

Esses metadados de proveniência capturados são baseados no modelo de dados de proveniência definido na Seção 4.3. Esse modelo de dados estende o modelo de dados W3C PROV-DM para descrever informações detalhadas sobre os dados e processos envolvidos no ciclo de vida dos objetos de dados.

Com base nessas informações é possível entender o ciclo de vida de publicação dos conjuntos de dados através da definição de um modelo de dados PROV-DM. A Figura 4.5 mostra o grafo direcionado que representa a proveniência de dados de um conjunto de dados gerado pela execução do *workflow* de publicação de dados abertos governamentais. Conforme definido no modelo PROV-DM, as atividades são representadas por retângulos, agentes por pentágonos, e entidades ou coleções por elipses.

Neste grafo o agente Software de Extração inicia a atividade ETL, que realiza a extração do conjunto de dados da base de dados dos Sistemas de Informação. Estes dados, posteriormente, são limpos e tratados para que sejam importados para um DW, onde

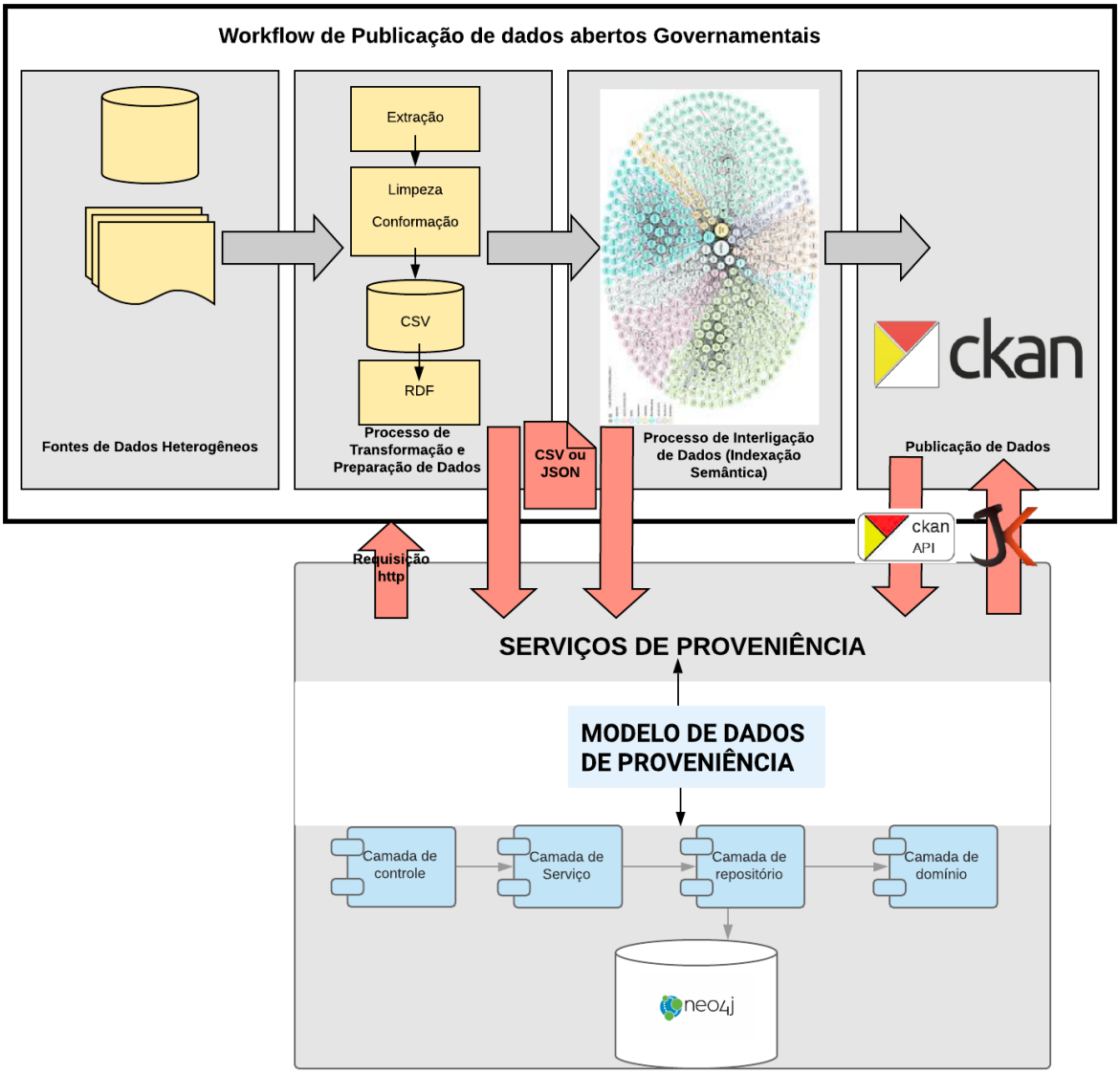

Figura 4.4: *Workflow* de publicação de dados abertos governamentais com captura da proveniência.

ficam disponíveis para consulta. Em seguida, o agente UnBGOLD realiza a atividade Requisição de Dados que utiliza dados extraídos do DW, gerando um arquivo em formato CSV aberto e já estruturado. Para realizar a indexação semântica, o agente UnBGOLD seleciona um vocabulário composto de metadados e ontologias que possam descrever semanticamente os dados. Neste processo os dados indexados são descritos em formato de triplas RDF, e armazenados no banco de dados. Por fim, configuram-se os parâmetros que proporcionam que os dados sejam publicados automaticamente no portal de dados abertos, que é uma instância da plataforma CKAN.

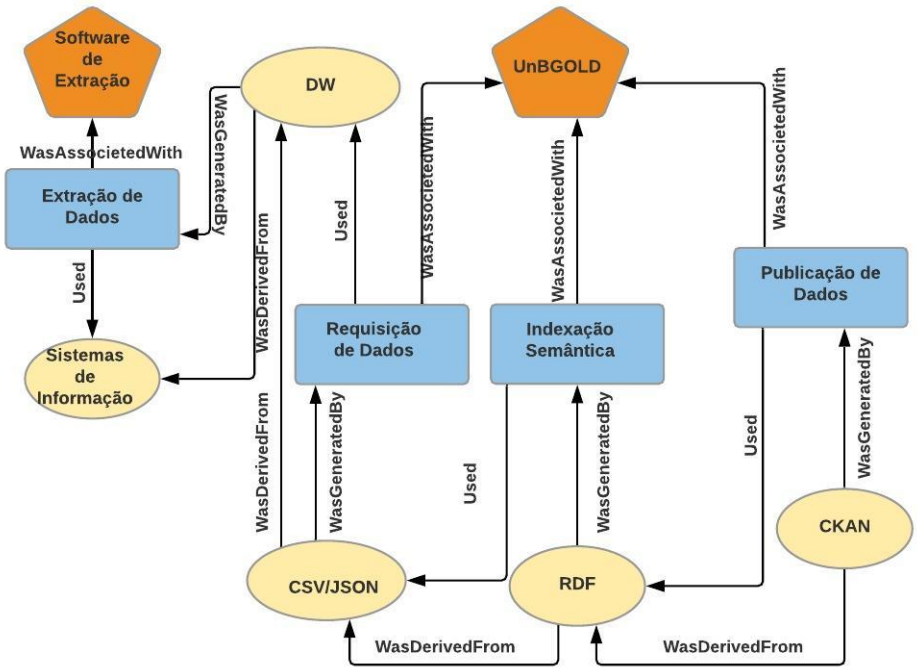

Figura 4.5: Grafo de proveniência da arquitetura de publicação de dados abertos.

4.6 Armazenamento no Banco de Dados Neo4J

O banco de dados Neo4j, de código aberto, possui uma estratégia nativa para armazenamento físico de dados, que suporta mais de uma forma de acesso, sendo o mais popular dentre os bancos de dados orientado a grafos. O Neo4j oferece suporte ao *Tinkerpop*. Dessa forma, tanto a definição dos esquemas das bases quanto a manipulação dos dados (carregamento e consulta) podem ser feitas por meio da linguagem *Gremlin*. Além disso, o Neo4j possui a linguagem de consulta proprietária *Cypher* e também da suporte a *SPARQL* (SQL for *linked data*). O armazenamento físico pode ser tanto em memória quanto em disco e o modelo físico e baseado em repositórios chave-valor. O Neo4j é transacional e oferece os dois tipos de arquitetura, a centralizada e a distribuída com suporte a replicação [68] [69].

O Sistema de Gerência de Bancos de Dados de Grafos Neo4j [68] utiliza o modelo de grafos de propriedade. Mais especificamente, o grafo de propriedades é um multigrafo atribuído, direcionado e rotulado. É multigrafo porque permite múltiplas arestas entre dois nós. Apresenta como características os nós possuírem propriedades (pares de valores-chave), podendo possuir um ou mais rótulos. Além disso, cada aresta tem um rótulo que é usado para especificar o seu tipo e sempre possuem um nó inicial e final, podendo também conter propriedades. E ele é multigrafo porque permite múltiplas arestas entre dois nó [70].

O armazenamento da proveniência de dados em um banco de dados orientado a grafos permite uma modelagem mais direta do armazenamento de estruturas de dados, uma vez que o PROV-DM é baseado em grafos [71]. Essas informações são evidenciadas examinando o modelo de dados PROV-DM do estudo de caso (Figura 4.5), e o grafo de proveniência dos dados armazenado no banco de dados Neo4J da (Figura 4.6).

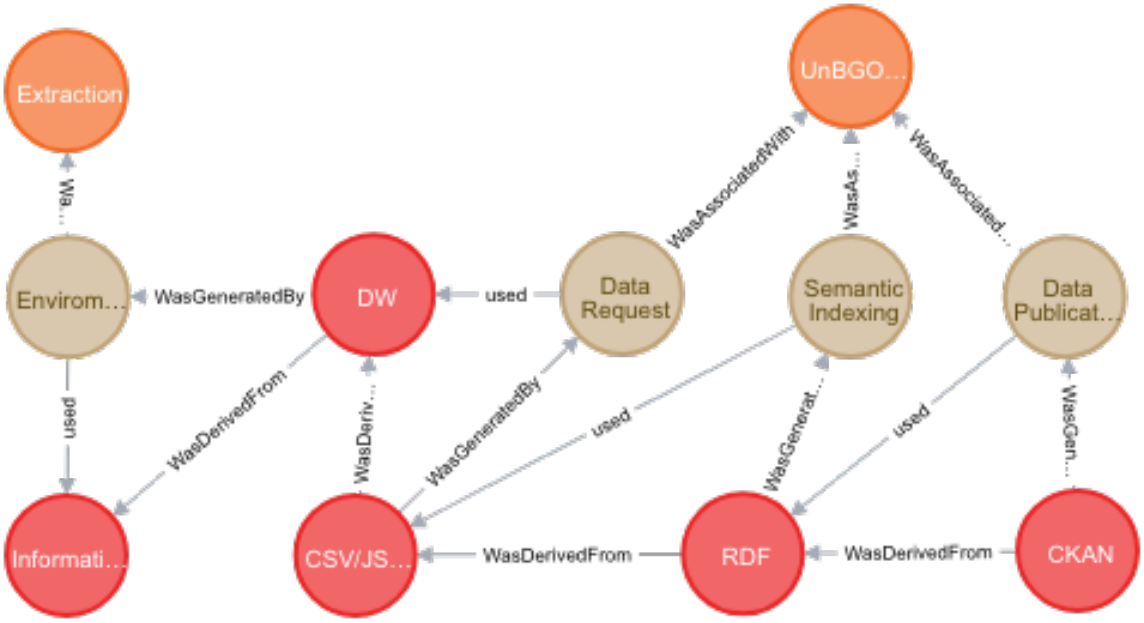

Figura 4.6: Grafo de proveniência de dados da arquitetura de publicação de dados abertos governamentais retirado da interface web do Neo4J.

Capítulo 5

Estudo de Caso

Este capítulo apresenta uma avaliação da abordagem em um cenário real de um conjunto de dados extraído de um dos sistemas de informação da UnB, chamado SIGRA. A Seção 5.1 descreve o contexto e o cenário de realização do estudo de caso, ou seja, a captura, o gerenciamento e a publicação da proveniência de dados de departamento. Por fim, a Seção 5.2 apresenta os resultados acadêmicos desta pesquisa.

5.1 Captura e Publicação da Proveniência de Departamentos

Cada vez mais, os governos estão mantendo catálogos de dados listando os conjuntos de dados que compartilham com o público. Esses catálogos desempenham um papel vital no aprimoramento da visibilidade e da localização dos conjuntos de dados do governo. O UnBGOLDProv agrega valor ao catálogo de dados abertos da UnB, pois permite a descoberta da proveniência de maneira fácil e eficiente, possibilitando ao usuário revisar conclusões e fazer descobertas.

O estudo de caso foi implementado com o objetivo de demonstrar a validade da arquitetura proposta no Capítulo 4, que captura a proveniência do *workflow* de publicação de dados abertos governamentais, gerencia e publica a proveniência aberta conectada dos dados abertos governamentais conectados publicados na plataforma CKAN.

Neste estudo de caso, um conjunto de dados sobre os cursos universitários foi extraído de um dos sistemas de informação da UnB, chamado SIGRA. Os dados extraídos foram:

- Departamentos: dados referentes aos departamentos internos da UnB;
- Cursos: dados referentes às cursos de graduação da UnB, estes cursos são ofertados pelos departamentos da UnB;

- Disciplinas: dados referentes as disciplinas que são ofertadas pelos departamentos;
- Professores: dados referentes ao corpo docente da UnB. Este conjunto de dados apresenta apenas informações que já são de domínio público, não apresentando dados pessoais que configurem quebra de privacidade;
- Ofertas: os dados das ofertas são referentes às disciplinas que são oferecidas pelos departamentos em um determinado período letivo, obrigatoriamente elas possuem um professor vinculado.
- Fluxos de Curso: é denominado o fluxo de um curso as disciplinas que são obrigatórias para que um aluno passa concluir a curso.

O *workflow* de publicação disponibiliza esses dados, extraídos do SIGRA, na plataforma CKAN em formato aberto e indexado semânticamente. Para descrever o caminho percorrido pelo dado e permitir a reprodutibilidade da publicação dos dados, o UnBGOLDProv captura a proveniência desses dados, armazena no Neo4J e gera o grafo de proveniência. Uma instância do grafo de proveniência que representa o processo de publicação do conjunto de dados de departamento é demonstrada na Figura 5.1. Posteriormente, esses metadados de proveniência são disponibilizados na instância de dados abertos da UnB na plataforma CKAN.

Conforme definido no modelo PROV-DM, as atividades são representadas por retângulos, agentes por pentágonos e coleções por círculos. A instância do grafo de proveniência da Figura 5.1 mostra que a coleta dos dados de proveniência inicia-se no momento em que o Agente, nomeado de Extração_SIGRA_01, se relaciona com a atividade de ETL por meio da aresta *WasAssociatedWith* entre esse agente e essa atividade do grafo.

Este processo ETL extraiu os dados dos Sistemas de Informação, que no caso desse estudo de caso foi extraído do SIGRA, realizou as transformações necessárias e carregou os dados dentro de um *Data Warehouse*, que foi chamado de DW_SIGRA_01. A atividade ETL de extração dos dados da coleção sistemas de informação está representada pelo relacionamento da aresta *Used*, enquanto que a aresta *WasGeneratedBy* indica que a atividade ETL gerou a coleção DW. A aresta *WasDerivedFrom* indica que o conjunto de dados disponíveis na coleção DW, deriva do conjunto de dados da coleção Sistemas de Informação.

Em seguida, uma atividade de Requisição de Dados de departamento é realizada pelo agente UnBGOLD, representada pela aresta *WasAssociatedWith* entre UnBGOLD e a Requisição_Departamento. Essa requisição é feita sobre o DW_SIGRA_01, conforme descreve o relacionamento representado pela aresta *Used*. Essa atividade de Requisição de Dados gera o conjunto de dados da coleção departamento.csv descrito pelo relacionamento *WasGeneratedBy*.

Posteriormente, o agente UnBGOLD realiza a atividade de indexação semântica de dados de departamento, conforme pode ser visto por meio da aresta *WasAssociatedWith* que liga os dois nós. A atividade Indexação_Departamento é feita sobre a coleção departamento.csv, descrita pela aresta *Used*, com o objetivo de gerar a coleção departamento.rdf, conforme descrito pela aresta *WasGeneratedBy*. Por meio da aresta *WasDerivedFrom* a modelo de dados deixa claro que a coleção departamento.rdf foi gerada com base na coleção departamento.csv, assim como a coleção departamento.csv deriva da coleção DW_SIGRA_01.

Por fim, o Agente UnBGOLD realiza a atividade de publicação de Dados de departamento, como mostra a aresta *WasAssociatedWith*, sendo que esta atividade Publicação_Departamento gera a coleção CKAN_Departamento, como mostra a aresta *WasGeneratedBy*, utiliza a coleção departamento.rdf por meio da aresta *Used*, sendo que a coleção CKAN_Departamento, deriva da coleção departamento.rdf, conforme demonstra a aresta *WasDerivedFrom*.

A captura e o armazenamento dos metadados de proveniência no banco de dados Neo4J, conforme mostrado na Figura 5.2, que exibe o *Neo4J Browser*[1], permite que sejam obtidas informação importantes das etapas do processo, promovendo uma maior reprodutibilidade da publicação. Dessa forma, pode-se verificar na instância do grafo de proveniência da Figura 5.1, na parte destacada em vermelho, tem-se por exemplo, que o operador que realizou a Extração de Software foi o Luiz Martins, que essa atividade de ETL teve início às 20:21:27 horas do dia 10 de março de 2018, extraindo dados do banco de dados SIGRA e gerando o DW DW_SIGRA_01. Que foi realizada uma requisição de dados de departamento por meio do UnBGOLD, e foi gerado um arquivo departamento.rdf que passou pela atividade de Indexação Semântica, e utilizou o arquivo departamento.csv e os vocabulários FOAF, DC, dentre outros. Esse arquivo departamento.rdf foi publicado na instância do CKAN de dados abertos da UnB.

Em seguida, as informações de proveniência são publicadas na instância de dados abertos da CKAN em um formato aberto, de modo que outras pessoas possam reutilizá-las e que esteja legível por máquina, ou seja, com uma semântica bem definida para permitir não apenas aos usuários humanos, mas também aos programas, acessarem, processarem e utilizarem as informações. Assim sendo, os metadados de proveniência foram disponibilizados, no formato CSV e no formato JSON, como mostra a Figura 5.3, que exibe apenas o conjunto de dados de departamento, em formato aberto (CSV, JSON e RDF), e os metadados de proveniência de departamento em formato aberto (CSV e JSON), visto que no estudo de caso, demonstrado na Figura 5.1, é possível verificar que o UnBGOLD fez a requisição apenas dos conjuntos de dados de departamento para o *Data Warehouse*.

[1] `https://neo4j.com/developer/neo4j-browser/`

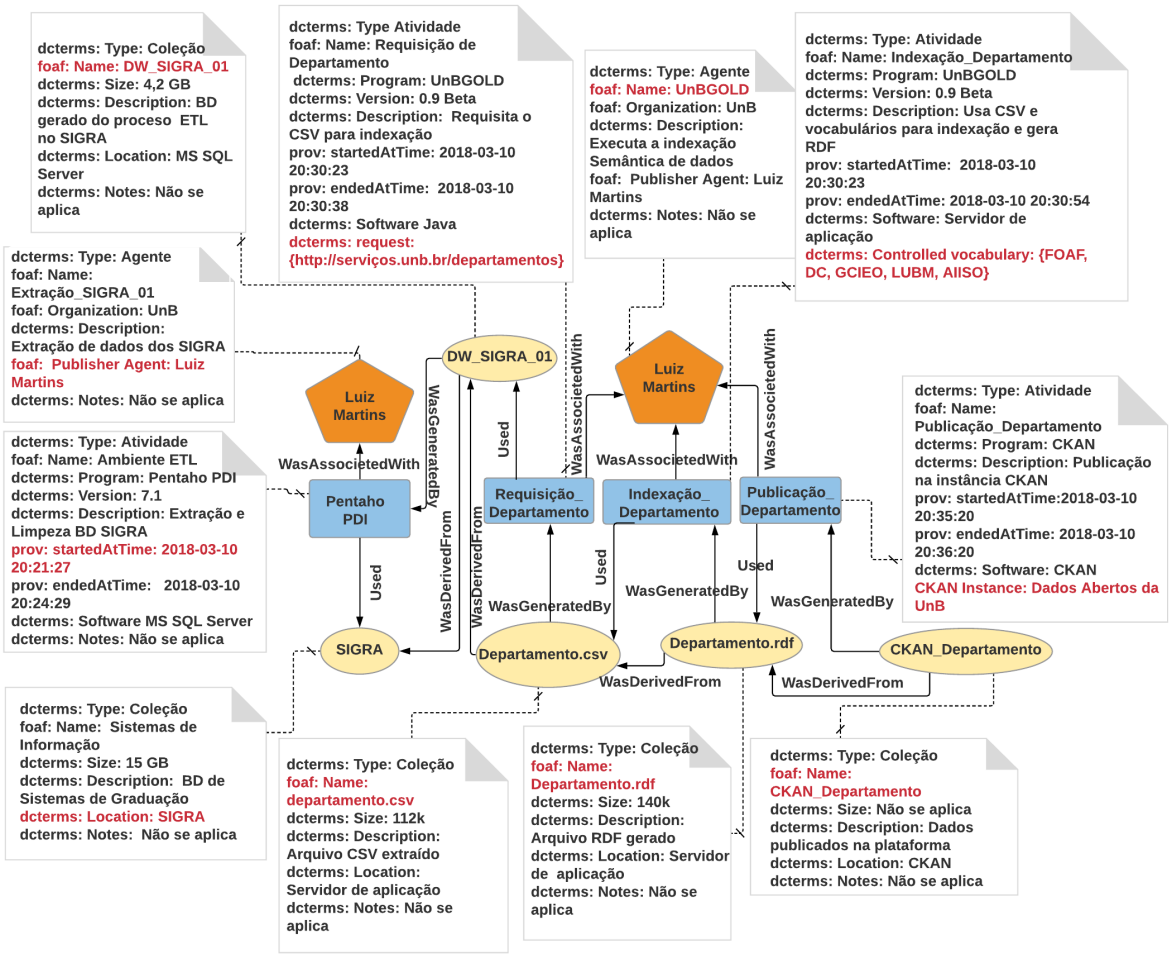

Figura 5.1: Instância do grafo de proveniência do *workflow* de publicação de dados abertos governamentais.

A instância de dados abertos da UnB, como mostra a Figura 5.4, armazena também o modelo de dados de proveniência para que seja possível o usuário entender melhor o fluxo de dados do *workflow* de publicação de dados abertos governamentais.

Os conjuntos de dados proveniência, publicados na plataforma CKAN nos formatos JSON e CSV, são exibidos nos Apêndices A e B, respectivamente. Os conjuntos de dados podem ser publicados de forma integrada, sendo possível que os dados sejam conectados uns aos outros criando assim uma web de dados abertos. A conexão dos dados se dá identificando a relação entre recursos diferentes por meio de um dado em comum.

Com a extensão CKAN, cada conjunto de dados de proveniência publicado é vinculado ao seu arquivo de origem. Ao vincular os dados de proveniência à sua origem e ao histórico de operações do *workflow* de publicação de dados abertos governamentais conectados, um determinado usuário pode examinar e reproduzir todas essas operações a partir dos dados originais, e terminar com uma cópia exata dos dados convertidos publicados.

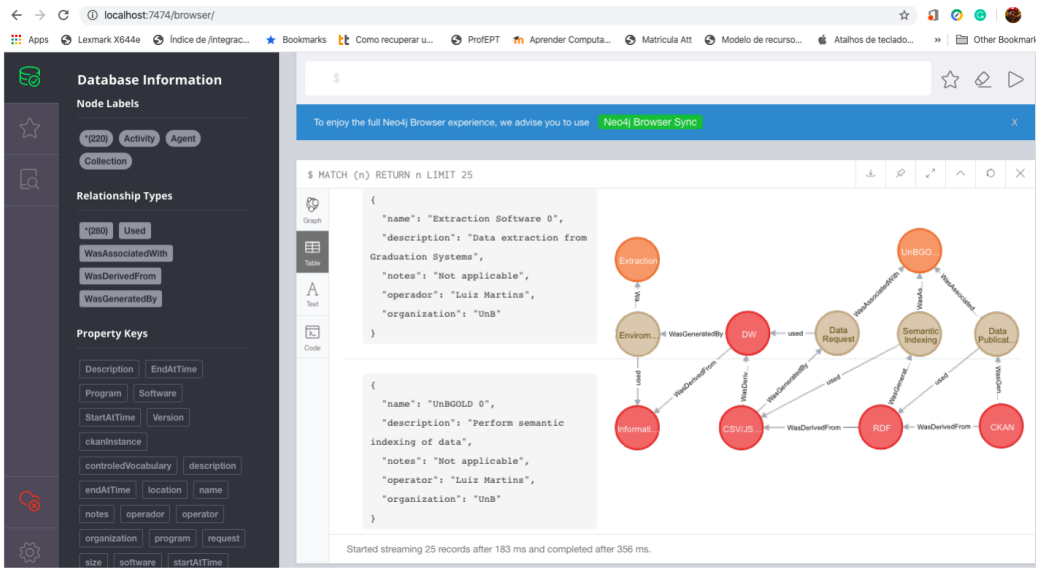

Figura 5.2: Neo4J Browser - Utilizado para consultas, visualizações e interações de dados.

5.2 Resultados Acadêmicos

Durante a pesquisa foram publicados dois trabalhos relacionados com o tema de gerenciamento de proveniência em *workflow* de publicação de dados abertos governamentais conectados. O artigo Modelo de Dados de Proveniência para uma Arquitetura de Dados Abertos Governamentais [72] foi selecionado e apresentado no VII *Workshop* de Transparência em Sistemas (VII WTRANS) realizado em Belém - Pará, no dia 15 de julho de 2019, no XXXIX Congresso da Sociedade Brasileira de Computação (CSBC 2019). Já o artigo *Enhancing Open Government Data With Data Provenance* [73] foi selecionado e apresentado no *The 11th International ACM Conference on Management of Digital EcoSystems* (MEDES'19) realizado na cidade de Limassol na República de Chipre, no dia 13 de novembro de 2019.

Em [72] foi definido um modelo de dados de proveniência, utilizando o PROV-DM, para uma arquitetura de dados abertos conectados. Para validar o estudo, foram utilizados um conjunto de dados, extraídos de uma das bases de dados dos sistemas de informação da UnB.

Uma solução tecnológica no contexto de Dados Abertos Governamentais Conectados para aprimorar a publicação de dados públicos abertos governamentais foi apresentada em [73]. A arquitetura proposta fornece a proveniência de dados abertos governamentais públicos, usando o PROV-DM e um banco de dados de grafos. Além disso, também apresenta um simulador que permite recuperar a proveniência dos dados no *workflow* de publicação de dados abertos governamentais conectados.

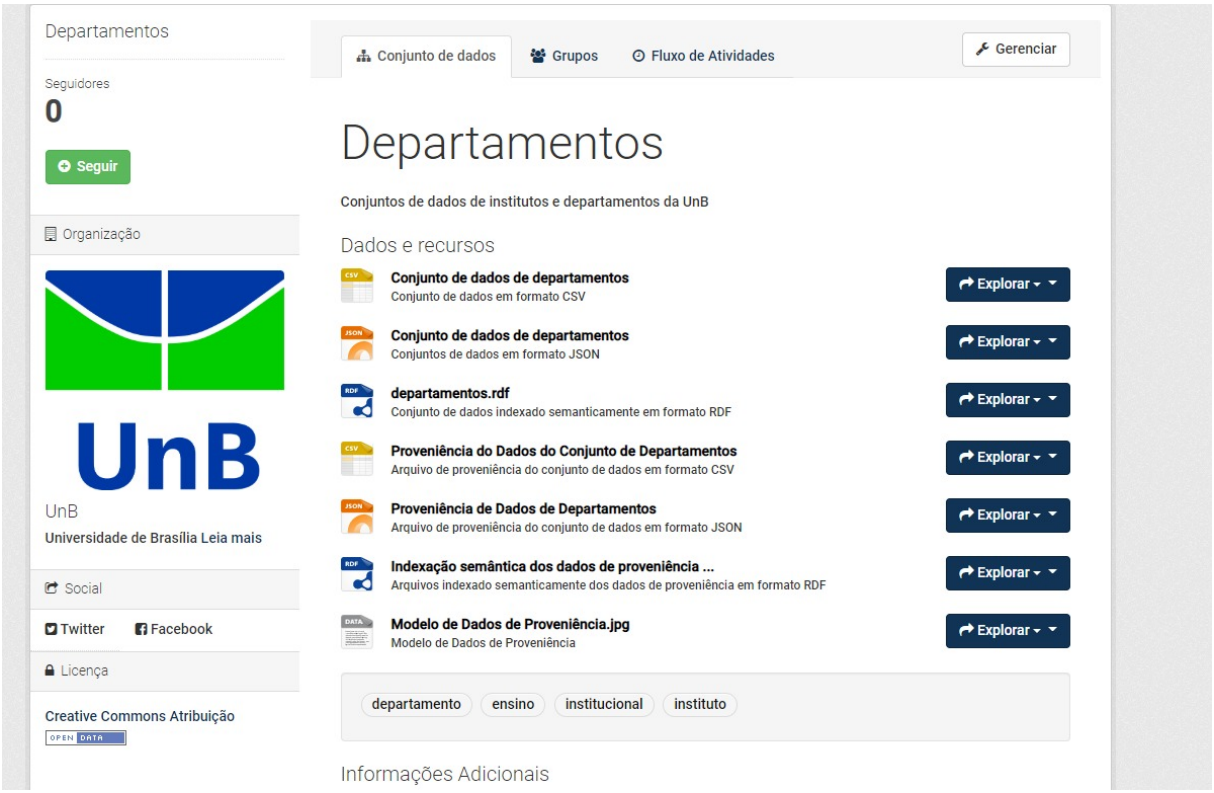

Figura 5.3: Instância CKAN do portal de dados abertos da UnB com os dados de proveniência de departamento.

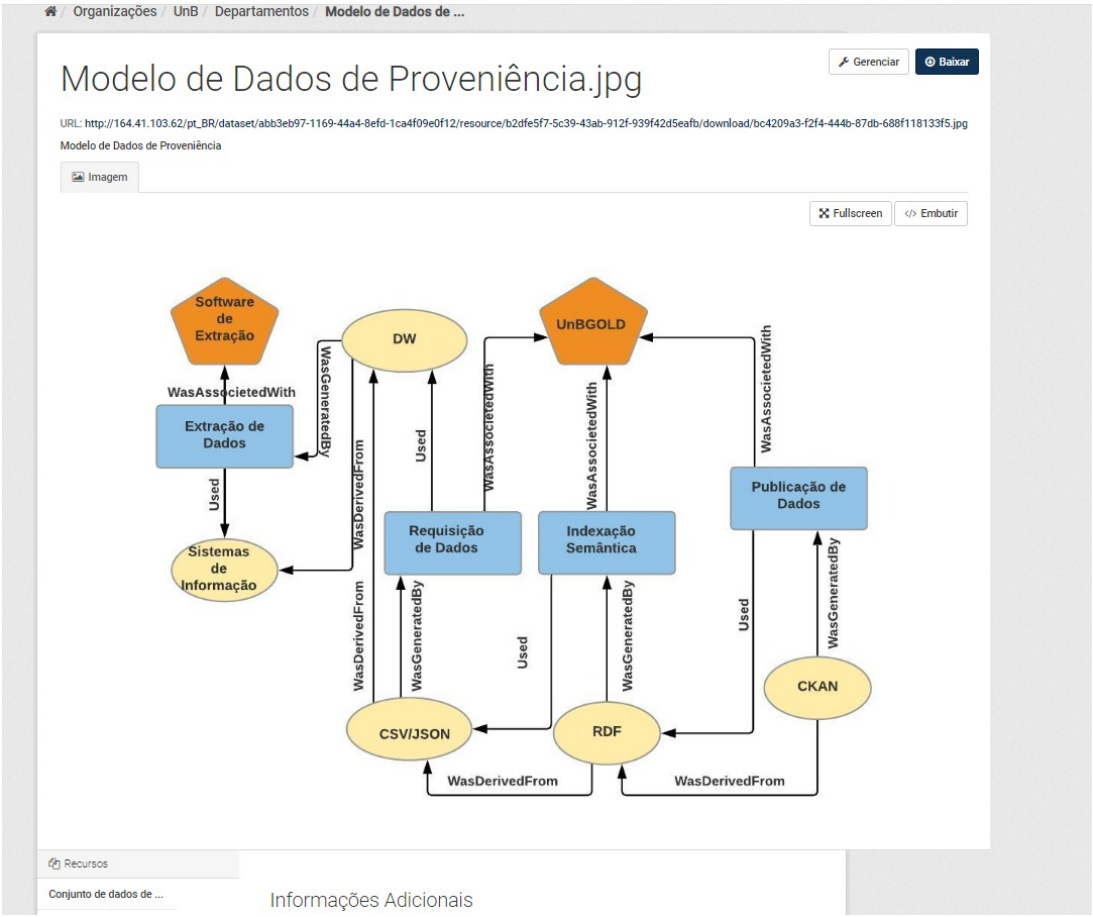

Figura 5.4: Modelo de dados PROV-DM disponível na instância CKAN.

Capítulo 6

Conclusão

Este trabalho apresenta uma arquitetura de proveniência para um *Workflow* de publicação de dados abertos governamentais conectados e um modelo de dados de grafos baseado no PROV-DM. Os resultados apontam para um mecanismo adequado que pode ajudar a melhorar a reprodutibilidade dos dados da Universidade de Brasília. A arquitetura da informação foi desenvolvida de forma modular, permitindo a criação de camadas independentes, marcadas pela capacidade de anexar outra interface sem alterar as outras camadas. Também facilita a manutenção e a adição de recursos, bem como a reutilização do código desenvolvido.

Para suportar a captura, a gravação, a consulta e o gerenciamento dos metadados de proveniência, foi definido um conjunto mínimo de metadados, que serviu de suporte para implementação de um modelo de dados, baseado no W3C PROV-DM. A proveniência foi capturada, de forma automática, por meio de uma requisição HTTP feita ao *workflow* de publicação, na qual foram capturadas a proveniência prospectiva, retrospectiva e de dados definidos pelo usuário. Após persistida, a proveniência é publicada na plataforma CKAN, em conjunto com dados abertos publicados na instância de dados abertos da UnB.

Para validar nossa abordagem, um estudo de caso possibilitou a coleta da proveniência da publicação de dados extraídos dos sistemas de informação da UnB. Os resultados mostram que nossa abordagem é eficaz na coleta e armazenamento e publicação dos metadados de proveniência.

Assim sendo, este trabalho contribui para melhorar o gerenciamento da implementação e dos resultados da publicação de dados, fornece a capacidade de armazenar e reconstruir cada fase realizada de publicação de dados, aumenta a confiabilidade do processo de publicação, além de permitir aos usuários revisar conclusões e fazer descobertas.

Como trabalho futuro, indica-se um estudo para verificar se a adequação de diversos bancos de dados de grafos e outros modelos de banco de dados para armazenar a proveniência dos dados, comparando seu desempenho. Além disso, pode ser desenvolvido um

plano de estudo para analisar a aceitação e a usabilidade da ferramenta UnBGOLD, com o mecanismo de coleta de proveniência para desenvolver recursos e funcionalidades que melhoram a rotina diária dos usuários. Pode-se ainda adicionar o recurso de publicação da instância do grafo de proveniência a partir do banco de dados Neo4J, no lugar do grafo genérico que representa a proveniência do *workflow* de publicação. Indica-se também a publicação dos metadados de proveniência no formato RDF na plataforma CKAN. E por fim, pode ser criada uma interface, para que o conjunto de metadados de proveniência a ser capturado possa ser alterado e adaptado de acordo com a necessidade.

Referências

[1] Isotani, Seiji e IgIbert Bittencourt: *Dados Abertos Conectados: Em busca da Web do Conhecimento*. Novatec Editora, 2015. x, 4, 6, 11, 12

[2] W3C: *Prov-dm: The prov data model.*, 2013. `https://www.w3.org/TR/prov-dm/`, acesso em 2018-10-11. x, 19, 20, 21, 32

[3] Robinson, I., J. Webber e E Eifrem: *Graph Databases*, volume 1. Ed. California, USA, O'Reilly, 2013. x, 23

[4] Martins, Luiz C. B., Márcio C. Victorino, Maristela Holanda, George Ghinea e Tor Morten Gronli: *Unbgold: Unb government open linked data: Semantic enrichment of open data tool*. Em *Proceedings of the 10th International Conference on Management of Digital EcoSystems*, MEDES, páginas 1–6, New York, NY, USA, 2018. ACM, ISBN 978-1-4503-5622-0. `http://doi.acm.org/10.1145/3281375.3281394`. x, 2, 27, 36

[5] Obama, Barack: *Memorandum on transparency and open government*, 2009. `http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf`. 1

[6] BRASIL: *Lei de acesso à informação*, 2011. `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`, acesso em 2018-10-11. 1

[7] Berners-Lee, Tim: *Linked data-design issues*, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`. 1

[8] Bizer, Christian, Heath Tom e Tim. Berners-Lee: *Linked data - the story so far.* International Journal on Semantic Web and Information Systems, 5:1–22, 2009. 1

[9] Hartig, O. e J. Zhao: *Publishing and consuming provenance metadata on the web of linked data.* International Provenance and Annotation Workshop, 6378:78–90, 2010. 2

[10] Golbeck, Jennifer: *Weaving a web of trust.* Science, 321(5896):1640–1641, 2008. 2

[11] Berneers-Lee, Tim, James Hendler e Ora Lassila: *The semantic web.* Scientific American, 284(5):34–43, 2001. 4, 26

[12] Nowack, B: *The semantic web technology stack (not a piece of cake...).* Linked Data Developer, página 15, 2009. `http://linkeddatadeveloper.com/Projects/Linked-Data/media/fig11.2.png`. 5

[13] Golbeck, Jennifer e James Hendler: *A semantic web approach to the provenance challenge*. Concurrency and Computation: Practice and Experience, 20(5):431–439, 2008. 5

[14] Core., Metadata Iniciative Dublin: *Metadata basics.*, 1995. `http://dublincore.org/metadata-basics/`. 5

[15] Gilliland, Anne J: *Setting the stage*. Introduction to metadata, 2:1–19, 2008. 7

[16] McGuinness, Deborah L, Frank Van Harmelen *et al.*: *Owl web ontology language overview*. W3C recommendation, 10(10):2004, 2004. 8

[17] Baca, Murtha: *Introduction to metadata*. Getty Publications, 2016. 8

[18] GRUBE R, T.: *A translation approach to portable ontology specifications.* Knowledge Acquisition, 5(2):199–220, 1993. 9

[19] Fensel, D.: *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce.* Springer-Verlag, 2003, ISBN 3540003029. 9, 10

[20] BRASIL, W3C: *Melhorando o acesso ao governo com o melhor uso da web.*, 2009. `http://www.w3c.br/GT/GrupoDadosAbertos`, acesso em 2018-10-11. 10, 14, 17

[21] Guarino, Nicola: *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, volume 46. IOS press, 1998. 10

[22] BERNERS-LEE, T.: *Design issues*, 2006. `http://www.w3.org/DesignIssues`, acesso em 2018-10-11. 11

[23] Bizer, Christian, Tom Heath e Tim Berners-Lee: *Linked data - the story so far.* International Journal on Semantic Web and Information Systems, 5:1–22, 2009. 11

[24] Group, Open Definiton: *The open definition.*, 2015. `http://opendefinition.org/`, acesso em 2018-10-11. 14

[25] Foundation, Open Knowledge: *Open data handbook.*, 2010. `http://opendatahandbook.org/guide/en/`, acesso em 2018-10-11. 14

[26] J. Hendler, J. Holm, C. Musialek e G. Thomas: *Us government linked open data: Semantic.data.gov.* IEEE Intelligent Systems, 27(3), 2012. 14

[27] S. Nigel, K. O'Hara, T. Berners Lee N. Gibbins H. Glaser W. Hall e M. Schraefel: *Linked open government data: Lessons from data.gov.uk.* IEEE Intelligent Systems, 27(3), 2012. 14

[28] Berners-Lee, Tim: *Putting government data online*, 2009. `https://www.w3.org/DesignIssues/GovData.html`. 14, 16

[29] República, Presidência da: *Decreto 8.777, de 11 de maio de 2016.*, 2016. `http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm`, acesso em 2018-10-11. 15

[30] Malamud, Carl, Tim O'Reilly, Greg Elin e et al: *8 principles of open government.*, 2007. `https://public.resource.org/8_principles.html`, acesso em 2019-02-11. 15, 26

[31] Ding, Li, Timothy Lebo, John S Erickson, Dominic Difranzo, Gregory Todd Williams, Xian Li, James Michaelis, Alvaro Graves, Jin Guang Zheng, Zhenning Shangguan *et al.*: *Twc logd: A portal for linked open government data ecosystems.* Journal of Web Semantics, 9(3):325–333, 2011. 16

[32] Maali, Fadi, Richard Cyganiak e Vassilios Peristeras: *A publishing pipeline for linked government data.* Em Simperl, Elena, Philipp Cimiano, Axel Polleres, Oscar Corcho e Valentina Presutti (editores): *The Semantic Web: Research and Applications*, páginas 778–792, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg, ISBN 978-3-642-30284-8. 16, 30, 32, 34

[33] Sheridan, John e Jeni Tennison: *Linking uk government data.* Em *Ldow*, 2010. 16

[34] Davidson, S. B. e J. Freire: *Provenance and scientific workflows: challenges and opportunities.* In Proceedings of the 2008, ACM SIGMOD international conference on Management of data., página 1345–1350, 2008. 16, 17

[35] Goble, C.: *Position statement: Musings on provenance, workflow and (semantic web) annotations for bioinformatics.* Workshop on Data Derivation and Provenance, 2002. 17

[36] Tan, W. C.: *Research problems in data provenance.* IEEE Data Eng. Bull., 27(4):45–52, 2004. 18

[37] Freire, J., D. Koop, E. Santos e C. T. Silva: *Provenance for computational tasks: A survey.* Computing in Science & Enginnering., 10(3):11–21, 2008. 18, 21, 22, 33

[38] Paula, Renato de. e MT de Holanda: *Proveniência de dados em workflows de bioinformática.*, 2012. `http://repositorio.unb.br/bitstream/10482/12699/1/2012_RenatodePaula.pdf`, acesso em 2018-10-11. 21

[39] Dominguez-Sal, David, Norbert Martinez-Bazan, Victor Muntes-Mulero, Pere Baleta e Josep Lluis Larriba-Pey: *A discussion on the design of graph database benchmarks.* Em *Technology Conference on Performance Evaluation and Benchmarking*, páginas 25–40. Springer, 2010. 23

[40] Angles, Renzo: *A comparison of current graph database models.* Em *2012 IEEE 28th International Conference on Data Engineering Workshops*, páginas 171–177. IEEE, 2012. 23

[41] Brasil: *Lei número 3.998, de 15 de dezembro de 1961.*, 1961. `http://www.planalto.gov.br/ccivil_03/LEIS/1950-1969/L3998.htm`. 25

[42] Simmhan, Yogesh L, Beth Plale e Dennis Gannon: *A survey of data provenance in e-science.* ACM Sigmod Record, 34(3):31–36, 2005. 29

[43] Cruz, Sérgio Manuel Serra da, Maria Luiza M Campos e Marta Mattoso: *Towards a taxonomy of provenance in scientific workflow management systems*. Em *2009 Congress on Services-I*, páginas 259–266. IEEE, 2009. 29

[44] Buneman, Peter e Susan B Davidson: *Data provenance — the foundation of data quality*, 2010. 29

[45] Herschel, Melanie e Marcel Hlawatsch: *Provenance: On and behind the screens*. Em *Proceedings of the 2016 International Conference on Management of Data*, páginas 2213–2217. ACM, 2016. 29

[46] Herschel, Melanie, Ralf Diestelkämper e Houssem Ben Lahmar: *A survey on provenance: What for? what form? what from?* The VLDB Journal—The International Journal on Very Large Data Bases, 26(6):881–906, 2017. 29

[47] Oinn, Tom, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat *et al.*: *Taverna: a tool for the composition and enactment of bioinformatics workflows*. Bioinformatics, 20(17):3045–3054, 2004. 30

[48] Oliveira, Daniel de, Eduardo Ogasawara, Fernanda Baião e Marta Mattoso: *Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows*. Em *2010 IEEE 3rd International Conference on Cloud Computing*, páginas 378–385. IEEE, 2010. 30

[49] Paula, Renato de, Maristela Holanda, Luciana SA Gomes, Sergio Lifschitz e Maria Emilia MT Walter: *Provenance in bioinformatics workflows*. BMC bioinformatics, 14(11):S6, 2013. 30

[50] Guimaraes, Valeria, Fernanda Hondo, Rodrigo Almeida, Harley Vera, Maristela Holanda, Aleteia Araujo, Maria Emilia Walter e Sergio Lifschitz: *A study of genomic data provenance in nosql document-oriented database systems*. Em *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, páginas 1525–1531. IEEE, 2015. 30

[51] Asghar, Muhammad Rizwan, Mihaela Ion, Giovanni Russello e Bruno Crispo: *Securing data provenance in the cloud*. Em *Open problems in network security*, páginas 145–160. Springer, 2012. 30

[52] Braun, Uri Jacob, Avraham Shinnar e Margo I Seltzer: *Securing provenance*. Em *Proceedings of the 3rd USENIX Workshop on Hot Topics in Security (HotSec'08)*. USENIX Association, 2008. 30

[53] Liang, Xueping, Sachin Shetty, Deepak Tosh, Charles Kamhoua, Kevin Kwiat e Laurent Njilla: *Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability*. Em *Proceedings of the 17th IEEE/ACM international symposium on cluster, cloud and grid computing*, páginas 468–477. IEEE Press, 2017. 30

[54] Faria Cordeiro, K. de, F. F. de Faria, B. de Oliveira Pereira, A. Freitas, C. E. Ribeiro, J. V. V. B. Freitas, ... e M. L. M. Campos: *An approach for managing and semantically enriching the publication of linked open governmental data.* In Proceedings of the 3rd workshop in applied computing for electronic government (WCGE), SBBD, 7(2):82–95, 2011. 30, 32, 34

[55] Maali, Fadi, Richard Cyganiak e Vassilios Peristeras: *Enabling interoperability of government data catalogues.* Em Wimmer, Maria A., Jean Loup Chappelet, Marijn Janssen e Hans J. Scholl (editores): *Electronic Government*, páginas 339–350, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg, ISBN 978-3-642-14799-9. 31

[56] Moreau, Luc, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers *et al.*: *The open provenance model core specification (v1. 1).* Future generation computer systems, 27(6):743–756, 2011. 31

[57] Mendonça, R. R. de, S. M. S. da Cruz, J. F. De La Cerda, M. C. Cavalcanti, K. F. Cordeiro e M. L. M. Campos: *Lop - capturing and linking open provenance on lod cycle.* Proc. of the 5th SWIM, ACM, 2013. 31, 32, 33, 34

[58] Mendonça, Rogers Reiche de, Sérgio Manuel Serra DA CRUZ e Maria Luiza Machado. CAMPOS: *Etl4linkedprov: Managing multigranular linked data provenance.* Journal of Information and Data Management., 7(2):70, 2016. 31, 32, 33, 34

[59] Campos, M.L.M. e G. Guizzardi: *Gt-linkeddatabr – exposição,compartilhamento e conexão de recursos de dados abertos na web (linked open data).*, 2010. `www.rnp.br/pd/gts2010-2011/gt_linkeddatabr.html`. 31

[60] Silva, Daniel L da, André Batista e Pedro LP Corrêa: *Data provenance in environmental monitoring.* Em *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, páginas 337–342. IEEE, 2016. 32, 33, 34

[61] Trinh, T. D., P. R. Aryan, B. L. Do, F. J. Ekaputra, E. Kiesling, A. Rauber, ... e A. M. Tjoa: *Linked data processing provenance: towards transparent and reusable linked data integration.* In Proceedings of the International Conference on Web Intelligence, páginas 88–96, 2017. 32, 33, 34

[62] Software, Pivotal: *Spring boot.*, 2014. `https://spring.io/projects/spring-boot`, acesso em 2019-08-11. 36

[63] Association, CKAN: *Api guide*, 2013. `https://docs.ckan.org/en/ckan-2.7.3/api/m`, acesso em 2019-06-17. 36

[64] Van Erven, Gustavo Cordeiro Galvão, Rommel Novaes Carvalho, Waldeyr Mendes Cordeiro da Silva, Sergio Lifschitz, Harley Vera-Olivera e Maristela Holanda: *Designing graph databases with graphed.* Journal of Database Management (JDM), 30(1):41–60, 2019. 39

[65] Dijkstra, Edsger W: *On the role of scientific thought.* Em *Selected writings on computing: a personal perspective*, páginas 60–66. Springer, 1982. 39

[66] Evans, Eric: *Domain-driven design: tackling complexity in the heart of software.* Addison-Wesley Professional, 2004. 44

[67] Vargas Agilar, Everton de: *Uma abordagem orientada a serviços para a modernização de sistemas legados.*, 2016. `https://repositorio.unb.br/handle/10482/22250`. 44

[68] Partner, J., A. Vukotic e N. Watt: *Neo4j in Action.* O'Reilly Media., 2013. 47

[69] Penteado, Raqueline RM, Rebeca Schroeder, Diego Hoss, Jaqueline Nande, Ricardo M Maeda, Walmir O Couto e Carmem S Hara: *Um estudo sobre bancos de dados em grafos nativos.* X ERBD-Escola Regional de Banco de Dados, 2014. 47

[70] Rodriguez, M. A. e P. Neubauer: *The graph traversal pattern.* IGI Global, páginas 29–46, 2012. 47

[71] Almeida, Rodrigo, Waldeyr Mendes Cordeiro da Silva, Klayton Castro, Aletéia Patricia Favacho De Araújo, Maria Emília Machado Telles Walter, Sergio Lifschitz e Maristela Holanda: *Managing data provenance for bioinformatics workflows using aprovbio.* International Journal of Computational Biology and Drug Design, 12(2):153–170, 2019. 48

[72] Reis Jr, Cleyton, Luiz Martins, Marcio Victorino e Maristela Holanda: *Modelo de dados de proveniência para uma arquitetura de dados abertos governamentais.* Em *Anais do VII Workshop de Transparência em Sistemas*, páginas 11–20. SBC, 2019. 53

[73] Reis, Cleyton P. dos, Waldeyr M. C. da Silva, Luiz C. B. Martins, Rodrigo Pinheiro, Márcio C. Victorino e Maristela Holanda: *Enhancing open government data with data provenance.* Em *Proceedings of the 11th International Conference on Management of Digital EcoSystems*, MEDES, página 142–149, New York, NY, USA, 2019. Association for Computing Machinery, ISBN 9781450362382. `https://doi.org/10.1145/3297662.3365791`. 53

Apêndice A

Conjunto de dados de proveniência de departamento no formato JSON

```
[ {
  "extractionSoftware" : {
    "id" : null,
    "name" : "Extracao_SIGRA_01",
    "organization" : "UnB",
    "description" : "Extração de dados dos SIGRA",
    "operador" : "Luiz Martins",
    "notes" : "Não se aplica"
  },
  "unbGOLD" : {
    "id" : null,
    "name" : "UnBGOLD",
    "organization" : "UnB",
    "description" : "Executa a indexação Semântica de dados",
    "operator" : "Luiz Martins",
    "notes" : "Não se aplica"
  },
  "etl" : {
    "id" : null,
    "name" : "Ambiente ETL",
    "program" : "Pentaho PDI",
    "version" : "7.1",
    "description" : "Extração e Limpeza BD SIGRA",
    "startAtTime" : "2018-03-10 20:21:27",
    "endAtTime" : "2018-03-10 20:24:29",
    "software" : "MS SQL Server",
    "notes" : "Não se aplica",
    "extractionSoftwareWasAssociatedWith" : null,
    "informationSystemsUsed" : null
  },
  "dataRequest" : {
    "id" : null,
    "name" : "Requisição de Departamento",
    "program" : "UnBGOLD",
    "version" : "0.9 Beta",
    "description" : "Requisita o CSV para indexação",
    "startAtTime" : "2018-03-10 20:30:23",
    "endAtTime" : "2018-03-10 20:30:38",
    "software" : "Java",
    "request" : [ "http://servicos.unb.br/dadosabertos/departamentos"],
    "dwUsed" : null,
    "unBGOLDWasAssociatedWith" : null
  },
  "semanticIndexing" : {
    "id" : null,
    "name" : "Indexação_Departamento",
    "controledVocabulary" : ["FOAF", "DC", "GCIEO", "LUBM", "AIISO" ],
    "unBGOLDWasAssociatedWith" : null,
    "csvUsed" : null,
    "program" : "UnBGOLD",
```

```
  "version" : "0.9 Beta",
  "startAtTime" : "2018-03-10 20:32:23",
  "endAtTime" : "2018-03-10 20:32:54",
  "software" : "Application Server",
  "description" : "Usa CSV e vocabulários para indexação e gera RDF"
},
"dataPublication" : {
  "id" : null,
  "name" : "Publicação_Departamento",
  "program" : "CKAN",
  "version" : "0.9 Beta",
  "description" : "Publicação na instância CKAN",
  "startAtTime" : "2018-03-10 20:35:20",
  "endAtTime" : "2018-03-10 20:36:20",
  "software" : "CKAN",
  "ckanInstance" : "Dados Abertos da UnB",
  "unBGOLDWasAssociatedWith" : null,
  "rdfUsed" : null
},
"informationSystems" : {
  "id" : null,
  "name" : "Sistemas de Informação",
  "size" : "15 GB",
  "description" : "BD de Sistemas de Graduação",
  "location" : "SIGRA",
  "notes" : "Não se aplica"
},
"dw" : {
  "id" : null,
  "name" : "DW_SIGRA_01",
  "size" : "4,2 GB",
  "description" : "BD gerado do proceso  ETL  no SIGRA",
  "location" : "MS SQL Server",
  "notes" : "Não se aplica",
  "etlWasGeneratedBy" : null,
  "informationSystemsWasDerivedFrom" : null
},
"csv" : {
  "id" : null,
  "name" : "departamento.csv",
  "size" : "112 k",
  "description" : "Arquivo CSV extraído",
  "location" : "Servidor de aplicação",
  "notes" : "Não se aplica",
  "dataRequestWasGeneratedBy" : null,
  "dwWasDerivedFrom" : null
},
"rdf" : {
  "id" : null,
  "name" : "Departamento.rdf",
```

```
  "size" : "140 k",
  "description" : "Arquivo RDF gerado",
  "location" : "Servidor de  aplicação",
  "notes" : "Não se aplica",
  "semanticIndexingWasGeneratedBy" : null,
  "csvWasDerivedFrom" : null
},
"ckan" : {
  "id" : null,
  "name" : "CKAN_Departamento",
  "size" : "Não se aplica",
  "description" : "Dados publicados na plataforma",
  "location" : "CKAN",
  "notes" : "Não se aplica",
  "dataPublicationWasGeneratedBy" : null,
  "rdfWasDerivedFrom" : null
}]
```

Apêndice B

Conjunto de dados de proveniência de departamento no formato CSV

n
"{name:Extracao_SIGRA_01,description:Extração de dados dos SIGRA,notes:Não se aplica,operador:Luiz Martins,organization:UnB}"
"{name:UnBGOLD,description:Executa a indexação Semântica de dados,notes:Não se aplica,operator:Luiz Martins,organization:UnB}"
"{software:CKAN,name:Publicação_Departamento,description:Publicação na instância CKAN,program:CKAN,ckanInstance:Dados Abertos da UnB,version:0.9 Beta,startAtTime:2018-03-10 20:35:20,endAtTime:2018-03-10 20:36:20}"
"{request:[http://servicos.unb.br/dadosabertos/departamentos],software:Java,name:Requisição de Departamento,description:Requisita o CSV para indexação,startAtTime:2018-03-10 20:30:23,version:0.9 Beta,endAtTime:2018-03-10 20:30:38}"
"{Program:UnBGOLD,Description:Usa CSV e vocabulários para indexação e gera RDF,Version:0.9 Beta,controledVocabulary:["FOAF", "DC", "GCIEO", "LUBM", "AIISO"],name:Indexação_Departamento,Software:Application Server,StartAtTime:2018-03-10 20:32:23,EndAtTime:2018-03-10 20:32:54}"
"{notes:Não se aplica,software:MS SQL Server,name:Ambiente ETL,description:Extração e Limpeza BD SIGRA,program:Pentaho PDI,startAtTime:2018-03-10 20:21:27,version:7.1,endAtTime:2018-03-10 20:24:29}"
"{name:Sistemas de Informação,description:BD de Sistemas de Graduação,location:SIGRA,notes:Não se aplica,size:15 GB}"
"{name:Departamento.rdf,description:Arquivo RDF gerado,location:Servidor de aplicação,notes:Não se aplica,size:140 k}"
"{name:departamento.csv,description:Arquivo CSV extraído,location:Servidor de aplicação,notes:Não se aplica,size:112 k}"
"{name:DW_SIGRA_01,description:BD gerado do proceso ETL no SIGRA,location:MS SQL Server,notes:Não se aplica,size:4,2 GB}"