

Modelo de Dados de Proveniência para uma Arquitetura de Dados Abertos Governamentais

Cleyton P. dos Reis Jr.¹, Luiz Carlos Barbosa Martins¹, Márcio de Carvalho Victorino¹, Maristela Holanda¹

¹Departamento de Ciência da Computação – Universidade de Brasília (UnB)
Caixa Postal 4446 – 70.910-900 – Brasília – DF – Brazil

²Faculdade de Ciência da Informação – Universidade de Brasília (UnB)
Caixa Postal 4446 – 70.910-900 – Brasília – DF – Brazil

cleytonreisl@gmail.com, {luizmartins, mcvictorino, mholanda}@unb.br

Abstract. *The Brazilian Government adhered to the policy of publication of linked open data, thus promoting a more transparent and open administration, allowing greater participation of society, strengthening democracy and combating corruption. In order to do so, it is necessary to collect information from provenance to know, for example, when, how, and why the data were created and published. In this paper, a provenance data model for a connected open data architecture was defined. For validation of the proposal, a real use case scenario was described in which the provenance data model was used.*

Resumo. *O Governo Brasileiro aderiu à política de publicação de dados abertos conectados, promovendo assim, uma administração mais transparente e aberta, possibilitando a participação da sociedade, fortalecendo a democracia e o combate à corrupção. Para tanto, faz-se necessária a coleta de informações de proveniência para que se saiba, por exemplo, quando, como, e porque os dados foram criados e publicados. Neste artigo, foi definido um modelo de dados de proveniência para uma arquitetura de dados abertos conectados. Para validação da proposta, foi descrito um cenário de caso de uso real em que o modelo de dados de proveniência foi utilizado.*

1. Introdução

Os dados abertos têm sido estimulados por movimentos globais, sobretudo após um memorando do presidente Barack Obama sobre Transparência e Dados Governamentais [Obama 2008] em 2008 e pela criação do portal de dados abertos do governo Norte-Americano. Neste cenário, em 2011 a *Open Government Partnership* foi formalmente lançada quando os 8 governos fundadores (Brasil, Indonésia, México, Noruega, Filipinas, África do Sul, Reino Unido e Estados Unidos) endossaram a Declaração do Governo Aberto e anunciaram sua ação no país. No Brasil, esta política foi consolidada por meio da Lei de Acesso à Informação (LAI) [Brasil 2011a], que garante o acesso a qualquer informação que possua interesse público, desde que esta informação não seja imprescindível à segurança da sociedade e do Estado.

Com o desenvolvimento e adoção da web semântica, vieram padrões e formatos para integrar dados e informações oriundos de diferentes fontes. Desta forma, o padrão de

dados conectados permite que qualquer instituição publique os dados de uma maneira que possam ser lidos por pessoas e processados por máquinas. O termo Dados Conectados surgiu em 2006, com a publicação do documento Design Issues [Berners-Lee 2006], com uma subseção de Web Semântica exclusiva para Dados Conectados. Dados Conectados (do inglês, *Linked Data*) refere-se a um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na Web, com o intuito de criar uma “Web de Dados” [Bizer and Berners-Lee 2009].

A adoção da web de dados [Bizer and Berners-Lee 2009] pode trazer diversos problemas, como o mesmo conjunto de dados conectados podem ser replicados e hospedados em diferentes locais na Web, através de diferentes URIs (URI – Uniform Resource Identifier). Conjuntos de dados podem ser conectados por diferentes triplas RDF (Resource Description Framework), criadas de maneira distinta e mantidas por diferentes editores. Como confiar em dados e links publicados? Qual dos objetos de dados conectados fornece informações mais confiáveis ou mais atualizadas sobre a entidade? Para responder a essas perguntas, precisa-se não apenas de dados sobre a entidade, mas também de informações sobre como os dados se tornaram disponíveis. É neste contexto que se fazem necessárias as informações sobre a proveniência dos dados conectados [Hartig and Zhao 2010].

A proveniência de dados visa descrever os acontecimentos e insumos utilizados na geração de uma determinada informação. Segundo [Buneman et al. 2001], proveniência de dados é “... a descrição das origens de uma peça de dados e do processo pelo qual ela chegou em um banco de dados.” (livre tradução), ou seja, para garantir a proveniência de dados se faz necessário guardar tanto a origem dos dados utilizados como matéria-prima, quanto os processos que transformaram estes dados no produto final.

Dentre os modelos de proveniência presentes na literatura, o PROV-DM tem se destacado, já que o principal objetivo é descrever as pessoas, recursos e atividades envolvidas na produção de uma peça de dado, criando condições para que a proveniência seja demonstrada e trocada entre diferentes sistemas[W3C 2013a]. Este artigo tem por objetivo definir um modelo de dados de proveniência, utilizando o modelo de dados PROV-DM para uma arquitetura de dados públicos abertos governamentais que prevê o armazenamento desses dados em um banco de dados de grafo. A abordagem de proveniência de dados públicos abertos governamentais amplia trabalhos anteriores ao adicionar o recurso de coleta dos dados de proveniência na ferramenta UnBGOLD (UnB Government Linked Open Data) [Martins et al. 2018].

Este artigo está organizado nas seguintes seções. Na Seção 2, tem-se os trabalhos relacionados. A Seção 3 apresenta a ferramenta *UnBGOLD*, no qual foi realizado o estudo de caso. A Seção 4 descreve a arquitetura de publicação de dados governamentais com captura dos metadados de proveniência. A Seção 5 propõe o modelo de dados da arquitetura de publicação de dados abertos. A Seção 6 relaciona o estudo de caso realizado para validação do modelo. Na Seção 7, concluímos e sugerimos trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção, são apresentados alguns trabalhos relacionados com a proveniência de dados em *workflows* de dados conectados e alguns trabalhos que abordam a proveniência de dados e a utilização do modelo PROV-DM.

No contexto de dados abertos conectados, existem trabalhos que abordam

questões de proveniência. Em [de Faria Cordeiro et al. 2011], foi proposta uma abordagem para estimular a publicação de dados governamentais, que inclui uma ferramenta de Extração, Transformação e Carga (ETL) para gerenciar o processo de publicação; *plugins* projetados para facilitar a extração, transformação e mapeamento de dados brutos em RDF (triplificação) e vinculação; e um repositório para gerenciar e armazenar dados no formato RDF. Para promover uma integração e 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. Esta arquitetura não aborda a gestão de proveniência e mecanismos para auxiliar no processo de curadoria de dados.

Em [de Mendonça et al. 2013], foi apresentado uma ferramenta para coleta de proveniência em dados abertos conectados. A proveniência foi capturada por mecanismo de publicação de dados conectados apoiado por um *workflow* ETL. A coleta da proveniência é realizada em três níveis, em que a camada inferior utiliza o *Open Provenance Model Vocabulary* (OPMV) [Zhao 2010] para descrever a proveniência básica para o *workflow*, a segunda camada usa o vocabulário COGS [Freitas et al. 2014] 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.

Outro estudo relevante foi realizado em [de Mendonça et al. 2016], que amplia trabalhos anteriores [de Mendonça et al. 2013] e [Campos and Guizzardi 2010], no qual foi introduzido *workflows* ETL de dados para publicar dados conectados. Neste estudo realiza a coleta e publicação de metadados de proveniência 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 proveniência por meio de triplas RDF anotadas em ontologias preexistentes, PROV-O[W3C 2013b], OPMW[Garijo and Gil 2014] e COGS, sendo que as duas últimas foram utilizadas como extensão da primeira.

Para [Trinh et al. 2017], 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 através 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, a maior parte do vocabulário foi reutilizado da ontologia PROV-O, além disso, vários conceitos foram fornecidos pelos vocabulários VOID[W3C 2011] e FOAF[Brickley and Miller 2014].

Em [Maali et al. 2012], foi definido um *pipeline* de publicação, centrado em torno do Google Refine, que suporta uma conversão de ponta a ponta de dados governamentais brutos para a dados governamentais conectados. O *pipeline* permite a publicação dos dados na plataforma CKAN de registro de dados abertos. As informações de proveniência foram representadas de acordo com o *Open Provenance Model Vocabulary* (OPMV) e compartilhando-as junto com os dados no CKAN. Para representação dos dados do go-

verno, foi desenvolvido o vocabulário DCAT [Maali et al. 2010]. Além disso, foram utilizados também, na representação de dados, diversos vocabulários como o VoID e o Data Cube Vocabulary [W3C 2014]. 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.

Foi apresentado em [d. Silva et al. 2016] 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 [DCMI 1995], 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.

Diferentemente desses estudos, nosso trabalho apresenta um modelo de dados de proveniência, utilizando o W3C PROV-DM para descrever informações detalhadas sobre os dados e processos envolvidos no ciclo de vida de uma arquitetura de publicação de dados abertos governamentais. Este artigo propõe o armazenamento dos dados de proveniência em um banco de dados grafo, permitindo uma modelagem mais natural do armazenamento de estruturas de dados, visto que o PROV-DM é baseado em grafos. Além disso, o banco de dados de grafo foi desenvolvido para dados com muitos links e nos dados de proveniência existem muitos links. E por fim, as estruturas de dados do grafo são visíveis para o usuário, permitindo expressar as consultas em um nível mais alto de abstração, facilitando a implementação de algoritmos eficientes para executar operações específicas.

3. UnBGOLD

A ferramenta UnBGOLD faz parte da arquitetura desenvolvida pela Universidade de Brasília (UnB) para melhorar a qualidade de publicação de seus dados. Trata-se de uma arquitetura fracamente acoplada que permite que outros órgãos possam utilizar essa ferramenta para também publicar seus dados, independentemente da fonte, além de oferecer uma interface na qual seja possível agilizar o processo através da automação da publicação.

O UnBGOLD oferece a opção de enriquecer os dados por meio de vocabulário controlado, utilizando metadados e ontologias que o representem semanticamente, transformando o dado aberto em dados abertos conectados. A arquitetura do UnBGOLD está dividida nas seguintes camadas:

- Extração dos dados: nesta camada são definidos os procedimentos de seleção, curadoria e extração dos dados, regras para acesso e disponibilização dos dados;
- 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;
- Publicação dos Dados: os dados são publicados pela ferramenta CKAN 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.

Essas camadas da arquitetura da publicação de dados abertos da UnB possui diferentes funções, onde 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 desejamos publicar, tornando-os assim dados abertos conectados. Já a segunda parte, consiste em indexar a fonte de dados através 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. A UnB utiliza a solução CKAN, que é uma plataforma Web projetada pela *Open Knowledge Foundation (OKF)* para a publicação e compartilhamento de dados abertos. Além disso, uma ferramenta de busca semântica está disponível com dados já indexados.

4. Captura e Publicação dos Metadados de Proveniência

O processo de publicação dos dados abertos governamentais conectados envolvem a execução de um *workflow*, definido como uma seqüência sistemática de atividades, com o objetivo de facilitar e automatizar o processo de configuração e uso das ferramentas existentes, orquestradas com o conjunto de ferramentas criadas no âmbito do próprio projeto.

Neste artigo, apresentamos uma estratégia de coleta de proveniência em diferentes etapas do processo de interligação dos dados abertos: as etapas de pré-processamento das fontes de dados, preparação e transformação, e também o processo de indexação semântica, quando as decisões sobre correspondência de esquema e fusão de dados devem ser registradas para uso futuro.

O mecanismo de coleta dados de proveniência é executado ao longo dessas atividades, registrando-a em diferentes níveis de granularidade. Para o Processo de Preparação e Transformação de Dados, dois tipos de proveniência são gerados, prospectiva (relacionada à composição do processo) e retrospectiva (relacionados à execução do processo).

No processo de interligação de dados também registra fontes de dados e metadados de mapeamentos como dados de proveniência. No entanto, nesses processos as associações de dados são o foco principal e as transformações de dados não são cobertas. Sendo assim, a proveniência coletada é do tipo retrospectiva. Os dados de proveniência coletados após a indexação semântica desempenham um papel fundamental no enriquecimento do contexto em torno dos dados abertos conectados, e podem ajudar a apoiar a avaliação de atributos como fidedignidade e qualidade.

A Figura 1 mostra o processo de Preparação e Transformação de Dados apoiado por uma abordagem de *workflow* de ETL para coletar e vincular dados de proveniência prospectivos e retrospectivos durante a execução do processo de publicação e também, a coleta da proveniência no processo de interligação dos dados. Posteriormente, a proveniência coletada é mapeada para o formato RDF e publicada, para que os dados abertos

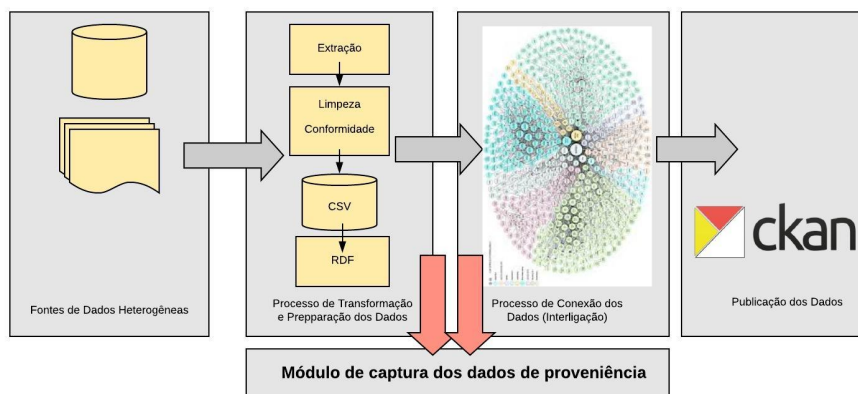


Figura 1. Workflow de publicação de dados abertos governamentais conectados com captura da proveniência

conectados e sua respectiva proveniência possam ser explorados conjuntamente por meio da ferramenta UnBGOLD e da plataforma CKAN.

5. Mapeamento das Entidades do PROV-DM para um Modelo de Grafos

Banco de dados de grafos permite o armazenamento de entidades e relacionamentos entre essas entidades. As entidades são conhecidas como nós, os quais possuem propriedades. Os relacionamentos são conhecidos como arestas, que também podem ter propriedades. As arestas têm significância direcional; nós comunicam-se por relacionamentos, os quais permitem encontrar padrões entre eles.

Sendo assim, os três tipos básicos do modelo PROV-DM, atividade, agente e entidade são representados como nós no banco de dados de grafo, e são diferenciados pela propriedade Tipo. Já as relações são representadas pelas arestas no banco de dados de grafo, sendo diferenciadas também por uma propriedade Tipo. Os principais componentes do modelo PROV-DM descritos a seguir são apresentados na Figura 2, onde os nós são:

- **Atividade:** representa os possíveis processos executados durante a execução do *workflow* de publicação dos dados abertos governamentais conectados. Vale ressaltar que, dos elementos que compõe o grafo (coleção, agente, grupo e atividade), este é o único que possui características temporais (hora inicial e hora final);
- **Coleção:** representa um arquivo ou *script* de banco de dados utilizado ou gerado durante a execução de uma atividade;
- **Agente:** representa qualquer entidade (pessoa, organização, software, serviço, etc...) que possa ter algum tipo de ação sobre uma Atividade ou possa ter algum tipo de responsabilidade sobre uma Coleção;
- **Grupo:** Este atributo está presente nos possíveis nós do grafo, ou seja, os elementos coleção, agente e atividade. Os grupos são definidos pelo usuário durante o cadastramento de cada elemento e, por ser um atributo múltiplo, cada elemento pode fazer parte de diferentes grupos. Dessa forma o usuário poderá solicitar a visualização de parte do grafo escolhendo os grupos desejados. Esse recurso é especialmente importante para grafos muito grandes onde o usuário pode querer visualizar somente parte dele.

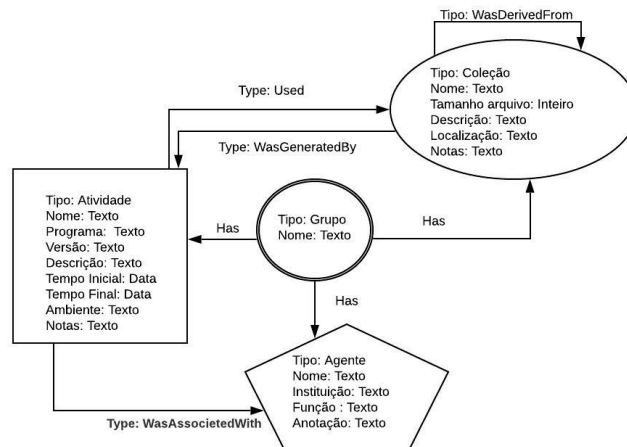


Figura 2. Mapeamento das Entidades PROV-DM para o Modelo de Grafos.

Já quanto as arestas tem-se:

- **Used:** Esta relação liga uma coleção ou uma entidade a atividade que está usando esta coleção. Como uma aresta direcionada ela sempre vai da atividade para a coleção. Cada coleção ou entidade poderá ser usada em diversas atividades. Da mesma forma que cada atividade poderá usar diversas coleções.
- **WasAssociatedWith:** indica que um agente teve algum tipo de ação sobre uma atividade. Seu direcionamento parte da atividade para o agente. Também são permitidas múltiplas ligações entre diferentes agentes e atividades.
- **WasDerivedFrom:** indica a ligação de derivação entre duas entidades. A aresta que representa a relação `wasDerivedFrom` é direcionada e aponta da coleção derivada para a respectiva coleção original. Também são permitidas múltiplas ligações entre coleções.
- **WasGeneratedBy:** indica qual atividade gerou uma determinada coleção. Seu direcionamento, como aresta, parte da coleção para a atividade que a gerou. Uma atividade poderá gerar diversas coleções ou atividades, porém cada coleção somente pode ser gerada por uma única atividade.

6. Estudo de Caso: Uso do Modelo Para publicação dos Dados Abertos

Para validar o modelo de dados proposto, foram utilizados um conjunto de dados reais, extraídos de uma das bases de dados dos sistemas de informação da Universidade de Brasília, o SIGRA. A base de dados do SIGRA possui dados referentes aos cursos de graduação existentes na universidade.

A Figura 3 mostra uma instância do grafo de proveniência gerado a partir da execução da arquitetura de publicação de dados abertos governamentais. Conforme definido no modelo PROV-DM, as atividades são representadas por retângulos, agentes por pentágonos e 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 do SIGRA.

disso foi realizado o mapeamento das entidades do PROV-DM para um modelo de dados de Grafos.

O modelo prevê a captura da proveniência em dois momentos diferentes da arquitetura de publicação de dados abertos. No workflow ETL está prevista a captura da proveniência prospectiva e retrospectiva. Na fase de indexação de dados temos apenas a proveniência retrospectiva. Essa proveniência deverá ser publicada e disponibilizada juntamente com os dados abertos conectados.

Como trabalho futuro tem-se a elaboração de uma arquitetura de gerência dos metadados de proveniência que será responsável pela captura automática, armazenamento e visualização dos metadados de proveniência em formato de um grafo utilizando o modelo de dados PROV-DM. .

Referências

- (2013a). Prov-dm: The prov data model. w3C. Disponível em: <https://www.w3.org/TR/prov-dm/> Acessado em: 07/10/2018.
- Berners-Lee, T. (2006). Linked data-design issues (2006). URL <http://www.w3.org/DesignIssues/LinkedData.html>, 10(11).
- Bizer, Christian, H. T. and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5:1–22.
- Brasil (2011a). Lei de acesso à informação. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm Acessado em: 28/01/2019.
- Brasil (2011b). Vocabulário controlado do governo eletrônico. Disponível em: <https://www.governodigital.gov.br/transformacao/compras/orientacoes/interoperabilidade/vocabulario-controlado-do-governo-eletronico-vcge> Acessado em: 20/12/2018.
- Brickley, D. and Miller, L. (2014). Foaf vocabulary specification 0.99. Disponível em: <http://xmlns.com/foaf/spec/> Acessado em: 14/01/2019.
- Buneman, P., Khanna, S., and Wang-Chiew, T. (2001). Why and where: A characterization of data provenance. *International Conference on Database Theory*, 1:316–330.
- Campos, M. and Guizzardi, G. (2010). Gt-linkeddatabr – exposição, compartilhamento e conexão de recursos de dados abertos na web (linked open data). Disponível em: www.rnp.br/pd/gts2010-2011/gt_linkeddatabr.html.
- d. Silva, D. L., Batista, A., and Corrêa, P. L. P. (2016). Data provenance in environmental monitoring. In *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pages 337–342.
- DCMI (1995). Dublin core metadata initiative. Disponível em: <http://www.dublincore.org/> Acessado em: 20/12/2018.
- de Faria Cordeiro, K., de Faria, F. F., de Oliveira Pereira, B., Freitas, A., Ribeiro, C. E., Freitas, J. V. V. B., ..., and Campos, M. L. M. (2011). 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.
- de Mendonça, R. R., DA CRUZ, S. M. S., and CAMPOS, M. L. M. (2016). Etl4linkedprov: Managing multigranular linked data provenance. *Journal of Information and Data Management.*, 7(2):70.
- de Mendonça, R. R., da Cruz, S. M. S., De La Cerda, J. F., Cavalcanti, M. C., Cordeiro, K. F., and Campos, M. L. M. (2013). Lop - capturing and linking open provenance on lod cycle. *Proc. of the 5th SWIM, ACM*.
- Freitas, A., Kämpgen, B., and Lebo, T. (2014). Cogs vocabulary. Disponível em: <http://vocab.der.iie/cogs> Acessado em: 20/12/2018.
- Garijo, D. and Gil, Y. (2014). The opmw-prov ontology. Disponível em: <http://www.opmw.org/model/OPMW/> Acessado em: 20/12/2018.
- Hartig, O. and Zhao, J. (2010). Publishing and consuming provenance metadata on the web of linked data. *International Provenance and Annotation Workshop*, 6378:78–90.
- Maali, F., Cyganiak, R., and Peristeras, V. (2010). Enabling interoperability of government data catalogues. In Wimmer, M. A., Chappellet, J.-L., Janssen, M., and Scholl, H. J., editors, *Electronic Government*, pages 339–350, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Maali, F., Cyganiak, R., and Peristeras, V. (2012). A publishing pipeline for linked government data. In Simperl, E., Cimiano, P., Polleres, A., Corcho, O., and Presutti, V., editors, *The Semantic Web: Research and Applications*, pages 778–792, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Martins, L. C. B., Victorino, M. C., Holanda, M., Ghinea, G., and Grønli, T.-M. (2018). Unbgold: Unb government open linked data: Semantic enrichment of open data tool. In *Proceedings of the 10th International Conference on Management of Digital EcoSystems*, MEDES '18, pages 1–6, New York, NY, USA. ACM.
- Obama, B. (2008). *Transparency and Open Government*. Washington, United States of America: The White House.
- Trinh, T. D., Aryan, P. R., Do, B. L., Ekaputra, F. J., Kiesling, E., Rauber, A., ..., and Tjoa, A. M. (2017). Linked data processing provenance: towards transparent and reusable linked data integration. *In Proceedings of the International Conference on Web Intelligence*, pages 88–96.
- W3C (2011). Describing linked datasets with the void vocabulary. Disponível em: <https://www.w3.org/TR/void/> Acessado em: 20/01/2019.
- W3C (2013b). Prov-o: The prov ontology. Disponível em: <https://www.w3.org/TR/prov-o/> Acessado em: 20/12/2018.
- W3C (2014). The rdf data cube vocabulary. Disponível em: <https://www.w3.org/TR/vocab-data-cube/> Acessado em: 20/12/2018.
- Zhao, J. (2010). Open provenance model vocabulary specification. Disponível em: <http://open-biomed.sourceforge.net/opmv/ns.html> Acessado em: 20/12/2018.