METADADOS DE NEGÓCIO E DADOS ABERTOS CONECTADOS: SEMÂNTICA NA ARQUITETURA DA INFORMAÇÃO PARA PROCESSOS DE NEGÓCIO

Business metadata and linked open data: semantics in information architecture for business processes
Metadatos de negocio y datos abiertos conectados: semántica en la arquitectura de la información para procesos de negocio

Mariana Baptista Brandt
Doutora em Ciência da Informação, Universidade Estadual Paulista (UNESP), Marília, SP, Brasil.
Analista Legislativo, Câmara dos Deputados, Brasília, DF, Brasil.
Lattes: http://lattes.cnpq.br/3761037263199030
ORCID: http://orcid.org/0000-0001-8119-7527

Silvana Aparecida Borseti Gregorio Vidotti
Doutora em Educação, Universidade Estadual Paulista (Unesp), Marília, SP, Brasil.
Assessora de gabinete da Pró-reitoria de Graduação da Universidade Estadual Paulista (Unesp), São Paulo, SP, Brasil.
Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (Unesp), Marília, SP, Brasil.
Lattes: http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4782033H4
ORCID: orcid.org/0000-0002-4216-0374

Resumo

Introdução: A publicação de dados em formato aberto e conectado não ocorre de forma ampla, impedindo que a web atinja todo seu potencial. **Objetivo**: Assim, objetiva-se propor uma forma de estruturar os dados diretamente nos sistemas de informação em que são gerados, por meio da inclusão de uma etapa para representação dos metadados de negócio em formato RDF, na metodologia de arquitetura da informação para processos de negócio (AIPN). **Metodologia**: Para tanto, procedeu-se à análise da literatura e documentação da W3C, além de pesquisa aplicada. **Resultados**: Observou-se que há elementos análogos entre a estrutura RDF e a AIPN. **Conclusão**: Há viabilidade para a inclusão de estrutura semântica na produção dos dados por meio de uma etapa de descrição RDF dos metadados de negócio na metodologia AIPN.

Palavras-chave: metadados de negócio; dados abertos conectados; arquitetura da informação para processos de negócio; web semântica.

Abstract

***Introduction**: Publication of linked open data is not widespread, preventing the web from reaching its full potential. **Objective**: Thus, the objective is to propose structuring the data directly in the information systems in which they are generated, through the inclusion of a step for representing the business metadata in RDF format, in the methodology of information architecture for business processes (AIPN). **Methods**: An analysis of the literature and documentation of the W3C was carried out, in addition to applied research. **Results**: It was observed that there are analogous elements between the RDF structure and the AIPN, **Conclusion**: The inclusion of a semantic structure in the production of data is possible through an RDF description step of the business metadata.*

***Keywords**: business metadata; connected open data; information architecture for business processes; semantic web.*

Resumen

***Introducción**: La publicación de datos en formato abierto y conectado no está muy extendida, lo que impide que la web alcance todo su potencial. **Objetivo**: Así, el objetivo es proponer una forma de estructurar los datos directamente en los sistemas de información en los que se generan, mediante la inclusión de una etapa de representación de metadatos empresariales en formato RDF, en la metodología de arquitectura de la información para procesos de negocio (AIPN). **Metodología**: Para ello se realizó un análisis de la literatura y documentación del W3C, además de una investigación aplicada. **Resultados**: Se observó que existen elementos análogos entre la estructura RDF y el AIPN. **Conclusión**: Es factible incluir una estructura semántica en la producción de datos a través de un paso de descripción RDF de los metadatos empresariales en la metodología AIPN.*

***Palabras clave:** metadatos de negocio; datos abiertos conectados; arquitectura de la información para procesos de negocio; web semántica..*

DOI: 10.22477/vi.widat.30

1. Introdução

Para que sejam aproveitados e reutilizados com todo seu potencial, os dados disponibilizados na *web* devem seguir padrões mínimos de estruturação. Para isso, o Consórcio *World Wide Web* (W3C) publica recomendações, diretrizes e guias com as melhores práticas relacionadas à *web*, incluindo para a estruturação e publicação de dados de forma a integrá-los à Web Semântica (WS). O elemento fundamental para estruturação de dados para a WS, ou Web de Dados, é o formato *Resource Description Framework* (RDF).

Porém, muitos dos conjuntos de dados disponibilizados na *web* não estão estruturados em RDF, não se integrando à WS. Na esfera governamental, a presença de dados abertos estruturados de forma semântica é praticamente nula: no portal de dados1 do governo federal, foram encontrados apenas 3 conjuntos de dados no formato RDF. Assim, em uma escala de 1 a 5 criada por Tim Berners-Lee (2006) e utilizada pela W3C para classificação de dados abertos, em que o padrão 5 significa o "ideal" de estruturação de dados para WS, a maioria dos dados governamentais chegam no máximo ao padrão 3: disponíveis na *web*, licença aberta, de forma estruturada e em formato não proprietário.

A Arquitetura da informação para processos de negócio (AIPN), proposta por Brandt (2020), é uma metodologia que pode ser implementada em qualquer processo de negócio, da área pública ou privada, de todos os setores. A AIPN tem como base os métodos, as práticas e os princípios de Biblioteconomia e de Ciência da Informação e tem como elemento principal o metadado de negócio.

Tais metadados de negócio, mapeados nesta metodologia, abrigam dados, ou seja, valores reais originados nos processos de trabalho. Esses dados podem ser de interesse público, em especial quando são dados governamentais. Para que ocorra um amplo acesso a esses dados, é recomendada sua publicação na *web*, por ser um meio de disseminação abrangente, além de promover a transparência.

O objetivo deste trabalho é propor uma nova etapa na AIPN para descrever os metadados de negócio em formato RDF, de modo a possibilitar a publicação dos dados dos processos de negócio no formato de dados abertos conectados.

1.1 Dados abertos conectados e RDF

Dados Abertos Conectados ou *Linked Open Data* (LOD) é a uma forma de publicação de dados na *web* em que os dados abertos são estruturados de acordo com padrões pré-estabelecidos, com a utilização de sintaxes, linguagens e vocabulários próprios, que se conectam com outros dados e são publicados com licença aberta, ou seja, os dados são públicos e podem ser reutilizados. Com isso, a WS é formada: "A Web Semântica não é

1 https://dados.gov.br/, em 19 de maio de 2023

só colocar dados na web. É fazer conexões (*links*), para que uma pessoa ou máquina possa explorar a web de dados." (BERNERS-LEE, 2006, online, tradução nossa).

O conceito de Dados Abertos Conectados surge de outros dois conceitos: Dados Abertos e Dados Conectados (*linked data - LD*). Assim, conjuntos de dados conectados publicados sob licença aberta configuram conjuntos de dados abertos conectados. O conceito de LD surgiu em 2006 como proposta de Tim Berners-Lee, que definiu regras para que um conjunto de dados seja considerado LD, entre elas o uso do padrão RDF.

O RDF é um padrão para modelagem de dados na *web* e foi criado pelo Grupo de Estudos RDF da W3C em 2004, com atualização em 2014. A W3C (2014) define RDF como um modelo padrão para representação da informação e intercâmbio de dados na *web*, com características que facilitam a fusão de dados, mesmo que estruturados com esquemas diferentes.

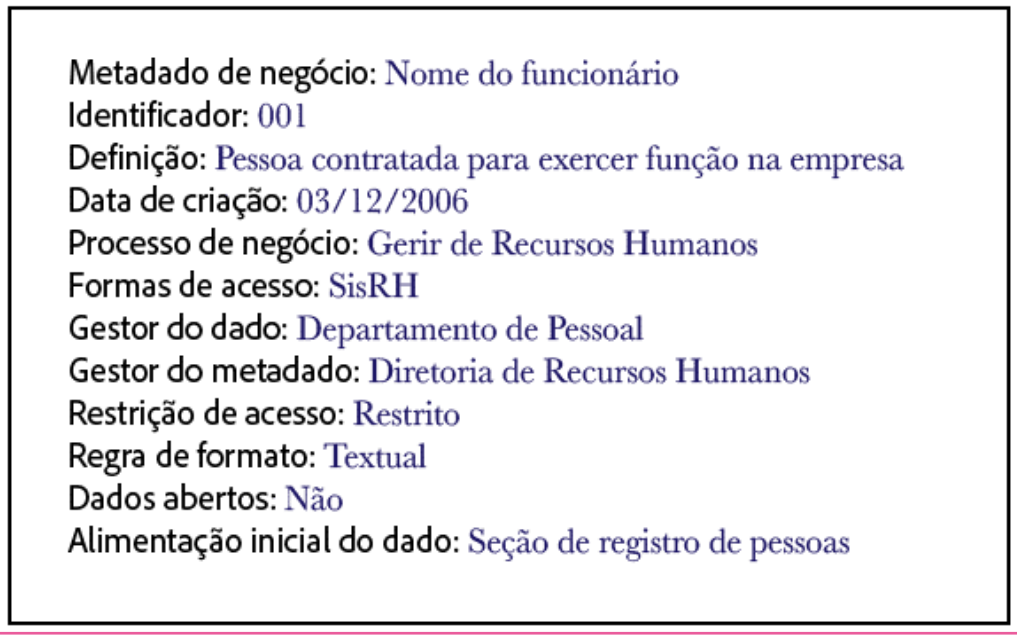
Isotani e Bittencourt (2015) explicam que o RDF é como uma linguagem de representação de informação na *web*, permitindo que recursos possam ser descritos e sejam acessíveis. A representação da informação em RDF é feita com base em triplas com a sintaxe sujeito, predicado e objeto (ou recurso, propriedade e valor). Ou seja, segue um modelo semelhante a sentenças gramaticais, em que o sujeito é o recurso que está sendo descrito, o predicado é uma propriedade do recurso e o valor, que é o dado em si, corresponde ao objeto da sentença.

1.2 Arquitetura da Informação para Processos de Negócio

A metodologia AIPN elabora um modelo de descrição das informações dos processos de negócio de uma instituição para guiar a construção de sistemas de informação, além de viabilizar a gestão da informação e a governança de dados. A AIPN mapeia as informações relevantes para o negócio, as quais devem ser gerenciadas. São os chamados metadados de negócio, e irão abrigar os dados do negócio (BRANDT, 2020). Esse mapeamento pode ser usado como base para a modelagem de bancos de dados em sistemas de informação da instituição.

Os metadados de negócio consistem no principal elemento da AIPN, pois são eles que abrigam os dados mais relevantes do negócio e nos quais deverá haver interesse em que sejam publicados na *web*. A metodologia prevê, para cada um dos metadados de negócio, o registro de uma descrição que contém elementos como: definição, gestor, forma de acesso, formato, restrição de acesso, entre outros que podem ser incluídos conforme as necessidades de cada instituição. Assim, cria-se uma catalogação para cada metadado de negócio, que funciona como uma espécie de manual de instruções para a gestão e a governança dos dados do processo de negócio e suas características para armazenamento nos sistemas (Figura 1).

Figura 1 - Catalogação do metadado de negócio "Nome do funcionário"



Fonte: Elaborado pelas autoras (2023)

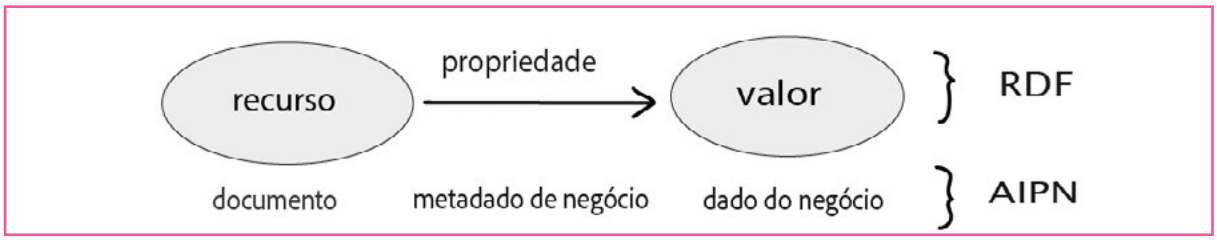
Entre esses elementos de descrição dos metadados de negócio, propõe-se a inclusão de um atributo que contenha a informação referente à estrutura do dado para publicação na *web*, em formato RDF e com a utilização de vocabulários da WS e outros dados ligados. Essa estrutura estaria registrada no banco de dados do sistema de informação de origem, atrelado ao registro dos dados de negócio que podem vir a ser publicados, constituindo assim uma camada semântica.

2. Procedimentos Metodológicos

O estudo caracteriza-se como exploratório, com base em literatura especializada, pesquisa documental e pesquisa aplicada. Foi utilizado um conjunto de metadados de negócio de um processo fictício para aplicar e exemplificar as etapas do procedimento metodológico.

Partiu-se da estrutura básica do RDF, que é a formação de triplas no formato recurso-propriedade-valor (W3C, 2014), onde buscou-se identificar a correlação entre tais elementos e os elementos da AIPN, com foco nos metadados de negócio. Visto que os metadados de negócio se referem a informações **a respeito dos dados** do negócio, pode-se concluir que se trata de uma propriedade, ou seja, um predicado do dado. A correlação do modelo RDF com a AIPN pode ser representada conforme a Figura 2.

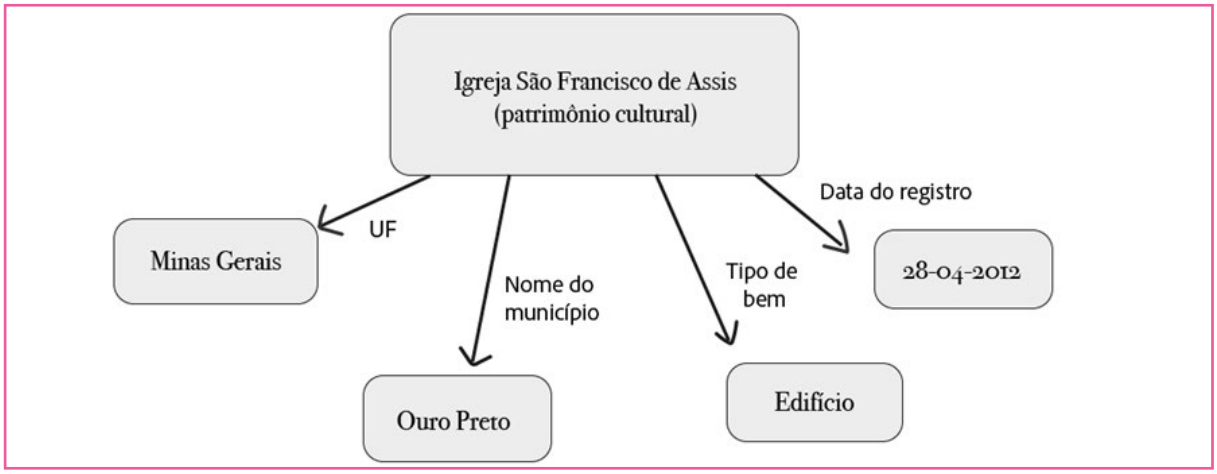
Figura 2 - Elementos do RDF e AIPN



Fonte: Elaborado pelas autoras (2023)

Assim, o metadado de negócio pode ser escrito em formato RDF, compatível com a WS. Para isso, propõe-se a construção de um diagrama de modelo conceitual (Figura 3) para verificar como os metadados de negócio se configuram como propriedades de um recurso (entidade).

Figura 3 - Modelo conceitual do recurso

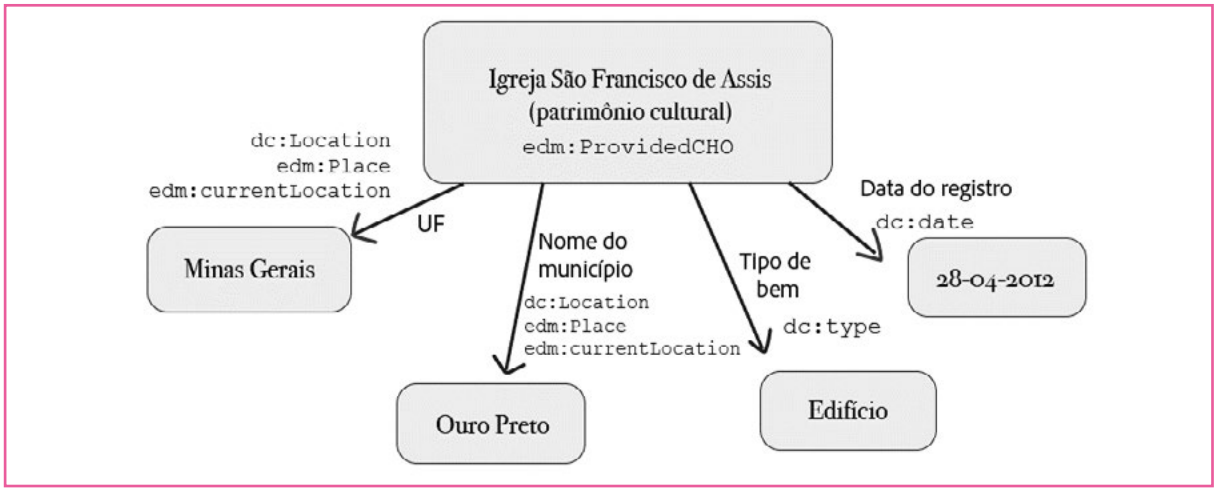


Fonte: Elaborado pelas autoras (2023)

Com a identificação desses relacionamentos, os metadados de negócio podem ser descritos como propriedades, utilizando vocabulários da WS. Os vocabulários são os elementos que inserem os conjuntos de dados na WS, pois trazem significado às relações entre os dados, e, consequentemente, aos dados.

O reuso de vocabulários é uma recomendação da W3C (LÓSCIO, BURLE, CALEGARI, 2017), portanto, a etapa de seleção de vocabulários deve ser iniciada com uma busca em vocabulários já existentes na WS. Recorreu-se ao *Linked Open Vocabularies*2 (LOV), repositório de vocabulários abertos e conectados. Escolhidos os vocabulários, as propriedades identificadas, ou seja, os metadados de negócio, podem ser descritos utilizando elementos padronizados: termos e classes dos vocabulários, conforme apresentado na Figura 4. Utilizou-se o vocabulário *Europeana Data Model* (EDM), do domínio de bens culturais, ao qual pertencem os dados desse exemplo. Ainda, termos do Dublin Core utilizados pelo EDM foram adicionados visando aumentar a integração à WS.

Figura 4 - Modelo conceitual com vocabulários



Fonte: Elaborado pelas autoras (2023)

Além das propriedades, representadas por vocabulários, os próprios valores também podem ser representados por outros conjuntos de dados, trazendo enriquecimento semântico e tornando o conjunto conectado a outros conjuntos de dados. Com isso, é atendido mais um princípio do *linked data* proposto por Berners-Lee (2006, online): "Incluir links para outras URIs, para que possam ser descobertas mais coisas". Segundo Heath e Bizer (2011, online, tradução nossa) "Links externos em RDF são fundamentais para a Web de Dados, pois conectam dados isolados num espaço global interconectado, além de possibilitar que as aplicações descubram novas fontes de dados".

Assim, os valores do conjunto de dados trabalhados devem ser analisados para identificar se também são recursos, ou seja, podem possuir URI. Feito isso, recomenda-se verificar se há outros conjuntos de dados do mesmo âmbito (instituição, órgão) que representam tais recursos. Caso existam, podem ser utilizados na modelagem do

2 Disponível em: http://lov.okfn.org/.

valor da tripla RDF. Outra forma de encontrar recursos que possam representar valores de conjuntos de dados é realizando busca em repositórios LOD: lod-cloud.net, https://datahub.io/, wikidata.org etc.

Para o exemplo deste trabalho, foi selecionado o Geonames, que contém lugares geográficos, como conjunto de dados conectados para representar os dados de UF e município. A vantagem de utilizar LOD em vez de literais (valor do dado) é que o dado representado pelo LOD trará informações adicionais sobre o dado, aumentando o nível de significado deste dado na *web*:

> Cada um dos aproximadamente 10 milhões de registros geográficos do Geonames é representado com um conjunto básico de elementos, sendo eles: número ID ou código Geonames, nome geográfico, nomes alternativos, latitude, longitude, código de classe (classes listadas anteriormente), código de categorização, código de país (baseado na ISO-3166), código alternativo de país, até quatro códigos administrativos (regiões e subregiões), população, elevação (em metros), área no fuso horário e última data de modificação. (SIMIONATO, SANTARÉM SEGUNDO, 2017, p. 124-125)

A conexão desses conjuntos de dados traz informações adicionais sobre os dados, gerando um ganho de significado chamado de enriquecimento semântico.

3. Resultados e discussão

O Quadro 1 resume os elementos mapeados neste exemplo, na nova etapa incluída na AIPN.

Quadro 1 - Recursos para enriquecimento semântico

Recurso	Propriedade	Metadado de negócio	Valor/dado	LOD / URI
Igreja São Francisco de Assis (Patrimônio cultural) edm:ProvidedCHO	edm:Place dc:Location edm:currentLocation	UF	Minas Gerais	http://www.geonames.org/3457153/minas-gerais.html
	edm:Place dc:Location edm:currentLocation	Nome do município	Ouro Preto	http://www.geonames.org/3455671/ouro-preto.html
	dc:description dc:type	Tipo de bem	edifício	Literal
	dc:date	Data de registro	29/05/2012	Literal

Fonte: Dados da pesquisa (2023).

Devem-se utilizar ainda as estruturas de RDF *Schema rdfs:label e rdf:value*, que não foram repetidas no Quadro 1, pois são usadas em todas as propriedades. Essas estruturas definem, respectivamente, o rótulo da propriedade (metadado de negócio) e seu valor literal (valor do dado). Com isso, os dados podem ser escritos em algum formato de serialização (*JSON, Turtle*, XML-RDF) para sua publicação, utilizando as propriedades de vocabulários identificados, as quais já estarão armazenadas no sistema de informação de origem pois foram mapeadas como atributos nos metadados de negócio de cada dado. Como o exemplo, o dado "MG" do metadado de negócio UF, poderia ser publicado como:

```
dc:Location[rdfs:label"UF";rdf:value"MG";edm:Place;edm:currentLocation;"http://www.geonames.org/3457153/" ];
```

Os procedimentos descritos na seção anterior demonstram a inclusão de um novo atributo aos metadados de negócio da AIPN, que podem ser mapeados desde a concepção do sistema de informação e incluídos na descrição dos metadados de negócio.Com isso, os conjuntos de dados produzidos nos processos de negócio podem ser publicados já no formato padrão LOD/RDF, pois a estrutura está presente desde a sua origem nos sistemas de informação da instituição. Ou seja, a modelagem em RDF para cada dado originado no sistema estaria armazenada no banco de dados, relacionada ao metadado de negócio correspondente.

Assim, no momento da publicação dos dados do processo de negócio na *web*, poderia ser incluída uma camada semântica que já foi mapeada e definida, de forma padronizada, sem a necessidade de uma nova estruturação.

Este estudo mostra como inserir semântica na metodologia AIPN, por meio de uma etapa que estrutura o modelo recomendado internacionalmente para publicação de dados na *web*, seguindo as boas práticas da W3C.

4. Considerações Finais

A W3C, instituição fundada pelo criador da *Web* para desenvolvê-la em seu maior potencial, tem trabalhado no sentido de fornecer várias recomendações, padrões, tutoriais, cursos e demais elementos necessários para realizar sua missão. Apesar disso, observa-se que a quantidade de pessoas e instituições que seguem essas diretrizes não é suficiente para levar a *web* a seu maior potencial. Apenas uma pequena parte dos dados publicados está inserida na WS.

Este trabalho demonstrou como os dados podem ser gerados já com estrutura semântica para publicação na *web*, no formato RDF para LOD, por meio da inclusão desta etapa na metodologia de AIPN, conforme objetivo da pesquisa. Com isso, criam-se condições para que mais conjuntos de dados sejam publicados na *web* em formato aberto e conectado e ainda com enriquecimento semântico.

Assim, as instituições que publicam seus dados na *web* podem se beneficiar da utilização da metodologia AIPN não somente para gestão da informação e governança de dados, mas também para a estruturação de dados em formato semântico e conectado. Com isso, seus dados já nascem estruturados conforme as melhores práticas da W3C, prontos para sua publicação na *web*.

Referências

BERNERS-LEE, Tim. **Linked data**: design issues 2006. 2016. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 07 jun. 2016.

BRANDT, Mariana Baptista. **Modelagem da informação legislativa**. 2020. Tese. (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista (Unesp), Marília, 2020. Disponível em: https://repositorio.unesp.br/handle/11449/191740. Acesso em: 17 maio 2023.

HEATH, Tom; BIZER, Christian. **Linked Data**: evolving the web into a global data space. EUA: Morgan & Claypool, 2011. Disponível em: http://linkeddatabook.com/editions/1.0/. Acesso em: 09 jun. 2017.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados abertos conectados**: em busca da web do conhecimento. São Paulo: Novatec, 2015.

LÓSCIO, Bernadette Farias; BURLE, Caroline; CALEGARI, Newton (Eds.). **Data on the web best practices.** 2017. Disponível em: https://www.w3.org/TR/dwbp. Acesso em: 20 jun. 2017.

SANTARÉM SEGUNDO, José Eduardo; SIMIONATO, Ana Carolina. Uma abordagem sobre a estrutura do geonames e suas contribuições para o linking open data. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais** [...]. Salvador: UFBA, 2017. Disponível em: https://brapci.inf.br/index.php/res/v/191935. Acesso em: 18 maio 2023.

WORLD WIDE WEB CONSORTIUM (W3C). *RDF 1.1* **Concepts and Abstract Syntax**. 2014. Disponível em: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/Overview.html . Acesso em: 18 maio 2023.