Arquitetura da informação para publicação de dados abertos conectados

Mariana Baptista Brandt
Doutora em Ciência da Informação, Universidade Estadual Paulista (UNESP), Marília, SP, Brasil.
Analista Legislativo, Câmara dos Deputados, Brasília, DF, Brasil.
http://lattes.cnpq.br/3761037263199030
Email: mariana.brandt@unesp.br

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.
http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4782033H4
Email: silvana.vidotti@unesp.br

Submetido em: 05/09/2023. Aprovado em: 30/04/2024. Publicado em: 18/07/2024.

RESUMO

A publicação de dados em formato aberto e conectado não ocorre de forma ampla, impedindo que a web atinja todo seu potencial. 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). Para tanto, procedeu-se à análise da literatura e documentação da W3C, além de pesquisa aplicada. Observou-se que há elementos análogos entre a estrutura RDF e a AIPN, o que viabiliza a inclusão de estrutura semântica na produção dos dados por meio de uma etapa de descrição RDF dos metadados de negócio.

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

INTRODUÇÃO

Para serem 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 para 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, portanto, não se integram à WS. Na esfera governamental, a presença de dados abertos estruturados de forma semântica é praticamente nula: no portal de dados 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.

Os dados publicados na web são gerados nos sistemas de informação que automatizam os processos de negócio das instituições. Esses processos onde se originam os dados podem ser modelados para a implementação dos sistemas de informação por meio da metodologia de Arquitetura da informação para processos de negócio (AIPN). A 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 ocorrer um amplo acesso a esses dados, é recomendada sua publicação na web, por ser um meio de disseminação abrangente, além de promover a transparência:

> A publicação de informações na web é vista, então, como um meio de excelência para publicar, e, assim, publicizar informações. A publicação dessas informações deve ser feita a partir de representações compatíveis com normas, padrões e requisitos próprios deste ambiente para que possam ser acessadas com facilidade e reutilizadas para os mais diversos fins. (Brandt, 2020, p. 136).

Assim, o objetivo deste trabalho é propor uma nova etapa na AIPN para descrever os metadados de negócio em formato RDF, conforme recomendação da pesquisa de Brandt (2020, p. 132): "[...] o atributo "Dados abertos" pode ser expandido, com a informação

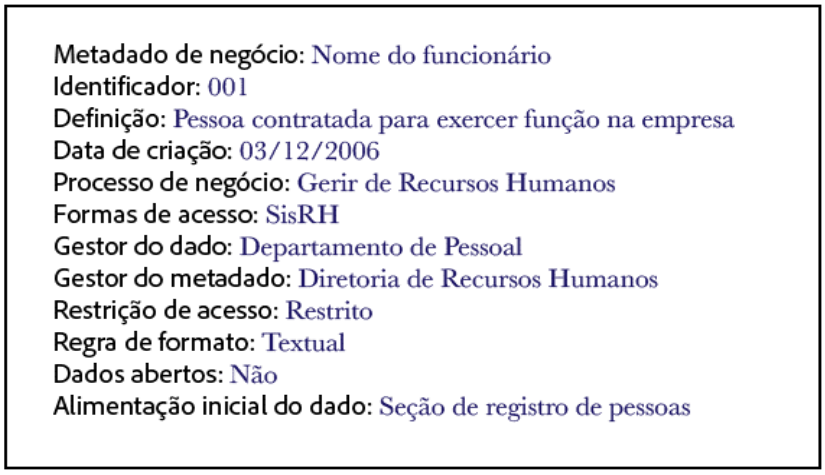
da estrutura do dado para publicação na web [...]". Essa estruturação visa possibilitar a publicação dos dados dos processos de negócio na Web Semântica em formato de dados abertos conectados.

Arquitetura da Informação para Processos de Negócio

A metodologia AIPN elabora um modelo de descrição das informações dos processos de negócio de instituições 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, que 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 das instituições.

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 do dado, gestor do metadado, forma de acesso, formato, restrição de acesso, entre outros que podem ser incluídos conforme as necessidades de cada instituição e de cada processo de negócio. 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, além de abarcar suas características para armazenamento nos sistemas de informação (**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 indicação de vocabulários da web semântica e outros conjuntos de

dados ligados que podem ser utilizados. Essa estrutura poderia estar registrada no banco de dados do sistema de informação de origem, atrelado aos registros dos dados de negócio que podem ser publicados, constituindo, assim, uma camada semântica compatível com as recomendações para publicação de dados. Essas recomendações para publicação de dados na web serão abordadas na seção 1.2 a seguir.

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 Web Semântica é formada: "A Web Semântica não é 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)[1].

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

METODOLOGIA

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. O recurso contém os seguintes dados:

1 Original: "The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data" (Berners-Lee, 2006, *online*).

Quadro 1 – Conjunto de dados utilizado na aplicação da metodologia

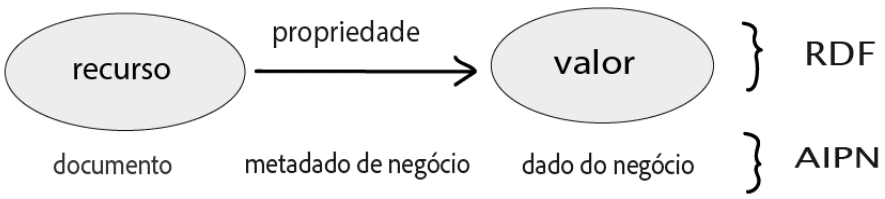
Patrimônio cultural	UF	Nome do município	Tipo de bem	Data de registro
Igreja São Francisco de Assis	**Minas Gerais**	**Ouro Preto**	**Edifício**	**28-04-2012**
...	...	...	...	...
...	...	...	...	...

Fonte: Elaborado pelas autoras (2023).

Este formato de estrutura de tabela para os dados é como se encontra grande parte dos conjuntos de dados abertos governamentais, e, como citado anteriormente, contempla o nível 3 estrelas da escala de *linked open data* da W3C.

Partiu-se da estrutura básica do RDF, que é a formação de triplas no formato recurso-propriedade-valor (W3C, 2014), onde se buscou 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**.

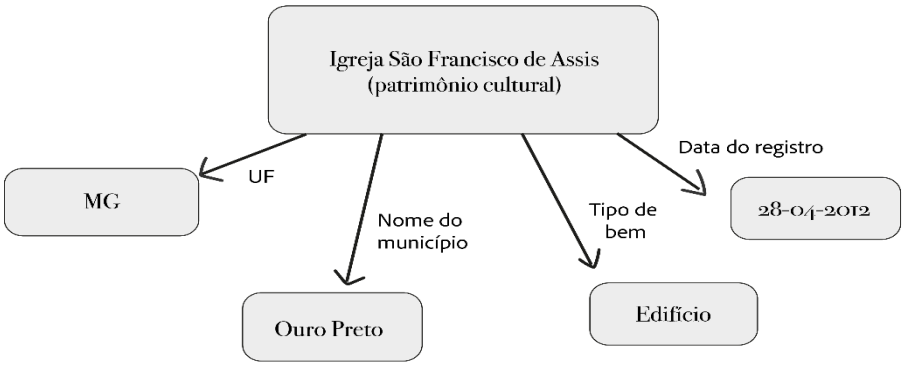
Figura 2 – Elementos do RDF e AIPN



Fonte: Adaptado de Brandt (2020).

Assim, o metadado de negócio pode ser escrito em formato RDF, compatível com a Web Semântica. Para isso, propõe-se, como primeira etapa, 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



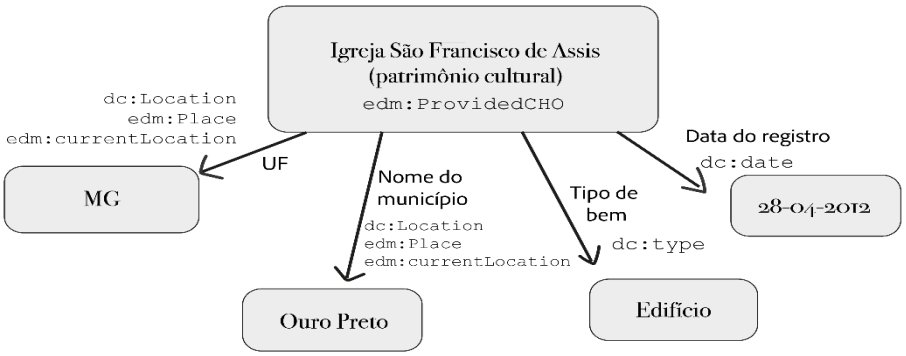
Fonte: Elaborado pelas autoras (2023).

A partir da identificação desses relacionamentos, os metadados de negócio (UF, Nome do município, Tipo de bem e Data de registro) podem ser descritos como propriedades, utilizando vocabulários da Web Semântica. 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 Web Semântica. Recorreu-se ao *Linked Open Vocabularies*[2] (LOV), repositório de vocabulários abertos e conectados que reúne definições de classes e propriedades (termos do vocabulário) já expressos em *linked open data*, no formato RDF. "Em suma, os vocabulários fornecem a cola semântica que permite que os dados se tornem dados significativos [...]" (Lov, 2023). A partir de buscas por termos relacionados ao conjunto de dados da presente pesquisa, foram encontrados alguns vocabulários que poderiam ser utilizados. 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**. Assim, optou-se por utilizar o vocabulário *Europeana Data Model* (EDM), do domínio de bens culturais, ao qual pertencem os dados desse exemplo. Além disso, termos do Dublin Core utilizados pelo próprio EDM foram adicionados, visando aumentar a integração à Web Semântica.

2 https://lov.linkeddata.es/dataset/lov/

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 poderem ser descobertas mais coisas". Segundo Heath e Bizer (2011, online, tradução nossa)[3]

> [...] *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, se podem possuir URI. Feito isso, recomenda-se verificar se há outros conjuntos de dados do mesmo âmbito (instituição, órgão) que representem tais recursos. Caso existam, podem ser utilizados na modelagem do valor da tripla RDF. Outra forma de encontrar recursos que possam representar valores de conjuntos de dados é realizando busca em repositórios de *Linked Open Data*: 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:

3 Original: "[...] external RDF links are fundamental for the Web of Data as they are the glue that connects data islands into a global, interconnected data space and as they enable applications to discover additional data sources in a follow-your-nose fashion [...]" (Heath; Bizer, 2011, *online*).

> 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 (Santarem Segundo; Simionato, 2016, p. 124).

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

ANÁLISE E DISCUSSÃO DOS RESULTADOS

O **Quadro 2** resume os elementos mapeados neste exemplo, na nova etapa incluída na AIPN.

Quadro 2 – 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	MG	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

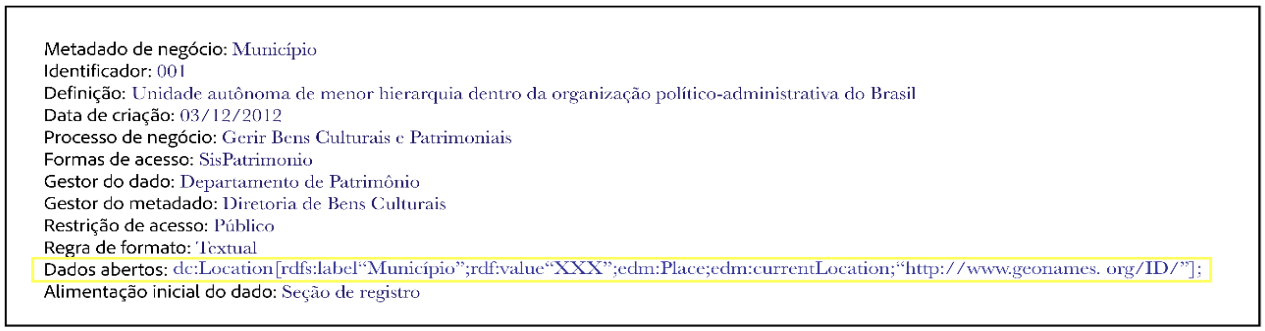
Fonte: Dados da pesquisa (2023).

Devem-se utilizar ainda as estruturas rdfs:label e rdf:value, do RDF *Schema*, 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, conforme a figura 5 a seguir:

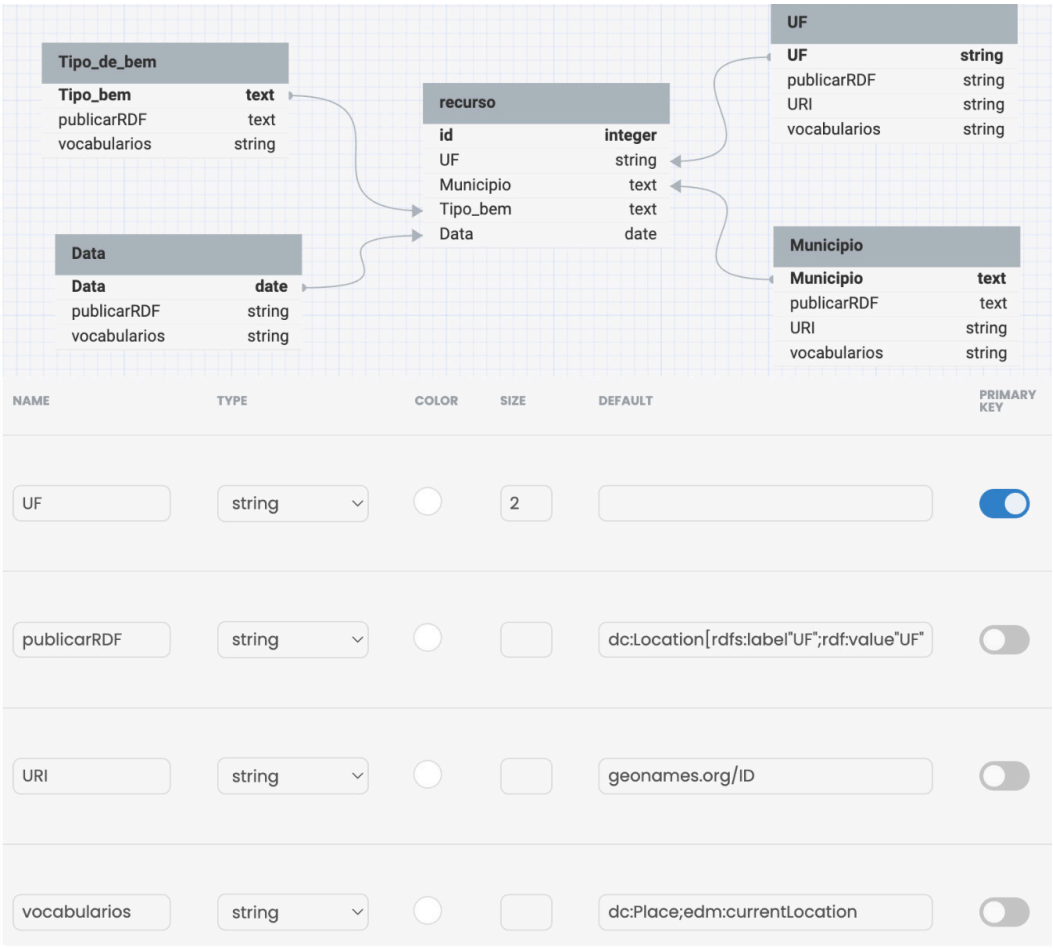
FIGURA 5 – Catalogação do metadado de negócio com detalhamento do atributo "Dados abertos"



Fonte: Elaborado pelas autoras (2023).

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. A modelagem do banco de dados poderia ser realizada conforme a sugestão ilustrada na figura 6:

FIGURA 6 – Sugestão de modelagem de banco de dados



Fonte: Elaborado pelas autoras no aplicativo Dbdesigner (2023).

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.

CONCLUSÕES

A W3C, instituição fundada pelo criador da Web, Tim Berners-Lee, 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 são 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 de estruturação dos dados contidos na metodologia de Arquitetura da Informação para Processos de Negócio, 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, T. **Linked data**: design issues 2006. [*S. l.*], 2016. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 7 jun. 2023.

BRANDT, M. B. **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, T.; BIZER, C. **Linked Data**: evolving the web into a global data space. EUA: Morgan & Claypool, 2011. Disponível em: http://linkeddatabook.com/editions/1.0/. Acesso em: 14 maio 2023.

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

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (eds.). **Data on the Web Best Practices**. [*S. l.*], 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 16 maio 2023.

LINKED OPEN VOCABULARIES. **About LOV**. [*S. l.*], 2023. Disponível em: https://lov.linkeddata.es/dataset/lov/about. Acesso em: 4 set. 2023.

SANTAREM SEGUNDO, J. E.; SIMIONATO, A. C. Uma abordagem sobre a estrutura do geonames e suas contribuições para o linking open data. **Informação & Tecnologia**, João Pessoa, v. 3, n. 1, p. 117-137, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/40979. Acesso em: 4 set. 2023.

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