

Metaprocesso para transformação de dados educacionais em dados conectados

Bruno Elias Penteado¹, Ig Ibert Bittencourt², Seiji Isotani¹

¹Instituto de Ciências Matemática e Computação – Universidade de São Paulo (USP)
13566-590 – São Carlos – SP – Brasil

²Instituto de Computação – Universidade Federal de Alagoas (UFAL) – 57072-970 –
Maceió – AL - Brasil

brunopenteado@usp.br, ig.ibert@ic.ufal.br, sisotani@icmc.usp.br

Abstract. *Open government data has been published in the last few years, aiming for its economic exploitation and social control. However, given the administrative autonomy of public bodies, the data show discrepancies of formats, identifiers and descriptions, even for same entities. Linked open data seek to fill this gap by creating links for different contexts. This paper proposes a metaprocess for publishing linked data as well as a framework to assist the choices of appropriate tools for this task. The proposal is evaluated with two examples from the literature, with relevant contexts for the educational domain.*

Resumo. *Dados abertos governamentais têm sido publicados nos últimos anos, buscando sua exploração econômica e de controle social. No entanto, com a autonomia de cada órgão produtor de dados, ocorre a discrepância entre formatos, identificadores e descrições, mesmo que para as mesmas entidades. Os dados abertos conectados buscam explorar essa lacuna, ao criar conexões entre diferentes contextos. Este artigo traz uma proposta de metaprocesso para a publicação de dados conectados bem como um framework para a escolha de ferramental apropriado para a tarefa. A proposta é avaliada com dois exemplos da literatura em cenários relevantes para o domínio educacional.*

1. Introdução

Governos de todo o mundo, principalmente na última década, têm concentrado esforços na disponibilização de dados coletados por eles próprios para com a sociedade, seguindo a filosofia dos dados abertos - ou seja, acessíveis para qualquer cidadão, sem restrições de uso. Esse movimento tem por motivação fatores como: aumento da transparência e da responsabilização democrática; apoiar o crescimento econômico, ao estimular a criação de novos serviços e produtos baseados em dados; e aprimorar como os serviços públicos são disponibilizados [Janssen et al., 2012]. No entanto, diversos fatores contribuem para que essa visão ainda não tenha sido alcançada. Do ponto de vista tecnológico, podemos citar os diferentes formatos de dados, a ausência de dados legíveis por máquinas, a ausência de metadados, dentre outros [Janssen et al., 2012; Neumaier et al., 2016].

A abordagem de dados conectados (*linked data*) viabiliza tratar problemas desse tipo. No contexto de administração pública, a intenção é termos esses dados também no formato aberto, tendo assim os dados abertos conectados (*linked open data, LOD*) [Isotani e Bittencourt, 2015] e que se trata de um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na Web, formando assim uma “Web de Dados”, reaproveitando a infraestrutura da Web e também tecnologias semânticas já consolidadas [Bizer, et al., 2009]. Berners-Lee (2009) propôs um esquema com 5 níveis, para denotar o grau de abertura dos dados (Figura 1):

- 1 estrela: dados de qualquer formato, com licença de uso definida (ex.: PDF, imagens, HTML);
- 2 estrelas: dados estruturados legíveis por máquina (ex.: planilhas Excel);
- 3 estrelas: como o anterior, mas em formato não-proprietário (ex.: XML, CSV...);
- 4 estrelas: como o anterior, mas usando padrões da Web para representá-los (ex.: URIs, RDF, SPARQL);
- 5 estrelas: como o anterior, mas que esteja ligado a fontes externas, fornecendo maior contexto aos dados (LOD).



Figura 1. Esquema 5-estrelas proposto por Berners-Lee (2009) para descrever grau de abertura dos dados abertos na Web.

No domínio educacional, Pereira et al. (2018) apontam os seguintes objetivos da publicação de dados conectados: disponibilizar dados educacionais como dados conectados; integração e interoperabilidade de dados e sistemas educacionais e consumo de dados conectados para diferentes propósitos educacionais. No Brasil, a publicação de dados abertos educacionais pelo governo é uma das pioneiras no mundo e vem produzindo novos conjuntos de dados ao longo dos últimos anos, principalmente pelos órgãos federais do MEC e do INEP [Penteado e Isotani, 2017], seja de modo centralizado no Portal Brasileiro de Dados Abertos (dados.gov.br) ou nos respectivos websites dos órgãos. Um grande esforço tem sido feito para o atendimento de padrões mínimos de qualidade, tanto em nível de maturação dos dados, metadados e de formato de arquivos, conforme estabelecido pelo e-PING¹ (Padrões de Interoperabilidade de Governo Eletrônico) e pela INDA² (Infraestrutura Nacional de Dados Abertos). No entanto,

¹ <http://eping.governoeletronico.gov.br>

² <https://www.governodigital.gov.br/transformacao/cidadania/dados-abertos/inda-infraestrutura-nacional-de-dados-abertos>

embora haja grande disponibilidade de dados abertos, não se encontram disponíveis dados abertos conectados publicados oficialmente, seja qual for o canal. Penteado et al. (2019) mapeiam quais práticas foram adotadas em relação aos dados abertos educacionais no Brasil, incluindo as relacionadas a dados abertos conectados e observaram que a publicação desses dados não atende aos requisitos de legibilidade por máquina.

Este trabalho tem por objetivo propor um framework para a produção de dados educacionais abertos e conectados. Para isso, propomos uma metodologia, que descreve os passos necessários para obter dados em nível 5 e também um framework de decisão sobre quais ferramentas usar para essa finalidade. O framework auxilia diretamente produtores de dados abertos educacionais a liberarem dados em formato conectado e, indiretamente, os consumidores desses dados, que podem explorá-lo semanticamente, integrando a outros sistemas e permitindo inferências por agentes de software, possibilitando extrair mais vantagens dos dados disponíveis.

2. Trabalhos relacionados

Bandeira et al. (2014) trouxeram os potenciais benefícios e limitações atuais do uso de dados conectados na educação, dentre eles o custo de publicação, o reuso de vocabulários e a qualidade dos dados. Penteado (2016) apresentou um método que combina dados de desempenho no IDEB, em nível municipal, e cruza com dados de outras fontes, tanto do próprio governo brasileiro quanto da Dbpedia e apresenta uma ontologia para modelar esses dados, no entanto sem focar no método necessário para sua produção. Pereira et al. (2018) conduziram uma revisão sistemática do uso de dados conectados para educação, mapeando ontologias, vocabulários e ferramentas neste contexto, apontando como um dos desafios a grande diversidade de vocabulários e de ferramentas adotadas.

Diversas metodologias para a produção de dados abertos conectados governamentais foram propostas e algumas delas basearam a criação da recomendação de boas práticas da W3C (2014), como Hyland & Wood (2011) e Villazón-Terrazas et al. (2011). No entanto, existem muitas críticas quanto a elas [Laessig et al., 2019; Jovanovik, M., Trajanov, 2017], por exemplo, por serem muito genéricas e por não apresentarem um conjunto de ferramentas que possa servir para auxiliar a publicação dos dados e, como resultado, a maioria dos estudos não especifica o uso de metodologias sistemáticas para a publicação dos dados [dos Santos, 2018]. Neste sentido, este trabalho estende os anteriores e busca trazer uma metodologia integrada para a publicação de dados abertos governamentais no domínio educacional.

Além disso, houve trabalhos que buscaram mapear e desenvolver ferramentas para o ciclo de vida dos dados abertos conectados como o LOD Project³ e o OpenGovIntelligence⁴. Barbosa et al. (2017) trazem um mapeamento sistemático de ferramentas para publicação e consumo de dados conectados, independentemente do

³ <http://lod2.eu/>

⁴ <http://www.opengovintelligence.eu/>

domínio de conhecimento. Muitas ferramentas são levantadas e aplicadas em diferentes cenários; no entanto, sem estabelecer um racional de como selecioná-las. Neste sentido, este trabalho busca fornecer um conjunto de práticas que auxiliem o responsável por publicar dados abertos conectados a escolher melhor seu ferramental.

3. Metodologia

Adotamos neste trabalho a metodologia de pesquisa qualitativa de *design science research* (DSR) [Hevner, 2010]. Ela foi selecionada por oferecer *framework* metodológico para a criação e avaliação de artefatos tecnológicos. Trata-se de um paradigma de soluções de problemas, prescritivo em vez de explicativo, nas quais os limites das capacidades organizacionais são estendidas junto com o conhecimento do domínio do problema por meio da construção e aplicação de artefatos projetados. A DSR é composta por dois eixos [March & Smith, 1995]: as *atividades* de pesquisa e os *produtos* da pesquisa. As atividades de pesquisa compreendem os passos necessários para definir o problema (definido na introdução), projetar uma solução (seção 4), avaliá-la (seção 5) e aumentar a base de conhecimento sobre o problema baseado nos resultados (seção 6). Como produto da pesquisa, o artefato pode assumir alguma das seguintes formas: construtos, modelos, métodos e implementações de sistemas.

Neste trabalho foram adotados como tipo de artefato o *método*, que consiste em algoritmos ou práticas para realizar determinada tarefa, e, como método de avaliação, foi adotada a técnica de *cenários descritivos*, na qual cenários detalhados são construídos aplicando o artefato de modo a se demonstrar sua utilidade. Como fundamento para validar o grau de abertura dos dados conectados usamos o esquema de 5-estrelas proposto por Berners-Lee (2009), já que se trata de um dos principais trabalhos que definem o que é dado conectado na Web. Usamos trabalhos já desenvolvidos, de modo a demonstrar a aplicabilidade da proposta para a análise de pesquisas já feitas em diferentes contextos e com os dois cenários mais comuns de aplicações em contextos educacionais (Pereira et al., 2018): disponibilidade de dados educacionais conectados, para análises diversas e integração de sistemas a partir de diferentes fontes. Assim, adotamos a técnica de casos diversos, com casos representativos de acordo com as principais categorias levantadas.

4. Proposta

4.1 Metaprocesso

Baseado nas metodologias presentes na literatura, apresentamos o modelo de processo ilustrado na Figura 1, composto por 5 fases: especificação, modelagem, conversão, publicação e exploração. Na primeira fase, de especificação, deve-se levantar quais os dados serão publicados em formato conectado, os metadados obrigatórios e opcionais que ele deve conter (dependendo do contexto do domínio), os vocabulários (ou ontologias, taxonomias) que deverão descrever esses dados e o padrão de URI a ser adotado pela organização.

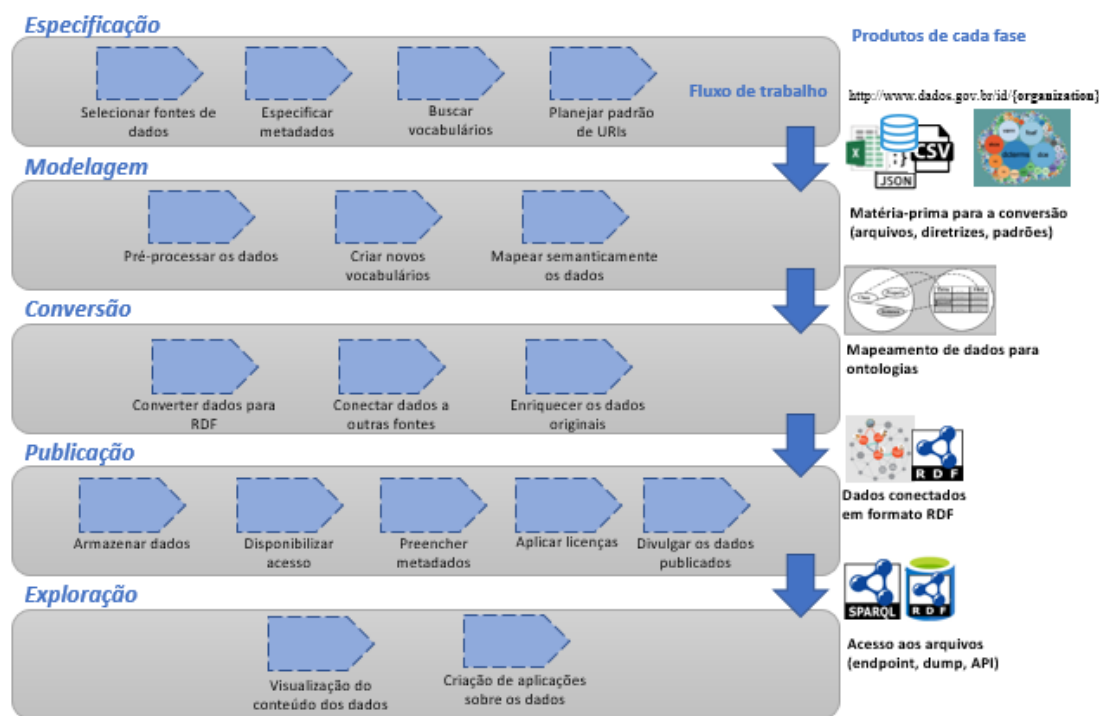


Figura 1. Meta-processo para publicação de dados abertos conectados.

Na segunda fase, dá-se a modelagem dos dados. Primeiro, é necessário pré-processá-los, de modo a filtrar dados de baixa qualidade (faltando, mal formatado, etc.). Em seguida, são levantadas as necessidades de anotações semânticas que ainda não foram atendidas pelos vocabulários existentes e são criados novos vocabulários. Assim, o próximo passo é anotar os dados originais com sua respectiva anotação semântica. Com todos os mapeamentos feitos (produto desta fase), se inicia a fase de conversão.

Na fase de conversão, os dados originais, já mapeados, são transformados para o modelo de dados semântico (geralmente, no formato RDF). Além disso, nesta fase é possível fazer a ligação com outras bases de dados externas, apontando para URIs desta outra base, usando o mesmo modelo de dados. Uma vez feita a ligação, é possível enriquecer os dados originais, ou seja, importar informações das bases externas, tornando-os mais contextualizados. A saída desta fase é um grafo RDF, com todos os relacionamentos mapeados, conectados externamente e enriquecidos.

Em posse desse grafo se inicia a fase de publicação. Nela o conteúdo do grafo gerado é armazenado para uso, seja por meio de upload em um website de catálogo de dados (ex.: CKAN), um arquivo *dump* ou de um endpoint SPARQL. Além disso, deve-se divulgar o ponto de acesso a essa fonte de dados, seja a URL do arquivo ou do *endpoint* SPARQL. Por fim, esse grafo deve ser versionado, preenchidos seus metadados e a ele deve ser atribuída uma licença de uso. A divulgação em mecanismos de busca ou catálogos de dados (*data hubs*) também é um passo importante para que os dados sejam reutilizados. O produto desta fase é o acesso a um ou mais desses meios.

Por fim, é possibilitada a fase de exploração, em que aplicações são desenvolvidas ou aplicadas aos dados recém criados. Trata-se de uma fase importante também para o publicador, uma vez que por meio dela será feita a validação da qualidade do processo como um todo. Outro ponto importante recomendado é o uso de aplicações existentes que possam ajudar o usuário consumidor a explorar e compreender o que foi publicado.

4.2 Framework de seleção de ferramentas

Um passo complementar ao processo de publicação de dados abertos conectados é o de seleção de ferramentas que dêem suporte às tarefas propostas, de forma mais concreta. Este trabalho busca especificar um framework de tomada de decisão para determinar a escolha de ferramentas apropriadas para a consecução das tarefas propostas na metodologia. Para isso, foram mapeados pontos de design que influenciam a escolha das mesmas. A Tabela 1 apresenta os pontos identificados e as respectivas ferramentas.

Tabela 1. Mapeamento de ferramentas e diretrizes de acordo com os requisitos do projeto de abertura e conexão de dados.

Fase	Fator	Níveis	Ferramentas / diretrizes ⁵
Especificação	URI	Modelagem	Cool URI
	Metadados	Obrigatórios e opcionais	ePING (para governo brasileiro)
	Vocabulário	Obrigatórios e opcionais	ePING (governo brasileiro), LOV, Swoogle
Conversão	Dinamicidade da fonte de dados	Arquivos estáticos	OpenRefine
		Bases de dados	D2RQ, Triplify
Modelagem	Especificação	Vocabulários existentes	Busca: LOV, Swoogle; Anotação: OpenRefine
		Vocabulários novos - Colaborativo	OntoWiki, Protégé Web
		Vocabulários novos - Individual	Protégé Desktop, TopBraid Composer
	Critério de conexão	Mesmo identificador	OpenRefine
		Identificador - Heurísticas	SILK, LIMES
	Formato da conexão	(Semi) Estruturado	OpenRefine
		Não estruturado	Dbpedia Spotlight, OpenCalais
Geográfico		geometry2rdf, ESRI2Open	
Publicação	Infraestrutura	Acesso por download	CKAN
		Acesso por SPARQL	Virtuoso, Fuseki, Dydra
		Acesso por API	Elda, Puelia (Linked Data API)

⁵ Referências disponíveis em: <https://github.com/brunopenteado/sbie2019/blob/master/ferramentas.txt>

	Licenças de uso	Recomendações das diretrizes de e-gov ou Creative Commons	ePING, Creative Commons Choose, Open Data Commons
Exploração	Propósito	Visualização de propriedades	Pubby, Disco
		Visualização do conteúdo	CubeViz, SemMap
		Conectores a outros ambientes	RapidMiner, Sparql R
		Novas aplicações	-

A Tabela 1 não traz uma lista extensiva de todos os pontos ou mesmo todas as ferramentas que abordam cada ponto. Além disso, algumas ferramentas podem até mesmo atender outros caminhos que não o indicado. No entanto, o mapeamento é construído baseado no uso mais comum das ferramentas na literatura para a publicação de dados conectados e na experiência dos autores.

5. Avaliação

Como primeiro exemplo de avaliação da proposta, consideramos o estudo de Penteadó (2016) com dados abertos conectados da educação brasileira. Trata-se de uma aplicação para que pesquisadores possam cruzar dados de fontes e formatos distintos para responder a uma questão relevante. O autor utilizou fontes de dados do INEP (dados do IDEB, em nível municipal), do IPEA (índice de vulnerabilidade social, coeficiente Gini e IDH municipal) e da Dbpedia (dados contextuais dos municípios), para possibilitar uma análise correlacional em ferramenta externa (software R). Como se tratam de medições complexas, realizadas em intervalos longos de tempo, os arquivos são estáticos e em diferentes formatos. Foi desenvolvida uma ontologia para este contexto, com reaproveitamento de campos de outras ontologias. Os dados originais foram mapeados e convertidos para RDF usando a ferramenta OpenRefine e publicados em um servidor SPARQL gratuito (Dydra). A conexão entre as fontes foi híbrida: usou tanto identificadores comuns (identificadores de municípios, criado pelo IBGE) quanto o mapeamento manual com a Dbpedia. A Figura 2 traz o instanciamento do meta-processo para este contexto, com os passos atendidos e não atendidos e os fatores levados em consideração na escolha das ferramentas.

Outro exemplo no escopo educacional é a aplicação apresentada por Juanes et al. (2014). Ela envolve a finalidade de integração de sistemas acadêmicos - internos e externos à universidade (com identificadores híbridos) beneficiando administradores de sistemas dentro e fora da universidade trocarem dados de modo uniforme e estender o conhecimento a partir dessa relação. Assim, um de seus requisitos foi o de ter dados *online*, ou seja, que os dados conectados reflitam em tempo real as mudanças na base de dados, fazendo com que as consultas sejam retornadas como se fossem grafos RDF. Desta maneira, as bases foram modeladas usando toda a pilha D2RQ, com a D2RQ Engine modelando os dados e o D2R Server servindo-os como um *endpoint* SPARQL. Por fim, o servidor de interfaces Pubby foi usado para ajudar na navegação pelos dados disponíveis, de modo a possibilitar vislumbrar novas conexões.

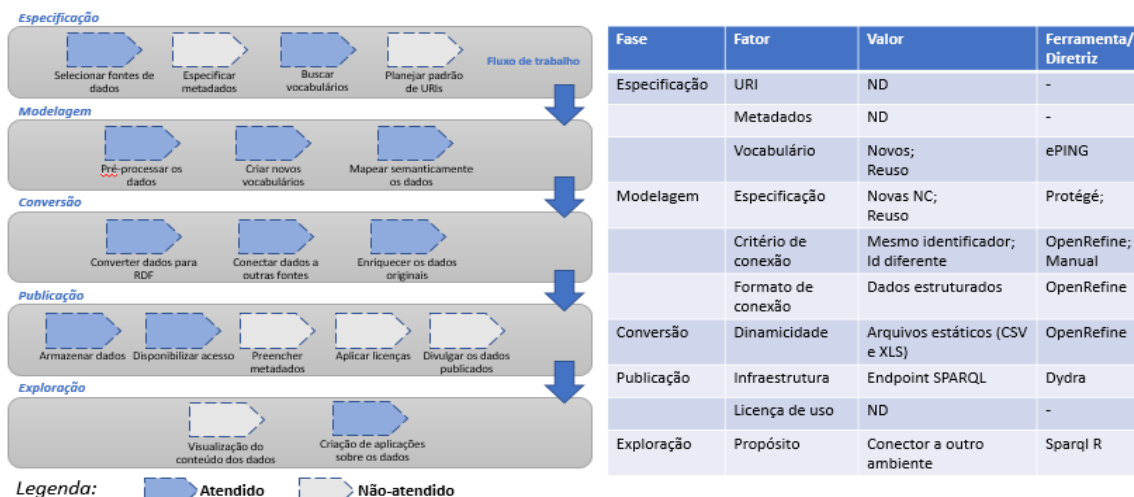


Figura 2. Instanciação do metaprocesso e do framework de seleção de ferramentas para Penteadó (2016).

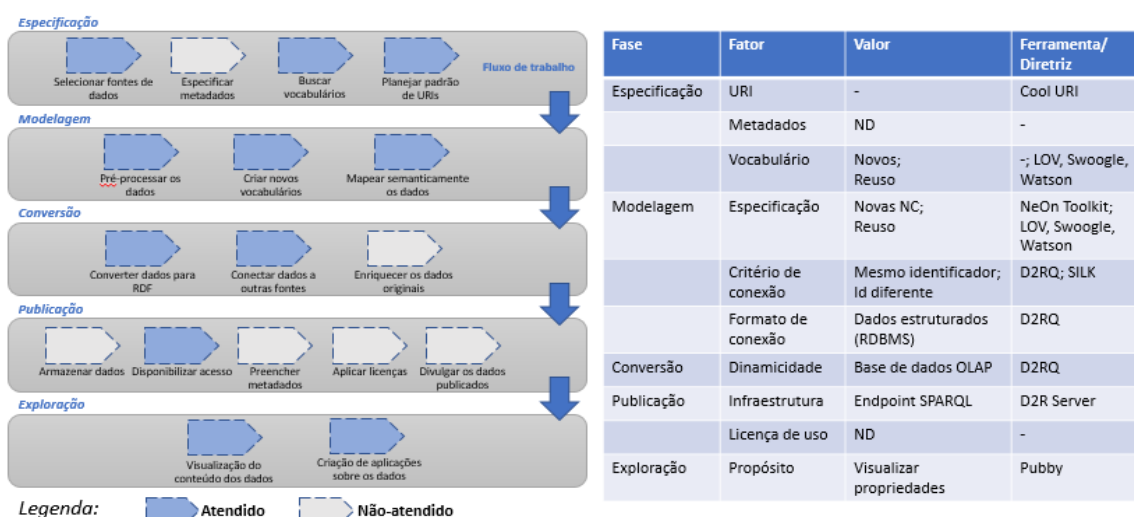


Figura 3. Instanciação do meta-processo e do framework de seleção de ferramentas para Juanes (2014).

6. Discussão

Neste trabalho, o metaprocesso proposto traz uma compilação das principais atividades relacionadas à publicação de dados conectados, dentro das diferentes metodologias e processos *ad-hoc* disponíveis na literatura. Seu atendimento completo faz com que os dados publicados tragam características que melhorem sua qualidade sob diferentes aspectos (descoberta, completude, disponibilidade, etc.), o que, por sua vez, contribuiria para que os casos selecionados pudessem atender melhor a esses requisitos. Deste modo, o metaprocesso pode ser usado como uma matriz que estabelece os passos possíveis para atender esse alto grau de qualidade, o que é desejável principalmente no escopo de dados governamentais, com alto potencial de reuso pela comunidade. No entanto, fatores contextuais como falta de recursos (financeiros, humanos ou de tempo) ou nível baixo de

necessidade de formalismo podem fazer com que o metaprocessos seja instanciado de forma mais simples, com menos passos, conforme a disponibilidade desses fatores e, ainda assim, atender ao seu objetivo original. Embora os casos selecionados sejam representativos das principais aplicações educacionais de dados conectados, outras validações são necessárias para garantir sua generalização a outras aplicações.

7. Conclusão

Neste artigo trouxemos contribuições para que a comunidade de pesquisa, membros técnicos do governo e praticantes na área de informática na educação possam publicar dados abertos em formato conectado. Por meio dos cenários ilustrativos, representativos das principais aplicações de dados conectados educacionais, buscamos validar nossa proposta e tornar mais concreto o processo, para que esses atores possam avaliar seus contextos e guiar seu processo de decisão. Como diferentes contextos exigem diferentes níveis de formalização do processo, este modelo serve como *template* para a geração de novos processos que possam ser adequados a diferentes contextos de publicação de dados. Além disso, ao fornecer um mapeamento de ferramentas de acordo com fatores que modulam o contexto dos dados, otimiza-se o espaço de buscas e experimentações necessárias para atingir os objetivos da publicação dos dados conectados. Questões como ética, privacidade, gestão de dados, propriedade intelectual e posse dos dados também devem ser contempladas em trabalhos futuros e levar esses aspectos em consideração. Por exemplo, houve recentemente a promulgação da Lei Geral de Proteção de Dados [Brasil, 2018], que versa sobre a proteção de dados pessoais de pessoas, tanto para entes públicos quanto privados. Na educação, isso pode implicar tanto na coleta e divulgação desses dados quanto para o processamento por parte de entes privados para fins econômicos e, por isso, consideradas fundamentais dentro do processo.

Agradecimentos

Os autores agradecem à CAPES - Código de Financiamento 001, CNPq e FAPESP (Processo 2015/24507-2) pelo apoio financeiro dado ao projeto.

Referências

- Bandeira, J., Alcantara, W., Barbosa, A., Ávila, T., Bittencourt, I., Isotani, S. (2014). Dados Abertos Conectados na Educação Brasileira. Jornada de Atualização em Tecnologia da Informação. Simpósio Brasileiro de Tecnologia da Informação. 2014
- Barbosa, A., Bittencourt, I. I., Siqueira, S. W. M., Silva, R. D. A., Calado, I. (2017). The Use of Software Tools in Linked Data Publication and Consumption. Intl. Journal on Semantic Web and Information Systems, v. 13, n. 4, p. 68–88.
- Berners-Lee, T. (2009). Linked Data. Acesso em: 06/06/2019. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>.
- Brasil, (2018). Lei nº 13.709, 14 de agosto de 2018. Acesso em: 09/06/2019. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm

- Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, v. 5, n. 3, p. 1–22.
- Dos Santos, H. D. A., Oliveira, M. I. S., Lima, G. F. A. B., Da Silva, K. M., Muniz, R. I. V. C. S., Lósico, B. F. (2018). Investigations into data published and consumed on the Web: a systematic mapping study. *Brazilian Computer Society* vol. 24 (14).
- Hevner, A., Chatterjee, S. (2010). Design Science Research in Information Systems. *Design Research in Information Systems*, p. 9–22.
- Hyland, B., Wood, D. (2011). The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web. *Linking Government Data*, p. 3–26.
- Isotani, S., Bittencourt, I. I. (2015). Dados abertos conectados. Editora Novatec.
- Janssen, M., Charalabidis, Y., Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, v. 29, n. 4, p. 258–268.
- Jovanovik, M., Trajanov, D. (2017). Consolidating Drug Data on a Global Scale Using Linked Data. *Journal of Biomedical Semantics*, 8, 3.
- Laessig, M., Jacob, B., AbouZahr, C. (2019). Opening data for global health. *The Palgrave Handbook of Global Health Data Methods for Policy and Practice*.
- Juanes, G. G., Barrios, A. R. (2014). Linked Data Strategy to Achieve Interoperability in Higher Education. *Int Conf on Web Information Systems and Technologies*, p. 57-64.
- March, S. T., Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), p. 251–266.
- Neumaier, S., Umbrich, J., Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *Data and Information Quality*, v. 8 (1), p. 1–29.
- Penteado, B. E., Bittencourt, I. I., Isotani, S. (2019). Análise exploratória sobre a abertura de dados educacionais no Brasil: como torná-los prontos para o ecossistema da Web? *Revista Brasileira de Informática na Educação*, v. 27, n. 01, p. 175.
- Penteado, B. E., Isotani, S. (2017). Dados abertos educacionais: que informações temos disponíveis? VI Congresso Brasileiro de Educação, vol. 4, p. 1933-1938.
- Penteado, B. E. (2016). Correlational Analysis Between School Performance and Municipal Indicators in Brazil Supported by Linked Open Data. *Intl. Conf. on World Wide Web*.
- Pereira, C. K., Siqueira, S. W. M., Nunes, B. P., Dietze, S. (2018). Linked Data in Education: A Survey and a Synthesis of Actual Research and Future Challenges. *IEEE Transactions on Learning Technologies*, vol. 11 (3), p. 400–412.
- Vega-Gorgojo, G., Asensio-Pérez, J. I., Gómez-Sánchez, E., Bote-Lorenzo, M. L., Muñoz-Cristóbal, J. A., Ruiz-Calleja, A. (2015). A Review of Linked Data Proposals in the Learning Domain. *Universal Computer Science*, vol. 21 (2), p. 326-364.
- Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O. and Gómez-Pérez, A. (2011). Methodological Guidelines for Publishing Government Linked Data. *Linking Government Data*, p. 27–49.
- W3C (2014). Linked Data Best Practices. <https://www.w3.org/TR/ld-bp/>.