Framework para produção de dados educacionais conectados

Bruno Elias Penteado, Seiji Isotani
Instituto de Ciências Matemáticas e da Computação,
Universidade de São Paulo, São Carlos/SP
4º ano de doutorado
brunopenteado@usp.br, sisotani@icmc.usp.br

Resumo. *Dados abertos governamentais têm sido publicados nos últimos anos, buscando sua exploração econômica e de controle social. No entanto, diversos fatores fazem com que esses dados não sejam explorados em todo seu potencial. Esta tese traz uma proposta de framework para a publicação de dados conectados, seguindo a abordagem metodológica do Design Science Research. A proposta trouxe resultados parciais na exploração dos dados abertos educacionais no Brasil e na geração dos artefatos que viabilizam esse processo. Ainda resta a avaliação dos artefatos gerados em casos de estudo reais para refinar os artefatos e apontar caminhos futuros.*

1. Motivação

Governos de todo o mundo têm concentrado esforços na disponibilização de dados coletados para com a sociedade, seguindo a filosofia dos dados abertos - acessíveis para qualquer cidadão, sem restrições de uso, motivado por fatores como: aumento da transparência e da responsabilização democrática; apoio ao crescimento econômico, ao estimular a criação de novos serviços e produtos baseados em dados; e aprimoramento dos serviços públicos [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*) visa tratar problemas desse tipo e disponibiliza-los também no formato aberto, tendo assim os *dados abertos conectados* (*linked open data*, *LOD*) [Isotani e Bittencourt, 2015], 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: i) dados de qualquer formato, com licença de uso (ex.: PDF, HTML); ii) dados estruturados legíveis por máquina (ex.: planilhas Excel); iii) como o anterior, mas em formato não-proprietário (ex.: XML, CSV); iv) usando padrões da Web (ex.: URIs, RDF, SPARQL); v) ligado a fontes externas, fornecendo maior contexto aos dados (LOD).

Diversas metodologias genéricas para a produção de dados abertos conectados governamentais foram propostas, como Hyland & Wood (2011), Villazón-Terrazas et al.

DOI: 10.5753/cbie.wcbie.2019.1422

(2011) e Laessig et al. (2019). 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 ferramental que possa servir para auxiliar a publicação dos dados. Como resultado, a maioria dos estudos nesta área não especifica o uso de metodologias sistemáticas para a publicação dos dados conectados [dos Santos, 2018].

No contexto educacional, mesmo com a grande quantidade de dados disponibilizados (Penteado & Isotani, 2017a), poucos trabalhos têm sido desenvolvidos para dados abertos governamentais - mais voltados para o monitoramento e embasamento de políticas públicas e, em sua maioria, são mais focados na conversão de dados para uma aplicação semântica específica. Nesta tese propomos um *framework* que estende as metodologias anteriores, endereçando os problemas existentes na produção de dados conectados ao embutir atividades de validação de qualidade em cada etapa e que atenda às potencialidades da Web como meio para a troca de dados.

2. Questões de pesquisa

Neste sentido, a pergunta de pesquisa principal deste trabalho é: *como produzir dados abertos educacionais governamentais em formato conectado?*

Foi conduzida uma fase exploratória, em que foram levantados: i) os dados educacionais atualmente disponíveis; ii) as atividades envolvidas para criação de dados conectados de alta qualidade e iii) artefatos (ferramentas, metodologias, ontologias e práticas) usados em cada atividade. Em seguida, é proposto um artefato (modelo de processo), combinando as diferentes metodologias e estendendo-as com atividades que incorporem as práticas sugeridas pela DWBP (W3C, 2017) e atividades de validação, visando auxiliar a produção de dados conectados de alta qualidade. Este modelo de processo pode ser instanciado de acordo com o contexto e recursos necessários.

3. Metodologia

Para um adequado desenvolvimento do artefato, adotamos a metodologia de pesquisa *design science research* (DSR) [Hevner, 2010; Peffers et al. 2008]. Ela foi selecionada por oferecer um *framework* metodológico para a criação e avaliação de artefatos tecnológicos. Trata-se de um paradigma prescritivo de soluções de problemas, em que o conhecimento do domínio do problema é estendido por meio da aplicação de artefatos projetados para sua resolução. O arcabouço metodológico é composto de um processo com 6 fases, aqui abordados da seguinte maneira:

Ponto de entrada: solução centrada no objetivo

A maior parte dos trabalhos recentes da literatura em dados abertos governamentais usa métodos *ad-hoc* para produção de dados abertos conectados, mesmo havendo diversas metodologias propostas, não adequadas para os contextos envolvidos; as metodologias existentes somente definem em alto nível os passos e não definem requisitos específicos. Deste modo, existe a necessidade de haver uma metodologia integradora para este fim.

Identificação do problema e motivação

A produção de dados abertos vem crescendo em todo o mundo na última década. No entanto, existe a limitação de que tanto humanos como agentes de software possam reutilizar esses dados, espalhados em diferentes locais da Web e em diferentes tipos de arquivo. Uma maneira de se aumentar a consciência (*awareness*) da disponibilidade desses dados é disponibiliza-los como dados conectados. Embora haja metodologias para essa finalidade, a comunidade cientifica não vem reutilizando-as, por considerarem-nas muito genéricas e que resultam em dados de baixa qualidade.

Objetivo da solução

O objetivo foi de desenvolver um *framework* para a produção de dados abertos conectados, que contemple um metamodelo de processo de produção e um referencial de ferramentas a serem utilizadas para esta finalidade. Um grande desafio é o de contemplar passos abstratos no modelo, ao mesmo tempo que oferecer diretrizes que possam ser aplicadas em diferentes aplicações ou contextos, como o educacional. A aplicação desse metamodelo permite adequar os passos à complexidade exigida para determinado cenário, ao passo que o referencial oferece um guia de ferramenta para conduzi-lo.

Projeto e desenvolvimento

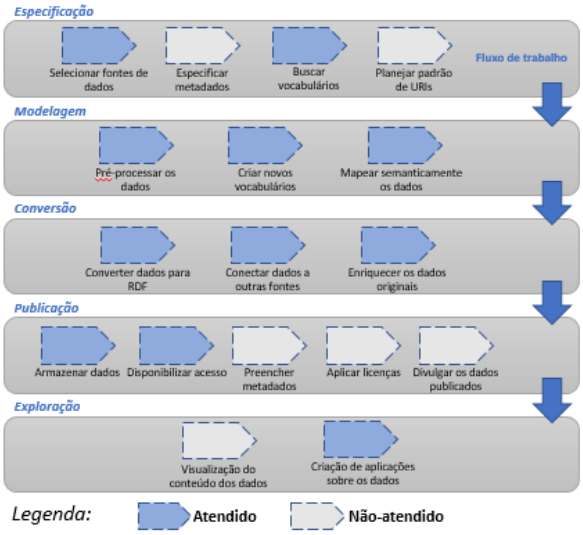
O metamodelo foi derivado a partir de revisão sistemática da literatura em metodologias de produção de dados abertos conectados. O resultado foi a existência de diversas metodologias, com números variados de atividades, mas com alto grau de abstração, sem proposição de ferramentas nem tarefas de validação em cada fase. A proposta compilou todos os passos adotados pelas metodologias e sintetizou em um esquema simplificado, dividido em fases e seus respectivos passos (Figura 1). O referencial foi sintetizado em uma matriz de decisões, pertinentes ao contexto da publicação, em que cada caminho leva a um conjunto diferente de ferramentas. Para o caso específico da educação foi também desenvolvida uma ontologia de domínio que descreve os dados educacionais disponibilizados pelo governo brasileiro.

Demonstração

Após o desenvolvimento destes artefatos, será desenvolvido um portal que contenha dados abertos educacionais de nível federal em formato conectado, de modo a demonstrar a eficácia do metamodelo na produção desse tipo de dados. Com isso, será possível tanto analisar a eficácia dos componentes do framework quanto os possíveis usos desses resultados. Serão utilizadas as ferramentas levantadas anteriormente e apontadas possíveis lacunas para novas ferramentas.

Avaliação

Para a avaliação deste framework, espera-se avaliar sua aderência a uma organização de publicação de dados abertos conectados, de preferência uma organização educacional pública, para que seus atores possam avaliar criticamente sua adesão às necessidades organizacionais, por meio de técnicas de pesquisa-ação, a ser ainda realizada.



Fase	Fator	Valor	Ferramenta/ Diretriz
Especificação	URI	ND	-
	Metadados	ND	-
	Vocabulário	Novos; Reuso	ePING
Modelagem	Especificação	Novas NC; Reuso	Protégé;
	Critério de conexão	Mesmo identificador; Id diferente	OpenRefine; Manual
	Formato de conexão	Dados estruturados	OpenRefine
Conversão	Dinamicidade	Arquivos estáticos (CSV e XLS)	OpenRefine
Publicação	Infraestrutura	Endpoint SPARQL	Dydra
	Licença de uso	ND	-
Exploração	Propósito	Conector a outro ambiente	Sparql R

Figura 1. Instanciação do meta-processo e do framework de seleção de ferramentas.

Comunicação

As contribuições deste trabalho têm sido disseminadas em publicações científicas de diferentes comunidades: educação (Penteado & Isotani, 2017a), informática na educação (Penteado & Isotani, 2017b; Penteado et al., 2019), internet (Penteado, 2016) e sistemas de informação (Penteado & Isotani., 2019), dentre outras em escrita.

Contribuição

Os artefatos de pesquisa resultantes deste trabalho incluem o metamodelo para a produção de dados abertos conectados, que pode ser contextualizado para diferentes cenários e complexidades organizacionais e o referencial de ferramentas que podem ser usadas para guiar concretamente este processo, além da ontologia para dados abertos em nível macro. Elas serão úteis para a comunidade em geral, não só educacional; no caso da comunidade de informática na educação permitirá a possível exploração desses dados educacionais para o desenvolvimento de tecnologias e análises sobre os dados enriquecidos pelo processo de conexão dos dados.

4. Discussão

O objetivo desta tese é o de facilitar a produção de dados abertos conectados no domínio macroeducacional. Trata-se de um problema complexo, que envolve diferentes atores, passos e conhecimentos. O processo proposto é genérico, mas será validado com dados educacionais. Esperamos contribuir para a comunidade em geral apresentando um metamodelo, uma ontologia e um referencial de ferramentas que possa ser ajustado conforme as exigências e tecnologias disponíveis para cada situação – já que o modelo de processo não prescreve todas as atividades como necessários, mas sim oferece um guia para a sua instanciação conforme os requisitos e os recursos disponíveis para o publicador dos dados. Além disso, existem poucos trabalhos em ontologias para dados educacionais em nível governamental (macro), dificultando o reuso de vocabulários existentes – um

dos requisitos da Web semântica. Com a formação de uma Web de dados educacionais, novas tecnologias semânticas ou de mineração de dados podem ser desenvolvidas sobre esses dados, permitindo o desenvolvimento de novos produtos ou pesquisas nesta área.

Referências

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., Lóscio, 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. 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.

Neumaier, S., Umbrich, J. and Polleres, A. (2016), Automated quality assessment of metadata across open data portals. Data and Information Quality, Vol. 8 No. 1, pp. 1-29.

Peffers, K., Tuunanen, T., Rothenberger, M. A., S. Chatterjee, S. (2008). A design science research methodology for information systems research. Management Information Systems, vol. 24 (3), p. 45-77.

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. (2017a). 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. World Wide Web - WWW '16.

Penteado, B. E., Isotani, S., Bittencourt, I. I. (2017b). Dados abertos educacionais no Brasil e sua preparação para os dados abertos na Web. Brazilian on Informatics in Education. Recife, Brazil.

Penteado, B. E., Isotani, S. (2019). The Brazilian open data scenario: a sociotechnical approach (revisão).

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. (2017). Data on the Web Best Practices (DWBP). Disponível em: https://www.w3.org/TR/dwbp/