

UM ESTUDO DOS CICLOS DE VIDA DE DADOS ABERTOS CONECTADOS

A STUDY OF LINKED OPEN DATA LIFE CYCLES

Lidiane Visintin⁽¹⁾, Murilo Silveira Gomes⁽¹⁾, José Leomar Todesco⁽²⁾

(1) Programa de Pós Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, {lidiane.visintin, lilo.flp}@gmail.com.

(2) Programa de Pós Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, titetodesco@gmail.com.

Resumo: Houve um aumento significativo do volume de dados e informações produzidas e disponibilizadas na web, nos últimos anos. Como consequência alguns conceitos têm emergido, sendo um destes: Dados Abertos Conectados. Este conceito tem ganhado ênfase em pesquisas, devido aos benefícios que os dados podem oferecer, tanto para indivíduos que publicam, quanto para aqueles que consomem estes dados. Este artigo objetiva-se em apresentar e discutir dois trabalhos que abordam ciclos de vida de publicação de dados abertos conectados, assim como fornecer uma análise preliminar do que a literatura aborda sobre as práticas e tratamentos de dados, com o intuito de encontrar características nestas pesquisas, assim como identificar possibilidades para pesquisas futuras. Com esta análise foi possível discutir e concluir que os ciclos de vida são descritos como um processo sequencial e unidimensional de fases, que um grupo muitas vezes não especificado realiza para fornecer dados a um público geral, sem uma preocupação em relação as fases de pós publicação, sendo que, ainda foi possível identificar que os ciclos de vida de dados abertos conectados analisados são complementares.

Palavras-chave: Dados; Ciclo de Vida; Dados Conectados Abertos

Abstract: There has been a significant increase in the volume of data and information produced and made available on the web in recent years. As a consequence, some concepts have emerged, being one of these is: Linked Open Data. This concept has gained an emphasis on research because of the benefits that data can offer both to individuals who publish and who consume this data. This article aims at analyzing and discussing two papers that address linked open data life cycles, in order to find characteristics in these surveys, as well as to identify possibilities for future research. With this analysis, it was possible to discuss and conclude that life cycles are described as a sequential and one-dimensional process of phases, which an often-unspecified group performs to provide data to a general audience without a concern regarding post-publication phases. And it was possible to identify that the linked open data life cycles analyzed are complementary.

Keywords: Data; Life cycle; Linked Open Data

1 Introdução

O volume de dados produzidos nos últimos anos, no mundo teve um aumento significativo. Estima-se que cerca de 90% dos dados presentes na Web foram criados em anos recentes sendo que estes não obtiveram um aumento apenas no volume, mas também no nível de detalhamento, tudo isso devido às novas tecnologias (DATA REVOLUTION GROUP, 2014).

As novas tecnologias são oriundas de movimentos como: IoT (BARNAGHI, 2012), Big Data (HITZLER; JANOWICZ, 2013), Web Semântica (BERNERS-LEE et al, 2001), Governo Aberto (MEIJER; CURTIN; HILLEBRANDT, 2012), entre

outros. Estes movimentos geraram também necessidades, como o maior nível de detalhamento dos dados, este contexto alavancou a Web semântica e com isso surgiram conceitos como: dados conectados (Linked Data) e dados abertos conectados (Linked Open Data), fortalecendo e aprimorando diversas pesquisas.

O movimento de Governo Aberto possibilitou que alguns países adotassem a abertura de dados principalmente no setor público, em busca de transparência, *accountability*, participação social e colaboração (MEIJER; CURTIN; HILLEBRANDT, 2012). Através deste movimento e

influenciado pelo conceito de *Openess*, surgiu o conceito de dados abertos (Open Data) que por sua definição são "dados que qualquer um pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, exigências que visem preservar sua proveniência e sua abertura" (OPEN KNOWLEDGE, 2017b), como consequência houve um crescimento considerável na quantidade de dados que são disponibilizados na Web.

Já se tem diversos *datasets* publicados na Web, em formato aberto (ABELE; MCCRAE, 2017). No entanto, o grande desafio de publicar estes dados está em como publicar, para que indivíduos possam fazer uso de dados confiáveis e facilmente ligar a outros dados. Para isso se faz necessário conhecer quais são as fases ou estágio da publicação de dados como um todo para descrever o desenvolvimento deste fenômeno, assim como para prever os próximos passos a serem realizados quando deseja-se publicar dados.

Observando o contexto evidenciado, o objetivo deste artigo é apresentar e discutir dois ciclos de vida de publicação de dados abertos conectados encontrados na literatura, assim como fornecer uma análise preliminar do que a literatura aborda sobre as práticas e tratamentos de dados.

Na sequência, são apresentadas as seções deste trabalho: na seção 2 aborda-se o contexto de dados abertos; na seção 3 é apresentado o contexto de dados abertos conectados; na seção 4 é apresentada uma breve descrição dos procedimentos metodológicos; na seção 5 apresenta-se uma breve discussão sobre dois dos ciclos de vida de dados abertos conectados, assim como uma análise preliminar do que já foi identificado na literatura até o momento; e por fim, na seção 6 são apresentados os encaminhamentos para trabalhos futuros e conclusões.

2 Dados Abertos

Um grupo de trabalho da Califórnia definiu oito princípios, para a abertura dos dados (DADOS.GOV 2017), são

estes: completos, primários, atuais, acessíveis, processáveis por máquina, acesso não discriminatório, formato não proprietário e licença livre. Sendo que, compreender como os dados são licenciados e como os mesmos podem ser disponibilizados são fatores fundamentais para se publicar dados.

O Open Knowledge (2017a) apresenta três regras para aqueles que desejam abrir seus dados:

- Manter simples: Começar pequeno, simples e rápido, pois não há exigências quanto ao tamanho do conjunto de dados (*dataset*) a ser aberto.

- Envolver-se cedo e com frequência: Envolver-se com os usuários dos dados sejam estes cidadãos, empresas, desenvolvedores, entre outros.

- Abordar medos e mal-entendidos comuns: pois, ao abrir os dados sempre surgirão dúvidas e medos.

Estas três regras viabilizam avaliar, assim como apreender com maior agilidade, com base na experiência. Para isso, ainda são apresentadas quatro etapas chave para a abertura dos dados, sendo que estas podem ser realizadas simultaneamente muitas vezes (OPEN KNOWLEDGE, 2017a):

Etapa 1: Escolha do conjunto de dados ou quais partes deste conjunto de dados pretende-se disponibilizar.

Etapa 2: Aplicar uma licença aberta especificando os direitos de propriedade intelectual existem sobre os dados.

Etapa 3: Disponibilizar os dados brutos e em formato que possibilite o processamento por máquina.

Etapa 4: Torna os dados visíveis, disponibilizando-os na Web.

No entanto, quando busca-se inserir dados na web de modo que máquinas ou pessoas possam explorar estes dados através das ligações entre os mesmos, deve-se seguir os princípios de dados conectados (BERNERS-LEE, 2006). Sendo que, dados conectados podem ser liberados sobre uma licença aberta (dados conectados abertos), fomentando assim a sua livre reutilização.

3 Dados Abertos Conectados

O conceito de dados conectados foi estendido, com o objetivo de construir a Web de Dados (HEATH; BIZER, 2011). A Web de dados é constituída de *datasets* com licença aberta, que fazem uso do formato RDF seguindo as recomendações de dados conectados, dando origem a dados conectados abertos (BIZER; HEATH; BERNERS-LEE, 2009).

Dados Conectados e Dados Conectados Abertos fazem uso de tecnologias presentes na arquitetura da Web Semântica (RDF e URIs), com o intuito de padronizar os dados, possibilitando assim o processamento por máquina, e a conexão dos dados e dos *datasets*. Deste modo, os *datasets* em RDF podem ser associados a outros conjuntos de dados para formar uma grande base de dados conectados, ou seja, materializando assim a Web de Dados (ALSHEHHI et al. 2013). Desde o início do projeto dados abertos conectados nota-se um aumento significativo no número de *datasets* que são disponibilizados, bem como um aumento no número de indivíduos que publicam seus dados (ABELE; MCCRAE, 2017).

A motivação por trás de dados abertos conectados é fornecer dados brutos, estruturados e ligados através da Web que possam ser acessados universalmente e ser facilmente compartilhados (BIZER; HEATH; BERNERS-LEE, 2009). Porém, o que é encontrado na literatura para guiar e auxiliar os indivíduos que desejam abrir seus dados, seja isto para dados abertos, dados conectados ou ainda para dados abertos conectados.

4 Procedimentos Metodológicos

No estudo de Broek et al. (2014) os autores apresentam os ciclos de vida existentes, para a publicação de dados. Desta forma, foi realizada uma busca por estes trabalhos na literatura, sendo que

os trabalhos que foram encontrados são apresentados no Quadro 1, incluindo o trabalho de Broek et al. (2014).

Quadro 1: Trabalhos que abordam ciclos de vida de publicação de dados.

Autor / Ano	Título
Hyland; Wood (2011)	The joy of data-a cookbook for publishing linked government data on the web.
Villazón Terrazas et al. (2011)	Methodological guidelines for publishing government linked data
Hausenblas (2011)	Linked data lifecycles.
Auer et al. (2011)	Introduction to linked data and its lifecycle on the web.
Janssen; Zuiderwijk (2012)	Open data and transformational government.
Scharffe et al. (2012)	Enabling linked-data publication with the datalift platform.
Broek et al. (2014)	Walking the extra byte: A lifecycle model for linked open data.

Fonte: Autores, (2017)



Com base nos artigos apresentados no Quadro 1, dois destes ciclos de vida foram selecionados para serem discutidos por este artigo, sendo estes: o ciclo de vida proposto por Auer et al. (2011) e o ciclo de vida proposto por Broek et al. (2014), devido a apresentarem focos distintos quanto a publicação de dados abertos conectados, sendo que os demais ciclos não abordam claramente o contexto de dados abertos conectados.

5 Resultados e Discussões Parciais

Esta seção analisa e discute dois ciclos de vida, que conceituam práticas e tratamentos para a publicação de dados abertos conectados.

A Quadro 2 apresenta os autores, o objetivo proposto por cada trabalho, bem como seu foco e o diagrama dos ciclos de vida analisados.

Quadro 2 – Ciclos de Vida de publicação de dados abertos conectados

Autor/Ano	Objetivo do trabalho	Foco	Diagrama
Auer et al. (2011)	O objetivo é facilitar a distribuição e instalação de ferramentas e componentes de software que suportem o ciclo de vida de publicação de dados conectados.	Tecnológico	
Broek et al. (2014)	Com base na literatura apresentam as etapas de um ciclo de vida onde busca-se considerar a visão estratégica de publicação de dados conectados.	Nas orientações básicas para a publicação de dados a nível estratégico.	

Fonte: Autores, (2017)

O primeiro ciclo analisado, foi proposto por Auer et al. (2011), o objetivo do projeto que construiu este ciclo é desenvolver e elencar ferramentas presentes na literatura, para apoiar a criação de dados conectados, sendo que o ciclo é composto por oito estágios, sendo estes: Armazenamento, Autoria, Fusão, Enriquecimento, Qualidade, Evolução, Exploração e Extração; Os autores deixam claro que estas etapas não devem ser tratadas separadamente, porém salientam que devem ser investigados métodos que facilitem benefícios mútuos para os estágio desenvolvidos, com o intuito de resolver os desafios que são apresentados.

Já o ciclo de vida proposto por Broek et al. (2014), foi desenvolvido baseado nos ciclos de vida encontrados pelos autores na literatura, sendo que este ciclo foi aperfeiçoado com base nas lições aprendidas em um estudo de caso desenvolvido em uma empresa semi-pública na Holanda. O ciclo de vida

apresenta cinco fases, sendo estas: (1) identificação, (2) preparação, (3) publicação, (4) reuso, (5) avaliação.

Observa-se nos demais trabalhos presentes na literatura, assim como no ciclo de vida proposto por Auer et al. (2011), que os mesmos levam em conta exclusivamente os processos operacionais de publicação de dados abertos conectados (como por exemplo, extração, limpeza, publicação e manutenção de dados), ignorando os processos estratégicos (como tomada de decisão, alinhamento com a alta gerência e fatores legais). Assim, as decisões sobre quais dados serão publicados, quem irá extrair os dados, como os dados são editados, como os dados podem ser acessados, quais licenças estão disponíveis e questões como: privacidade e responsabilidade não são tratadas na maioria dos trabalhos encontrados na literatura. Estes processos estratégicos mais gerais, são abordados apenas pelo ciclo

de vida proposto por Broek et al. (2014), sendo que os autores também definem *stakeholders* que auxiliam no processo de publicação de dados abertos conectados.

O trabalho de Broek et al. (2014), apresenta orientações a nível estratégico. No entanto, observa-se que os autores não explicitam quais critérios que foram utilizados no agrupamento das fases genéricas identificadas na literatura para o desenvolvimento das fases do ciclo de vida proposto. Outro ponto observado no trabalho é que não está claro como foram identificados os cinco *stakeholders* apresentados, sendo que ao longo do texto observa-se um sexto *stakeholder* o gerente de projeto, que não é identificado junto ao ciclo de vida.

Observa-se ainda no mesmo trabalho que o ciclo proposto possui um foco diferenciado. Sendo que, através das fases de reuso e avaliação tem-se a possibilidade de obter *feedbacks*, possibilitando melhorar os processos estratégicos das organizações, referente aos dados que são disponibilizados.

Nota-se através desta análise que os dois ciclos de vida, são complementares, ou seja, o ciclo proposto por Auer et al. (2011) fornece ferramentas, focado em como tratar os dados, com o intuito de minimizar o árduo trabalho de publicar dados abertos conectados e o ciclo proposto por Broek et al. (2014) apresenta as orientações a nível estratégico, assim como os *stakeholder* envolvidos em todo o processo.

Realizando uma análise preliminar dos demais trabalhos encontrados na literatura, percebe-se que em sua maioria os ciclos de vida de publicação de dados são descritos como um processo sequencial e unidimensional de fases, que um grupo muitas vezes não especificado de *stakeholders* realiza frequentemente para fornecer uma quantidade de dados à um público geral. Percebe-se também que estes trabalhos em sua maioria focam nos processos operacionais, ignorando muitas vezes a mensuração do consumo dos dados e o *feedback* dos usuários, sendo que isto

possibilitaria não somente melhorias quanto a qualidade dos dados disponibilizados, mas também poderia potencializar novos negócios. Visto que, a qualidade, a disponibilidade e a utilização dos dados têm um papel central no sucesso daqueles que publicam seus dados (HAIDER; HAIDER, 2013).

Outro item observado é que os ciclos de vida para a publicação de dados conectados e dados conectados abertos engloba a maioria dos trabalhos citados por este artigo. No entanto só um dos trabalhos abrange o contexto somente de dados abertos, sendo que não é abordado como um ciclo de vida, mas sim como um processo (JANSSEN; ZUIDERWIJK, 2012). Também observa-se que pouco se fala em licença aberta quando abordado o contexto de dados abertos e dados abertos conectados.

6 Conclusão ou Considerações Finais

Este artigo apresentou resultados parciais de um estudo, sobre os trabalhos encontrados na literatura que discutem o tratamento dos dados para publicação. O mesmo foi desenvolvido com o propósito de encontrar características referente as pesquisas da área, assim como fornecer possibilidade de futuras pesquisas.

Com a realização deste estudo observou-se que a maioria dos ciclos de vida encontrados na literatura abordam apenas processos operacionais, ou seja, seu foco é a nível tecnológico, e que em sua maioria ignoram as fases de pós-publicação, sendo que a mensuração dos dados utilizados, assim como os *feedbacks* fornecidos podem gerar benefícios. Com base neste itens elencado, também percebe-se que há uma falta de alinhamento estratégico do “Por que e para quem estamos publicando dados?”, o que leva a pensar que as organizações apenas publicam dados por publicar, sem pensar nos benefícios que podem ser obtidos.

Conclui-se também que não há estudos que guiem ou auxiliem tanto em nível estratégico, quanto em nível tecnológico, aqueles que desejam começar a abrir seus dados, sem pensar

em dados conectados. Para isso sugere-se como possibilidade de pesquisas futuras a estruturação de um ciclo de vida para a publicação de dados abertos, com foco tecnológico e estratégico, baseado nas recomendações do Open Knowledge e dos ciclos de vida de publicação de dados.

Referências

- ABELE, A., MCCRAE, J. Linking open data cloud diagram. **LOD Community**. 2017. Disponível em: <<http://lod-cloud.net/>>. Acesso em: 04 ago. 2017.
- ALSHEHHI, Maryam et al. Visual analytics in the web of data. In: **Electronics, Circuits, and Systems (ICECS), 2013 IEEE 20th International Conference on**. IEEE, 2013. p. 102-103.
- AUER, Sören; LEHMANN, Jens; NGOMO, Axel-Cyrille Ngonga. Introduction to linked data and its lifecycle on the web. In: **Reasoning Web. Semantic Technologies for the Web of Data**. Springer Berlin Heidelberg, 2011. p. 1-75.
- BARNAGHI, Payam et al. Semantics for the Internet of Things: early progress and back to the future. **International Journal on Semantic Web and Information Systems (IJSWIS)**, v. 8, n. 1, p. 1-21, 2012.
- BERNEERS-LEE, Tim. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, v.284, n. 5, p. 34-43, 2001.
- BERNERS-LEE, Tim. Linked data-design issues. 2006. Disponível em: <<http://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 04 ago. 2017.
- BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. **Information Science Reference (an imprint of IGI Global), USA**, p. 205-227, 2009.
- BROEK, Tijs Adriaan Van Den; VAN VEENSTRA, A. F. E.; FOLMER, Erwin Johan Albert. Walking the extra byte: A lifecycle model for linked open data. 2013.
- DADOS.GOV. Portal Brasileiro de Dados Abertos. **O que são dados abertos**. 2017. Disponível em: <http://dados.gov.br/dados-abertos/> Acesso em: 28 ago. 2017.
- DATA REVOLUTION GROUP. **A World That Counts: Mobilising the Data Revolution for Sustainable Development**. 2014. Disponível em: <<http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-ThatCounts.pdf>>. Acesso em: 01 ago. 2017.
- AIDER, Waqar; HAIDER, Abrar. Governance structures for engineering and infrastructure asset management. In: **Technology Management in the IT-Driven Services (PICMET), 2013 Proceedings of PICMET'13**. IEEE, 2013. p. 1229-1238.
- HAUSENBLAS, M. Linked data lifecycles, presentation from DERI research institute, Galway, Ireland, July 2011.
- HEATH, Tom; BIZER, Christian. Linked data: Evolving the web into a global data space. **Synthesis lectures on the semantic web: theory and technology**, v. 1, n. 1, p. 1-136, 2011.
- HITZLER, Pascal; JANOWICZ, Krzysztof. Linked Data, Big Data, and the 4th Paradigm. **Semantic Web**, v. 4, n. 3, p. 233-235, 2013.
- HYLAND, Bernadette; WOOD, David. The joy of data-a cookbook for publishing linked government data on the web. **Linking government data**, p. 3-26, 2011.
- JANSSEN, M. and ZUIDERWIJK, A. Open data and transformational government. In **Transforming Government Workshop**. Brunel University, United Kingdom. 2012.
- MEIJER, Albert J.; CURTIN, Deirdre; HILLEBRANDT, Maarten. Open government: connecting vision and voice. **International Review of Administrative Sciences**, v. 78, n. 1, p. 10-29, 2012.
- OPEN KNOWLEDGE. How to Open Data. Disponível em: <<https://okfn.org/opendata/how-to-open-data/>>. Acesso em: 29 jul. 2017a.
- OPEN KNOWLEDGE. **Open Definition**. Disponível em: <<http://opendefinition.org/od/2.1/en/>>. Acesso em: 02 ago. 2017b.
- SCHARFFE, François et al. Enabling linked-data publication with the datalift platform. In: **Proc. AAAI workshop on semantic cities**. 2012.
- VILLAZÓN-TERRAZAS, Boris et al. Methodological guidelines for publishing government linked data. **Linking government data**, p. 27-49, 2011.