PRÁTICAS PARA PUBLICAÇÃO DE DADOS CONECTADOS: UMA REVISÃO SISTEMÁTICA

Jefferson de Oliveira Chaves[1];

José Leomar Todesco[2].

Abstract: *Maintaining connected datasets is a costly activity that involves using multiple resources, obeying principles, and respecting best practices. Thus, this systematic review aims to perform an analysis of the practices used for the maintenance of connected data. In a universe of ninety-one articles, we have reached a clipping of fifteen works on which this review is based. It is possible to observe that there is great adherence to the principles and some difficulties with good practices, indicating a relative consolidation of practices in some points, while there are others that still require special attention in the construction of connected data.*

Keywords: linked data; semantic web; linked data workflow; linked data life cycle.

Resumo*: A manutenção de conjuntos de dados conectados é uma atividade onerosa que envolve a utilização de diversos recursos, obedecer princípios e respeitar boas práticas. Assim esta revisão sistemática tem como objetivo realizar uma análise sobre as práticas utilizadas para a manutenção de dados conectados. Num universo de noventa e um artigos, chegou-se a um recorte de quinze trabalhos, nos quais esta revisão está baseada. É possível observar que há grande aderência aos princípios e algumas dificuldades com as boas práticas, indicando uma relativa consolidação de práticas em alguns pontos, enquanto há outros que ainda demandam uma atenção especial na construção de dados conectados.*

Palavras-chave: dados conectados; web semântica; fluxo de dados conectados; ciclo de vida de dados conectados.

1 INTRODUÇÃO

Atualmente, a *web* passa por um constante processo de evolução, revolucionando a maneira de se produzir, obter e compartilhar informações na internet. Desde sua criação, a *web* tem seus documentos organizados por meio de representações baseadas em hipertexto. Ainda que essa organização tenha facilitado o acesso por usuários humanos, essa representação sem qualquer semântica e pouco expressiva, dificultou a extração de

1 Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina - (UFSC) Florianópolis – Brasil. Correo electrónico: jefferson.chaves@ifpr.edu.br

2 Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina - (UFSC) Florianópolis – Brasil. Correo electrónico: jose.todesco@ufsc.br

informações destes documentos (Cunha, Lóscio, Souza, 2009; Berners-Lee, Bizer, Heath, 2006).

Dessa forma, a informação contida nesses documentos, embora possam ser compreendidos por humanos, tornam-se inadequados para o uso por meio de agentes de *software*. Neste cenário, tornou-se necessário a atribuição de significados para os elementos, dados e expressões, além de interligá-los com outros conjuntos de dados ou outros domínios de conhecimento, de forma a se criar uma relação de significância entre os conteúdos publicados na internet. Essa abordagem foi definida como *web* semântica (Cunha, Lóscio, Souza, 2009; Berners-Lee *et al*., 2001). De acordo com Berners-Lee (2001, p. 34) "[…] a web semântica é uma extensão da *web* atual, na qual a informação recebe um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação".

Como um componente da evolução da *web* semântica, surgiu o conceito de *Linked Data* ou Dados Conectados[3], proposto por Tim Berners-Lee, que pode ser entendido como um conjunto de práticas para publicação, conexão e compartilhamento de conjuntos de dados estruturados. Como resultados da aplicação desses padrões, temos a denominada *web* de dados (Bizer *et al*. 2009 *apud* Cunha, Lóscio, Souza, 2009).

Entretanto, a publicação desses dados de forma conectada constitui-se como um novo desafio, uma vez que as ferramentas existentes são insuficientes para apoiar tal processo. Além disso, a própria execução da tarefa de publicação é complexa, uma vez que domínios e contextos distintos exigem infraestrutura, ferramentas e configurações específicas para cada caso. Assim, esforços substanciais são gastos em reproduzir conjuntos de dados ao longo do tempo (Rautenberg, *et al*., 2016).

Nesse sentido, a questão de pesquisa que norteia este trabalho é: quais são as práticas para publicação de Dados Conectados? Cabe destacar, que por práticas compreendem-se o conjunto de variáveis utilizadas para a publicação de Dados Conectados, tais como: aderência aos princípios e as boas práticas, tecnologias utilizadas e o ciclo de vida definido. O objetivo do trabalho, portanto, é o de identificar as práticas definidas em trabalhos acadêmicos encontrados na plataforma Scopus.

3 Na literatura especializada poderão ser encontradas outras nomenclaturas para Dados Conectados, tais como: *Linked Data*, Dados Ligados e Dados vinculados, entretanto, para fins de padronização optou-se por utilizar a nomenclatura de Dados Conectados.

Para tal, este trabalho encontra-se dividido da seguinte forma: na seção 2 será apresentado de forma breve as discussões sobre publicação de Dados Conectados; na seção 3 será apresentado a metodologia que estruturou a pesquisa para esta revisão sistemática; na seção 4 será apresentada uma breve análise sobre os resultados da pesquisa no Scopus, bem como uma discussão sobre as pesquisas levantadas e, por fim, a seção 5 traz considerações finais acerca do tema.

2 DADOS CONECTADOS

Dados Conectados referem-se a um conjunto de melhores práticas para publicação e conexão de dados estruturados na *web*, utilizando padrões internacionais , que permitem estabelecer conexões entre itens de diferentes fontes de dados para formar um único espaço de dados global (Heath; Bizer, 2011). Este conjunto de melhores práticas são sumarizados em quatro princípios: i) Use URI's (Uniform Resource Identifier) para nomear as coisas; ii) Use URI's HTTP para que as pessoas possam procurar o desejado; iii) Quando alguém olha para um URI, forneça informações úteis, usando os padrões (RDF, SPARQL); iv) Incluir links para outros URI's, para que eles possam descobrir explorar mais as coisas. Cabe destacar que esses quatro princípios constituem a base norteadora para a definição das tecnologias, atividades e ferramentas do Ciclo de Vida para publicação de Dados Conectados.

2.1 BOAS PRÁTICAS PARA PUBLICAÇÃO DE DADOS CONECTADOS

Conforme apontado, as práticas são compreendidas como um conjunto amplo de variáveis, dentre elas destacamos as boas práticas que são definidas pela *World Wide Web Consortium* (W3C). Assim, além dos princípios supracitados que devem ser seguidos para a publicação de Dados Conectados, A W3C, por meio de um grupo de trabalho criado para o estudo de Dados Conectados Governamentais, estabeleceu uma série de boas práticas para facilitar o desenvolvimento e fornecimento de Dados Conectados. Apesar destas práticas terem sido concebidas inicialmente para o ambiente governamental, é possível estendê-las a outros segmentos. Dentre as boas práticas enumeradas pela W3C (2014), para fins deste trabalho nos restringiremos as seguintes: **1)** Selecionar um conjunto de dados: seleção de

dados que serão publicados, de forma que possibilitem o uso para distintas finalidades; **2)** Modelar os dados: construir a melhor representação dos dados e a forma como serão utilizados por distintas aplicações, independentemente da origem desses dados; **3)** Especificar uma licença apropriada: definição da licença mais apropriada para as condições de uso, com o intuito de definir termos sobre a origem, propriedade e outras condições de uso dos dados; **4)** Construir boas URIs para Dados Conectados: a implementação de URIs deve ser feita considerando o referenciamento dos dados baseados em URIs HTTP. Esse planejamento deve conter: nomes de objetos, o suporte para múltiplos idiomas, o suporte a mudança de dados ao longo do tempo e a estratégia de persistência; **5)** Utilizar um vocabulário padrão: sempre que possível, devem ser usados vocabulários existentes para identificar os objetos. Tais vocabulários podem ser estendidos se necessário. Novos vocabulários podem ser criados, desde que necessários, seguindo as boas práticas; **6)** Converter dados para representações de Dados Conectados: esta etapa objetiva transformar dados em uma representação de Dados Conectados. Essa etapa é tipicamente apoiada por *scripts* ou *softwares* que automatizam essa tarefa; **7)** Prover acesso automatizado para os dados: devem ser implementados meios que permitam o acesso automatizado aos dados por motores de busca ou outros mecanismos de processamento e consumo de dados; **8)** Anunciar para o público: essa etapa envolve a tarefa de tornar público o conjunto de dados ligados. Destaca-se que a publicação desse conjunto de dados, implicitamente, gera um efeito de contrato social com o público.

Assim é de interesse desse trabalho observar se estas boas prática são aplicadas quando da abertura de dados.

2.2 CICLO DE VIDA DE DADOS CONECTADOS

Existem distintas definições de ciclos de vida na literatura. Para esta revisão sistemática optou-se por utilizar a formulada por Auer (2014), por compreender que esta representa uma visão holística sobre o ciclo de vida dos processos de publicação de Dados Conectados.

Essas etapas são delineadas da seguinte forma (Auer, 2014, p. 3-4): **1)** Extração (*Extraction*): esta etapa consiste na obtenção de dados de fontes e formatos de arquivos diversos; **2)** Armazenamento e Consulta (*Storage/Querying*): esta etapa tem o objetivo de

fornecer meios para armazenamento, principalmente em formato RDF, empregando técnicas que visam otimizar o desempenho das consultas por meio do uso de cache, junções e outras técnicas de processamento otimizado; **3)** Revisão manual e autoria (*Manual Revision/Authoring*): esta etapa deve validar a informação e enriquecer o valor semântico dos dados. Técnicas como a Wiki Semântica e o paradigma WYSIWYM (o que você vê é o que você quer dizer) podem ser utilizadas para ampliar as redes sociais de colaboração; **4)** Interconexão (*Interlinking*): nesta etapa devem ser estabelecidas conexões com outras fontes de dados a fim de estabelecer um único espaço de dados global. O pressuposto básico por trás de dados conectados é de que o valor e a utilidade dos dados aumentam proporcionalmente ao número de ligações que eles estabelecem com outros dados; **5)** Classificação e Enriquecimento (*Classification/Enrichment*): os dados extraídos na primeira etapa são costumeiramente dados brutos. Assim, para que eles sejam transformados em dados conectados (integração, fusão, pesquisa e outras aplicações), devem passar necessariamente por um processo de classificação e enriquecimento, por meio da vinculação e integração com ontologias de nível superior; **6)** Análise de Qualidade (*Quality Analysis*): devem ser empregadas técnicas para avaliação da qualidade dos dados. Tais técnicas podem basear-se em variáveis tais como: proveniência, contexto, cobertura e estrutura dos dados; **7)** Evolução e Reparar (*Evolution/Repair*): uma das características dos dados na *web* é a dinamicidade. Assim, uma das etapas fundamentais do ciclo de vida para publicação de Dados Conectados consiste na manutenção da estabilidade dos dados. Assim, é necessário que toda mudança, todo vocabulário e toda ontologia seja transparente e passível de observação; **8)** Pesquisa/Navegação/Exploração (*Search/Browsing/Exploration*): considerando a "invisibilidade" dos dados da *web* para muitos usuários, o intuito desta etapa é o de desenvolver técnicas de pesquisa, navegação, exploração e visualização para distintos dados conectados, tornando assim, os dados da *Web* acessíveis ao usuários.

2.3 FERRAMENTAS E TECNOLOGIAS PARA PUBLICAÇÃO E CONSUMO DE DADOS CONECTADOS

A complexidade que envolve a publicação de Dados Conectados exige ferramentas de apoio que aplicadas às etapas do ciclo de vida supracitadas, garantam a execução da tarefa.

Ainda, tecnologias que permitam a transformação, representação, visualização e disponibilização de dados Conectados são necessários para este processo. Nesse sentido, para a manutenção de dados, são utilizadas ferramentas como, geradores, armazenadores e publicadores de triplas RDF. Ainda, podem ser empregadas outras ferramentas como conversores de dados e ambientes de programação. Para o provimento de acesso aos dados por máquina, são utilizados servidores de triplas que oferecem acesso a *endpoints* em linguagem SPARQL (Bandeira, 2015).

Assim, é possível observar que inúmeras ferramentas e tecnologias são necessárias para as distintas etapas, com o intuito de viabilizar a publicação dos Dados Conectados. Identificar tais ferramentas e tecnologias aplicadas durante esse processo é fundamental, já que diferentes contextos e domínios exigem configurações específicas.

3 METODOLOGIA

Para a construção desta revisão sistemática, o ponto de partida consistiu em compreendê-la como uma forma de pesquisa que utiliza como fonte de dados a literatura sobre determinado tema, da mesma forma que outras técnicas. Nesse sentido, uma revisão sistemática deve ter uma pergunta de pesquisa clara, a definição de uma estratégia de busca, critérios de inclusão e exclusão de referências e, sobretudo, uma análise criteriosa da literatura selecionada (Mancini, Sampaio, 2007).

A busca sistemática foi feita na plataforma *Scopus*, tomando por base os textos publicados a partir do ano 2010 até o presente. Considerando-se a pergunta de pesquisa e o objetivo do presente trabalho, foram definidas as seguintes palavras-chave para a busca inicial: (1) *linked data; linked data workflow* e *linked data lifecycle*.

A escolha dos termos assinalados acima foi construída partindo da compreensão de que *Linked Data* seria equivalente ao termo Dados Conectados e que *lifecycle* e *workflow* atenderiam a ideia de uma série de procedimentos que ocorrem em uma determinada ordem e que são necessários para a publicação de Dados Conectados.

Com o intuito de selecionar as publicações com maior relevância para a revisão, estabelecemos os seguintes critérios para o refinamento da pesquisa, além das palavras-chave: (2) limitação por área; (3) exclusão de áreas; (4) limitação de outras palavras-chave definidas

pelo(s) autor(es), (5) exclusão por meio da leitura dos títulos, resumos e palavras-chave e por fim, (6) seleção dos 15 artigos mais citados. Assim, os critérios 1, 2, 3 e 4, resultaram na *string* conforme a Tabela 1:

Tabela 1. *String* de busca

TITLE-ABS-KEY ("linked data" AND "workflow" OR "lifecycle") AND PUBYEAR > 2009 AND PUBYEAR > 2009 AND (LIMIT-TO (SUBJAREA , "COMP ") OR LIMIT-TO (SUBJAREA , "MATH ") OR EXCLUDE (SUBJAREA , "SOCI ") OR EXCLUDE (SUBJAREA , "ENGI ") OR EXCLUDE (SUBJAREA , "DECI ") OR EXCLUDE (SUBJAREA , "ARTS ") OR EXCLUDE (SUBJAREA , "MEDI ") OR EXCLUDE (SUBJAREA , "BIOC ") OR EXCLUDE (SUBJAREA , "CHEM ") OR EXCLUDE (SUBJAREA , "EART ") OR EXCLUDE (SUBJAREA , "PHYS ") OR EXCLUDE (SUBJAREA , "AGRI ") OR EXCLUDE (SUBJAREA , "BUSI ") OR EXCLUDE (SUBJAREA , "ENVI ") OR EXCLUDE (SUBJAREA , "MATE ")) AND (LIMIT-TO (EXACTKEYWORD , "Linked Data ") OR LIMIT-TO (EXACTKEYWORD , "Life Cycle ") OR LIMIT-TO (EXACTKEYWORD , "lifecycle ") OR LIMIT-TO (EXACTKEYWORD , "Work-flows ") OR LIMIT-TO (EXACTKEYWORD , "Workflow "))

Fonte: autoria própria.

A Tabela 2 traz o número de textos retornados de acordo com o critério de exclusão aplicado:

Tabela 2. Filtragem da pesquisa

Ordem	**Critério**	**Resultado**
1	Palavras-chave: *linked data; workflow* e *lifecycle*	255
2	Limitando as áreas: Computação e Matemática	220
3	Excluindo as áreas:	178
4	Limitando as palavras-chave*: linked data; life cycle; lifecycle; workflow; work flows*	91
5	Leitura do título, resumo e palavras-chave	25
6	Seleção dos mais citados	15

Fonte: autoria própria.

Consolidado o quarto critério, chegamos ao total de 91 trabalhos relacionados pela plataforma. A partir desse ponto, iniciamos a leitura dos títulos, resumos e palavras-chave desses trabalhos, com o intuito de identificarmos àqueles que mais se aproximavam do tema pesquisado. A leitura do título, do resumo e das palavras-chave permitiu o refinamento dos textos, de tal forma que, ainda que aparecessem na pesquisa, foram excluídos àqueles textos que estavam fora de contexto. Assim, a amostra consolidou-se em 15 trabalhos que versam sobre a temática e que serão analisados e discutidos nos tópicos seguintes.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Uma análise da proveniência da literatura selecionada demonstrou que no período delimitado os anos de 2014 e 2017 são os que mais tiveram publicações sobre a área com três publicações cada. Seguindo, os anos de 2011, 2013 e 2016 possuem duas publicações cada.

Os trabalhos concentram-se, essencialmente, em treze *Conference Papers* e dois artigos. Estes por sua vez, estão distribuídos nas seguintes publicações: a) *Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial IntelligenceAndLecture Notes In Bioinformatics:* três publicações; b) *Ceur Workshop Proceedings*: trêspublicações; c) *ACM International Conference Proceeding Series*: 2 publicações; d) *Communications In Computer And Information Science*: 1 publicação; e) *Future Generation Computer Systems*: 1 publicação; f) *IEEE International Symposium On Parallel And Distributed Processing Workshops And Phd Forum*: 1 publicação; g) *Journal Of Web Semantics*: 1 publicação; h) *Proceedings International Computer Software And Applications Conference*: 1 publicação; i) *Works 11 Proceedings Of The 6th Workshop On Workflows In Support Of Large Scale Science Co Located With Sc 11*: 1 publicação; j) *Www 2014 Companion Proceedings Of The 23rd International Conference On World Wide Web*: 1 publicação.

A distribuição por autores observa um empate entre Auer, Garijo, Gil, Marshall e Zhao, com duas publicações cada, seguidos por Aoyama, Baierery, Bechhofer, Bizer e Boyce, com uma publicação cada.

Por fim, um ponto a ser destacado é que as publicações concentram-se em termos territoriais, no eixo estadunidense e europeu, em países com alto índice de desenvolvimento econômico, com destaque para Alemanha e Estados Unidos com cinco publicações cada, seguidos por Espanha e Reino Unido com três publicações cada, Austria, Irlanda e Holanda, com 2 publicações e por fim, Bélgica, Brasil e Canadá com 1 publicação cada.

Assim, a partir das informações apresentadas, nosso próximo passo consiste em apresentar as principais contribuições dos textos.

4.1 PRINCIPAIS CONTRIBUIÇÕES DE CADA ARTIGO

Tabela 3. Principais contribuições de cada artigo

Trabalho	Contribuição

A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data	Apresentou uma nova abordagem para a publicação de *workflows* científicos utilizando dados conectados representados por ontologias. Deixou explícito o método utilizado, permitindo a reutilização do fluxo de publicação.
Emerging practices for mapping and linking life sciences data using RDF - A case series	Representou os resultados de experimentos de em um formato padronizado (RDF), o que possibilitou a montagem de consultas que testam hipóteses sobre drogas potencialmente úteis para o tratamento da Doença de Alzheimer, com dados integrados. Além disso, trouxeram um conjunto de boas práticas para criação e publicação de fontes de Dados Conectados.
Semantically Enhanced Quality Assurance in the JURION business use case	Apresentou-se a arquitetura e o ciclo de vida de uma aplicação que funde e interliga mais de um milhão de dados. Essa arquitetura, apresentada em detalhes, diminuiu a lacuna entre o desenvolvimento de *software* e o desenvolvimento de dados por meio da integração de controles de qualidade na no conjunto de ferramenta de *softwares* existentes.
Identifying Web Tables – Supporting a Neglected Type of Content on the Web	Apresentou um *framework* para extrair e analisar dados extraídos de forma automatizada de tabelas HTML. Além disso, foi apresentado algoritmos de aprendizado para análise da estrutura de uma tabela bem como a geração automática de ontologias e publicação do conjunto de dados extraídos.
RESTful Open Workflows for Data Provenance and Reuse	Apresentou uma arquitetura que integrou todas as fases de um ciclo de vida típico de trabalho, incluindo a especificação de serviços, sua composição para os trabalhos, bem como a execução dos trabalhos. Por meio do uso de uma ontologia, foram especificados todos os recursos do ciclo de vida.
Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services	Desenvolveu uma ferramenta para facilitar a publicação e a reutilização de cubos de dados vinculados: *OpenCubeToolkit* Essa ferramenta integra,de forma facilitada, componentes separados que lidam com diferentes etapas do ciclo de vida do cubo de Dados Conectadosque permitem ao usuário um conjunto de funcionalidades para trabalhar com dados semânticos estatísticos.
A Linked Data Approach to Sharing Workflows and Workflow Results	Apresentou uma aplicação para gerenciamento de *workflows* de trabalho para preservação e criação de um equivalente digital da seção materiais e métodos descritos em artigos científicos na área de bioinformática.
Abstract, link, publish, exploit: An end to end framework for workflow sharing	Apresentou uma implementação de estrutura para publicação de fluxos de trabalho, com base em padrões, tais como: OWL, RDF e PROV. A abordagem apresentada ainda permite a publicação de outros sistemas de fluxo de trabalho.
Design Management: a Collaborative Design Solution	Apresentou uma aplicação para gestão de projetos. O armazenamento foi implementado em um repositório central, vinculados a outros repositórios, facilitando a colaboração no desenvolvimento dos conjuntos de dados. Além disso, descreveu o ciclo de vida de toda a geração

	dos conjunto de dados.
Designing the Cloud-based DOE Systems Biology Knowledgebase	Apresentou uma aplicação em nuvem que permitiu o enriquecimento semântico por meio de anotações, além da publicação e compartilhamento dos dados. Ainda, trouxe a análise do fluxo de publicação de tais dados.
A Linked Data Lifecycle for Smart Cities in Spain	Descreveu como são aplicados os ciclos de vida de dados vinculados, da especificação à exploração, dentro dos domínios de Cidade Inteligente. Foram apresentadas algumas abordagens relacionadas às Cidades Inteligentes que seguem o conceito de Dados Conectados.
PROMIS: A Management Platform for Software Supply Networks Based on the Linked Data and OSLC	Contribuiu com a apresentação de uma arquitetura de *software*(PROMIS) que forneceu uma solução para a troca de dados de gerenciamento de projetos de diferentes domínios.
LinkedPipes ETL in use: Practical publication and consumption of Linked Data	Apresentou o LinkedPipes ETL, uma ferramenta para publicação de Dados Abertos Conectados, que concentrou-se principalmente no suporte a *workflows* de publicação de Dados Abertos Conectados de maneira amigável. Além disso, a ferramenta também facilita o consumo de fontes de dados já existentes.
A Life-cycle Workflow Architecture for Linked Data	Apresentou uma proposta de arquitetura para o ciclo de vida de publicação de Dados Conectados. Apresentou uma visão geral, sistemática, que descreveu os principais componentes da arquitetura.
LDWPO – A Lightweight Ontology for Linked Data Management	Apresentou um modelo de conhecimento para fluxos de trabalhos suportados por uma ontologia. Essa ontologia contempla: o processo metodológico que orienta o ciclo de vida dos conjuntos de dados RDF, o plano completo de produção do conjunto de dados, e a documentação das execuções do *workflow*. Como resultado, a abordagem permitiu a reprodutibilidade e repetibilidade de etapas de processamento de dados vinculados ao longo do tempo.

Fonte: autoria própria.

Com base na tabela acima é possível observar que: **a)** 11 trabalhos trouxeram abordagens do Ciclo de Vida de publicação de Dados Conectados; **b)** 3 trabalhos apresentaram abordagens por meio do desenvolvimento de aplicações; **c)** 4 trabalhos apresentaram abordagens por meio da construção de uma arquitetura de publicação; **d)** 6 trabalhos apresentaram abordagens para domínios organizacionais e 5 trabalhos apresentaram abordagens para domínios da área da saúde.

4.2 ADERÊNCIA AOS PRINCÍPIOS E AS BOAS PRÁTICAS

Com base na discussão apresentada no item 2 desta revisão este ponto buscará analisar se os trabalhos aderiram aos princípios e as boas práticas apresentadas pela W3C.

4.2.1 Aderência aos princípios

Para fins de didáticos optamos por definir três resultados possíveis para análise dos trabalhos: i) **sim**, quando o trabalho definia explicitamente a aderência ao princípio; ii) **não descrito**, quando o trabalho não faz menção ao princípio, e iii) **não**, quando o trabalho explicitou a não aderência ao princípio.

A seguir, apresentamos a Tabela 4 em que é possível observar se o trabalho atendeu aos quatro princípios estabelecidos por Berners-Lee e apresentados na seção 2 desta revisão.

Tabela 4. Aderência aos princípios

Trabalho	Princípio 1	Princípio 2	Princípio 3	Princípio 4
A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data	Sim	Sim	Sim	Sim
Emerging practices for mapping and linking life sciences data using RDF - A case series	Sim	Sim	Sim	Sim
Semantically Enhanced Quality Assurance in the JURION business use case	Sim	Sim	Sim	Sim
Identifying Web Tables – Supporting a Neglected Type of Content on the Web	Não descrito	Não descrito	Sim	Sim
RESTful Open Workflows for Data Provenance and Reuse	Não descrito	Não descrito	Sim	Não
Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services	Não descrito	Não descrito	Sim	Sim
A Linked Data Approach to Sharing Workflows and Workflow Results	Sim	Sim	Sim	Sim
Abstract, link, publish, exploit: An end to end framework for workflow sharing	Sim	Sim	Sim	Sim
Design Management: a Collaborative Design Solution	Não descrito	Não descrito	Sim	Sim
Designing the Cloud-based DOE Systems Biology Knowledgebase	Não descrito	Não descrito	Sim	Sim
A Linked Data Lifecycle for Smart Cities in Spain	Não descrito	Não descrito	Sim	Sim
PROMIS: A Management Platform for Software Supply Networks Based on the Linked Data and OSLC	Não descrito	Não descrito	Sim	Sim
LinkedPipes ETL in use: Practical publication and consumption of Linked Data	Sim	Sim	Sim	Sim
A Life-cycle Workflow Architecture for Linked Data	Sim	Sim	Sim	Sim
LDWPO – A Lightweight Ontology for Linked Data Management	Sim	Sim	Sim	Sim

Fonte: autoria própria.

A análise dos trabalhos nos traz o seguinte quadro: **a)** Dos quinze trabalhos, oito aderiram aos quatro princípios; **b)** Seis trabalhos não mencionaram a aderência aos princípios

1 e 2, contudo, aderiram aos princípios 3 e 4; **c)** Um trabalho não menciona a aderência ao princípio 1 e 2, adere ao princípio 3, e não adere ao princípio 4.

4.2.2 Aderência às boas práticas

Para a análise da aderência às boas práticas optamos por definir quatro resultados possíveis para análise dos trabalhos: i) **sim**, quando o trabalho definia explicitamente a aderência; ii) **não descrito**, quando o trabalho não faz menção; iii) **não**, quando o trabalho explicitou a não aderência ao princípio, e iv) **parcialmente**, quando foi evidenciado de que não foi atendido totalmente. Assim, de acordo com a Tabela 5, temos a seguinte distribuição:

Tabela 5. Aderência às boas práticas

Boas práticas	Sim	Não descrito	Não	Parcialmente
Selecionar um conjunto de dados	15	-	-	-
Modelar os dados	15	-	-	-
Especificar uma licença apropriada	4	11	-	-
Construir boas URIs para Dados Conectados	8	-	7	-
Utilizar um vocabulário padrão	9	5	-	1
Converter dados para representações de Dados Conectados	15	-	-	-
Prover acesso automatizado para os dados	15	-	-	-
Anunciar para o público	12	-	3	-

Fonte: autoria própria.

4.2.3 Ciclo de vida

De acordo com o item 4.1, 11 dos 15 trabalhados analisados explicitaram ou descreveram ciclos de vida para publicação de Dados Conectados. A Tabela 6 descreve os ciclos de vida apresentados por cada um dos trabalhos.

Tabela 6. Ciclos de vida apresentados

Trabalho	Ciclo de Vida apresentado
A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data	i) Geração de fluxo de trabalho; ii) Conversão; iii) Publicação; iv) Compartilhamento; v) Reutilização;
Emerging practices for mapping and linking life sciences data using RDF - A case series	i) Selecionar as fontes de dados ou partes dos mesmos para ser publicados como RDF; ii) Identificar URLs persistentes; iii) Personalizar o mapeamento manualmente, se necessário; iv) Fazer a ligação conceitos no novo mapeamento RDF para conceitos em outras fontes de dados RDF; v) Colocar os dados RDF através de um *endpoint* ou como

	SPARQL; vi) Publicar aplicativos da *web*semântica usando os dados publicados.
Semantically Enhanced Quality Assurance in the JURION business use case	i) Extração; ii) Armazenamento; iii) Criação; iv) Interligação; v) Enriquecimento; vi) Análise da qualidade; vii) Reparação; viii) Publicação.
Identifying Web Tables – Supporting a Neglected Type of Content on the Web	i) Procurar páginas da Web relevantes para serem processados; ii) Extrair das informações para trabalhar com; iii) Determinar relevante da tabela; iv) Revelar a estrutura da informação encontrada; v) Identificar o intervalo de dados da tabela; vi) Mapear dos resultados extraídos para vocabulários e ontologias existentes.
A Linked Data Approach to Sharing Workflows and Workflow Results	i) Recuperar; ii) Rever; iii) Repetir; iv) Reusar; v) Reposicionar; vi) Conservar.
Abstract, link, publish, exploit: An end to end framework for workflow sharing	i) Especificação, onde as fontes de dados a serem usadas são identificadas e uma convenção de nomenclatura URI é projetada; ii) Modelagem, onde os usuários decidem quais vocabulários devem ser usado para representar os dados adequadamente de acordo com as requisitos e cenários; iii) Geração, ou seja, o processo de transformar os dados de seus formatos heterogêneos para um formato estruturado (RDF); iv) Publicação, em que o conjunto de dados resultante e seus *metadados* são disponibilizado por meio de um loja *triple store*; v) Exploração, onde os benefícios do conjunto de dados são clarificados através de aplicações ou consultas que consomem.
Designing the Cloud-based DOE Systems Biology Knowledgebase	i) Descoberta; ii) Acesso; iii) Integração de dados orientados por semântica.
A Linked Data Lifecycle for Smart Cities in Spain	i) Especificação; ii) Modelagem; iii) Geração; v) Publicação; vi) Exploração.
PROMIS: A Management Platform for Software Supply Networks Based on the Linked Data and OSLC	i) Obter; ii) transformar; iii) publicar.
A Life-cycle Workflow Architecture for Linked Data	i) Extração de dados de origem; ii) Transformação em RDF; iii) alinhamento com vocabulários comumente usados; iv) vinculação a outros conjuntos de dados; v) publicação na web; vi) carregamento em uma *triplestore*; vii) registro do conjunto de dados em um catálogo de dados como CKAN.
LDWPO – A Lightweight Ontology for Linked Data Management	i) Busca, Pesquisa; ii) Extração; iii) Armazenamento; iv) Revisão Manual; v) Interligação, combinação; vi) Enriquecimento; vii) Avaliação de qualidade; viii) Evolução/reparo;

Fonte: autoria própria.

A análise dos 11 trabalhos que apresentaram um ciclo de vida mostra que: **a)** 10 descrevem terem executado a etapa de Extração ou atividade análoga; **b)** 4 descrevem terem executado a etapa de Armazenamento e Consulta ou atividade análoga; **c)** 6 descrevem terem executado a etapa de Revisão manual e autoria ou atividade análoga; **d)** 7 descrevem terem executado a etapa de Interconexão ou atividade análoga; **e)** 2 descrevem terem executado a etapa de Classificação e Enriquecimento ou atividade análoga; **f)** 2 descrevem terem executado a etapa de Análise de Qualidade ou atividade análoga; **g)** 2 descrevem terem

executado a etapa de Evolução e Reparo ou atividade análoga; **h)** 9 descrevem terem executado a etapa de Pesquisa/Navegação/Exploração ou atividade análoga.

5 CONSIDERAÇÕES FINAIS

Esta revisão sistemática teve como principal objetivo a apresentação de práticas para a publicação de Dados Conectados. Para tal, realizou uma pesquisa na plataforma Scopus utilizando palavras-chaves previamente definidas e que possibilitaram um recorte de trabalhos publicados acerca do tema. Foi possível observar que há maior produção sobre o tema nos Estados Unidos e na Alemanha e Estados Unidos e que os picos – dentro do recorte realizado aqui – encontram-se nos anos de 2014 e 2017, com três publicações cada.

Em relação a aderência aos princípios a grande maioria dos trabalhos aderiu aos mesmos, o que indica uma prática que vem se consolidando dentro das pesquisas sobre Dados Conectados. Por sua vez, a aderência as boas práticas é consenso em quatro, dos oito princípios aqui selecionados. Entretanto, constituem desafios a consolidação na comunidade das seguintes práticas: a especificação de uma licença apropriada, a construção de bons URIs para Dados Conectados e, por fim, a utilização de um vocabulário padrão.

Os dados aqui consolidados, ainda que parcialmente e que norteiam para pesquisas futuras, indicam que há um caminho que vem sendo consolidado, mas que ainda deve percorrer um trecho para consolidar e efetivar as práticas para publicação de Dados Conectados.

REFERÊNCIAS

AOYAMA, M. et al. *PROMIS: A management platform for software supply networks based on the linked data and OSLC.* Proceedings - International Computer Software and Applications Conference. 2013

AUER, S. *Introduction to LOD2.* [s.l: s.n.]. v. 8661.

BANDEIRA, J. et al. Dados abertos conectados para a Educação. *Jornada de Atualização em Informática na Educação*, v. 4, n. 1, p. 47–69, 2015.

BERNERS-LEE, T. *Linked Data.* W3C, 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso: 20 de Jul. 2020.

BERNERS-LEE, T.; HENDLER J.; LASSILIA O. *The semantic web.* Scientific American, 284. p. 34–44, Mai. 2001.

BERNERS-LEE, T.; BIZER, C.; HEATH, T. *Linked Data - The story so far.* [ed.] Tim Heath, M. Hepp and Christian Bizer. International Journal on Semantic Web and Information System, Special Issue on Linked Data, 2006.

CUNHA, Danusa RB; LÓSCIO, B. F.; SOUZA, D. *Linked Data: da Web de Documentos para a Web de dados.* Livro Texto dos Minicursos ERCEMAPI, AM Santana et al., SBC: Teresina, BR, p. 79-99, 2011.

ECKERT, K. et al. *RESTful open workflows for data provenance and reuse.* WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web. 2014

ELAASAR, M.; CONALLEN, J. *Design management: A collaborative design solution.* [s.l: s.n.]. v. 7949 LNCS. 2013.

GALKIN, M.; MOUROMTSEV, D.; AUER, S. *Identifying web tables: Supporting a neglected type of content on the web.* [s.l: s.n.]. v. 518. 2015.

GARIJO, D.; GIL, Y. *A new approach for publishing workflows: Abstractions, standards, and linked data.* WORKS'11 - Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science, Co-located with SC'11. 2011

GARIJO, D.; GIL, Y.; CORCHO, O. Abstract, link, publish, exploit: An end to end framework for workflow sharing. *Future Generation Computer Systems*, v. 75, p. 271–283, 2017.

GONZÁLEZ, A.; VILLAZÓN-TERRAZAS, B.; GÓMEZ, J. M. *A linked data lifecycle for smart cities in Spain.* CEUR Workshop Proceedings. 2014

HEATH, T; BIZER C. Linked Data: *Evolving the Web into a Global Data Space*. Disponível em: <https://info.sice.indiana.edu/~dingying/Teaching/S604/LODBook.pdf>. Acesso em: 09 de Ago. de 2020.

KALAMPOKIS, E. et al. *Creating and utilizing linked open statistical data for the development of advanced analytics services.* CEUR Workshop Proceedings. 2014

KLÍMEK, J.; ŠKODA, P. *LinkedPipes ETL in use: Practical publication and consumption of Linked Data.* ACM International Conference Proceeding Series. 2017

KONTOKOSTAS, D. et al. *Semantically enhanced quality assurance in the JURION business use case.* [s.l: s.n.]. v. 9678. 2016.

KRAEMER et al. Maturidade de Gestão do Conhecimento: uma revisão sistemática da literatura para apoiar o desenvolvimento de novos modelos de avaliação. *Perspectivas em Gestão & Conhecimento*, João Pessoa, v. 7, Número Especial, p. 66-79, mar. 2017.

LANSING, C. et al. *Designing the cloud-based DOE systems biology knowledgebase.* IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum. 2011

LEE, Y. *A life-cycle workflow architecture for Linked Data.* ACM International

Conference Proceeding Series. 2017

MANCINI, MC.; SAMPAIO, RF. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. *Revista brasileira de fisioterapia*, v. 11, n. 1, p. 83-89, 2007. Disponível em: < https://www.scielo.br/pdf/rbfis/v11n1/12.pdf > Acesso em: 02 de Jul de 2020.

MARSHALL, M. S. et al. Emerging practices for mapping and linking life sciences data using RDF - A case series. *Journal of Web Semantics*, v. 14, p. 2–13, 2012.

RAUTENBERG, S. et al. *LDWPO - A lightweight ontology for linked data management*. CEUR Workshop Proceedings. 2016.

W3C. Disponível em: <https://www.w3.org/TR/ld-bp/>. Acesso em: 09 de Jul. de 2020.