Criação e uso de artefatos no processo de avaliação de qualidade de dados Linked Data

Creation and use of artifacts in the Linked data quality assessment process

Creación y uso de artefactos en el proceso de evaluación de la calidad de los datos Linked Data

Ananda Fernanda de Jesus
Doutoranda em Ciência da Informação, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. Bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)
http://lattes.cnpq.br/8553935355036462
https://orcid.org/0000-0001-7873-6040

José Eduardo Santarem Segundo
Livre Docente em Informação e Tecnologia pela Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brasil. Professor Associado, Universidade de São Paulo (USP)
http://lattes.cnpq.br/5562746387565465
https://orcid.org/0000-0002-0702-7586

Resumo

Introdução: O processo de avaliação de qualidade geralmente é baseado na seleção de critérios, dimensões e métricas que permitem quantificar e qualificar os níveis de qualidade de um conjunto de dados, podendo ser utilizados artefatos para auxiliar nesse processo. Essa pesquisa tem por objetivo identificar e discutir os artefatos disponíveis para avaliação de qualidade de dados *Linked Data* e ainda responder ao questionamento: existem artefatos voltados para auxiliar na seleção de dados *Linked Data* para ligação? **Metodologia:** Recorte em revisão sistemática da literatura e elaboração de formulário para coleta de informações. **Resultados:** Foram identificados 75 artigos que discutem a elaboração e/ou aplicação de artefatos para a avaliação de qualidade de dados *Linked Data*, esses artefatos foram classificados em 11 categorias estabelecidas a *posteriori*. **Conclusão:** Em resposta à pergunta apresentada, conclui-se que não foram identificados artefatos voltados para seleção de dados *Linked Data* para ligação com fontes externas. Em relação ao objetivo da pesquisa, conclui-se que entre as categorias se destacam os artefatos relacionados com vocabulários, ontologias, tesauros e metadados descritivos; baseados em adaptação de técnicas de outros domínios; focados em um domínio específico e ainda artefatos que buscam explorar características estruturais de dados *Linked Data*. Destaca-se ainda a importância dos vocabulários semânticos no processo de avaliação de qualidade de dados *Linked Data*.

Palavras-chave: *linked data*; avaliação de qualidade; qualidade de dados.

Abstract

Introduction: *The quality assessment process is generally based on the selection of criteria, dimensions and metrics that allow quantifying and qualifying the quality levels of a dataset, and artifacts can be used to assist in this process. This research aims to identify and discuss the artifacts available for evaluating the quality of Linked Data datasets and also answer the question: Are there*

artifacts aimed at assisting in the selection of Linked Data d for linking? **Methodology:** *Systematic literature review and preparation of a form to collect information. Results: 75 articles were identified that discuss the creation and/or application of artifacts for evaluating the quality of Linked Data, these artifacts were classified into 11 categories established a posteriori.* **Conclusion:** *In response to the question presented, it is concluded that no artifacts aimed at selecting Linked Data datasets for connection with external sources were identified. In relation to the objective of the research, it is concluded that among the categories, artifacts related to vocabularies, ontologies, thesauruses and descriptive metadata stand out; based on adaptation of techniques from other domains; focused on a specific domain and also artifacts that seek to explore structural characteristics of Linked Data datasets. The importance of semantic vocabularies in the Linked Data data quality assessment process is also highlighted.*

Keywords: linked *data; quality evolution; qualidade de dados*

Resumen

Introducción: *El proceso de evaluación de la calidad generalmente se basa en la selección de criterios, dimensiones y métricas que permiten cuantificar y calificar los niveles de calidad de un conjunto de datos, y se pueden utilizar artefactos para ayudar en este proceso. Esta investigación tiene como objetivo identificar y discutir los artefactos disponibles para evaluar la calidad de los datos vinculados y también responder a la pregunta: ¿Existen artefactos destinados a ayudar en la selección de datos vinculados para vincularlos?* **Metodología:** *Revisión sistemática de la literatura y elaboración de un formulario para recolectar información. Resultados: Se identificaron 75 artículos que discuten la creación y/o aplicación de artefactos para evaluar la calidad de los datos de Linked Data, estos artefactos se clasificaron en 11 categorías establecidas a posteriori.* **Conclusión:** *En respuesta a la pregunta presentada, se concluye que no se identificaron artefactos destinados a seleccionar datos de Linked Data para su conexión con fuentes externas. En relación al objetivo de la investigación, se concluye que entre las categorías destacan artefactos relacionados con vocabularios, ontologías, tesauros y metadatos descriptivos; basado en la adaptación de técnicas de otros dominios; enfocados en un dominio específico y también artefactos que buscan explorar características estructurales de los datos vinculados. También se destaca la importancia de los vocabularios semánticos en el proceso de evaluación de la calidad de los datos vinculados.*

Palabras clave: *cualidad de datos; evaluación de calidad de datos; datos vinculados.*

1 INTRODUÇÃO

As inquietações sobre como mensurar e melhorar a qualidade de dados não são recentes e as bases sobre como realizar esse processo de avaliação de qualidade, como ainda é conduzido atualmente, se estabeleceram na década de 1970. (Langer *et al.*, 2018).

O processo de avaliação de qualidade permiti mensurar os níveis de qualidade de um conjunto de dados, verificando o quão aptos para uso esses dados estão. Para a realização desse processo é necessário elaborar ou selecionar dimensões, critérios e métricas de qualidade. Essas dimensões agrupam as características dos dados que serão observadas para permitir mensurar a sua

qualidade (Wang; Strong, 1996). As dimensões podem ser fragmentadas em critérios que descrevem os atributos específicos a serem avaliados.

As métricas podem ser entendidas como indicadores que permitem mensurar a qualidade de dados, podendo ser quantitativas ou qualitativas, subjetivas ou objetivas. Cada critério pode possuir mais de uma métrica. (Wang; Strong, 2013; Assaf; Senart; Troncy, 2016; Melo, 2017).

O processo de avaliação de qualidade geralmente ocorre por meio de uma abordagem contextual, onde para a elaboração ou seleção de critérios e métricas é necessário considerar o contexto em que esses dados serão aplicados.

Nesse sentido, o presente estudo foca suas discussões em conjuntos de dados publicados como *Linked Data*, que tem por objetivo orientar a publicação de dados estruturados e conectados na *Web*, respeitando os seguintes princípios: 1) Use *Uniform Resource Identifier* (URIs) como nomes para as coisas; 2) Use *Hypertext Transfer Protocol* (HTTP) URIs, para que as pessoas possam procurar esses nomes; 3) Utilize o padrão *Resource Description Framework* para fornecer informações sobre os recursos; e 4) Inclua *links* para outros URIs, para que eles possam descobrir mais coisas. (Berners-Lee, 2006). Um dos desafios para a publicação de dados como *Linked Data* é justamente o processo de seleção de dados para ligação, baseado em processos de avaliação de qualidade.

O processo de avaliação de qualidade pode se beneficiar do uso de artefatos, que podem auxiliar em uma ou em todas as etapas desse processo. Esses artefatos são muito diversos entre si, podendo ser automáticos ou semiautomáticos, de uso geral ou específico para determinado contexto de aplicação.

Partindo da compreensão de que dados publicados como *Linked Data* possuem características próprias que tem impacto no processo de avaliação de qualidade, e que um dos desafios para sua adoção é a escolha de fontes para ligação. Compreendendo ainda que o uso de artefatos pode auxiliar nesse processo de avaliação e seleção de fontes, a presente pesquisa tem como objetivo identificar e discutir os artefatos disponíveis para avaliação de qualidade de dados *Linked*

Data. Parte ainda do questionamento: Existem artefatos voltados para auxiliar na seleção de dados *Linked Data* para ligação?

Para embasar a pesquisa foi realizado um recorte em uma Revisão Sistemática da Literatura (RSL), cujo objetivo foi discutir de maneira geral e abrangente o estado da arte sobre qualidade de dados publicados como *Linked Data*. Para instrumentalizar a realização do mencionado recorte, estabeleceu-se um protocolo de pesquisa específico para o mesmo. Elaborou-se ainda um formulário de sistematizar a coleta de informações relacionadas com os artefatos identificados. Apresentados o objetivo e o problema de pesquisa que embasaram a elaboração do presente estudo, a próxima seção apresenta os procedimentos metodológicos adotados, os protocolos de pesquisa e o formulário coleta elaborados visando atingir o objetivo proposto.

2 PROCEDIMENTOS METODOLÓGICOS

As Revisões Sistemáticas da Literatura são revisões baseadas na documentação das decisões do pesquisador e na adoção de um protocolo de pesquisa, buscando garantir a sua reprodutibilidade e auditabilidade. As RSLs permitem ainda que o *corpus* selecionado possa ser analisado com outras perspectivas, gerando novos resultados. A condução da RSL inicial teve como propósito identificar artigos que abordassem a qualidade de dados publicados como *Linked Data*. Entre os artigos aceitos foram identificados artefatos que auxiliam no processo de avaliação de dados *Linked Data*. Com base na observação de que existe uma grande diversidade entre os artefatos disponíveis para a avaliação de qualidade de dados *Linked Data*, identificou-se a necessidade de um aprofundamento maior a respeito desses artefatos, que são objeto de estudo da presente pesquisa.

Visando uma maior compreensão das etapas da pesquisa, os procedimentos metodológicos adotados foram divididos em: 1) Condução da RSL – onde estabeleceu-se *corpus* inicial e 2) Nova rodada de seleção e de coleta de dados,

focada em artigos com objetivo de discutir artefatos de avaliação de qualidade, baseada em formulário de extração.

O Quadro 1 apresenta uma versão resumida das informações utilizadas para estruturar a primeira etapa da pesquisa.

Quadro 1 - protocolo de pesquisa resumido (1° rodada)

Protocolo de busca	
Pergunta de pesquisa (principal)	Como tem sido abordada a questão da qualidade de dados no contexto do Linked Data?
Objetivos	Identificar os principais enfoques temáticos através dos quais se discute qualidade de dados publicados como Linked Data.
Estratégia de busca	("Linked Data" OR "Linked Open Data") AND ("Data Quality")
Bases de dados consultada	1ª rodada Web of Science; 2ª rodada ISTA; LISTA; 3ª rodada BRAPCI (busca simples pelo termo "Qualidade de dados")
Período abrangido	Sem restrição temporal.
Idiomas	Português, inglês e espanhol.
Critérios de Inclusão	(I) Foco principal é voltado para discutir qualidade de dados publicados de acordo com os princípios do Linked Data
Critérios de exclusão	(E) Não está nos idiomas estabelecidos para a pesquisa; (E) Apenas menciona a temática de interesse; (E) Não aborda a temática de interesse; (E) Não foi possível obter acesso ao documento completo;
Formulário de extração	Desafios de qualidade
Data da coleta	Entre dezembro de 2021 e maio de 2022

Fonte: Autores (2023)

Para a seleção das bases de dados, optou-se pelo uso da *Web of Science* por sua representatividade internacional e caráter interdisciplinar, também foram consideradas as principais bases temáticas da Ciência da Informação, incluindo a Base de Dados em Ciência da Informação (BRAPCI), por sua representatividade a nível nacional.

Na primeira rodada foram aceitos apenas documentos cujo objetivo principal foi promover discussões a respeito da qualidade de dados publicados de acordo com os princípios do *Linked Data*. Os resultados da estratégia de busca aplicadas nas diferentes bases de dados foram submetidos ao *State of the Art through Systematic Review* (StArt), [1]que auxilia na condução de Revisões Sistemáticas da Literatura, inclusive permitindo a identificação semiautomática de documentos duplicados.

O quadro 2 apresenta o protocolo da segunda rodada, realizada com a

[1] Disponível em: https://www.lapes.ufscar.br/resources/tools-1/start-1. Acesso em: 20 fev. 2024.

finalidade de identificar, no corpus de pesquisa estabelecido com base no quadro 1, artigos que discutem/apresentam artefatos disponíveis para avaliação de qualidade de dados *Linked Data*.

Quadro 2 - protocolo de busca da RSL (2° rodada)

Protocolo de busca	
Pergunta de pesquisa (principal)	Como tem sido abordados os artefatos para avaliação de qualidade de dados Linked Data na literatura disponível?
Pergunta de pesquisa (secundária)	Existem artefatos voltados para auxiliar na seleção de dados *Linked Data* para ligação?
Objetivos	Objetivo identificar e discutir os artefatos disponíveis para avaliação de qualidade de dados Linked Data.
Estratégia de busca	("Linked Data" OR "Linked Open Data") AND ("Data Quality")
Bases de dados consultada	1ª rodada Web of Science; 2ª rodada ISTA; LISTA; 3ª rodada BRAPCI
Período abrangido	Sem restrição temporal.
Idiomas	Português, inglês e espanhol.
Critérios de Inclusão	(I) Foco em apresentar um artefato para avaliação de qualidade de dados
Critérios de exclusão	(E) Foco principal na discussão de qualidade de dados publicados de acordo com os princípios do *Linked Data* (E) Não está nos idiomas estabelecidos para a pesquisa; (E) Apenas menciona a temática de interesse; (E) Não aborda a temática de interesse; (E) Não foi possível obter acesso ao documento completo;

Fonte: Autores (2024)

Na segunda rodada, foram considerados para o processo de seleção os documentos aceitos na primeira rodada. Esses documentos foram submetidos aos procedimentos do quadro 2, sendo aceitos apenas documentos que apresentavam ou discutiam um artefato para avaliação de qualidade de dados *Linked Data*. Para essa pesquisa também foi elaborado um formulário para coletar informações desses artefatos, no qual foi incluído os seguintes campos de coleta:

- Nome do artefato
- Contexto de sua criação, com histórico, responsáveis e motivação que embasam a sua necessidade de criação
- Tipo de artefato
- problema inicial que levou ao desenvolvimento do artefato
- Objetivos de criação do artefato
- Atividade que desempenha
- Forma como desempenha essa atividade

- Domínio ao qual é direcionado
- Público a que se destina
- Categorias, dimensões e critérios englobados pelo artefato
- Status (proposta, em elaboração, concluído, disponível para uso)
- Se foi realizada uma prova de conceito e em que contexto
- Possibilidade de customização (considerando inclusive se o artefato é aberto, gratuito ou livre)

Apresentados os principais procedimentos metodológicos que embasam a condução do presente estudo, a próxima seção apresenta os resultados da pesquisa.

3 RESULTADOS

Na primeira rodada de busca e seleção da RSL foram recuperados 225 documentos. Desses, 30 foram identificados como documentos duplicados, 100 documentos foram rejeitados e 95 artigos foram aceitos para compor o *corpus* teórico da pesquisa.

Os artigos então foram categorizados por meio dos enfoques dos estudos, levando em consideração seu objetivo, pergunta de pesquisa e resultados obtidos.

Para compor o corpus de coleta a respeito dos artefatos, foram considerados os artigos aceitos na RSL e incluídos na categoria "Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*".

O corpus foi submetido então a uma leitura flutuante, ou análise exploratória, onde se observou que os artefatos, embora tenham em comum a busca pela avaliação de qualidade de dados *Linked Data*, são diversos entre si, desde o objetivo até as atividades que realizam e como realizam essas atividades.

Procedeu-se então a etapa de exploração do material, onde os artefatos foram, por meio das informações fornecidas nos artigos, categorizados.

As categorias foram estabelecidas *a posteriori*, tendo como base o problema inicial que levou ao desenvolvimento do artefato, seus objetivos e a as atividades que desempenham, sendo sempre priorizado o objetivo central do artefato para a

categorização. Foram considerados na análise 75 artigos, que foram distribuídos em 11 categorias. O quadro 3 apresenta as categorias e o número de artigos incluído em cada categoria.

Quadro 3 - categorização dos artefatos de avaliação de qualidade de dados Linked Data

N° da categoria	Categoria	N° de artigos incluídos
1	Artefatos focados em vocabulários, ontologias, tesauros e metadados descritivos	13
2	Artefatos que são baseados em adaptação de técnicas de outros domínios	13
3	Artefatos focados em um domínio específico	13
4	Artefatos que buscam explorar características estruturais de dados *Linked Data*	12
5	Artefatos focados em categorias e dimensões específicas	9
6	Artefatos voltados para seleção e curadoria de fontes para a ligação	4
7	Artefatos para avaliação geral de qualidade de dados *Linked Data*	4
8	Artefatos voltados para a conversão de dados legados	3
9	Artefatos voltados para a identificação e solução de conflito de objetos	2
10	Artefatos que realizam a adaptação de um outro artefato já existente	1
11	Artefatos focados na Integração de dados	1

Fonte: Autores (2024)

A análise do quadro permite observar que existe um destaque para artefatos de representação, como vocabulários, ontologias, tesauros e metadados descritivos. Dentro dessa categoria os vocabulários são abordados de três maneiras distintas: 1) como um aspecto da qualidade – onde o objetivo do artefato é verificar a aplicação correta dos vocabulários no processo de formação de triplas em RDF 2) Como um meio para checagem de outros aspectos de qualidade – onde esses vocabulários e metadados fornecem informações a respeito dos conjuntos de dados, que podem ser exploradas na realização do processo de avaliação e 3) Como um meio para compartilhamento formal dos resultados do processo de avaliação de qualidade – o que permite que esses resultados sejam reaproveitados e processados com maior facilidade por usuários máquina.

Também se destacam os artefatos que buscam a adaptação de técnicas já consolidadas em outros domínios para viabilizar o processo de avaliação, sendo exemplos de técnicas: o *test-drive* proveniente da engenharia de *software*, uso de

técnicas de *machine learning*, Aplicação de técnicas de *data profile*, exploração das possibilidades de aplicação de *crowdsourcing,* exploração da *blockchain* e etc.

Outro destaque entre as categorias é a criação de artefatos para atender às necessidades de domínios específicos, com dados provenientes de áreas médicas e da saúde, dados relacionados a contextos biológicos, geoespaciais e do domínio enciclopédico.

Também foram muito explorados os aspectos estruturais de dados *Linked Data*, onde os artefatos buscam avaliar problemas de qualidade que ocorrem por conta das especificidades estruturais de dados publicados como *Linked Data* (como a utilização do RDF e a estrutura de *links*) ou se utilizam dessas características para a criação do artefato, como na utilização do SPARQL para identificar problemas de qualidade ou da propriedade owl:sameAs para identificar anomalias.

Apresentados e discutidas as principais categorias de artefatos elaborados para a avaliação de qualidade de dados *Linked Data*, a próxima seção apresenta as considerações finais.

4 Considerações Finais

Essa pesquisa consistiu em um recorte de uma Revisão Sistemática da Literatura, e teve como finalidade identificar e discutir os artefatos disponíveis para avaliação de qualidade de dados *Linked Data*.

Por meio de uma análise dos objetivos e das atividades desempenhadas por esses artefatos foi possível categorizá-los. Dentre essas categorias observou-se que se destacam os artefatos relacionados com vocabulários, ontologias, tesauros e metadados descritivos; baseados em adaptação de técnicas de outros domínios; focados em um domínio específico e ainda artefatos que buscam explorar características estruturais de dados *Linked Data*.

Destaca-se a importância dos vocabulários semânticos no processo de avaliação de qualidade de dados *Linked Data*, mesmo quando os vocabulários não são o foco principal dos artefatos, parte dos artefatos automáticos e

semiautomáticos se utilizam de vocabulários semânticos em alguma etapa de seu funcionamento, seja para agilizar o processo de análise, coletando metadados que auxiliem no processo de avaliação, seja para exportar os relatórios gerados.

Devido à diversidade dos artefatos existentes e das diferentes contribuições que podem agregar ao processo de seleção de fontes e de gestão de dados *Linked Data*, entende-se que a seleção desses artefatos é desafio que precisa ser agregado aos processos de ciclo de vida de dados *Linked Data*.

O presente estudo também foi pautado na pergunta de pesquisa: Existem artefatos voltados para auxiliar na seleção de dados *Linked Data* para ligação?

Até o presente momento não foi identificado nenhum artefato elaborado especificamente para essa demanda. Entretando, destaca-se a existência de artefatos incluídos nas categorias: "Seis (6) - Artefatos voltados para seleção e curadoria de fontes para a ligação" e "Sete (7) -Artefatos para avaliação geral de qualidade de dados *Linked Data*". A primeira por sua proximidade com essa finalidade e segunda pela possibilidade de serem adaptados para os objetivos da pesquisa.

Como estudos futuros pretende-se explorar o funcionamento desses artefatos com base nas atividades que desempenham e na forma como as desempenham. Pretende-se ainda se aprofundar no potencial de adaptação dos artefatos das categorias seis e sete para atuarem na seleção de dados *Linked Data* para ligação, levando em consideração o processo de avaliação de qualidade.

AGRADECIMENTOS

Agradecemos à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento recebido para o desenvolvimento dessa pesquisa. N° 2021/03349-0

REFERÊNCIAS

ASSAF, Ahmad; SENART, Aline; TRONCY, Raphaël. Towards an objective assessment framework for linked data quality. *International Journal On Semantic Web And Information Systems*, [*s.l.*], v. 12, n. 3, p. 111-133, jul. 2016.Disponível em: http://dx.doi.org/10.4018/ijswis.2016070104. Acesso em: 27 maio 2022.

BERNERS-LEE, Tim. *Linked data*. [*S.l.*], 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 14 mar. 2021.

LANGER, André. *et al*. SemQuire: assessing the data quality of linked open data sources based on dqv. *Current Trends in Web Engineering*, [*s.l.*], p. 163-175, 2018. Disponível em: http://dx.doi.org/10.1007/978-3-030-03056-8_14. Acesso em: 17 ago. 2022.

MELO, Jessica Oliveira de Souza Ferreira. *Metodologia de avaliação de qualidade de dados no contexto do linked data*. 2017. 111 f. Dissertação (Mestrado) - Curso de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências de Marília, Universidade Estadual Paulista, Marília, 2017.

WANG, Richard; STRONG, Diane. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, [*s.l.*], v. 12, n. 4, p 5–33, jan. 1996. Disponível em: http://www.jstor.org/stable/40398176?origin=JSTOR-pdf. Acesso em: 18 ago. 2023.