XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT-8 – Informação e Tecnologia

A QUESTÃO DA QUALIDADE EM DADOS PUBLICADOS COMO LINKED DATA: UM MAPEAMENTO SISTEMÁTICO DA LITERATURA

THE ISSUE OF QUALITY IN DATA PUBLISHED AS LINKED DATA: A SYSTEMATIC MAPPING OF THE LITERATURE

Ananda Fernanda de Jesus. UNESP.

José Eduardo Santarem Segundo. UNESP.

Modalidade: Trabalho Completo

Resumo: A qualidade de dados é intrínseca à capacidade de atuação satisfatória nas atividades ou aplicações nas quais esses dados vão ser empregados, podendo ser avaliada através de dimensões e métricas específicas para cada domínio. Os dados disponibilizados de acordo com o *Linked Data* seguem um conjunto de princípios que visam a sua publicação estruturada e conectada no ambiente *Web*, entretanto, esses também são afetados por questões de qualidade. Nesse sentido, o presente trabalho objetiva compreender os principais enfoques temáticos por meio dos quais se discute a qualidade de dados publicados como *Linked Data*. Realizou-se um Mapeamento Sistemático da Literatura, no qual foram recuperados 89 artigos. Esses artigos foram agrupados em três categorias temáticas, sendo elas: 1) Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*; 2) Realiza um estudo de avaliação de qualidade em um ou mais *datasets*; 3) Levantamentos e estudos teóricos sobre qualidade de dados e *Linked Data*. Conclui-se que a maioria dos artigos tem como foco a elaboração e a discussão de artefatos que permitam avaliar e promover melhorias em diversos aspectos da qualidade de dados publicados de acordo com os princípios do *Linked Data*. Conclui-se ainda que são poucos os estudos cujo foco é promover discussões teóricas aprofundadas sobre a temática.

Palavras-Chave: Qualidade de Dados. Web Semântica. Linked Data.

Abstract: The data quality is intrinsic to the ability to perform satisfies the activities or applications in which the data are employed and can be assessed through specific dimensions and metrics for each domain. The data published in accordance with the Linked Data follows a set of principles that aim at its structured and connected publication in the Web environment, however, these data are also affected by quality questions. In this sense, the purpose of this work is understanding the main thematic approaches through which data quality Linked Data is discussed. For this, a Systematic Mapping of Literature was carried out, a total of 89 studies were identified. These articles are grouped into three thematic categories, being them: 1) Propose an artifact to assess or improve data quality published as Linked Data; 2) Perform a quality assessment study on one or more datasets; 3) Surveys and theoretical studies on data quality and Linked Data. It is concluded that most of the articles are focused on the elaboration and discussion of artifacts to assess and promote improvements in various aspects of the quality of data published in accordance with the principles of Linked Data. It is also

concluded that there are few studies whose main focus is to promote in-depth theoretical discussions on the subject.

Keywords: Data Quality. Semantic Web. Linked Data.

1 INTRODUÇÃO

No contexto tecnológico vigente o funcionamento eficaz e eficiente dos mais diversos setores da sociedade é permeado pela utilização estratégica de dados, tanto no processo de tomada de decisão como em atividades operacionais. A qualidade desses dados, ou seja, a sua capacidade de atuarem de maneira satisfatória nas atividades ou aplicações nas quais são empregados, tem impacto na vida cotidiana, no funcionamento e no faturamento das organizações. (BATINI; SCANNAPIECO, 2016).

Uma proposta para tornar os dados compreensíveis e aplicáveis tanto para usuários humanos como para agentes computacionais é o *Linked Data,* que consiste na 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).

Mesmo baseados em uma estrutura clara, os dados publicados como *Linked Data* não estão livres de questões relacionadas à qualidade. Considerando as especificidades dos dados publicados como *Linked Data*, essa pesquisa foi delineada em torno do seguinte questionamento: De que forma tem sido discutida a qualidade de dados no contexto do *Linked Data*? O objetivo da pesquisa é identificar os principais enfoques temáticos por meio dos quais se discute qualidade de dados publicados como *Linked Data*.

Para atingir o objetivo proposto realizou-se um Mapeamento Sistemático da Literatura (MSL), cujos procedimentos são apresentados na seção 3. Na próxima seção discute-se a relação entre qualidade de dados e o *Linked Data*.

2 A QUALIDADE EM DADOS PÚBLICADOS COMO *LINKED DATA*

A preocupação com a questão da qualidade dos dados não é algo recente, remetendo à década de 1970, com a busca por estabelecer formas de analisar a qualidade dos dados (LANGER; SIEGERT; GÖPFERT; GAEDKE, 2018). As questões de qualidade de dados atualmente

afetam muitas áreas da sociedade, tendo influência em atividades operacionais de instituições e empresas, nos lucros e no funcionamento eficiente dessas organizações.

Fisher e Kingma (2001) apresentam exemplos onde a qualidade resultaram na perda de vidas humanas. Um exemplo é o caso do lançamento do ônibus espacial *Challenger*, onde graves problemas de qualidade na base de dados da *National Aeronautics and Space Administration (NASA*) levaram à explosão do ônibus espacial, a um prejuízo de mais de um bilhão de dólares e a morte de sete tripulantes.

O conceito de qualidade de dados pode ser abordado por meio de perspectivas distintas, sendo elas: intrínseca, contextual e representacional. (NOOGHABI; DASTGERDI, 2016). Em uma perspectiva intrínseca, a qualidade de um conjunto de dados (*datasets*) está relacionada diretamente com as características desses dados, preocupando-se com o quão livre de erros eles estão, independentemente do contexto no qual serão aplicados. Na perspectiva contextual, a qualidade também leva em consideração as características dos dados, mas o foco está nas necessidades dos usuários, nas tarefas e aplicações que se pretende realizar com esses dados.

A perspectiva representacional, como apontado por Nelson, Todd e Wixom (2005), discute a apresentação de informações necessárias para a interpretação, compreensão e aplicação dos dados. Na literatura, qualidade de dados é frequentemente descrita por meio do conceito de "*fitness for use*", ou seja, em uma perspectiva contextual, onde para serem considerados de qualidade os dados devem atender satisfatoriamente a demanda para determinada aplicação. (JURAN, 1988; WANG; STRONG, 1995). Nesse sentido, um conjunto de dados pode ter qualidade suficiente para determinadas aplicações, enquanto que seus problemas de qualidade podem comprometer sua aplicação em outros contextos. (ZAVERI; RULA; MAURINO; PIETROBON; LEHMANN; AUER, 2015).

Wang e Strong (1996) organizam o processo de avaliação em categorias, dimensões de qualidade e critérios de qualidade. As categorias refletem as perspectivas mencionadas anteriormente, os autores acrescentam a categoria acessibilidade dos dados, cujo foco é a capacidade de acesso dos mesmos. Dimensão é entendida pelos autores como um aspecto mais abrangente de características dos dados, cada dimensão é composta por um conjunto de critérios que descrevem a qualidade dos dados com base em um atributo específico.

Alguns critérios podem ser mensurados quantitativamente, enquanto outros necessitam de uma análise qualitativa.

Os critérios também diferem no aspecto de sua avaliação, onde alguns podem ser avaliados objetivamente enquanto outros necessitam de uma análise subjetiva, especialmente os que dependem da percepção do usuário em relação a algo. Para mensurar a qualidade dos dados em relação a um critério são elaborados indicadores, quantitativos ou qualitativos, denominados de métricas. (WENG; STRONG, 2013; ASSAF; SENART; TRONCY, 2016; FÄRBER; BARTSCHERER; MENNE; RETTINGER, 2017; MELO, 2017).

Não é de interesse do presente estudo apresentar a descrição aprofundada de cada critério e métrica aplicado no processo de avaliação de qualidade, bastando a compreensão de que cada uma das dimensões apresentadas a seguir possuí um conjunto de critérios e que cada critério pode ser avaliado por mais de uma métrica, sendo elas quantitativas, qualitativas ou quali-quantitativas.

Färber, Bartscherer, Menne e Rettinger (2017) realizaram uma adaptação das dimensões apresentadas por Wang e Strong (1996) ao contexto do *Linked Data*. Com base em Färber, Bartscherer, Menne e Rettinger (2017) e Wang e Strong (1996), o Quadro 1 apresenta a sistematização dessas dimensões, organizadas de acordo com as categorias de qualidade.

Quadro 1 – Dimensões de qualidade organizadas com base em categorias de qualidade

Categoria	Dimensão	Definição
Intrínseca Qualidade inerente dos dados, não influenciável pelo cenário de aplicação ou pelas necessidades de seus potenciais usuários.	Precisão	Verifica o quão adequados estão os dados em relação a sintática e a semântica, ou seja, quanto às regras do esquema seguido e a validade semântica dos valores atribuídos.
	Confiabilidade	Avalia a credibilidade, reputação, objetividade e verificabilidade do conjunto de dados.
	Consistência	Caracterizada pela ausência de valores conflitantes, podendo estar relacionada aos valores dos triplos ou a sua estrutura, sendo dividida em inconsistências de classe e de relacionamento. Quando relacionada aos valores, é subjetiva, considerando diferentes níveis de conhecimento e diferentes visões do mundo.
Contextual	Relevância	Verifica em que medida os dados são úteis e atendem satisfatoriamente as necessidades da tarefa na qual serão aplicados.

Relativa à percepção do usuário e a adequação dos dados para determinada aplicação	Completude	Verifica se os dados possuem a granularidade necessária e estão dentro do escopo correto.
	Temporalidade	Verifica se os dados não estão obsoletos para a atividade que será realizada, não sendo medida pela data de criação do *dataset* e sim pela sua última atualização.
Representacional Diz respeito a presença de informações de interesse relativas aos dados	Facilidade de compreensão	Verifica a capacidade de compreensão dos dados por usuários humanos, sendo claros e livres de ambiguidades
	Interoperabilidade	Verifica a facilidade de compreensão e troca de informações entre agentes computacionais
Acessibilidade Diz respeito a como os dados podem ser acessados	Acessibilidade	Verifica se os dados estão disponíveis, podem ser obtidos e acessados tanto por usuários humanos como agentes computacionais. Se divide em abordagens objetivas e subjetivas.
	Licença	Verifica a provisão de informação de licença legível por usuários humanos e agentes computacionais
	Interlinking	Verifica em que medida as entidades que representam um mesmo conceito estão conectadas, dentro de um mesmo *dataset* ou promovendo a ligação com fontes externas

Fonte: Autores (2022)

A avaliação de dados publicados como *Linked Data* demanda o uso de dimensões específicas, tendo em vista que estes se diferem dos demais dados tanto em seu contexto de publicação como em suas estruturas.

Com o objetivo de tornar os dados publicados na *Web* mais adequados às necessidades dos usuários foi elaborado pelo *W3C* (consorcio responsável pelo desenvolvimento da *Web)* um guia para a publicação de dados como *Linked Data*, composto por 10 Melhores Práticas (MPs) (W3C, 2014). As MPs orientam todo o processo de publicação de dados como *Linked Data*, abordando a preparação das partes interessadas, seleção dos dados a serem convertidos, do modelo de dados, da licença, do vocabulário a ser adotado, o estabelecimento de bons *URIs*, o processo de conversão dos dados e sua disponibilização ao público.

Também foi elaborado pelo W3C (2017) 35 MPs para a publicação de dados na *Web*. Essas MPs são mais detalhadas e focadas na publicação geral de dados na *Web*, embora direcionem-se para uma estrutura de dados adequadas ao *Linked Data.* Uma das 35 MPs, a MP6, está voltada para a qualidade de dados, entretanto o foco dessa Melhor Prática é garantir o registro de informações de qualidade, e não fornecer instrumentos para a realização

da avalição ou apresentar dimensões e métricas para esse processo. O documento direciona ao vocabulário criado pelo W3C (2016), estruturado para viabilizar o registro formal de qualidade dos dados, o *VOCAB-DQV*. Esse vocabulário fornece propriedades e classes para que as dimensões de qualidade de um *dataset* possam ser registradas.

Embora ambos os documentos reconheçam a importância da qualidade de dados para garantir a reutilização dos mesmos, e que o seguimento das orientações impacte em questões de qualidade, nenhum dos guias é focado especificamente em qualidade dados, não sendo indicadas, por exemplo, dimensões e métricas que embasem o processo de avaliação.

Além das recomendações focadas na estrutura dos dados, também existe a preocupação com questões de acesso e uso desses dados, visando sua disponibilização de maneira aberta, ou seja, sem barreiras legais de acesso e reuso. Desde sua criação, destaca-se o projeto da *Linked Open Data Cloud (LOD Cloud)*, [1]plataforma que reúne dados abertos publicados como *Linked Data*. Em 2007 a *LOD Cloud* contava com 12 datasets, atualmente a nuvem conta com 1301 datasets. Esses dados são provenientes de diversos domínios, como dados governamentais, provenientes de *crowdsourcing* como a *DBpedia*, de bibliotecas e diversas fontes da saúde. (LOD CLOUD DIAGRAM, 2022).

A velocidade com que foram publicados esses *datasets* é um dos complicadores para a avaliação de sua qualidade. Para além da quantidade, uma grande preocupação, apontada na literatura, é heterogeneidade dessas fontes. Os dados advêm de diversos domínios, sendo publicados com objetivos diferentes.

As próprias características da publicação de dados na *Web*, como um ambiente aberto e complexo, criam desafios para avaliação de dados *Linked Data*, tendo em vista que usuários não especializados podem criar, publicar e recuperar dados, levando a problemas como imprecisão, falta de confiança nas fontes, desatualização dos dados, informações inconsistentes, incompletas ou mal interpretadas (ASSAF; SENART; TRONCY, 2016; HADHIATMA, 2017).

Outra preocupação é com a forma de criação desses dados, que podem ter sido criados de acordo com os princípios, ou podem ser provenientes da conversão de dados legados, tendo como fonte original dados não estruturados ou semiestruturados. Debattista, Auer e Lange (2016) apontam que não é incomum o cenário onde os *datasets* são provenientes de

[1] Disponível em: https://lod-cloud.net/

dados convertidos de forma automática de fontes semiestruturadas, e que os dados provenientes desse processo são mais predispostos a inconsistências, informações falsas e incompletas.

Um conjunto de dados publicados com uma preocupação menor de qualidade pode ser perfeitamente adequado a uma determinada aplicação, entretanto podem causar problemas caso aplicados em outro contexto. Assaf, Senart e Troncy (2016) trazem como exemplo o caso da *DBpedia*, dados criados através de colaboração entre usuários, e convertidos de maneira automática de fonte semiestruturada. Os autores afirmam que esses dados são adequados para enriquecer resultados de busca na *Web*, mas podem causar problemas quando aplicados em cenários mais críticos, como aplicações na área da saúde.

Assim sendo, a avaliação desses dados de maneira intrínseca, apenas com base em suas características, não é suficiente, tornando o processo de avaliação ainda mais complexo. Para além da heterogeneidade das fontes, dos domínios e das diferentes necessidades de qualidade dos usuários, outra questão relacionada a qualidade de dados *Linked Data* é a sua estrutura.

Ahmed (2017) aponta que as questões podem se concentrar em três vertentes: 1) esquema ou ontologia e vocabulário 2) recurso ou entidade 3) dado ou instância. Por advirem de domínios distintos esses dados requerem ontologias e vocabulários diferentes, em muitas situações vocabulários de criação própria, podendo levar a problemas de integração. De acordo com Hadhiatma (2017), tanto o recurso como as instâncias podem estar incorretos, com dados ruidosos, e incompletos, contendo dados ausentes.

Além desses três aspectos soma-se a questão dos *links* do *Linked Data*. Monika Rani; Sapna e Mishra (2018) indicam a existência de *links* inadequados como um problema recorrente em *datasets Linked Data***.** Hadhiatma (2017) fala que as relações entre os conjuntos de dados podem estar incorretas ou incompletas, tornando necessário ponderar se esses *links* são apropriados e úteis.

Haller *et al.* (2020) discutem que embora a estrutura de *links* seja o grande diferencial da proposta do *Linked Data*, ela também remete a uma série de questões como: referência a *URIs* inacessíveis, redundância ao copiar os dados do conjunto externo e não apenas promover ligação, alterações nos conjuntos de dados externos estão fora de controle do usuário.

Dados provenientes de fontes externas podem passar por atualizações sem aviso prévio, comprometendo assim a qualidade dos dados internos, fazendo com que a avaliação seja um processo continuo e não uma ação única. A própria seleção dos *datasets* para realizar a ligação leva a uma série de questões de qualidade.

Discutida a relação entre qualidade de dados e *Linked Data* a próxima seção apresenta os procedimentos metodológicos para realização de um mapeamento sistemático da literatura.

2 PROCEDIMENTOS METODOLÓGICOS

O presente trabalho resulta de uma pesquisa exploratória e descritiva, com resultados quantitativos e qualitativos, realizada através de um Mapeamento Sistemático da Literatura. Para Kitchenham, Pretorius, Budgen, Brereton, Turner, Niazi e Linkman (2010) o MSL é um método aplicado à levantamentos bibliográficos, com abordagem ampla, pautado em um protocolo de busca e no registro do processo decisório do pesquisador, tendo como principal objetivo classificar os estudos de uma temática específica em categorias bem definidas, que permitam um olhar estruturado em busca de assuntos comuns e assuntos pouco explorados.

O MSL foi realizado através das seguintes etapas: (1) Planejamento: Preenchimento do protocolo de busca que irá orientar a pesquisa; (2) Execução: Busca nas bases de dados e seleção dos documentos; e (3) Sumarização: agrupamento dos documentos por semelhança, criação de categorias para classificação dos resultados, sistematização das informações de interesse em imagens e quadros-resumo com resultados quantitativos e qualitativos. O Quadro 2, apresentado a seguir, apresenta as informações utilizadas para estruturar o levantamento:

Quadro 2 – protocolo de busca

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	1º rodada Web of Science 2º rodada ISTA; LISTA; 3° rodada BRAPCI
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	Enfoque do documento; Tipo de artefato proposto, Característica dos artefatos.
Data da coleta	entre dezembro de 2021 e maio de 2022

Fonte: Autores (2022)

As buscas na Base de Dados em Ciência da Informação (BRAPCI) foram realizadas por meio de busca manual, pelo termo "Qualidade de dados", seguida da aplicação dos critérios de exclusão. Os documentos recuperados foram sistematizados em formato de planilha, onde foram identificadas as duplicatas e aplicados os critérios de exclusão. Após a seleção, realizou-se a leitura dos documentos aceitos, agrupando-os em categorias temáticas construídas a *posteriori*, de acordo com a identificação de padrões nas temáticas dos documentos. A coleta dos enfoques temáticos dos documentos foi realizada através de uma análise de seus objetivos, metodologias e resultados obtidos.

3 A QUALIDADE DE DADOS E O *LINKED DATA*: RESULTADOS DO MSL

A presente seção apresenta os resultados obtidos com base nos procedimentos metodológicos discutidos na seção anterior. Primeiro serão apresentados os resultados quantitativos do processo de busca e seleção, e em seguida uma discussão mais aprofundada dos documentos aceitos, através da discussão dos enfoques temáticos desses documentos.

Considerando as três rodadas de busca realizadas foram recuperados 225 documentos, dos quais 30 (trinta) eram duplicados. Foram rejeitados 100 (cem) documentos, dos quais 42 (quarenta e dois) foram excluídos com base no critério "(E) Apenas menciona a temática de interesse". Como a questão da qualidade de dados perpassa de muitas formas a publicação de dados, boa parte dos estudos recusados mencionam a qualidade de dados como uma etapa dos processos de publicação e reutilização de *datasets Linked Data*, ou ainda citam a existência da preocupação com a qualidade, entretanto esse não foi o enfoque principal dos artigos. Foram recuperados ainda artigos que abordam qualidade de dados mencionando o contexto do *Linked Data* como um dos diversos contextos que ressaltam a preocupação com a qualidade, junto a outras temáticas como *Big Data*, Internet das Coisas e *Web* Semântica.

Ao final da seleção, foram aceitos 89 (oitenta e nove) artigos para compor o *corpus* teórico da pesquisa, desses artigos 80 (oitenta) advém da base multidisciplinar e apenas 8 (oito) são provenientes de bases temáticas internacionais da CI e 1 (um) proveniente da base temática nacional. [2]

Com base na leitura dos objetivos, dos procedimentos metodológicos e dos resultados dos artigos aceitos foram extraídos os enfoques temáticos desses documentos. Esses enfoques foram agrupados em um quadro-resumo, que permitiu a identificação de padrões e a elaboração de categorias temáticas. A relação entre as categorias elaboradas e o número de artigos incluído em cada categoria é apresentada no Quadro 3.

Quadro 3 – número de artigos por categorias temáticas

N°	Categoria	Quantidade de artigos incluídos
1	Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*	73
2	Realiza um estudo de avaliação de qualidade em um ou mais *datasets*	15
3	Levantamentos e estudos teóricos sobre qualidade de dados e *Linked Data*	8

Fonte: autores (2022)

As categorias foram apresentadas no Quadro 2 em ordem decrescente de número de artigos incluídos, sendo, portanto, a mais volumosa delas a categoria "1 - Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*." Essa categoria é composta pela maior parte dos artigos (82.02%) e nela foram incluídos todos os artigos que tinham por objetivo apresentar, propor ou discutir um artefato [3].

Os artefatos propostos nessa categoria divergem em muitos aspectos, algumas dessas divergências são descritas no quadro 4.

[2] Em casos em que a referência foi recuperada tanto na WoS como nas bases da CI optou-se por marcar como duplicatas os artigos da WoS, para um melhor panorama dos artigos aceitos relacionados à CI

[3] Para essa pesquisa um artefato é entendido enquanto instanciação física elaborada com a finalidade de representar determinada realidade, solucionar problemas ou promover melhorias/inovação, podendo se materializar em modelos, sistemas de informação, fluxos, *constructus*, métodos, entre outras formas (MARCH; SMITH, 1995; HEVNER et al., 2004).

Quadro 4 – características dos artefatos recuperados

Característica dos artefatos	Descrição
Tipo de artefato	Foram identificados entre as propostas artefatos como *frameworks*; *softwares*; métodos baseados em comparação probabilística de semelhança; ontologias; instrumentos baseados em inteligência artificial; modelos de dimensões e métricas; instrumentos baseados em *crowdsourcing;* guias para avaliação e melhoria de qualidade, fluxos de trabalho; métodos estatísticos para análise de métricas; metodologias complexas (compostas por mais de um artefato);
Atividade a ser realizada	Em relação às atividades esses instrumentos divergem especialmente entre três vertentes: avaliação – instrumentos que se propõe a identificar problemas de qualidade, mensurar os níveis de qualidade ou ainda comparar os níveis de qualidade entre dois ou mais *datasets*; promover melhorias – ferramentas voltadas para a correção de problemas específicos de qualidade de dados; e ainda ferramentas de avaliação e melhorias – que são ferramentas que desempenham as duas etapas, gerando relatórios de qualidade e corrigindo os problemas identificados
Finalidade do artefato	Foram identificados artefatos que variam entre finalidades gerais, ou seja, serem aplicados a todos os tipos de *datasets Linked Data*, visando a avaliação e melhorias de qualidade em uma serie de dimensões e os mais específicos – aqueles criados visando atender a dados provenientes de um domínio específico ou a avaliação e melhoria focada em dimensões específicas;
Forma como desempenha a atividade	Em relação a forma de desempenho dos artefatos segue-se as categorias mencionados por Acosta, Zaveri, Simperl, Kontokostas, Flöck e Lehmann (2015), que classificam os artefatos voltados para a qualidade em automáticos – que desempenham avaliação e melhorias de qualidade sem interferência humana; semiautomáticos – que realizam parte do processo de maneira automática mas necessitam de decisão humana em determinada etapa, geralmente na avaliação final; e ainda os manuais – guias para orientar a avaliação através de atividade de pessoas especializadas ou de usuários
Público a que se destina	Foram identificadas ferramentas voltadas para os publicadores de dados – ou seja para que esses realizem uma avaliação dos próprios dados e possam promover melhorias de qualidade visando atender de maneira mais satisfatória a demanda de seus consumidores; voltadas para os consumidores de dados – ferramentas pensadas para auxiliar na

	seleção de *datasets LD* tanto para aplicações como para ligação entre *datasets*

Fonte: Autores (2022)

A categoria "2 Levantamento sobre qualidade de dados e *Linked Data*" reúne estudos com a proposta de avaliar os níveis de qualidade de um ou mais *datasets*, aplicando dimensões e métricas. Foram identificados estudos voltados para avaliar apenas um *dataset*, como o ocorreu no estudo de Font, Zouaq e Gagnon (2015) onde se avaliou os dados provenientes do *DBpedia*.

Foram identificados estudos que avaliaram dois ou mais *datasets*, com a finalidade de realizar comparações entre eles e ainda estudos voltados para avaliação de *datasets* provenientes de domínios específicos como dados governamentais.

Parte desses estudos também realizaram uma checagem mais extensiva de *datasets*, como ocorreu no estudo de Debattista, Auer e Lange (2016) onde foram avaliados 130 *datasets* provenientes da *Linked Open Data Cloud*, os autores utilizaram 27 métricas de qualidade voltadas para o contexto do *Linked Data*, chegando à conclusão de que a avaliação de dados *Linked Data* não pode ocorrer de maneira pontual, devendo ser um processo contínuo. Também ocorreram casos onde os estudos focaram na avaliação de apenas algumas dimensões.

Já a última categoria, "3-Levantamentos e estudos teóricos sobre qualidade de dados e *Linked Data*" é uma categoria mais diversa que reúne diferentes levantamentos bibliográficos, como o exemplo do artigo de Zaveri, Rula, Maurino, Pietrobon, Lehmann e Auer (2015), que realizou um levantamento sistemático sobre métodos de melhoria de qualidade. Nessa categoria também foram incluídos levantamentos com abordagem mais específica, como o artigo de Possemato (2018) que discute o papel do *Resource Description and Access (RDA)*, um instrumento de padronização proveniente do domínio bibliográfico, na qualidade de dados publicados como *Linked Data*.

3 CONSIDERAÇÕES FINAIS

O MSL dessa pesquisa foi conduzido com o intuito de identificar os principais enfoques temáticos através dos quais se discute qualidade de dados publicados como *Linked Data* e ainda como essa temática tem sido discutida em artigos científicos indexados em bases temáticas da Ciência da Informação.

Em relação à análise dos enfoques temáticos, conclui-se que o principal enfoque está na elaboração e discussão de artefatos que permitam avaliar e promover melhorias em diversos aspectos da qualidade de dados publicados de acordo com os princípios do *Linked Data.* Esses artefatos são plurais em muitos aspectos, o que torna necessária uma avaliação mais aprofundada da proposta de cada artefato, visando permitir a seleção do artefato mais adequado para cada tarefa e ainda identificação de necessidade de criação/atualização desses artefatos para que atendam a uma demanda mais específica.

Observa-se que são poucos os estudos que visam promover discussões teóricas aprofundadas sobre as muitas facetas da qualidade em dados *Linked Data*, que se aprofundem em discussões sobre as especificidades desses dados e seus principais problemas de qualidade. Também são poucos os estudos que visam discutir a questão por meio de uma perspectiva de um domínio específico.

Em relação aos artigos recuperados em bases temáticas da CI, observa-se que ainda são recuperados poucos artigos voltados para essa temática, que boa parte desses possuem um caráter mais especifico e teórico, com estudos que buscam discutir a temática na perspectiva dos dados bibliográficos; das bibliotecas, arquivos e museus e ainda de identificar a influência de instrumentos de padronização provenientes da CI na qualidade de dados *Linked Data*.

Identificada a lacuna de discussões teóricas aprofundadas sobre as muitas facetas da qualidade em dados *Linked Data*, em especial por meio da perspectiva da Ciência da informação, como estudos futuros pretende-se apresentar um aprofundamento teórico do assunto, com destaque para discussão sobre as diferentes perspectivas de qualidade (intrínseca, contextual, representacional e acessibilidade dos dados), relacionando essas perspectivas com as e as contribuições potencias da Ciência da Informação pra o desenvolvimento das mesmas.

Pretende-se ainda analisar e discutir os artefatos identificados nesse Mapeamento, categorizando-os quanto ao tipo de artefato, atividade a ser realizada, finalidade do artefato, forma como desempenha a atividade, público a que se destina, dimensões e métricas abordadas.

REFERÊNCIAS

ACOSTA, M.; ZAVERI, A.; SIMPERL, E.; KONTOKOSTAS, D.; FLÖCK, F.; LEHMANN, J. Detecting Linked Data quality issues via crowdsourcing: a dbpedia study. **Semantic Web**, [s.l], v. 9, n. 3, p. 303-335, 12 abr. 2018. Disponível em: http://dx.doi.org/10.3233/sw-160239. Acesso em: 16 ago. 2022.

AHMED, H. H. Data quality assessment in the integration process of linked open data (lod). **International Conference On Computer Systems And Applications**, [*s.l.*], v. 1, n. 14, p. 1-6, out. 2017. Disponível em: http://dx.doi.org/10.1109/aiccsa.2017.178. Acesso em: 27 maio 2022.

ASSAF, A.; SENART, A.; TRONCY, R. 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, T. **Linked data**, 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 14 mar. 2021.

BATINI, C.; SCANNAPIECO, M. **Data quality: concepts, methodologies and techniques**. Berlin: Springer, 2006. 520 p.

DEBATTISTA, J.; AUER, S.; LANGE, C. Luzzu: a framework for linked data quality assessment. **Tenth International Conference On Semantic Computing**, [*s.l.*], p. 124-131, fev. 2016. Disponível em: http://dx.doi.org/10.1109/icsc.2016.48. Acesso em: 27 maio 2022.

FÄRBER, M.; BARTSCHERER, F.; MENNE, C.; RETTINGER, A. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. **Semantic Web**, [*s.l.*], v. 9, n. 1, p. 77-129, nov. 2017. Disponível em: http://dx.doi.org/10.3233/sw-170275. Acesso em: 26 maio 2022.

FISHER, C. W.; KINGMA, B. R. Criticality of data quality as exemplified in two disasters. **Information & Management**, [*s.l.*], v. 39, n. 2, p. 109-116, 2001.

FONT, L,.; ZOUAQ, A.; GAGNON, M. Assessing the quality of domain concepts descriptions in dbpedia. **Th International Conference On Signal-Image Technology & Internet-Based Systems,** [*s.l.*], v. 11, n. 1, p. 254-261, nov. 2015. Disponível em: http://dx.doi.org/10.1109/sitis.2015.104. Acesso em: 26 maio 2022.

HADHIATMA, A. Improving data quality in the linked open data: a survey. **Journal Of Physics**: Conference Series, [*s.l.*], v. 978, p. 12-26, mar. 2018. Disponível em: http://dx.doi.org/10.1088/1742-6596/978/1/012026. Acesso em: 26 maio 2022.

JURAN, J. M. **Quality Control Handbook**. New York: Mcgraw-Hill. 1988. 500 p.

KITCHENHAM, B.; PRETORIUS, R.; BUDGEN, D.; BRERETON, O. P.; TURNER, M.; NIAZI, M.; LINKMAN, S. Systematic literature reviews in software engineering: tertiary study.

Information And Software Technology, [*s.l*], v. 52, n. 8, p. 792-805, ago. 2010. Disponível em: http://dx.doi.org/10.1016/j.infsof.2010.03.006. Acesso em: 19 ago. 2022.

LANGER, A.; SIEGERT, V.; GÖPFERT, C.; GAEDKE, M. 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.

LOD CLOUD DIAGRAM. **About**. Disponível em: https://lod-cloud.net/#about. Acesso em: 26 maio 2022.

MELO, J.O.S F.**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.

MONIKA RANI, G.; SAPNA, R; MISHRA, S. An investigative study on the quality aspects of linked open data. **International conference on cloud computing and internet of things**, [*s.l.*], p. 33-39, 2018. Disponível em: http://dx.doi.org/10.1145/3291064.3291074. Acesso em: 26 maio 2022.

NELSON, R.R.; TODD, P.A.; WIXOM, B.H. Antecedents of information and system quality: an empirical examination within the context of data warehousing. **Journal of Management Information Systems**, v. 21 n. 4, p. 199–235.

NOOGHABI, M. Z.; DASTGERDI, A. F. Proposed metrics for data accessibility in the context of linked open data. **Program**, [*s.l.*], v. 50, n. 2, p. 184-194, 4 abr. 2016. Disponível em: http://dx.doi.org/10.1108/prog-01-2015-0007. Acesso em: 26 maio 2022.

POSSEMATO, T. How RDA is essential in the reconciliation and conversion processes for quality linked data. **Jlis**, [*s.l.*], v. 1, n. 9, p. 48-60, 2018. Disponível em: http://dx.doi.org/10.4403/jlis.it-12447. Acesso em: 26 maio 2022.

WANG, R. Y., STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **J. Manage. Inf. Syst**. v. 12, n. 4, p 5–33, jan. 1996.

W3C. **Data on the web best practices**. 2017. Disponível em: https://www.w3.org/TR/dwbp/#intro. Acesso em: 16 ago. 2022.

W3C. **Data on the web best practices**: Dataset Usage Vocabulary. 2016. Disponível em: https://www.w3.org/TR/vocab-duv/. Acesso em: 16 ago. 2022.

W3C. **Best practices for publishing linked data**. 2014. Disponível em: https://www.w3.org/TR/ld-bp/. Acesso em: 26 jan. 2021.

ZAVERI, A.; RULA, A.; MAURINO, A.; PIETROBON, R.; LEHMANN, J.; AUER, S. Quality assessment for Linked Data: a survey. **Semantic Web**, [*s.l.*], v. 7, n. 1, p. 63-93, 17 mar. 2015. Disponível em: http://dx.doi.org/10.3233/sw-150175. Acesso em: 19 ago. 2022.

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 intitulado FLUXO DE SELEÇÃO DE FONTES DE DADOS LINKED DATA PARA ENRIQUECIMENTO SEMÂNTICO DE DADOS DE COMUNICAÇÃO CIENTÍFICA.