em questão
E-ISSN 1808-5245

A descrição formal da qualidade de dados publicados na Web: análise do *Data Quality Vocabulary* (DQV)

Ananda Fernanda de Jesus[1]

[1]Universidade Estadual Paulista, Marília, SP, Brasil;
af.jesus@unesp.br; https://orcid.org/0000-0001-7873-6040

José Eduardo Santarem Segundo[1,2]

[1]Universidade Estadual Paulista, Marília, SP, Brasil;
[2]Universidade de São Paulo, Ribeirão Preto, SP, Brasil;
santarem@usp.br; https://orcid.org/0000-0003-3360-7872

Resumo: O processo de avaliação de qualidade desempenha um papel importante na reutilização dos dados disponibilizados na Web. Para garantir o uso e reuso desses dados faz-se necessária à sua descrição formal, de maneira compreensível à agentes computacionais. Uma das possibilidades para viabilizar essa descrição é o *Data Quality Vocabulary*, elaborado pelo Word Wide Web Consortium. Objetivou-se apresentar o *Data Quality Vocabulary* com base em sua relação com o processo de descrição formal da qualidade de dados publicados na Web, analisando os objetivos, características e a estrutura do vocabulário. A pesquisa possuí um caráter exploratório e descritivo, adotando como método um estudo da documentação oficial publicada pelo consórcio. Como resultados obteve-se um panorama do cenário que levou ao desenvolvimento do vocabulário, foi apresentada sua estrutura e discutido o seu potencial de aplicação. Conclui-se que o *Data Quality Vocabulary* disponibiliza uma estrutura descritiva formal, geral e customizável para o fornecimento de resultados do processo de avaliação de qualidade de dados, o que permite que esses resultados sejam compartilhados pelos seus fornecedores. Permite ainda que a comunidade participe do processo de avaliação e compartilhe os resultados obtidos de maneira formal, diminuindo assim o retrabalho. Conclui-se ainda que o vocabulário contribui para o reuso de dados no contexto da Web ao facilitar o uso de ferramentas automáticas e semiautomáticas de avaliação e seleção de fontes de dados para a aplicação.

Palavras-chave: qualidade de dados; avaliação de qualidade; DQV

1 Introdução

Com última atualização em janeiro de 2017, o documento intitulado *Data on the Web Best Practices*, elaborado pelo consorcio responsável pelo desenvolvimento

da Web, o Word Wild Web Consortium (W3C), oficializou um conjunto de 35 (trinta e cinco) *Melhores práticas* (MP) para a publicação de dados na Web.
O documento fornece recomendações para a disponibilização e o uso de dados publicados na Web, com objetivo de garantir a descoberta, compreensão e reuso dos mesmos, tanto por usuários humanos como por agentes computacionais. (W3C, 2017).

Dentre as recomendações desse documento encontra-se a MP 06 que orienta a disponibilização de informações sobre a qualidade do conjunto de dados de maneira formal, utilizando uma descrição em estrutura legível por agentes computacionais.

A qualidade de dados geralmente é conceituada pela literatura em uma abordagem contextual, como "*fitness for use*", ou seja, adequada a situação onde os dados serão empregados. Nesse contexto um mesmo conjunto de dados pode suprir a necessidade de um usuário, mas não ser adequado para outro. (JURAN, 1988; WANG; STRONG, 1996; ZAVERI *et al*., 2012).

O processo de avaliação de qualidade de dados é baseado nos conceitos de dimensões e métricas. As dimensões estão relacionadas às características dos dados, ou conjuntos de dados (*datasets*), que os usuários precisam avaliar para decidir se esses são adequados para a tarefa que pretendem realizar, sendo cada característica geral contemplada por uma dimensão.

Para mensurar o quão adequado um *dataset* está em relação a determinada dimensão, são elaboradas métricas que permitem quantificar e qualificar esses dados. A escolha das dimensões e o peso de cada métrica, nessa perspectiva contextual, precisam ser estabelecidas pelo usuário no início do processo de avaliação, pois estão condicionadas às suas necessidades e às tarefas nas quais os dados serão empregados.

Para reutilizar *datasets* em novas aplicações os usuários precisam realizar o processo de avaliação de qualidade, que pode ser feito tanto de forma manual como utilizando-se ferramentas que tornam o processo automático ou semiautomático. Eles também podem consultar processos de avaliação realizados por outros usuários da comunidade ou por agências de certificação.

Nesse sentido a descrição formal da qualidade dos *datasets* pode ter impacto direto na sua reutilização, auxiliando consumidores de dados no processo de seleção de fontes adequadas para suas aplicações, pois facilita o uso de agentes computacionais e consequentemente adoção de ferramentas automáticas e semiautomáticas de avaliação de qualidade.

Visando instrumentalizar a adoção da MP6 o W3C elaborou o *Data Quality Vocabulary* (VOCAB-DQV ou DQV), um vocabulário que provê termos e formaliza as relações existentes entre os conceitos utilizados para a descrição da qualidade de *datasets*. Entretanto, esse vocabulário não tem sido amplamente discutido pela comunidade científica, inclusive pela comunidade da Ciência da Informação. Nesse sentido, partiu-se do seguinte questionamento: Qual o papel do DQV no processo de descrição formal da qualidade de dados publicados na Web?

O estudo consiste em uma pesquisa de caráter exploratório e descritivo, pautado na realização de um estudo documental, com objetivo de apresentar e descrever os principais aspectos relacionados ao DQV, como objetivos, características e estrutura, bem como a sua relação com o processo de descrição formal da qualidade de dados publicados na Web.

2 Procedimentos metodológicos

Caracterizada como exploratória e descritiva, a presente pesquisa teve como base um levantamento documental, realizado no portal oficial do W3C. Foram buscados os termos VOCAB-DQV, DQV, *Data Quality Vocabulary* e considerados documentos que abordam diretamente o DQV ou documentos importantes para a compreensão de sua proposta e estrutura. Também foram considerados materiais bibliográficos indicados nas referências da documentação, sendo aceitos para compor as referências do artigo apenas aqueles necessários para promover uma maior compreensão do DQV. O quadro 1 apresenta as referências dos documentos aceitos e uma breve descrição dos mesmos.

Quadro 1 - Documentos e materiais bibliográficos aceitos para compor a pesquisa

Referência	Título do documento	Descrição da pertinência
W3C (2004)	*Primer RDF*	Explica a estrutura do modelo RDF. Foi incluído para explicar o papel do DQV no processo de descrição formal de dados publicados na Web.
W3C (2015a)	*Data Quality Vocabulary (DQV)*	Documento que apresenta uma visão geral do DQV.
W3C (2015b)	*Vocabularies*	Apresenta o conceito de vocabulários na perspectiva do W3C.
W3C (2016a)	*Data on the Web best practices: Data Quality Vocabulary*	Documentação oficial do DQV.
W3C (2016b)	*List of DQV implementations*	Documento que faz referência a algumas das aplicações do DQV.
W3C(2017)	*Data on the Web best practices*	Conjunto de Melhores Práticas (MPs) para a publicação de dados na Web. Incluído pois a elaboração do DQV foi feita para instrumentalizar essas MPs.
W3C (2020)	*Data Catalog Vocabulary (DCAT) - Version 2*	Documentação oficial do DCAT, vocabulário do qual o DQV é uma extensão.
Zaveri et al (2012)	*Quality assessment for Linked data: A Survey*	Referência considerada para apresentação de dimensões e métricas de qualidade que aparecem a critério de exemplo de aplicação do DQV.
ISO/IEC 25012	ISO/IEC 25012	Referência considerada para apresentação de dimensões e métricas de qualidade que aparecem a critério de exemplo de aplicação do DQV.
Debattista (2014)	*daQ, an Ontology for Dataset Quality Information.*	Referência que apresenta uma ontologia utilizada como base para a elaboração do DQV.

Fonte: Elaborado pelos autores.

Não foram aceitas as versões anteriores dos documentos citados, ou versões classificadas como "em desuso" pelo próprio portal.

3 A publicação de dados na Web: Melhores práticas e Data Quality Vocabulary

Data on the Web Best Practices é um documento oficial do W3C elaborado com o propósito de incentivar e orientar a publicação de conjuntos de dados na Web,

tradicionalmente focada na disponibilização de recursos informacionais. O documento é composto por 35 *Melhores práticas* (MPs) que orientam sobre os procedimentos necessários para ampliar a recuperação e reuso desses dados tanto por usuários humanos como, e especialmente, por agentes computacionais. Essas MPs abordam diversos aspectos da publicação dos dados, cada uma sendo relacionada diretamente com um conjunto de objetivos. O relatório das melhores práticas lista oito objetivos, sendo eles: reuso, compreensão, capacidade de ligação, capacidade de descoberta, confiabilidade, acessibilidade, interoperabilidade e processabilidade. Podendo cada MP ser relacionada a mais de um benefício.

Um dos aspectos mais abordados pelas MPs é o fornecimento de informações descritivas que permitam aos usuários recuperar, acessar e compreender o conjunto de dados, ressaltando a importância do fornecimento de diferentes tipos de metadados.

Entre essas recomendações voltadas para destacar a importância do fornecimento de metadados descritivos encontra-se a "Melhor prática 6: fornecer informações de qualidade de dados" ou MP6, que é associada aos benefícios de reuso e confiabilidade dos dados. O W3C (2017, não paginado, tradução nossa) afirma que:

> A documentação da qualidade dos dados facilita significativamente o processo de seleção do conjunto de dados, aumentando as chances de reutilização. Independentemente das peculiaridades específicas do domínio, a qualidade dos dados deve ser documentada e os problemas de qualidade conhecidos devem ser explicitamente declarados nos metadados.

Essa descrição explícita das limitações dos conjuntos de dados tem impacto direto na confiabilidade dos mesmos. Conhecendo os potenciais problemas de qualidade os consumidores passam a ter maior autonomia para verificar se esses dados atendem ou não às necessidades da sua aplicação.

Ao abordar a descrição formal das informações de qualidade dos *datasets* a MP apresenta como exemplo o *Data Quality Vocabulary*, vocabulário proposto pelo consórcio em 2016.

O objetivo do DQV é prover uma estrutura que permita a descrição da qualidade dos *datasets*, viabilizando que a qualidade seja comunicada pelos fornecedores. O vocabulário também possibilita que um usuário ou uma entidade que tenham realizado um processo de avaliação de qualidade comuniquem os resultados dessa avaliação para outros potenciais usuários, o que permitiria reuso dessas informações, evitando retrabalho. (W3C, 2016a).
Na documentação do vocabulário é indicada a perspectiva contextual de sua abordagem de qualidade:

> Alguns conjuntos de dados serão julgados como recursos de baixa qualidade por alguns consumidores de dados, enquanto atenderão perfeitamente às necessidades de outros. De acordo, damos muita importância em permitir que muitos atores avaliem a qualidade dos conjuntos de dados e publiquem suas anotações, certificados, opiniões sobre um conjunto de dados. (W3C, 2016a, não paginado, tradução nossa).

O vocabulário foi desenvolvido pelo W3C *Data on the Web best practices working group* que observaram a sua necessidade durante os trabalhos para desenvolvimento das MPs. A próxima seção irá se aprofundar no DQV, discutindo os principais conceitos necessários para a sua compreensão.

4 Os vocabulários na Ciência da Informação e no contexto Web

Os vocabulários são um aspecto intrínseco das relações entre Ciência da Informação (CI) e o contexto de publicação de dados na Web. Tomoyose, Triques e Simionato (2018) apontam que a definição do conceito de vocabulário está condicionada ao domínio no qual o termo é aplicado. No âmbito da Ciência da Informação, o conceito de vocabulário se estabelece inicialmente como um produto da Organização da informação e do Conhecimento, como um instrumento para padronização das atividades de controle de vocabulário. O controle de vocabulário pode ser entendido:

> [...] como um conjunto de técnicas e procedimentos aplicados a linguagem para resolver problemas de compreensão, ambiguidade, escopo e relacionamento entre os termos que expressam conceitos, e/ou entre que expressam nomes de pessoas, lugares, produtos ou instituições (BARITÉ, 2014, p. 97).

Existem diferentes nomenclaturas para se referir ao conjunto de instrumentos aplicados no controle de vocabulário, termos como linguagens documentárias "[...] listas controladas, índices controlados, linguagens controladas e finalmente vocabulários controlados." (BARITÉ, 2014, p. 101).

Os vocabulários controlados podem ser incluídos em um agrupamento mais abrangente, o de Sistema de Organização do Conhecimento (SOC). Os SOCs podem ser definidos como "[...] estruturas terminológicas que apresentam relações conceituais por meio de termos." (BISCALCHIN, 2019, p. 2). Inicialmente relacionados diretamente apenas com o conteúdo, ou os assuntos dos recursos informacionais, esses sistemas acompanharam os avanços tecnológicos,

> [...] pela evolução das tecnologias de informação e no contexto da internet, surgiram sistemas de organização do conhecimento mais adequados às necessidades de organização de documentos nativos digitais que desempenham um importante papel no gerenciamento e aplicações de informações digitais em geral. (FUJITA; TOLARE, 2019, p. 98).

Nessa pesquisa optou-se por abordar SOCs como um termo abrangente, que engloba diversos tipos de vocabulários controlados, mas também outros sistemas criados posteriormente, inclusive como uma resposta grandes mudanças tecnológicas.

Com essas alterações, os SOCs ganharam novos tipos e acumularam funções. Frente a complexidadde do ambiente Web, "Hoje trata-se não só de recuperar documentos (ou suas representações) mas, representações digitais de qualquer coisa […]". (MARCONDES, 2021, p. 251).

Em síntese, o conceito de vocabulário na Ciência da informação pode ser caracterizado como um termo abrangente, utilizado para se referir a diferentes tipos de SOCs, estudados na Ciência da informação também em uma perspectiva interdisciplinar, onde além dos vocabulários criados para controle de vocabulário, são objetos de estudo os vocabulários criados para estruturar os termos e permitir a descrição não só dos conteúdos, mas também de informações descritivas de recursos informacionais e de conjuntos de dados disponibilizados na Web de maneira formal, compreensível à agentes computacionais.

Nesse sentido, cria-se a necessidade de formalização do processo de descrição de recursos informacionais, mas também dos conjuntos de dados e metadados contidos na Web.

Segundo Isotani e Bittencourt (2015) para que os agentes computacionais possam utilizar as informações e os dados contidos na Web é necessário que sejam fornecidas características desses recursos e das ligações existentes entre eles. Quando essas informações são fornecidas em uma estrutura compreensível aos agentes computacionais, pode-se dizer que essa descrição foi feita de maneira formal, ou seja, legível por computadores.

Um dos caminhos para viabilizar a descrição formal dos recursos na Web é o *Linked data*, um conjunto de princípios propostos em 2006 e que de acordo com Berners-Lee (2006), consistem na conexão de dados representados com base na utilização de identificadores únicos (*Uniform Resource Identifier* - URIs) e na adoção de um formato padrão (*Resource Description Framework* - RDF), partindo da utilização de *links* entre diversas fontes, imbuídos de informações semânticas a respeito dessas relações.

O RDF é o modelo que orienta o processo de descrição dos recursos no contexto do *Linked data*, ele permite que a descrição possa ser feita utilizando uma base estrutural única em diferentes domínios, facilitando o reuso desses recursos e dos metadados que os descrevem, promovendo interoperabilidade. “O RDF baseia-se na ideia de que as coisas que estão sendo descritas têm propriedades que possuem valores e que esses recursos podem ser descritos ao se fazer declarações [...]”. (W3C, 2004, não paginado, tradução nossa).

As descrições em RDF são elaborados no formato de triplas estruturadas como “Recurso+Propriedade+ Valor”.

> O RDF usa uma terminologia específica para falar sobre as várias partes das declarações. Especificamente, a parte que identifica o assunto da declaração é chamada de recurso. A parte que identifica a característica do recurso que a instrução específica (criador, data de criação ou idioma nesses exemplos) é chamada de propriedade e a parte que identifica o objeto dessa propriedade é chamada de valor. (W3C, 2004, não paginado, tradução nossa).

A cada nova propriedade do assunto, ou cada novo objeto de uma mesma propriedade, é feita uma nova declaração, fragmentando assim o processo de descrição e facilitando a sua reutilização.

Para poder ser abrangente e atender a diferentes domínios o RDF se limita a orientar a estrutura das declarações. Para formalizar as características/propriedades dos assuntos descritos são elaborados e aplicados vocabulários. Esses vocabulários são criados de acordo com as necessidades terminológicas e de relacionamento de cada domínio, existindo, entretanto, um conjunto de vocabulários básicos.

Os vocabulários nesse contexto são compostos por um conjunto de potenciais propriedades, que posteriormente poderão ser empregados para indicar a relação existente entre um assunto e o seu objeto em um determinado domínio, como aponta o W3C (2015b, não paginado):

> [...] os vocabulários definem os conceitos e relacionamentos (também referidos como "termos") usados para descrever e representar uma área de interesse. Os vocabulários são usados para classificar os termos que podem ser usados em uma aplicação específica, caracterizar possíveis relacionamentos e definir possíveis restrições ao uso desses termos. Na prática, os vocabulários podem ser muito complexos (com vários milhares de termos) ou muito simples (descrevendo apenas um ou dois conceitos).

Esses termos também podem ser denominados como metadados, "[...] ou seja, dados responsáveis pela descrição de outros dados, cujo objetivo consiste na promoção de maior semântica aos recursos informacionais que representam". (TOMOYOSE, TRIQUE E SIMIONATO, 2018, p. 84).

Tomoyose (2021) aponta ainda que os vocabulários criados para padronização dos metadados são denominados pela Ciência da Informação como padrões de metadados:

> A necessidade de representar as informações existe em diversas áreas do conhecimento, e para atender a essa necessidade foi criada uma série de padrões de metadados que variam desde estruturas simples, passando por um tipo de padrão intermediário, até padrões de estruturas mais complexas de descrição (ALVES, 2010, p. 59).

Os vocabulários são objetos constantes nas discussões do W3C, existindo orientações próprias para a sua elaboração. Eles também aparecem nas

MPs, onde se orienta que sempre que possível sejam reutilizados vocabulários existentes e estabelecidos. Quando um novo vocabulário se faz necessário a recomendação é a de que esse seja elaborado com base em um vocabulário já estabelecido.

A elaboração do DQV ocorreu em torno desses princípios de reaproveitamento, sendo incluídos termos de outros vocabulários em sua estrutura. Além de outros vocabulários, também foram incluídos para o fornecimento de exemplos a International Organization for Standardization/ International Electrotechnical Commission (ISO/IEC) que orienta a qualidade de dados, a ISO/IEC 25012 e o trabalho de Zaveri *et al*. (2012). As próximas subseções discutem os vocabulários reutilizados para a elaboração do DQV e a presença da ISO/IEC 25012 e do trabalho de Zaveri *et al* (2012) na estrutura do vocabulário. Na última subseção será apresentada e discutida a estrutura de classes do DQV.

4.1 Os vocabulários presentes no DQV

Os vocabulários possibilitam a significação dos dados em nível semântico, conceituam e descrevem as propriedades, possibilitando a explicitação das relações entre assuntos e objetos, o que faz com que possuam um papel importante na disponibilização de dados na Web, em especial quando busca-se atingir os princípios do *Linked data*.

Como já mencionado, as 35 MPs enfatizam a importância do favorecimento de vocabulários existentes, especialmente os já estabelecidos no domínio a ser descrito e criados/mantidos por instituições de destaque, em detrimento da criação de novos vocabulários.

Uma das motivações que embasam essa recomendação é o seu impacto na interoperabilidade. Essa recomendação se relaciona também com a confiança na estrutura e na persistência dos vocabulários, já que eles influenciam diretamente na consistência das descrições. E por fim, relaciona-se ainda com evitar o retrabalho.

O que justifica a criação de um novo vocabulário é a ausência de termos que permitam representar as propriedades necessárias para a descrição em um

dado domínio ou a identificação de conceitos que podem contribuir para a descrição formal em domínios diversos. O DQV se enquadra na segunda situação.

O DQV reaproveita em sua estrutura termos de outros vocabulários para estabelecer entidades e conceitos, como pode ser observado na lista dos principais vocabulários cujos *namespace* aparecem no documento oficial do DQV, apresentados no quadro 2.

Quadro 2 - *Namespaces* presentes no DQV

Vocabulário	Namespace
Dataset Quality Ontology	daq
Data Catalog Vocabulary	dcat
Dublin Core Metadata Initiative (DCMI Metadata Terms)	dcterms
Dataset Usage Vocabulary	duv
Web Annotation Ontology	oa
PROV Family	prov
Simple Knowledge Organization System	skos

Fonte: Adaptado de W3C (2016a).

O *Data Catalog Vocabulary* (*Vocab-dcat* ou DCAT) se destaca na estruturação do DQV, já que ele foi elaborado como uma extensão do DCAT, que pode ser definido como:

> [..] um vocabulário direcionado à publicação de dados para catálogos, reutilizando-se de elementos advindos de outros vocabulários, promovendo a sustentabilidade. Dessa forma, o DCAT pode ser uma opção de vocabulário para a padronização da descrição de conjuntos de dados de pesquisa, bem como no contexto *Linked data.* (TOMOYOSE, 2021, p. 96).

O DCAT apresenta propriedades para permitir a descrição de *datasets*, orientando o fornecimento das informações que um usuário em potencial necessita para recuperar, analisar e selecionar *datasets* para serem reutilizados

em suas aplicações, inclusive utilizando-se de agentes computacionais para tornar essa atividade automática ou semiautomática.

Nesse sentido muitas das informações que podem ser representadas utilizando o DCAT são importantes para o processo de avaliação de qualidade, como informações sobre a proveniência dos dados, tipos de dados, frequência de atualização, licença de uso e idioma.

O DQV e o DCAT podem ser utilizados em conjunto para atuar em todo o ciclo da avaliação de qualidade. DQV complementa o DCAT na medida em que fornece propriedades para descrever as informações dos resultados do processo de avaliação de qualidade, classificando esses resultados de acordo com as dimensões e métricas que são elementos principais em um processo de avaliação. O DCAT, por outro lado, tem um papel maior em viabilizar a utilização de agentes computacionais no processo de avaliação de qualidade, fornecendo as informações que são necessárias para a realização dessa atividade.

Outra base importante na estrutura do DQV é o *Dataset Quality Ontology* (daQ).

> O Dataset Quality Ontology (daQ) é um vocabulário para anexar os resultados da avaliação de qualidade de um conjunto de dados Linked Open Data. A ideia por trás do daQ é fornecer um vocabulário central, que pode ser facilmente estendido com métricas adicionais para medir a qualidade de um conjunto de dados. O benefício de ter um esquema extensível é que as métricas de qualidade podem ser adicionadas ao vocabulário sem grandes mudanças, como a representação de novas métricas seguiriam as previamente definidas. (DEBATTISTA; LANGE; AUER, 2014, p. 2, tradução nossa).

O DQV adapta a estrutura conceitual dessa ontologia, que representa informações sobre as relações existentes entre os conceitos do processo de avaliação de qualidade e fornece orientações sobre como estender a ontologia para que ela compreenda novas métricas.

Ao longo da descrição das classes do DQV são inclusive destacadas as equivalências entre o vocabulário e a ontologia, quando essas existem. Também são acrescentadas notas explicativas quando ocorrem divergências entre os dois modelos.

Destaca-se ainda o papel do Simple Knowledge Organization System (SKOS) na elaboração do DQV. O SKOS é um vocabulário oficial do W3C, que fornece propriedades para a representação formal de sistemas de organização do conhecimento, elaborado em alinhamento com os princípios do *Linked data*. Sousa e Ramalho (2020, p. 14) apontam que "O SKOS é uma alternativa para elaboração de vocabulários controlados para aqueles que objetivam uma arquitetura mais amigável, devido à sua simplicidade e capacidade de estabelecer relacionamentos entre conceitos".

Esse vocabulário foi utilizado no âmbito do DQV especialmente para apresentação de uma tabela de compatibilidade existente entre os conceitos de qualidade da ISO/IEC 25012 que norteia a qualidade de dados, e no artigo de Zaveri *et al*. (2012). Esses dois instrumentos serão discutidos na próxima subseção.

4.2 Os conceitos na avaliação de qualidade representados pelo DQV: norma técnica e referencial teórico

Para maior compreensão da estrutura do DQV faz-se necessária a descrição dos principais conceitos que circundam o processo de avaliação de qualidade. Wang e Strong (1996), estruturam o processo de avaliação de qualidade em categorias, dimensões, critérios e métricas.

As categorias são a forma mais abrangente de agrupamento dos conceitos de qualidade, elas são constituídas pelas diferentes perspectivas por meio das quais a qualidade pode ser compreendida, discutida e avaliada. Essas categorias permitem organizar e agrupar as dimensões e as métricas. As principais perspectivas de qualidade, e consequentemente principais categorias são: intrínseca, contextual, representacional e acessibilidade (NOOGHABI; DASTGERDI, 2016).

A intrínseca seria uma perspectiva que visa avaliar a qualidade dos dados através de um conjunto de características inerentes aos mesmos, que não são afetadas pelo contexto no qual serão aplicados.

A perspectiva contextual agruparia dimensões que sofrem influência direta dos objetivos e das tarefas que o usuário pretende executar com esses

dados. Essa é a perspectiva mais adotada na literatura para definir o próprio conceito de qualidade, e muitas das aplicações existentes têm como foco de sua elaboração uma abordagem contextual. (JURAN, 1988; WANG; STRONG, 1996; ZAVERI *et al*., 2012; NELSON; TODD; WIXOM, 2005).

Esse foco em abordagens contextuais se justifica pelo impacto do contexto no desempenho final dos dados, que faz com que um conjunto de dados atue satisfatoriamente em um cenário, mas cause problemas quando aplicado a outro.

A perspectiva representacional seria mais recente, na qual busca-se avaliar o quão adequadamente representados os dados estão, se apresentam as informações necessárias para sua interpretação, compreensão, seleção e aplicação.

Wang e Strong (1996) acrescentam ainda a perspectiva de acessibilidade, cujo foco é abranger dimensões relacionadas à capacidade de acesso dos dados.
As dimensões podem ser entendidas como a expressão mais abrangente das características que serão observadas no processo de avaliação, cada dimensão é composta por um conjunto de critérios que descrevem a qualidade dos dados com base em um atributo específico.

No DQV uma dimensão é definida como "[...] uma característica relacionada à qualidade de um conjunto de dados relevante para o consumidor" (W3C, 2016a, não paginado, tradução nossa). Nesse sentido a definição de dimensão do vocabulário é abordada por meio de uma perspectiva contextual.

Os critérios seriam os atributos que descrevem os itens a serem observados para avaliar um *dataset* em relação a determinada dimensão. Cada critério deve ser relacionado a uma ou mais métricas. As métricas são indicadores elaborados para permitir quantificar e qualificar um conjunto de dados em relação a determinada dimensão (WANG; STRONG, 1996; ASSAF; SENART; TRONCY, 2016; FÄRBER *et al*., 2017; MELO, 2017).

O DQV define as métricas como responsáveis por fornecer "um procedimento para medir uma dimensão de qualidade de dados, que é abstrata, observando um indicador de qualidade concreto. Geralmente, há várias métricas por dimensão". (W3C, 2016a, não paginado, tradução nossa). O DQV não adota

o conceito de critério, referindo-se apenas à categorias, que seriam um conceito abstrato composto por dimensões, também um conceito abstrato composto por métricas, essas sim conceitos concretos e mensuráveis.

O vocabulário fornece meios para que sejam representados os conceitos de categorias, dimensões e métricas, viabilizando a criação de declarações em RDF para comunicar os resultados da avaliação de qualidade.

Para além da estrutura do processo de qualidade, Wang e Strong (1996) também apresentaram uma série de dimensões e métricas que atuaram como base no estabelecimento de muitos modelos de qualidade publicados posteriormente, como ocorre no caso da ISO/IEC 25012 e do modelo proposto por Zaveri *et al.* (2012).

Ambos os documentos integram o DQV com o propósito de apresentar um conjunto de dimensões e métricas, que atuam como exemplo e ponto de partida para a aplicação do vocabulário.

Ao apresentar esses exemplos o W3C (2016a) ressalta o caráter não normativo do vocabulário, ou seja, a sua falta de intenção em compilar, apresentar e definir quais as métricas que podem ou não ser adotadas no processo de avaliação de qualidade. Entretanto, essas referências fornecem uma visão bastante abrangente das principais dimensões e métricas que são adotadas para a avaliação de qualidade de dados publicados na Web.

De acordo com a ISO/IEC 25012 (2008, não paginado, tradução nossa):

> Define um modelo geral de qualidade de dados para dados retidos em um formato estruturado dentro de um sistema de computador. Pode ser usada para estabelecer requisitos de qualidade de dados, definir medidas de qualidade de dados ou planejar e executar avaliações de qualidade de dados.

Nesse sentido, a norma se constitui como uma fonte oficial para o estabelecimento de dimensões e métricas para a avaliação de qualidade, embora seja caracterizada como um modelo generalista, que não pode e nem pretende abranger as necessidades de todos os domínios. A principal contribuição da norma é fornecer um modelo de qualidade que pode atuar como base para novos modelos.

A ISO/IEC 25012 (2008) apresenta as dimensões agrupadas em duas categorias: inerentes aos dados e dependentes do sistema. Pode-se dizer que as dimensões são divididas então em intrínsecas e contextuais. Existem dimensões que ficam no limiar entre ambas as categorias, contendo critérios de qualidade inerentes e dependentes dos sistemas, essas dimensões podem ser agrupadas como "inerentes e dependentes dos sistemas". O quadro 3, elaborado com base em ISO/IEC 25012 (2008) apresenta as dimensões organizadas em categorias.

Quadro 3 - Categorias e dimensões da ISO/IEC 25012

Categoria	Dimensão
Inerentes aos dados Agrupa dimensões que avaliam características intrínsecas dos conjuntos de dados	**Acuracia** Busca-se verificar em que medida os atributos das declarações possuem valores corretos em relação ao conceito ou evento que representam, pode ser dividida em acurácia semântica (relacionada aos valores das descrições) e sintática (relacionada a estrutura).
	Completude Busca-se verificar em que medida todos os atributos e valores esperados em determinado contexto estão presentes.
	Consistência Busca-se verificar se o conjunto de dados está livre de contradições (lógicas/formais).
	Credibilidade Busca verificar em que medida o conjunto de dados pode ser considerado verdadeiro e crível por um conjunto de usuários para uma aplicação, incluindo conceitos como autenticidade e proveniência.
	Atualidade Busca verificar se os dados estão suficientemente atualizados e com a frequência de atualização necessária para a realização da tarefa pretendida
Dependentes do sistema Agrupa dimensões cuja avaliação depende de uma explicitação do uso pretendido do conjunto de dados	**Disponibilidade** Busca verificar se a descrição dos dados é compacta e bem formatada.
	Portabilidade Busca verificar em que medida os dados possuem a capacidade de integração com versões anteriores dos dados internos ou com dados advindos de fontes externas.
	Recuperabilidade Busca verificar em que medida os dados estão descritos utilizando uma estrutura formal bem formada que permite o uso de agentes computacionais na

Categoria	Dimensão
	interpretação dessas informações
Inerentes aos dados e dependentes do sistema Agrupa dimensões que avaliam características intrínsecas dos conjuntos de dados mas que também podem conter critérios que para serem avaliados precisam que o contexto de avaliação seja estabelecido.	**Acessibilidade** Busca verificar em que medida os dados podem ser acessados, incluindo a questão da acessibilidade para pessoas que necessitam de suporte tecnológico especial devido à alguma deficiência.
	Conformidade Busca verificar em que medida as informações contidas no conjunto de dados podem ser aceitas como corretas, verdadeiras, reais e críveis.
	Confidencialidade Busca verificar em que medida os dados são elaborados de forma a garantir que apenas pessoas autorizadas tenham acesso.
	Eficiência Busca verificar em que medida os dados desempenham as atividades propostas de maneira adequada às expectativas dos usuários
	Precisão Busca verificar em que medida os dados são exatos e precisos, tendo como base um contexto especifico de aplicação.
	Rastreabilidade Busca verificar em que medida as alterações e os responsáveis por essas alterações podem ser identificados.
	Compreensibilidade Busca-se verificar em que medida os dados são expressos utilizando linguagens, símbolos e unidades apropriadas a determinada aplicação e que permitam que esses dados sejam compreendidos. Alguns aspectos dessa dimensão dependem da descrição correta e suficiente, realizada com base em metadados adequados.

Fonte: Adaptado de ISO/IEC 25012 (2008).

Zaveri *et al.* (2012) realizaram uma revisão sistemática da literatura na qual um dos objetivos foi identificar as principais dimensões e métricas de qualidade relacionadas a publicação de dados como *Linked data*. Como resultado os autores publicaram um modelo de avaliação de qualidade que se estabeleceu como base na literatura sobre qualidade de dados *Linked data*. Zaveri inclusive aparece como um dos contribuidores para o desenvolvimento do DQV. O Quadro 4 apresenta as dimensões apresentadas por Zaveri *et al.* (2012), agrupadas de acordo com as categorias de qualidade

Quadro 4 - Categorias e dimensões de qualidade propostas por Zaveri *et al.*

Categoria	Dimensão
Intrínseca Agrupa dimensões que buscam avaliar características inerente dos dados, não influenciável pelo cenário de aplicação ou pelas necessidades de seus potenciais usuários.	**Acurácia sintática** Está relacionada com a conformidade dos dados em relação ao modelo adotado. Em caso de dados publicados como *Linked data* busca-se verificar em que medida as regras estruturais do RDF são respeitadas.
	Acurácia semântica Busca verificar em que medida os valores dos dados representam corretamente os fatos do mundo real.
	Consistência Busca verificar se o conjunto de dados está livre de contradições (lógicas/formais).
	Concisão Busca verificar em que medida o conjunto de dados está livre de redundâncias tanto em sua estrutura como no conteúdo dos dados.
	Completude Busca verificar se todas as informações necessárias estão presentes.
Contextual Agrupa dimensões que buscam avaliar características relativas à percepção do usuário e a adequação dos dados para determinada aplicação.	**Relevância** Busca verificar a presença de informações necessárias para a realização da tarefa ou aplicação pretendida pelo usuário.
	Confiabilidade Busca verificar em que medida as informações contidas no conjunto de dados podem ser aceitas como corretas, verdadeiras, reais e críveis.
	Compreensibilidade Busca verificar em que medida os dados podem ser facilmente compreendidos sem ambiguidade e usados por um consumidor humano de informações.
	Temporalidade Busca verificar se os dados estão suficientemente atualizados e com a frequência de atualização necessária para a realização da tarefa pretendida
Representacional Agrupa dimensões que buscam avaliar o fornecimento de informações que permitam recuperar e selecionar o conjunto de dados	**Concisão representacional** Busca verificar se a descrição dos dados é compacta e bem formatada.
	Interoperabilidade Busca verificar em que medida os dados possuem a capacidade de integração com versões anteriores dos dados internos ou com dados advindos de fontes externas.
	Interpretabilidade

Categoria	Dimensão
	Busca verificar em que medida os dados estão descritos utilizando uma estrutura formal bem formada que permite o uso de agentes computacionais na interpretação dessas informações.
	Versatilidade Busca verificar em que medida a descrição dos dados é disponibilizada em diferentes representações e de forma internacionalizada.
Acessibilidade Agrupa dimensões que buscam avaliar em que medida os dados podem ser obtidos e utilizados	**Disponibilidade** Busca verificar em que medida os dados estão presentes, podem ser obtidos e estão aptos a serem utilizados.
	Licenciamento Busca verificar informação sobre se/como o consumidor pode reutilizar um conjunto de dados, e quais as condições definidas para essa reutilização.
	Interligação Busca verificar em que medida as entidades que representam o mesmo conceito estão ligadas umas às outras (considerando as ligações internas externas).
	Segurança Busca-se verificar em que medida os dados são protegidos contra alterações e uso indevido.
	Desempenho Busca verificar o nível de eficientemente de um sistema ao processar o conjunto de dados.

Fonte: Adaptado de Zaveri *et al.* (2012).

Apresentados os principais conceitos e influencias necessários para a compreensão da estrutura do DQV, a próxima subseção apresenta e discute essa estrutura.

4.3 Estrutura do DQV

Como o DQV não fornece dimensões e métricas para a avaliação de qualidade, é necessário definir uma metodologia de avaliação de qualidade, na qual serão estabelecidas as dimensões e métricas a serem observadas e os pesos que cada métrica irá receber nessa avaliação.

Também é necessário estabelecer a técnica de avaliação, podendo essa ser ou não baseada em uma ferramenta de avaliação automática ou semiautomática. Nesse contexto atual da publicação de dados na Web, geralmente associado a grandes volumes de dados, um processo de avaliação de qualidade manual muitas vezes se mostra inviável. O DQV pode ser aplicado então na etapa de comunicação dos resultados obtidos desse processo de avaliação

O documento oficial é estruturado em oito seções e seis apêndices. O quadro cinco apresenta as seções do vocabulário seguidas da descrição de seu conteúdo.

Quadro 5 - Seções do DQV

Seção	Breve descrição do conteúdo
1.Introdução	São apresentadas as motivações que levaram ao desenvolvimento do modelo. É ressaltada sua perspectiva contextual e seu caráter geral e não prescritivo. São indicados os materiais utilizados como base na sua elaboração.
2. *Namespace*	São indicados os *namespaces* utilizados no vocabulário, incluindo os provenientes de outros vocabulários.
3. Visão geral do vocabulário	Apresenta um panorama da estrutura de dados, o modelo do vocabulário, as relações possíveis entre entidades e classes.
4. Especificações do vocabulário	Descreve as classes, propriedades e instâncias, fornecendo para cada uma delas uma definição, relação de dependência com outras classes, classes equivalentes em outros vocabulários e notas de uso.
5. Nota sobre documentação de recursos expressos com DQV	Consiste em uma nota explicativa sobre como fornecer informações adicionais legíveis para humanos utilizando outros vocabulários e destacando o foco na descrição formal do DQV.
6. Exemplo de uso	Apresenta exemplos de descrição para auxiliar na aplicação do vocabulário em diversos cenários.
7. Dicas de dimensões e métricas	Apresenta as dimensões e métricas, selecionadas com base na ISO/IEC 25012 e no artigo de Zaveri *et. al.* (2012).
8. Requisitos	Indicação dos requisitos de boas práticas que orientaram a elaboração do vocabulário, publicada pelo comitê do W3C que trabalhou na elaboração das Melhores Práticas para a Publicação de Dados na Web.

Fonte: Elaborado pelos autores.

O vocabulário lista dimensões e métricas que tiveram como base ISO/IEC 25012 e o modelo de avaliação de qualidade de Zaveri *et al.* (2012). Caso as métricas e dimensões disponibilizadas como exemplo não atendam à necessidade do usuário, é possível estender o vocabulário. Também é possível mesclar novas categorias, dimensões e métricas às sugeridas no modelo.

> Eles podem estender esses pontos de partida, criando seus próprios refinamentos de categorias e dimensões e, claro, suas próprias métricas. Eles podem misturar abordagens existentes — mostramos que as propostas da ISO e Zaveri *et al.* não são completamente incompatíveis. Os implementadores também podem adotar classificações completamente diferentes, se as existentes não se adequarem aos seus cenários de aplicação específicos. (W3C, 2016a, não paginado, tradução nossa).

A única ressalva feita no documento em relação ao uso de classificações, dimensões e métricas novos é a de que ela pode afetar a interoperabilidade com outras fontes, e ainda prejudicar a utilização de instrumentos voltados para identificação e comparação automática de qualidade de dados. O W3C indica, portanto que os usuários ao estenderem o DQV devem "[...] estar cientes de que confiar nas classificações e métricas existentes aumenta a interoperabilidade, ou seja, a chance de que agentes humanos e máquinas possam entender e explorar adequadamente suas avaliações de qualidade." (2016a, não paginado, tradução nossa).

Para auxiliar na aplicação do vocabulário para apresentação de informações de qualidade mais específicas, foram elaboradas uma série de descrições ilustrativas, que compõem a "seção 6 - Exemplo de uso". Essas representações abrangem temas como: proveniência e acessibilidade dos dados, como expressar resultados de avaliação utilizando métricas, documentar a proveniência dos dados e conjuntos de dados, documentar a proveniência da avaliação de qualidade, realizar questionamentos e fornecer *Feedbacks*, relacionar a qualidade do conjunto de dados em relação a uma classificação de qualidade, expressar a qualidade de um conjunto de links, conformidade dos dados com um padrão ou política de qualidade.

Os exemplos de descrição seguem o modelo apresentado na figura 1, que ilustra o exemplo de como informar a qualidade de um conjunto de dados tendo como parâmetro para avaliação o sistema de cinco estrelas[1].

Figura 1 - Exemplo apresentado pelo W3C na seção 6

```
<https://certificates.theodi.org/en/datasets/393> a dcat:Dataset ;
        dqv:hasQualityAnnotation :classificationQA .

:classification QA
    a dqv:UserQualityFeedback ;
    oa:hasTarget <https://certificates.theodi.org/en/datasets/393> ;
    oa:hasBody :four_stars ;
    oa:motivadoPor dqv:avaliação da qualidade, oa:classificando ;
    dqv:inDimension:disponibilidade .

:four_stars
   a skos:Conceito;
   skos:inScheme :OpenData5Star ;
   skos:prefLabel "Quatro estrelas"@en ;
   skos:definition "Conjunto de dados disponível na Web com estrutura não legível por máquina
   formato proprietário. Ele usa URIs para denotar coisas."@en .
```

Fonte: W3C (2016a).

Além das seções o documento oficial também possuí apêndices, onde são apresentados os agradecimentos, histórico de alterações e as referências utilizadas na construção do vocabulário.

Nos apêndices do documento é apresentada também uma tabela de compatibilidade entre as dimensões e métricas apresentadas na ISO/IEC 25012 e as apresentadas por Zaveri *et al.* (2012) e informações de compatibilidade com o vocabulário RDF *Data Cube*, cujo foco é expressar dados multidimensionais inclusive dados estatísticos, que se justifica, pois, parte das métricas são indicadores quantitativos de caráter estatístico.

A entidade central e principal das declarações a serem feitas utilizando o DQV é o *dataset*, ou seja, o conjunto de dados objeto do processo de avaliação. A indicação desse conjunto de dados pode ser realizada de duas maneiras, uma pela instância do DCAT dcat:Dataset, classe que se refere a "Uma coleção de dados, publicada ou com curadoria de um único agente, e disponível para acesso ou download em uma ou mais representações." (W3C, 2020, não paginado,

tradução nossa). A indicação também pode ser feita pela instancia dcat:Distribution, que "representa uma forma acessível de um conjunto de dados, como um arquivo para download." (W3C, 2020, não paginada tradução nossa).

Para representar as características de qualidade dos conjuntos de dados são estabelecidas classes, propriedades e instâncias. Na explicação de cada um desses elementos é apresentada uma definição, a indicação de relação com outros elementos, onde é indicado se o elemento é uma subclasse de outra classe do vocabulário. Também se indica a equivalência desse termo em outros vocabulários, quando essa equivalência se mostra pertinente. A estrutura geral do vocabulário é composta por oito classes e duas subclasses, apresentadas no quadro 6.

Quadro 6 - Classes e subclasses do DQV

Rótulo	Definição
dqv:QualityMeasurement	Classe que representa os resultados da avaliação de qualidade de um *dataset* em relação a uma métrica específica. A classe é relacionada a propriedades que permitem indicar a métrica que está sendo observada, o valor da avaliação, a unidade de medida e o tipo de dados que se espera obter com essa avaliação.
dqv:Metric	Representa os indicadores utilizados para mensurar as dimensões de qualidade
dqv:Dimension	Representa a característica que está em observação, cada dimensão obrigatoriamente precisa ser associada a uma ou mais métricas e agrupada em uma categoria.
dqv:Category	Representa a organização das dimensões de acordo com a perspectiva de qualidade adotada.
dqv:QualityMeasurementDataset	Categoria que permite representar um *dataset* de avaliação de qualidade, onde estariam armazenados os resultados de avaliação de um ou mais *datasets.*
Rótulo	Definição
dqv:QualityPolicy	Permite representar uma política ou acordo, adotado pelo provedor dos dados, que tenha orientado a elaboração, manutenção e disponibilização dos dados.
dqv:QualityAnnotation	Permite a representação de notas, como a indicação de selos de qualidade ou o registro de f*eedbacks*. É necessário vincular essa propriedade com uma indicação de motivação para especificar o propósito dessa anotação.

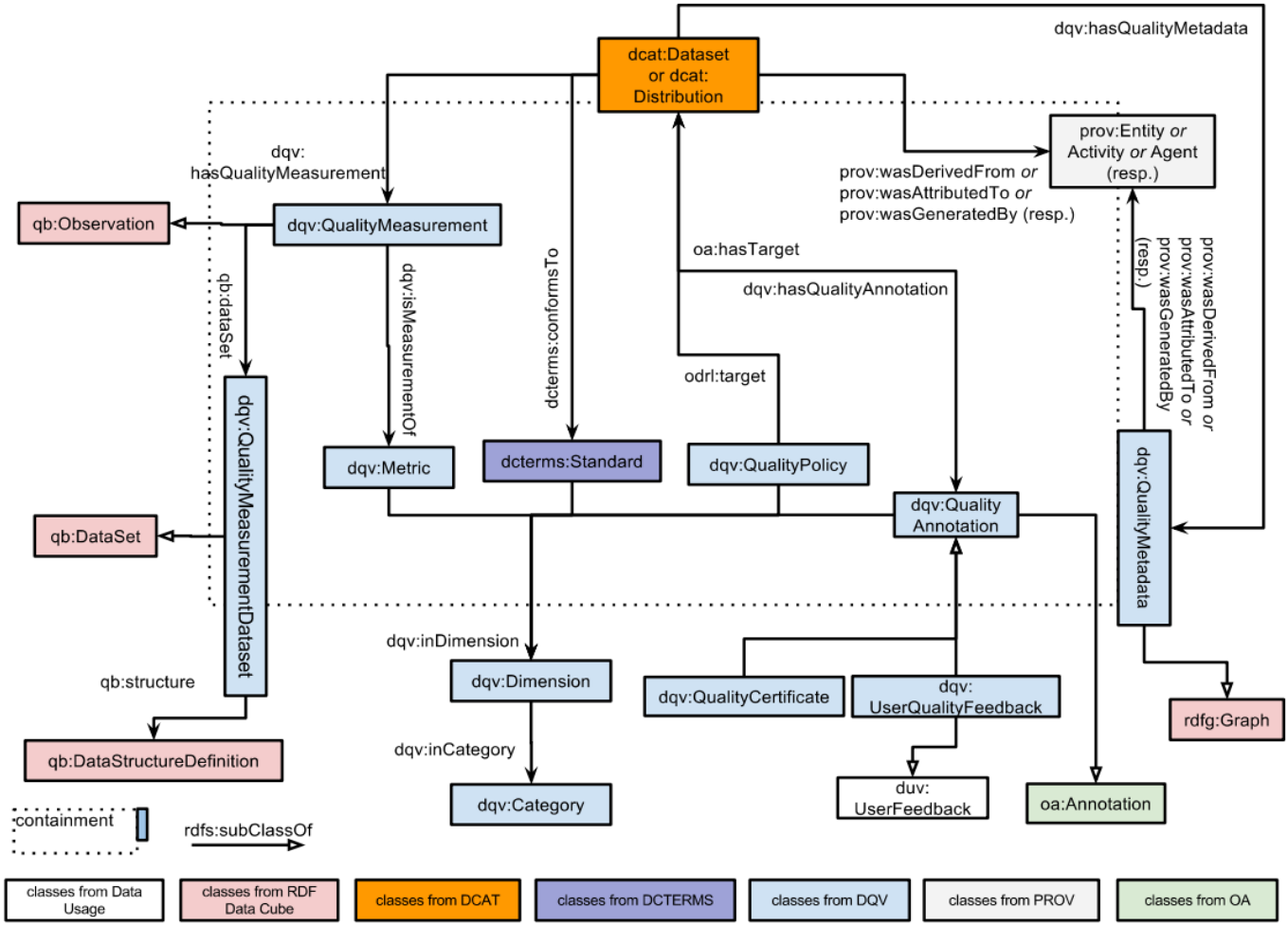
dqv:QualityCertificate (subclasse)	É uma subclasse de dqv:QualityAnnotation, focada especificamente na indicação de certificados ou selos de qualidade fornecidos por outras instituições.
dqv:UserQualityFeedback (subclasse)	Também é uma subclasse de dqv:QualityAnnotation elaborada para indicar um *feedback* dos usuários em relação a percepção de qualidade. Além da necessidade de indicar a motivação da nota, é necessário incluir uma descrição do tipo de feedback (Ex: questionamentos, classificações, sugestões e etc).
dqv:QualityMetadata	Representa o agrupamento de metadados de qualidade, onde deve-se relacionar os certificados de qualidade, a política, as dimensões, as métricas e anotações de um determinado *dataset*. Esse registro pode ser feito em forma de triplas RDF.

Fonte: Elaborado pelos autores.

O documento ainda apresenta o modelo conceitual do vocabulário, com as relações possíveis entre suas classes. Esse modelo é apresentado na figura 2. É possível observar que as cores no modelo são utilizadas para indicar o vocabulário original das classes que compõe o DQV, sendo as classes em azul originais do DQV.

O ponto inicial do modelo de dados, destacado em laranja, são as classes provenientes do DCAT dcat:Dataset (classes utilizadas para identificar o conjunto de dados em si) ou dcat:Distribution (para se referir a um ponto de acesso de um conjunto de dados). Toda estrutura é então relacionada a uma dessas classes, que representam aquilo que será objeto do processo de avaliação. Essa classe então pode ser relacionada a informações de proveniência, com classes originárias do PROV (em cinza) ou a classe dqv:QualityMeasuremen, que permite relacionar os resultados do processo de avaliação, que é discutido usando categorias, dimensões e métricas de qualidade.

Figura 2 - Modelo conceitual do DQV



Fonte: W3C (2016a).

A classe dqv:QualityAnnotation e suas subclasses tem um papel importante em ampliar alcance da descrição de qualidade. Ao discutir a inclusão dessa classe o W3C (2016a) enfatiza a importância de incentivar a participação da comunidade de usuários e de instituições responsáveis pelo fornecimento de certificações de qualidade no processo de avaliação. Essa participação faz com que os usuários não dependam exclusivamente dos fornecedores para ter acesso à avaliações dos *datasets*. Faz ainda com que as avaliações realizadas por outros usuários de maneira orgânica possam ser reutilizadas, evitando retrabalho.

> [...] damos muita importância em permitir que muitos atores avaliem a qualidade dos conjuntos de dados e publiquem suas anotações, certificados, opiniões sobre um conjunto de dados. O editor de um conjunto de dados deve procurar publicar metadados que ajudem os consumidores de dados a determinar se podem usar o

> conjunto de dados em seu benefício. No entanto, os editores não devem ser os únicos a opinar sobre a qualidade dos dados publicados em um ambiente aberto como a Web. Agências de certificação, agregadores de dados, consumidores de dados podem fazer avaliações de qualidade relevantes (W3C, 2016a, não paginado, tradução nossa).

A categoria dqv:QualityMeasurementDataset também se destaca ao permitir a representação de um *dataset* que agrupe os resultados de mais de uma avaliação de qualidade. Ela permite representar os resultados obtidos da análise de um conjunto de dados de determinado domínio, permite ainda o registro de relatórios provenientes de ferramentas automáticas e semiautomáticas de avaliação de qualidade. Contribuindo ainda mais para incentivar outros caminhos para o registro de qualidade, sem que esse esteja limitado as percepções do publicador.

Por fim a classe dqv:QualityMetadata permite representar esse conjunto de metadados provenientes do processo de qualidade, indicando de maneira formal para os agentes computacionais onde se encontram as informações pertinentes, o que poderia facilitar, por exemplo o uso desses agentes na comparação de *datasets* para uma determinada aplicação.

Outro aspecto importante sobre a estrutura do DQV, ressaltada pelo W3C, é o de que sua abrangência não se limita a descrição dos dados e abrange também a descrição da qualidade dos metadados.

> Os elementos DQV podem ser aplicados não apenas para expressar metadados sobre a qualidade dos conjuntos de dados; eles também podem ser usados para expressar declarações sobre a qualidade dos próprios metadados. Isso é especialmente verdadeiro quando se trata de representar a proveniência desses metadados ou sua conformidade com os padrões de metadados estabelecidos. (W3C, 2017, não paginado).

Mas uma vez destaca-se a importância dos metadados e seu papel tanto no processo de avaliação de qualidade como na recuperação e no reuso dos dados disponibilizados na Web.

4.4 aplicações do DQV

Como forma de ilustrar os potenciais usos do DQV em diferentes cenários o W3C (2016b) listou algumas das aplicações do vocabulário. A lista apresentada pelo consorcio fornece como exemplo 6 ferramentas que fazem uso do DQV, sendo elas: *LQTA;* LusTRE*;* RDFUnitç *LD* Sniffer*; LUZZU* e Qskos.

De acordo como o W3C (2016b) o LQTA é uma proposta para a implementação de métricas de qualidade em conjuntos de *links*. Já o LD Sniffer realiza a avaliação de qualidade de dimensões relacionadas a categoria acessibilidade, exportando os resultados do processo de avaliação utilizando o DQV (MIHINDUKULASOORIYA, 2017).

O LusTRE (*Linked Thesaurus fRamework for Environment*) é um portal de tesauros multilíngue, publicados como *Linked data*, o portal adota o DQV para exportar os resultados da avaliação de qualidade tanto dos tesauros em si, como da sua estrutura de ligação utilizando o RDF (ALBERTONI *et al*., 2018).

O RDFUnit é uma aplicação que permite exportar relatórios de qualidade utilizando o DQV. (W3C, 2016b). Já o LUZZU é uma ferramenta criada para avaliação de qualidade de conjuntos de dados publicados como *Linked Open Data*, a estrutura inicial do compartilhamento foi baseada no daQ, mas atualizações da ferramenta permitem exportar os resultados também em DQV. (DEBATTISTA; AUER; LANGE, 2016).

O Qskos é uma ferramenta elaborada para identificar problemas de qualidade em vocabulários publicados utilizando o SKOS, ela também utiliza o DQV para exportar os relatórios de qualidade gerados (MADER, 2018).

Embora as aplicações apresentadas pelo W3C sejam ferramentas distintas, com estruturas, funcionamento e propósitos muito diferentes, o DQV cumpre um proposito semelhante em todas as aplicações listadas. Nelas o vocabulário é incluído na etapa final do processo, fornece a estrutura formal para a comunicação dos resultados de processos de avaliação de qualidade.

5 Considerações finais

Esse estudo foi norteado pelo objetivo de apresentar o *Data Quality Vocabulary* com base em sua relação com o processo de descrição formal da qualidade de

dados publicados na Web, por meio de uma análise dos objetivos, características e estrutura do vocabulário, bem como da apresentação de alguns casos de aplicação.

Para que esse objetivo pudesse ser cumprido o documento oficial que orienta a aplicação do DQV foi analisado em profundidade, traçando a sua relação com o contexto de avaliação de qualidade de dados, da publicação de dados na Web, com a representação formal e com o papel dos vocabulários nesses cenários.

Em relação à análise da estrutura do DQV apontou-se que ele possui relações intrínsecas com outros vocabulários, em especial como o DCAT, que estrutura a descrição de *datasets*, e com daQ, que conceitua as principais relações existentes no processo de avaliação de qualidade.

Também foi possível observar que o DQV foi elaborado com base em uma perspectiva contextual de avaliação de qualidade, o que é ressaltado a todo momento por sua documentação oficial, sendo frisado seu caráter de modelo genérico e abrangente, sem objetivo de ditar quais dimensões e métricas devem ser utilizadas, mas que fornece um compilado de dimensões baseadas em uma norma técnica e um referencial teórico e que apresenta os caminhos para que o modelo seja customizado e estendido para atender as necessidades de domínios específicos.

Em relação ao impacto do vocabulário no reuso de dados no contexto Web, observa-se que esse modelo foi construído em consonância com os princípios do *Linked data* e com as MPs para a publicação de dados na Web, sendo inclusive objeto de uma dessas melhores práticas.

Já em relação a seu papel no processo de descrição formal da qualidade de dados, conclui-se que o DQV se caracteriza como uma opção oficial do consórcio responsável pelo desenvolvimento da Web para o fornecimento de resultados do processo de avaliação de qualidade de dados, o que facilita a sua integração com outros modelos e vocabulários do W3C. Esse papel também pode ser observado nos casos de aplicação citados, onde mesmo sendo elaboradas para atender a cenários diferentes de avaliação, as aplicações

utilizam o DQV para comunicar de maneira formal os resultados de processos de avaliação de qualidade.

O vocabulário viabiliza que os fornecedores de dados disponibilizem informações sobre seus conjuntos de dados, destacando suas vantagens e limitações, aumentando a confiança dos usuários e facilitando para esses o processo de seleção, tornando mais simples o emprego de agentes computacionais nesse contexto.

Entretanto, o maior destaque do vocabulário está em permitir, e inclusive incentivar, que para além dos fornecedores, a comunidade de usuários e as entidades de certificação de qualidade compartilhem os resultados de seus processos de avaliação.

Esse direcionamento possibilita que os usuários não dependam exclusivamente dos fornecedores de dados para se informar dos parâmetros de qualidade de um *dataset*. Isso é importante por que os fornecedores podem, por motivos diversos, acabar negligenciando a representação de metadados de qualidade de seus *datasets* ou ainda terem uma percepção diferente desse nível de qualidade da que pode ser observada por um usuário. Permite ainda que esses resultados sejam reutilizados, evitando assim o retrabalho da realização de um novo processo de avaliação.

Como estudos futuros, pretende-se aprofundar na relação entre DCAT, DaQ e DQV, para compreender as similaridades e diferenças existentes entre esses vocabulários. Também pretende-se ampliar o estudo das aplicações do DQV, identificando mais ferramentas e verificando a existência de potenciais outras formas de aplicação do vocabulário, como para criação de modelos de qualidade.

Outra questão a ser analisada em estudos futuros é a aplicabilidade do DQV e de outros vocabulários no processo de avaliação de qualidade realizados por artefatos automáticos e semiautomáticos. Em especial pretende-se verificar o impacto do uso desses vocabulários na geração de relatórios de qualidade que normalmente são emitidos por esses artefatos ao final do processo de avaliação.

Financiamento

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

Referências

ALBERTONI, R. *et al*. LusTRE: a framework of linked environmental thesauri for metadata management. **Earth Science Informatics**, Atlanta, v. 11, n. 4, p. 525-544, 2018. Disponível em: http://dx.doi.org/10.1007/s12145-018-0344-8. Acesso em: 18 abr. 2023.

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

ASSAF, A.; SENART, A.; TRONCY, R. Towards an objective assessment framework for linked data quality. **International Journal on Semantic Web and Information Systems**, Hershey, v. 12, n. 3, p. 111-133, 2016. Disponível em: https://doi.org/10.4018/ijswis.2016070104. Acesso em: 18 abr. 2023.

BARITÉ, M. Control de vocabulario: orígenes, evolución y proyección. **Ciência da Informação**, Brasília, v. 41, n. 1, p. 95-119, 2014. Disponível em: https://doi.org/10.18225/ci.inf.v43i1. Acesso em: 30 mar. 2023.

BERNERS-LEE, T. Linked data, **W3.Org**., Massachusetts, 27 Jul. 2006.

BISCALCHIN, R. Os sistemas de organização do conhecimento e os desafios frente a geração google. **Páginas a&b: arquivos e bibliotecas**, Porto, s. 3, n. 11, p. 3-9, 2019. Disponível em: https://doi.org/10.21747/21836671/pag11a1. Acesso em: 30 mar. 2023.

DEBATTISTA, J.; LANGE, C.; AUER, S. DaQ, an ontology for dataset quality information. **Ldow2014**, Seoul, v. 1, n. 1, p. 1-8, 2014.

DEBATTISTA, J.; AUER, S.; LANGE, C. Luzzu: a methodology and framework for *Linked data* quality assessment. **Journal of Data and Information Quality**, Estados Unidos, v. 8, n. 1, p. 1-32, 2016. Disponível em: http://dx.doi.org/10.1145/2992786. Acesso em: 18 abr. 2023.

FÄRBER, M. *et al. Linked data* quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. **Semantic Web**, Kansas, v. 9, n. 1, p. 77-129, 2017. Disponível em: http://dx.doi.org/10.3233/sw-170275. Acesso em: 26 maio 2022.

FUJITA, M. S. L; TOLARE, J. B. Vocabulários controlados na representação e recuperação da informação em repositórios brasileiros. **Informação &**

Informação, Londrina, v. 24, n. 2, p. 93-125, 2019. Disponível em: https://doi.org/10.5433/1981-8920.2019v24n2p93. Acesso em: 30 mar. 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION; INTERNATIONAL ELECTROTECHNICAL COMMISSION **(**ISO/IEC 25012**)**. **Software engineering -** software product quality requirements and evaluation (SQuaRE): data quality model, Switzerland, 2008.

ISOTANI, S.; BITTENCOURT, I. I. **Dados abertos conectados**. São Paulo: Novatec, 2015.

JURAN, J. M. **Quality control handbook**. New York: Mcgraw-Hill. 1988.

MADER, C. qSKOS. **GitHub**, [*s.l.*], 7 Sept. 2018.

MARCONDES, C. H. Fundamentos da organização do conhecimento. **Pontodeacesso**, Salvador, v. 15, n. 3, p. 249-282, 2021. Disponível em: https://doi.org/10.9771/rpa.v15i3.47468. Acesso em: 30 mar. 2023.

MELO, J. O. S. F. **Metodologia de avaliação de qualidade de dados no contexto do Linked data**. 2017. Dissertação (Mestrado em Ciência da Informação) - Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017.

MIHINDUKULASOORIYA, N. Linked data sniffer, **GitHub**, [*s.l.*], 16 Mar. 2017.

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**, United Kingdom, v. 21 n. 4, p. 199-235, 2005. Disponível em: https://doi.org/10.1080/07421222.2005.11045823. Acesso em: 30 mar. 2023.

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

SOUSA, J. L.; RAMALHO, R. A. S. SKOS para vocabulários controlados. **TPBCI**: **Tendências da Pesquisa Brasileira e Ciência da Informação,** Brasil, v. 13, n. 1, p. 1-16, 2020.

TOMOYOSE, K. **O data catalog vocabulary (dcat) para a publicação de dados de pesquisa nos princípios Linked Data**. 2021. Dissertação (Mestrado em Ciência da Informação) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de São Carlos, São Carlos, 2021.

TOMOYOSE, K.; TRIQUES, M. L.; SIMIONATO, A. C. Vocabulários controlados e linked open data: análise dos vocabulários getty. **Informação@Profissões**, Londrina, v. 7, n. 1, p. 77-91, 2018. Disponível em: http://dx.doi.org/10.5433/2317-4390.2018v7n1p77. Acesso em: 12 jan. 2022.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information Systems**, New York, v. 12, n. 4, p. 5-33, 1996. Disponível em: https://doi.org/10.1080/07421222.1996.11518099. Acesso em: 17 abr. 2023.

WORLD WIDE WEB CONSORTIUM (W3C). **RDF 1.1 Primer**, Massachusetts, June 24 2004.

WORLD WIDE WEB CONSORTIUM (W3C). **Data quality vocabulary (DQV)**, Massachusetts, June 25 2015a.

WORLD WIDE WEB CONSORTIUM (W3C). **Vocabularies**, Massachusetts, June 25 2015b.

WORLD WIDE WEB CONSORTIUM (W3C). **Data on the web best practices**: data quality vocabulary, Massachusetts, Dec. 15 2016a.

WORLD WIDE WEB CONSORTIUM (W3C). **List of DQV implementations**, Massachusetts, Dec. 15 2016b.

WORLD WIDE WEB CONSORTIUM (W3C). **Best practices for publishing *Linked data***, Massachusetts, Jan. 31 2017.

WORLD WIDE WEB CONSORTIUM (W3C). **Data catalog vocabulary (DCAT)**: version 2, Massachusetts, Feb. 4 2020.

ZAVERI, A. *et al*. Quality assessment methodologies for linked open data. **SWJ: Semantic Web Journal**, Kansas, v.1, p. 1-5, 2012.

The formal description of the quality data published on the Web: analysis of the Data Quality Vocabulary (DQV)

Abstract: The quality assessment process plays an important role in the reuse of data made available on the Web. To ensure the use and reuse of these data, it is necessary to formally describe them in a way that computational agents can understand. One of the possibilities to make this description viable is the Data Quality Vocabulary, elaborated by the World Wide Web Consortium. The objective was to verify the impact of the Data Quality Vocabulary in the process

of formal description of the quality of data published on the Web, analyzing the objectives, characteristics, and structure of the vocabulary. The research has an exploratory and descriptive character, adopting as a method a study of the official documentation published by the consortium. As a result, an overview of the scenario that led to the development of the vocabulary was obtained, its structure was presented and its potential application was discussed. It is concluded that the Data Quality Vocabulary provides a general and customizable descriptive structure for providing the results of the data quality assessment process, which allows these results to be shared by its providers. It also allows the community to participate in the evaluation process and formally share the results obtained, thus reducing rework. It is also concluded that the vocabulary contributes to the reuse of data in the context of the Web by facilitating the use of automatic and semi-automatic tools in the evaluation and selection of data sources for the application.

Keywords: data quality; quality assessment; DQV

Recebido: 16/01/2023
Aceito: 03/06/2023

Declaração de autoria

Concepção e elaboração do estudo: Ananda Fernanda de Jesus, José Eduardo Santarem Segundo

Coleta de dados: Ananda Fernanda de Jesus, José Eduardo Santarem Segundo

Análise e interpretação de dados: Ananda Fernanda de Jesus, José Eduardo Santarem Segundo

Redação: Ananda Fernanda de Jesus, José Eduardo Santarem Segundo

Revisão crítica do manuscrito: Ananda Fernanda de Jesus, José Eduardo Santarem Segundo

Como citar:

JESUS, Ananda Fernanda de; SANTAREM SEGUNDO, José Eduardo. A descrição formal da qualidade de dados publicados na Web: análise do Date Quality Vocabulary (DQV). **Em Questão**, Porto Alegre, v. 29, e-129415, 2023. DOI: https://doi.org/10.1590/1808-5245.29.129415

[1] Sistema adotado pelo W3C para indicar em que medida um conjunto de dados pode ser considerado aberto e conectado, onde para receber a primeira estrela basta que os dados estejam disponíveis sob uma licença aberta, sendo nível máximo relacionado à adoção do *Linked Data*.

Pareceres de avaliação
Os pareceres de avaliação deste artigo estão disponíveis em:

https://seer.ufrgs.br/index.php/EmQuestao/article/view/129415/89765

https://seer.ufrgs.br/index.php/EmQuestao/article/view/129415/89766

Ananda Fernanda de Jesus, José Eduardo Santarem

A descrição formal da qualidade de dados publicados na Web: análise do Data Quality Vocabulary (DQV)
The formal description of the quality data published on the Web: analysis of the Data Quality Vocabulary (DQV)

Em Questão
vol. 29, e-129415, 2023
Universidade Federal do Rio Grande do Sul,
ISSN: 1807-8893
ISSN-E: 1808-5245

DOI: https://doi.org/10.1590/1808-5245.29.129415