XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT – 8 – Informação e Tecnologia

METODOLOGIA DE AVALIAÇÃO DE QUALIDADE PARA DADOS CONECTADOS

Jessica Oliveira de Souza Ferreira Melo (Universidade Estadual Paulista - UNESP)

Leonardo Castro Botega (Centro Universitário Eurípides de Marília - UNIVEM)

José Eduardo Santarém Segundo (Universidade Estadual Paulista - UNESP)

LINKED DATA QUALITY ASSESSMENT METHODOLOGY

Modalidade da Apresentação: Comunicação Oral

Resumo: A Web Semântica sugere a utilização de padrões e tecnologias que atribuem estrutura e semântica aos dados, de modo que agentes computacionais possam fazer um processamento inteligente, automático, para cumprir tarefas específicas. Neste contexto, foi criado o projeto *Linking Open Data* (LOD), que consiste em uma iniciativa para promover a publicação de dados conectados. Com o evidente crescimento da publicação de dados conectados, a qualidade se tornou essencial para que tais conjuntos atendam os objetivos básicos da Web Semântica. Isso porque problemas de qualidade nos conjuntos publicados constituem em um empecilho não somente para a sua utilização, mas também para aplicações que fazem uso de tais dados. Considerando que os dados conectados possibilitam um ambiente favorável para aplicações inteligentes, problemas de qualidade podem dificultar ou impedir a integração dos dados provenientes de diferentes conjuntos de dados. A literatura apresenta a aplicação de diversas dimensões de qualidade para dados conectados, porém, é indagada a aplicabilidade de tais dimensões para avaliação de qualidade de dados conectados. Deste modo, esta pesquisa tem como objetivo propor uma metodologia para avaliação de qualidade nos conjuntos de dados conectados, bem como estabelecer um modelo do que pode ser considerado qualidade de dados no contexto da Web Semântica. Para isso, foi adotada uma abordagem exploratória e descritiva a fim de se estabelecer problemas, dimensões, requisitos de qualidade e métodos quantitativos na metodologia de avaliação, a fim de realizar a atribuição de índices de qualidade. O trabalho resultou na definição de 7 dimensões de qualidade e 14 fórmulas diferentes avaliar conjuntos de dados sobre publicações científicas. A metodologia proposta consiste em um meio viável para quantificação dos problemas de qualidade em dados conectados, e que apesar dos diversos requisitos, podem existir conjuntos que não atendam determinados requisitos de qualidade, e por sua vez, não deveriam estar inclusos no diagrama do projeto LOD.

Palavras-Chave: Dados Conectados; Gestão de Qualidade de Dados; Metodologia de Avaliação de Qualidade de Dados; Web Semântica.

Abstract: The Semantic Web suggests the use of patterns and technologies that assign structure and semantics to the data, so that computational agents can perform intelligent, automatic processing to

accomplish specific tasks. In this context, the Linking Open Data (LOD) project was created, which consists of an initiative to promote the publication of linked data. With the obvious growth of linked data publication, quality has become essential for such assemblies to meet the basic goals of the Semantic Web. This is because quality problems in published sets are a hindrance not only to their use, but also to applications that make use of them. Considering that linked data enables a favorable environment for intelligent applications, quality problems can hinder or prevent data integration from different datasets. The literature applies several quality dimensions to linked data, however the applicability of such dimensions for the linked data context is investigated. Thus, this research aims to propose a methodology for linked data quality assessment, as well as to establish a model of what can be considered data quality in the Semantic Web. For this, an exploratory and descriptive approach was adopted in order to establish problems, dimensions and quality requirements and quantitative methods in the evaluation methodology in order to perform the assignment of quality indices. The work resulted in the definition of 7 quality dimensions and 14 different formulas to evaluate datasets on scientific publications. It is concluded that the proposed methodology consists of a viable means for quantification of liked data quality problems, and that despite the different requirements, there may be sets that do not meet certain quality requirements, and in turn should not be included in the LOD diagram.

Keywords: Linked Data; Data quality; Assessment methodology; Semantic Web

1 INTRODUÇÃO

A Web Semântica foi idealizada por Berners-Lee et al (2001), que descreveram uma evolução da Web na qual os documentos disponibilizados possuíam dados e informações para computadores manipularem. De acordo com Shadbolt et al (2006) a Web Semântica consiste em uma Web de informações acionáveis provenientes de dados formatados em uma estrutura semântica. Tal estrutura dispõe de significados nos quais a conexão lógica entre os termos promove a interoperabilidade entre sistemas.

Para que as informações sejam acionáveis é necessário a publicação de dados, que por sua vez são transformados em informações por meio da conexão e estruturas semânticas entre os dados. A publicação desses dados utilizando as tecnologias da Web Semântica possibilita agregar informações de diferentes fontes e domínios para a construção de poderosas aplicações capazes de descobrir novas informações relacionadas. Por esta razão, a publicação e disponibilização de dados desempenha um papel fundamental para a Web Semântica.

Biezer et al (2009) descrevem dados conectados como a prática de criar *links* entre dados provenientes de diferentes fontes por meio de tecnologias que auxiliam a atribuir uma estrutura semântica entre os *links*.

O maior exemplo de adoção e aplicação das tecnologias da Web Semântica, a fim de disponibilizar dados possibilitando a construção de ambientes de dados conectados semanticamente é o *Linking Open Data* (LOD)[1]. O LOD consiste em um projeto fundado em 2007 que teve como objetivo

[1] http://linkeddata.org/

iniciar a web de dados por meio da identificação de conjuntos de dados disponíveis sob licença (aberta no cenário ideal), convertê-los de acordo com princípios da Web Semântica e publicá-los na Web (BIZER et al, 2009).

No entanto, estudos recentes mostram que a maioria dos conjuntos de dados inseridos no diagrama do projeto LOD apresentam problemas de qualidade como inconsistência, problemas representacionais, interoperabilidade, completude, etc. (HOGAN et al., 2012; RULA, 2011; ZAVERI et al, 2015). Deste modo, tratar e detectar problemas de qualidade é crucial para o sucesso das aplicações baseadas na Web Semântica, visto que quando não avaliada, os problemas são descobertos apenas quando as aplicações falham ou as buscas nos conjuntos de dados retornam resultados incorretos. Conforme evidenciado por Fürber e Hepp (2011), saber a qualidade do conjunto de dados é importante para poder realizar um processo de aprimoração dos dados, selecionar os dados apropriados, decidir o quanto se pode confiar na informação disponibilizada.

Uma das medidas mais conhecidas para evitar problemas de qualidade em dados conectados, independentemente do domínio dos dados, é descrita por Berners-Lee (2006), que define quatro regras que expõem expectativas de comportamento para publicação que, quando atendidas, promovem a interconexão dos dados. As regras estabelecidas são: (1) utilizar URI (*Uniform Resource Identifier*) para nomear recursos, (2) utilizar HTTP (*HyperText Transfer Protocol*) como URI de modo que tais dados possam ser encontrados, (3) prover informações úteis utilizando os padrões RDF, SPARQL (*Protocol and RDF Query Language*), e por fim (4) incluir *links* que guiem a outros recursos URIs, de modo que o usuário possa encontrar mais informações relacionadas.

A literatura aponta problemas de qualidade não somente nos dados, mas também na estrutura provida para sua publicação, fator que pode dificultar seu acesso e até mesmo inviabilizar sua utilização, evidenciando o fato de que a qualidade consiste em um fator de extrema importância.

O W3C (*World Wide Web Consortium*) fornece abrangente conteúdo visando orientar o processo de construção das informações a serem publicadas em bases de LOD, visando evitar erros de qualidade como: formatos de dados errados, *links* quebrados, criação de URIs, guias para utilização de padrões de metadados, ontologias etc. Porém, após quase uma década da criação do LOD ainda é possível encontrar conjuntos na rede de dados que apontam para *links* quebrados e problemas como os citados acima, alguns dos quais se propagam desde a criação do projeto.

Definições quanto às dimensões de qualidade para dados conectados e para os métodos de avaliação são também disponibilizadas na literatura, porém é indagado como tais problemas e dimensões acontecem nos conjuntos de dados sobre publicações científicas. Bem como, de que modo os problemas de qualidade podem ir contra os princípios de utilização dos dados ligados e qual é a proporção de tais problemas de acordo com a categoria específica de dados sobre Publicações Científicas. Deste modo, este trabalho tem como objetivo descrever uma metodologia para avaliação de

qualidade para dados publicados em um segmento específico de Dados Conectados, o de publicações científicas. Adicionalmente, é apresentado um modelo referencial de qualidade de dados no contexto da Web Semântica e dos Dados Conectados.

A metodologia foi desenvolvida com o propósito de avaliar não somente conjuntos de dados já publicados, mas também para auxiliar a verificação de problemas de qualidade antes da publicação de um conjunto de dados, especificamente no segmento de dados sobre publicações de Dados Conectados.

Para o desenvolvimento da pesquisa foi adotada uma metodologia de natureza qualitativa, na qual utilizou de abordagem exploratória e descritiva com intuito de estabelecer problemas, dimensões e requisitos de qualidade.

A fim de propor a atribuição de índices de qualidade na metodologia de avaliação criada no trabalho foram adotados métodos quantitativos.

2 A GESTÃO DA QUALIDADE DE DADOS

De modo geral, a qualidade pode ser definida como medidas para que o produto oferecido esteja de acordo com o que se espera dele, podendo este ser uma informação, um dado, um serviço ou um processo. Estando ele livre de problemas, possibilita que as atividades dependentes sejam executadas com sucesso. É notado que a forma como os dados, informações, produtos, etc., são manuseados influenciará na qualidade das atividades desempenhadas nos sistemas de diferentes domínios.

O significado de qualidade pode variar de acordo com o que cada domínio requer para que o dado atinja os objetivos necessários. Outro fator que influencia na qualidade são as exceções quanto a obrigatoriedade dos requisitos, que pode variar de acordo com o domínio. Desta forma, os problemas são definidos de acordo com um contexto específico e requisitos que podem variar de um domínio para outro. Assim, problemas com qualidade podem existir em diferentes áreas e quando o produto dela (relatórios, catálogos, banco de dados) possui problemas, afeta as atividades dessas áreas.

Qualidade aplicada a dados indica que se espera que eles atendam à medida de perfeição, precisão e conformidade no domínio no qual estão inseridos. Dados que não descrevem fielmente componentes do mundo real diminuem a efetividade dos sistemas, contribuem de forma negativa para as atividades envolvidas em sua utilização e seus impactos podem ser tanto sociais quanto econômicos. De acordo com Olson (2003) um dado é de qualidade se satisfaz os requisitos para o seu uso; consequentemente, carece de qualidade quando não os satisfaz.

Considerando que cada domínio possui diferentes requisitos de qualidade, a literatura aborda diferentes pontos de vista quanto à aplicação das dimensões de qualidade (WANG; STRONG, 1996; PAIM et al, 1996; LEE et al., 2002; OLETO, 2006; CALAZANS, 2008; BATINI et al., 2008). Por meio da análise realizada se nota que existem dimensões mais utilizadas, porém não há um padrão estabelecido de

dimensões; cada domínio utiliza dimensões que atendam aos seus requisitos específicos. Dentre as mais utilizadas na literatura constam: completude, precisão, relevância e consistência.

2.1 Qualidade de dados no Projeto *Linking Open Data*

Estudos recentes apontam a presença de problemas de qualidade em conjuntos de dados publicados de acordo com diferentes dimensões, tais como inconsistência, problemas representacionais, interoperabilidade, completude, etc. (HOGAN et al., 2012; RULA; ZAVERI, 2014).

Em uma análise realizada por Acosta et al (2013), foram identificados inúmeros problemas de qualidade no conjunto de dados do DBpédia, que é um dos maiores conjuntos disponibilizados sob licença livre, o qual possui mais de 3.64 milhões de recursos. Dentre os problemas identificados constam:

- 163 mil recursos com código postal no formato errado;
- 7 mil livros com formato ISBN errado;
- 40 mil pessoas com data de morte presente, porém sem a data de nascimento;
- 638 mil pessoas sem data de nascimento;
- 197 locais sem coordenadas geográficas;
- 242 mil recursos com a mesma coordenada não correspondem ao marcador correto (dbo:Place);
- 28 mil recursos com a mesma coordenada geográfica;
- 9 recursos com coordenadas de longitude inválidas.

Tais fatores evidenciam a existência de problemas de qualidade em conjuntos de dados publicados como Dados Conectados. Considerando que o DBpédia é um dos maiores conjuntos de dados do LOD e diversos conjuntos de dados, incluindo os classificados em outras categorias, possuem relacionamento com os dados deste conjunto, é provável que seus problemas de qualidade tenham se propagado e afetado também seus relacionamentos.

Visto que Dados Conectados possuem características singulares, algumas dimensões são únicas e outras são adaptadas de acordo com o contexto de dados conectados, dentre as dimensões comumente citadas constam: *interlinking*, licenciamento, consistência, precisão sintática e semântica, completude e avaliação temporal.

2.2 Metodologias para Avaliação de Qualidade em Dados Conectados

Em análise bibliográfica fica evidente que não há um padrão das dimensões avaliadas em cada metodologia. Cada metodologia foi desenvolvida para avaliar um conjunto de dados ou realizar alguma tarefa específica (AMICIS; BATINI, 2004; LEE et al 2002). Nem todas adotam dimensões exclusivas para o domínio de Dados Conectados, assim como não houve também um processo de definição de quais

seriam as prioridades e dimensões aplicáveis ao seu contexto (FÜRBER; HEPP, 2011; MENDES et al, 2012).

No contexto de Dados Conectados não há uma metodologia para avaliação de qualidade, comumente utilizada e/ou descrita na literatura. É possível encontrar métricas e métodos para avaliação de qualidade (ACOSTA et al., 2013; ZAVERI et al., 2015), ferramentas (KONTOKOSTAS et al., 2013), *frameworks* (BIZER; CYGANIAK, 2009; MENDES et al., 2012).

3 MODELO E METODOLOGIA DE AVALIAÇÃO DE QUALIDADE DE DADOS NO CONTEXTO DE DADOS CONECTADOS

Diferente das metodologias descritas na literatura, a metodologia proposta visa avaliar, além das dimensões interdisciplinares, as que são aplicáveis especificamente no domínio de Dados Conectados. A metodologia de avaliação proposta é dividida em três passos, descritos a seguir: sendo esses (1) Levantamento de requisitos de qualidade para Dados Conectados, (2) Definição das dimensões e métricas e, por fim, (3) Avaliação de qualidade.

Na primeira etapa, deve ser realizado o levantamento dos requisitos para Dados Conectados, que podem ser identificados por meio da definição dos objetivos a serem cumpridos com a utilização do conjunto de dados a ser avaliado.

A seguir, na segunda etapa, considerando o conjunto de dados a ser avaliado, deve ser realizada a definição das dimensões de qualidade a serem avaliadas, bem como as métricas para avaliação de cada dimensão.

A partir dos requisitos, dimensões e métricas de qualidade, na terceira etapa os problemas devem ser avaliados de acordo com fórmulas, resultando no índice de qualidade do conjunto de dados avaliado.

Um fator importante para a utilização da metodologia consiste em realizar uma especificação de um referencial de qualidade para os Dados Conectados. Tal processo resultou na definição um modelo composto por três pilares, fundamentais para a sustentação de cada fase da metodologia de avaliação proposta, sendo eles: (1) Requisitos da literatura, por meio da qual foram obtidas informações sobre problemas e dimensões de qualidade para Dados Conectados; (2) Padrões do W3C para o funcionamento tanto da Web Semântica como dos Dados Conectados e o (3) Princípios de qualidade do *Linking Open Data*, o qual reúne conjuntos de dados, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios.

O modelo foi estabelecido a fim de auxiliar a identificação dos requisitos e necessidades no domínio dos Dados Conectados. No primeiro pilar (W3C) foi realizado um levantamento dos requisitos estabelecidos por esta organização que dita os padrões a fim de que a Web Semântica atinja seu pleno potencial. O pilar LOD consiste no principal canal no qual conjuntos de dados são publicados de acordo

com os padrões recomendados pela Web Semântica e Dados Conectados, deste modo foi realizada uma análise de como funciona o processo de inclusão de conjuntos de dados no diagrama a fim de obter requisitos de qualidade.

O Quadro 1 apresenta os insumos e requisitos extraídos desses dois pilares, para a definição do modelo, dimensões e métricas do processo de avaliação.

Quadro 1: Requisitos de qualidade para conjuntos de Dados Conectados.

W3C	LOD
Ter URI para nomear recursos.	Os itens devem estar acessíveis via URIs referenciáveis
Ter HTTP como URI, de modo que tais dados possam ser encontrados.	Conjunto de dados deve possuir pelo menos 50 links RDF apontando para outros conjuntos de dados ou pelo menos um conjunto de dados com 50 links RDF apontando para ele.
Ter informações úteis utilizando os padrões RDF e SPARQL.	Deve possuir os seguintes metadados sobre o *dataset*: Nome, título, URL, autor, e-mail, *tag*, *taglod*, *link* para um exemplo RDF, URL para SPARQL *endpoint*, URL de download para cada arquivo RDF, URL para página com a lista de *downloads*, versão, notas, licença.
Ter *links* que guiem a outros recursos URIs de modo que o usuário possa encontrar mais informações relacionadas.	Disponibilizar *link* para *download* dos arquivos voID, XML Sitemap, RDF *Schema* e vocabulário; inserir as *tags*
Não ter nomes dos hosts, extensão de páginas, detalhes sobre o desenvolvimento na URI, visto que não apresentam informações sobre o recurso**.**	*Tags* obrigatórias: *limited-sparql-endpoint*, *format-<prefix>*, *no-proprietary-vocab*, *deref-vocab* ou *no-deref-vocab*, *vocab-mappings* ou *no-vocab-mappings*, *provenance-metadata* ou *no-provenance-metadata*, *license-metadata* ou n*o-license-metadata, published-by-producer* ou *published-by-third-party*, *lodcloud.nolinks*, *lodcloud.unconnected* e *lodcloud.needsfixing*
Não ter conjuntos ou identificadores numéricos para os recursos.	
Ter identificadores significativos para o domínio do conjunto de dados, como a combinação do nome e sobrenome do recurso.	

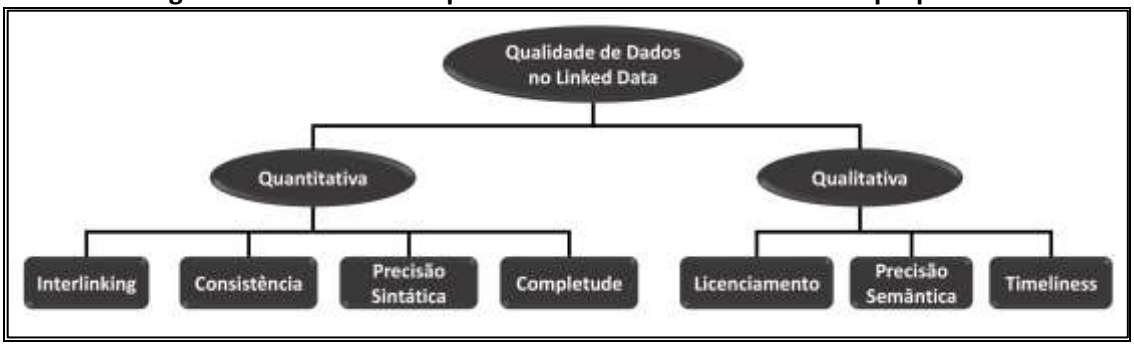
Fonte: Elaborado pelos autores.

Assim, no próximo pilar (literatura) foi realizada uma análise bibliográfica quanto às dimensões mais relacionadas com o domínio da pesquisa e o tipo dos problemas abrangidos em cada uma. Visto que a literatura aborda diferentes dimensões, e tendo em vista que em algumas situações a mesma dimensão foi aplicada para diferentes problemas de qualidade, dois fatores foram levados em consideração na definição das dimensões para a metodologia proposta: (1) quais das dimensões abordadas se relacionavam com os requisitos obtidos na análise realizada nos pilares 1 e 2; (2) das dimensões relacionadas, quais foram utilizadas em unanimidade ou pela maioria das metodologias descritas na literatura. Desse modo, foram definidas as seguintes dimensões de qualidade: *interlinking*, licenciamento, consistência, completude, avaliação temporal, precisão sintática e semântica (BIZER;

CYGANIAK, 2009; RULA, 2011; MENDES et al, 2012; ZAVERI et al., 2012; RULA; ZAVERI, 2014; ZAVERI et al., 2016; FÜRBER; HEPP, 2011; KONTOKOSTAS et al., 2014).

A aplicação do modelo sucedeu da seguinte maneira: os subsídios da primeira fase da metodologia (levantamento de requisitos) foram obtidos dos pilares LOD e W3C. Para a constituição do segundo passo da metodologia, a análise realizada no modelo resultou nas seguintes dimensões: *interlinking*, consistência, completude, licenciamento, avaliação temporal, precisão sintática e semântica. Para o processo de avaliação, que consiste no terceiro passo da metodologia, foram obtidas diferentes métricas descritas na literatura, onde os autores identificaram 69 métricas para avaliação de qualidade em Dados Conectados, que são aplicáveis em 18 dimensões diferentes. A Figura 2 apresenta as dimensões de acordo com o tipo da avaliação a ser realizada, que podem ser qualitativas (onde a pontuação não se aplica) e quantitativas e objetivas.

Figura 2: Dimensões de qualidade de acordo com o modelo proposto



Fonte: Elaborado pelos autores.

3.1 Levantamento de requisitos de qualidade para Dados Conectados

Os requisitos foram obtidos por meio do modelo, os quais foram estabelecidos por meio da análise das principais organizações que estabelecem padrões, princípios e regras para o funcionamento da Web Semântica e Dados Conectados (Quadro 1).

3.2 Definição das dimensões e métricas

Um fato de comum acordo na literatura da comunidade de qualidade de dados é que algumas dimensões e formas de avaliação precisam de auxílio humano para raciocinar sobre os dados dispostos. Então, por meio de tal auxilio os dados são classificados como de baixa qualidade ou não, diferentemente de outras dimensões nas quais a qualidade pode ser quantificada objetiva e automaticamente (BIZER; CYGANIAK, 2009).

Desse modo, as dimensões propostas abrangem tanto dimensões de avaliação subjetiva, quanto objetiva, nas quais duas das 15 métricas podem ser classificadas como subjetivas e qualitativas.

Tanto as dimensões quanto as métricas foram definidas por meio do modelo proposto, o qual resultou em 7 dimensões, sendo: *interlinking*, consistência, completude, licenciamento, avaliação temporal, precisão sintática e semântica.

3.3 Avaliação de Qualidade

As dimensões quantitativas possuem dois tipos de avaliação: local e geral. O índice local consiste na porcentagem individual de cada dimensão e o índice geral é composto pelos índices locais de cada dimensão quantitativa avaliada (*interlinking*, consistência, precisão sintática e completude).

Quando os componentes analisados em cada métrica das dimensões estiverem de acordo com os padrões de referência, a dimensão receberá um valor local entre 0 e 100%. Considerando que os índices locais possuem o valor dentro de um intervalo de 0 a 100% e serão utilizados para o índice global, no qual correspondem a 25% (visto que são quatro dimensões), o resultado será readaptado de modo que o valor entre 0 a 100 tenha o seu equivalente dentro do intervalo geral (0 a 25). Quando uma das métricas propostas na dimensão for invalidada, em vista da inexistência dos dados necessários e não pela falta e/ou não disponibilização, o valor da métrica inaplicável será distribuído entra as métricas aplicáveis.

São propostas duas fórmulas diferentes para avaliação de qualidade geral, que variam em razão da inexistência ou inviabilidade da aplicação de determinada dimensão no processo de avaliação. A Fórmula 1 apresenta um dos cálculos propostos, no qual todas as dimensões possuem o mesmo índice e importância, ou seja, cada uma das dimensões corresponde a 25% do índice geral, onde Ig corresponde ao índice geral, que é composto pela média dos resultados de cada dimensão, I representa o valor local da dimensão *interlinking*, $Cons$ representa o valor local de consistência, Pr o valor de precisão sintática e Com corresponde ao índice de completude.

Dentre as quatro dimensões, completude e *interlinking* são as únicas divididas entre mais de uma métrica. Então, o resultado da somatória de cada métrica é dividido por 4, assumindo que o peso de cada uma das dimensões corresponde a 25%.

$$Ig = \frac{I+Cons+Pr+Com}{4} \quad (1)$$

A Fórmula 2 apresenta o cálculo a ser conduzido, caso uma das dimensões avaliadas possuam importância maior para o avaliador, sendo possível redefinir o peso de cada dimensão, contanto que a somatória dos pesos seja igual a 100, onde *P* representa o peso, que por sua vez, é multiplicado pelo índice local de cada dimensão; então o resultado é divido dividido por 100, assumindo que P1+ P2+ P3+ P4 = 100.

$$Ig = \frac{P_1 I+P_2 Cons+P_3 Pr+P_4 Com}{100} \quad (2)$$

Em vista da ausência de fundamentação teórica para lidar com a inexistência ou inviabilidade da dimensão durante o processo de avaliação, tal problema pode ser encarado de duas maneiras: (1) por meio da definição de um índice positivo, ou seja, de 100% para o índice local ou (2) por realizar a exclusão da dimensão do índice geral. A fórmula 2 se aplica em ambos os casos, onde (1) o resultado da dimensão local será de 100% e no (2) será atribuído peso 0 em P, na dimensão excluída.

Dentre os pré-requisitos para aplicação da metodologia de avaliação proposta, constam os seguintes fatores:

- É requerida a disponibilização do arquivo de dados do conjunto de dados, independentemente da linguagem de descrição utilizada;
- É requerido que o avaliador tenha o entendimento ou conhecimento básico sobre as propriedades e classes utilizadas no conjunto de dados, a fim de definir quais devem ser avaliadas em cada dimensão, quando aplicável.

3.3.1 Interlinking

Dimensão dividida em duas métricas, ao final do cálculo de cada métrica a Fórmula 3 será utilizada para obter o resultado do índice local de *interlinking*. Se propõe uma avaliação quantitativa para definir um índice de porcentagem de quantos *links* RDF não foram afetados com problemas de qualidade.

$$I = \frac{m_1+m_2}{2} \quad (3)$$

Métrica 1 - Detectar *links* de boa qualidade: *links* de boa qualidade são aqueles que não fazem uso de identificadores numéricos para identificar recursos, que não utilizam detalhes de implementação (nome do host, extensão da linguagem de programação utilizada para desenvolver a página Web) e utilizam identificadores significativos para o domínio do conjunto de dados como o nome ou sobrenome do recurso, por exemplo.

A partir das verificações serão obtidos quatro resultados diferentes: o primeiro consiste na porcentagem de quantas URIs não possuem problemas de qualidade, que será o índice local; os próximos três valores serão informativos sobre o quanto cada requisito possui de URIs com problemas de qualidade. A Fórmula 4 apresenta o cálculo para o índice da métrica, onde *U* representa o total de URIs e *Up* a quantidade de URI presente com problema de qualidade.

$$m1 = \left(\frac{(\sum U)-(\sum Up)}{\sum U}\right) * 100 \quad (4)$$

Métrica 2 - Detectar existência de *links* para fontes externas: De acordo com a proposta do LOD o ideal é que conjuntos de dados tenham pelo menos 50 *links* para fontes externas e essa métrica será disposta em duas etapas: (1) será realizada uma verificação na página do conjunto de dados, se tais informações não forem disponibilizadas, (2) uma busca no arquivo dos dados será realizada visando encontrar o conteúdo sob a propriedade owl:sameAs.

Como exemplo foi citada a propriedade owl:sameAs, mas outras propriedades podem ser utilizadas para apontar links externos. Desse modo, a execução dessa métrica consiste em verificar a propriedade equivalente no vocabulário utilizado do conjunto de dados avaliado.

O cálculo dessa métrica é apresentado na Fórmula 5, onde *Ln* consiste no total de *links* necessários, subtraída a quantidade de *links* presentes *Lp* e então o resultado é transformado em porcentagem.

$$m2 = \left(\frac{(\sum Ln)-(\sum Lp)}{\sum Ln}\right) * 100 \quad (5)$$

A aplicação da fórmula é proposta quando a quantidade de *links* presentes for menor do que a quantidade necessária, visto que, quando for maior, o requisito foi cumprido, totalizando os 50% correspondentes dessa métrica.

De modo geral, cada métrica corresponde a 50% do índice local dessa dimensão; a pontuação da métrica 1 será realizada para verificar a porcentagem de URIs sem problema de qualidade e a métrica 2 para identificar se o conjunto de dados possui ao menos 50 *links*. E então será efetuada a soma dos valores das duas métricas para chegar no índice local de qualidade dessa dimensão. A avaliação de ambas as métricas resultam em um valor quantitativo.

3.3.2 Licenciamento

Métrica 1 - Detectar a existência de uma licença na documentação do conjunto de dados: Um requisito para os dados abertos e para que o conjunto de dados seja inserido no diagrama do projeto LOD é a disponibilização da licença na sua página de registro, independentemente do tipo. Visto que o licenciamento consiste em uma dimensão qualitativa, não será utilizada no cálculo quantitativo de qualidade. Será avaliada qualitativamente quanto à ausência ou à presença da licença.

Métrica 2 - Especificar a licença correta, se está atribuída sob a licença original: Considerando o fato de que existem tipos diferentes de licenças, essa métrica, também qualitativa, tem como objetivo que o avaliador distinga se a licença disponibilizada é de fato a licença que deveria ter sido disponibilizada, se é a correta, de acordo com o domínio do conjunto de dados.

Essa verificação também é realizada na página do conjunto de dados; ao clicar no link da licença disponibilizada, o avaliador é redirecionado para uma página de descrição da licença utilizada. Desse modo, o avaliador poderá então distinguir se o que foi disponibilizado é a licença correta ou não. A não disponibilização da licença invalida a aplicabilidade dessa métrica, porém, visto que é um requisito obrigatório, quando acontecer será avaliada de modo negativo, ou seja, como se não fosse a licença correta disponibilizada para o tipo de dados.

Métrica 3 - Disponibilizar uma licença aberta para o conjunto de dados: Considerando que a prática ideal para a publicação, de acordo com os princípios do LOD, é disponibilizar uma licença aberta para os conjuntos de dados*,* esta verificação terá como objetivo analisar se a licença disponibilizada é do tipo aberta.

3.3.3 Consistência

Métrica 1 - Verificar se os dados estão de acordo com a especificação da ontologia: Esta métrica tem como objetivo identificar valores contraditórios, ou seja, inconsistentes com a classe ou propriedade que o especifica. Considerando a possibilidade da utilização de diferentes vocabulários para a descrição dos dados, alguns exemplos de inconsistência são descritos a seguir, supondo a utilização do OWL (tais propriedades foram utilizadas para exemplificar uma situação de análise; no caso da utilização de vocabulários distintos, o avaliador deve identificar valores contraditórios com o que foi especificado na classe e/ou propriedade):

- Owl:nothing: representa uma classe vazia; assim, não deve conter membros e caso ocorra a utilização dessa classe, a avaliação consiste em verificar se membros foram inseridos;
- Owl:sameAs: utilizada para se referir ao mesmo recurso, é uma propriedade que relaciona duas URIs. A análise consiste em verificar se a relação apresentada está de fato consistente com o domínio que ela representa.
- Owl:differentFrom: se refere a diferentes indivíduos; será analisado se esta propriedade foi utilizada junto com owl:sameAs, sobrepondo assim tal propriedade;

Será atribuído um índice por meio do cálculo apresentado na Fórmula 6, onde: *TTi* consiste na somatória de classes/propriedades da amostragem, no qual é subtraída a somatória de *Ti*, que representa a quantidade de itens inconsistentes. O resultado é então divido dividido pelo total de itens *TTi*, que ao final é multiplicado por 100 para encontrar a porcentagem final.

$$CONm1 = \left(\frac{(\sum TTi)-(\sum Ti)}{\sum TTi}\right) * 100 \quad (6)$$

Métrica 2 - Verificar o tipo de dados permitido para a propriedade: Cada métrica representará 50% do índice local dessa dimensão, onde será efetuada a soma dos índices que será convertido para a porcentagem final de consistência. O cálculo dessa métrica é apresentado na Fórmula 7, onde *Tm* consiste no total de atributos que possuem um tipo pré-determinado de dados, *Tmi* representa o total de propriedades com problema de consistência.

$$CONm2 = \left(\frac{(\sum Tm)-(\sum Tmi)}{\sum Tm}\right) * 100 \quad (7)$$

3.3.4 Precisão Sintática

Métrica 1 - Detectar o uso de regras sintáticas: Essa verificação é conduzida para analisar o tipo de caracteres ou modelo de valores permitidos. Cada domínio possui tipos específicos de dados que devem conter uma quantidade ou modelo exato, como por exemplo, O DOI (*Digital Object Identifier*) consiste em um identificador único para identificar entidades físicas, digitais ou abstratas e prover um *link* persistente para sua localização na internet.

O cálculo do índice dessa dimensão é apresentado na Fórmula 8, onde *Tam* consiste no total de atributos com tipos e formatos específicos de dados e *Tap* no total de valores com problemas de precisão; o resultado será convertido para valor final de precisão local.

$$PSm1 = \left(\frac{(\sum Tam)-(Tap)}{\sum Tam}\right) * 100 \quad (8)$$

3.3.5 Precisão Semântica

Métrica 1 - Detectar a utilização de propriedades inexistentes: Verificação de classes e propriedades não existentes no conjunto do vocabulário/linguagem utilizada para descrição dos dados.

Métrica 2 - Detectar a utilização de propriedades não definidas: Essa métrica tem como objetivo verificar a utilização inapropriada de atributos entre dois recursos (ou seja, como uma relação) e de relações (propriedades) como valores literais. Para exemplificar, no caso da utilização do OWL, pode acontecer a utilização errada dos termos owl:DatatypeProperty e owl:ObjectProperty: datatypeProperty descreve uma propriedade (atributo) e objectProperty define uma propriedade de relacionamento entre dois recursos, ou seja, a relação.

3.3.6 Completude

Avaliação quantitativa objetiva dividida em duas etapas: a primeira tem como objetivo avaliar a completude dos meios de descrição do conteúdo, do esquema como um todo. Essa etapa é dividida em três métricas, sendo: completude de esquema, de propriedade e de população.

A segunda etapa da avaliação consiste em avaliar o quão completa está a descrição dos recursos, tanto para a descrição do conjunto de dados na sua página de registro, como para o conjunto de metadados utilizados para descrever um recurso. Tal avaliação depende de dois requisitos, que são: (1) definir o que é considerada uma descrição completa, (2) quais metadados de descrição são considerados prioritários para o domínio de publicações.

A primeira etapa da avaliação de completude é abordada nas métricas 1, 2, 3 e a segunda etapa é descrita nas métricas 4 e 5. O índice local de completude será composto pelas cinco métricas propostas conforme apresenta a Fórmula 9, cada uma equivalendo a 20% do valor total; assim, caso todas as métricas estejam completas, o índice local de completude será 100%. Esse 100% será então

adaptado para o seu valor equivalente dentro de um intervalo de 0 a 25% do domínio geral, que é quanto cada uma das cinco dimensões ocupa.

$$Com = \frac{m_1+m_2+m_3+m_4+m_5}{5} \quad (9)$$

No caso de exceções inesperadas, como a não disponibilização ou não utilização dos dados necessários para conduzir a avaliação, a métrica é invalidada e não aplicada; assim, o valor total da métrica é dividido pela quantidade de cada métrica passível de aplicação. Caso aconteça de não ser disponibilizado o necessário em todas as métricas, o valor de completude do conjunto de dados é de 0%.

Métrica 1 - Completude de esquema: Essa análise obterá como resultado o quão completo o esquema da ontologia está, em questão de propriedades e de classes utilizadas. O cálculo é apresentado na Fórmula 10, onde Cr consiste na soma das classes e propriedades representadas na amostragem a ser analisada e Tc o total de classes e propriedades.

$$COMm1 = \left(\frac{\sum Cr}{Tc}\right) * 100 \quad (10)$$

Métrica 2 - Completude de propriedade: Nessa métrica será obtida a quantidade de propriedades que foram utilizadas na ontologia. O cálculo é apresentado na Fórmula 11, onde Pr consiste no total de propriedades representadas e TP no total de propriedades.

$$COMm2 = \left(\frac{\sum Pr}{TP}\right) * 100 \quad (11)$$

Métrica 3 - Completude de população: Consiste em obter a completude das classes, ou seja, quantas foram utilizadas em relação ao total de classes disponíveis para utilização. O cálculo é apresentado na Fórmula 12, onde Tcp consiste no número de objetos do mundo real representados e TC no número total de objetos do mundo real.

$$COMm3 = \left(\frac{\sum Tcp}{TC}\right) * 100 \quad (12)$$

Métrica 4 - Quantidade de metadados necessários a serem disponibilizados página do registro do conjunto de dados: A avaliação dessa métrica será realizada por meio da verificação das informações disponibilizadas no registro. O cálculo é apresentado na Fórmula 13, em que Ip consiste na quantidade de informações presentes, de onde são subtraídas as adicionais Ia; o resultado obtido é dividido pelo total de informações necessárias In , de onde são subtraídas as adicionais Ia.

$$COMm4 = \left(\frac{(\sum Ip)-(Ia)}{(\sum In)-Ia}\right) * 100 \quad (13)$$

Métrica 5 - Completude considerando atributos prioritários: A fim de obter subsídios para definir um modelo de completude para informações sobre publicações foi realizada uma análise em repositórios de grandes universidades nacionais, internacionais e institutos de tecnologia. Foram verificados os metadados para a descrição de três tipos de publicações científicas, a saber: artigos de conferências, artigos de revistas, teses e dissertações. Foram analisados os metadados das seguintes

instituições: Universidade Estadual Paulista "Júlio Mesquita Filho", Universidade de São Paulo, Universidade Bielefeld, Universidade de Southampton e Instituto de Engenheiros Elétricos e Eletrônicos (IEEE).

Ao analisar a forma como cada instituição descreve o conhecimento científico produzido é observado que ela varia; também que as instituições nacionais fazem uma descrição mais abrangente sobre seus recursos do que as internacionais. A análise também possibilita afirmar que os metadados utilizados para descrição de conteúdo científico variam, dependendo da instituição; não existe um padrão. Alguns metadados essenciais são utilizados por todas as instituições, e a variação depende do tipo de publicação; assim, artigos de revistas compartilham os seguintes campos: autor, título e URI; artigos de eventos: autor, título, resumo, assunto, tipo e URI; teses e dissertações: autor, título, assunto, tipo e URI. Desse modo, é incentivado que, ao conduzir a avaliação, o avaliador defina quais são os atributos necessários no conjunto de dados que está avaliando, bem como quais são os prioritários.

Em vista disso, a utilização da fórmula para completude descrita por Botega (2016) é proposta, na qual a prioridade dos atributos descritivos pode ser atribuída pelo avaliador, visto que tanto o produtor quanto o consumidor do conjunto de dados podem definir quais são os metadados prioritários de acordo com seu contexto específico.

A Fórmula 14 apresenta o cálculo de completude proposto, o qual considera se os elementos essenciais, conforme a pesquisa realizada, estão presentes. Nela, S representa a presença do objeto, neste caso a tripla a ser avaliada. Quando ele está presente, S = 1 e quando está ausente, S = 0; β representa o atributo descritivo, que quando presente é igual a 1 e quando ausente é 0; e por fim γ representa o peso, que, quando considerado prioritário, tem valor igual a 2 e quando não prioritário é 1. Sendo assim, deve ser realizada a somatória da multiplicação de cada peso por presença, e o resultado deve ser dividido pelo total de atributos; o resultado da fórmula será um valor entre 0 e 100% de completude.

$$Com5 = S\left[\left(\frac{\sum \beta * \gamma}{\sum \gamma} * 0{,}9\right) + 0{,}1\right] \quad (14)$$

3.3.7 Avaliação Temporal (*Timeliness* e Volatilidade)

Métrica 1 - Verificar quão atuais são os dados: Essa métrica tem como objetivo obter duas informações temporais sobre o conjunto de dados, sendo que tais informações são classificadas em duas categorias: (1) *timeliness* verifica se o dado é atual para o contexto no qual será utilizado; (2) volatilidade caracteriza a frequência na qual os dados, no caso o dataset, variam no tempo (BOUZEGHOUB, 2004; BATINI; SCANNAPIECO, 2016).

A verificação quanto à *timeliness* será realizada do seguinte modo: a data da última atualização será subtraída da data atual, resultando na idade do conjunto de dados; a partir do resultado o

avaliador poderá verificar se, de acordo com suas tarefas e prioridades, o dado poderá ser considerado atual ou não. A volatilidade consiste em contabilizar a quantidade de vezes em que o conjunto de dados foi atualizado; quando o resultado é 0, significa que não foi atualizado nenhuma vez, o que, dependendo da sua data de criação, pode significar que o conjunto de dados estejam desatualizado. Consequentemente, quando o valor for maior que 0, significa que o conjunto de dados já foi atualizado.

3 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo definir um modelo do que é realmente qualidade de dados para o domínio de Dados Conectados, modelo o qual foi constituído sobre três pilares, sendo eles: (1) Requisitos da literatura, por meio da qual forram obtidas informações sobre problemas e dimensões de qualidade para Dados Conectados; (2) Padrões do W3C para o funcionamento tanto da Web Semântica como de Dados Conectados e o (3) Princípios de qualidade do *Linking Open Data*, o qual reúne conjuntos de dados, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios que estabelece princípios de qualidade, reúne conjunto de dados, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios.

Por meio do modelo definido foram obtidos insumos para construir uma metodologia composta por três etapas da metodologia de avaliação proposta: (1) levantamento de requisitos de qualidade para Dados Conectados, (2) definição das dimensões e métricas e, por fim, (3) avaliação de qualidade. Quanto aos requisitos, foi definido um modelo não somente de requisitos de avaliação para os dados e componentes da Web Semântica, como URI, utilização de vocabulários, mas também quanto aos metadados disponibilizados sobre os conjuntos de dados. Na segunda etapa da metodologia foram definidas as seguintes dimensões de qualidade: *interlinking*, consistência, completude, licenciamento, avaliação temporal, precisão sintática e semântica. Já na terceira etapa, foram definidas as métricas de avaliação para os problemas específicos de cada dimensão, bem como 14 fórmulas para realizar uma avaliação quantitativa do conjunto de dados.

A proposta da metodologia foi motivada pelo fato de não existir uma metodologia para avaliação de qualidade comumente utilizada e/ou descrita na literatura; é possível encontrar métricas e métodos para avaliação de qualidade, ferramentas e *frameworks*.

A metodologia foi desenvolvida a fim de avaliar não somente conjuntos de dados já publicados, mas também para auxiliar a verificação de problemas de qualidade antes da publicação de um conjunto de dados, especificamente no segmento de dados sobre publicações de Dados Conectados

Quanto ao processo de avaliação proposto contatou-se os seguintes fatos:

- É muito importante, no processo de avaliação de qualidade, que o avaliador tenha um conhecimento considerável, ou procure compreender os vocabulários utilizados para

descrição dos dados, visto que cinco das sete dimensões propostas fazem uso de tais informações para realizar a avaliação.

- Um dos meios de obtenção de requisitos e princípios de qualidade foi LOD, projeto que estabelece princípios de qualidade, reúne conjuntos de dados, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios. A visibilidade dos conjuntos de dados que atendem aos requisitos de qualidade é feita por meio da disponibilização de um diagrama composto pelos conjuntos de dados que atendem a tais requisitos, o qual é dividido em 9 categorias diferentes. Porém, ao aplicar a avaliação de qualidade é possível encontrar conjuntos de dados que não cumprem com dos requisitos de qualidade do LOD. Desse modo, pode-se concluir que podem existir diversos conjuntos de dados que não atendam a requisitos de qualidade, e, por sua vez, não deveriam estar inclusos no diagrama.

Apesar de muitas pesquisas, principalmente as aplicadas, estarem sendo desenvolvidas na área de Ciência da Computação, a Ciência da Informação tem assumido um papel importante nas questões relacionadas a qualidade de dados, em modelos de publicações de dados baseados nas melhores práticas de Dados Conectados, contribuindo para minimizar os problemas de qualidade nos processos de armazenamento e recuperação da informação.

Como trabalhos futuros propõe-se a investigação de problemas de qualidade em outros conjuntos de dados da categoria abordada nessa metodologia (publicações), a fim de verificar relações entre dimensões, problemas ou até mesmo novas dimensões aplicáveis no domínio de publicações, bem como a definição de um meio de representação dos resultados da avaliação de qualidade. Propõe-se também a investigação de problemas nas outras 8 categorias de conjuntos de dados do LOD, os quais, assume-se, cumprem os requisitos e princípios de qualidade; considere-se também a possibilidade de verificar quais dimensões afetam as diversas categorias e propor uma avaliação de acordo com os problemas específicos de cada dimensão. Considerando a necessidade do humano no processo de avaliação, um estudo conveniente consiste na incorporação de retinas para a redução da necessidade de tal participação humana e redução de subjetividade.

REFERÊNCIAS

ACOSTA, Maribel *et al*. Crowdsourcing Linked Data quality assessment. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 12., 2013, Sidney. **Proceedings**... Berlin: Springer, 2013. p.260-276.

BERNERS-LEE, Tim. Linked Data: design issues. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 29 jun. 2016

BERNERS-LEE, Tim; HENDLER, James; Lassila, Ora. The Semantic Web. **Scientific American**, New York, v.284, n.5, p.28-37, mai. 2001. Disponível em: <

https://www.scientificamerican.com/magazine/sa/2001/05-01/#article-clarifications>. Acesso em: 6 jun. 2016.

BIZER, Christian; CYGANIAK, Richard. Quality-driven information filtering using the WIQA policy framework. **Web Semantics: Science, Services and Agents on the World Wide Web**, Oxford, v.7, n.1, p.1-10, jan. 2009. Disponível em: < http://www.websemanticsjournal.org/index.php/ps/article/view/157>. Acesso em: 1 jan. 2016.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data: the story so far. **International Journal on Semantic Web and Information Systems**, Hershey, v.5, n.3, p.205-227, jan. 2009. Disponível em: < https://www.igi-global.com/gateway/article/37496>. Acesso em: 26 jan. 2016.

CALAZANS, Angélica Toffano. Qualidade da informação: conceitos e aplicações. **Transinformação.** Campinas, v.20, n.1, p.29-45, jan./abr. 2008.

FÜRBER, Christian; HEPP, Martin. SWIQA: a semantic web information quality assessment framework. In: EUROPEAN CONFERENCE ON INFORMATION SYSTEMS, 18., 2011, Roksilde. **Proceedings**... Helsink, 2011. p.19.

HOGAN, Aidan *et al*. An empirical survey of linked data conformance. **Web Semantics: Science, Services and Agents on the World Wide Web**, Oxford, v.14, p.14-44, jul. 2012. Disponível em: < http://www.websemanticsjournal.org/index.php/ps/article/view/287 >. Acesso em: 5 mar. 2016.

KONTOKOSTAS, Dimitris *et al*. TripleCheckMate: a tool for crowdsourcing the quality assessment of linked data. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND SEMANTIC WEB, 4., 2013, St. Petersburg. **Proceedings**... Heidelberg: Springer, 2013. p.265-272.

LEE, Yang *et al.* AIMQ: a methodology for information quality assessment. **Information & Management**, v.40, n.3, p.133-146, dec. 2002. Disponível em: < https://pdfs.semanticscholar.org/19f3/255f143b7e611731cca13ab5fccb33e03708.pdf >. Acesso em: 16 may. 2016.

MENDES, Paulo; MÜHLEISEN, Hannes; BIZER, Christian. Sieve: linked data quality assessment and fusion. In: INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY, 15., 2012, Berlin. **Proceedings**... New York: ACM, 2012. p.116-123

OLETO, Ronaldo Ronan. Percepção da qualidade da informação. **Ciência da informação**. Brasília, v. 35, n.1, p.57-62 jan./abr. 2006.

PAIM, Isis; NEHMY, Rosa Maria; GUIMARÃES César. Problematização do conceito" Qualidade" da Informação. **Perspectivas em Ciência da Informação**. Belo Horizonte, v.1, n.1, p.111-119, jan./jun. 1996.

RULA, Anisa; ZAVERI, Amrapali. Methodology for Asssessment of Linked Data Quality. In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 10., 2014, Leipzig. **Proceedings**... Leipzig: LDQ, 2011. p. 40

SHADBOLT, Nigel; BERNERS-LEE, Tim; HALL, Wendy. The semantic web revisited. **IEEE Intelligent Systems**, New York, v.21, n.3, p.96-101, jan./feb. 2006. Disponível em: < http://ieeexplore.ieee.org/abstract/document/1637364/>. Acesso em: 6 abr. 2016.

ZAVERI, Amrapali *et al*. Quality assessment for linked data: a survey. **Semantic Web**, Amsterdam, v.7, n.1, p.63-93, jan. 2016. Disponível em: <http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey >. Acesso em: 20 jun. 2016.

ZAVERI, Amrapali *et al*. User-driven quality evaluation of DBpedia. In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 9., 2013, Graz. **Proceedings**... New York: ACM, 2012. p. 97-104