Mapeamento de Problemas de Qualidade no Linked Data

Jéssica Oliveira de Souza, *Universidade Estadual Paulista - UNESP*;
José Eduardo Santarém Segundo, *Universidade de São Paulo - USP*

***Resumo*— A qualidade dos dados e informações tanto na web quanto em bases de dados atuais consiste num tópico de extrema relevância para utilização destes dados de forma efetiva e substancial. Visto que a Web Semântica (WS) foi criada com objetivo de aprimorar a experiência de uso da web atual e o Linked Data é o principal meio no qual sua aplicação ocorre de teoricamente de modo pleno, respeitando os devidos critérios e requisitos da WS, a qualidade dos dados e informações armazenadas nos conjuntos de dados do Linked Data (dados ligados) é essencial para cumprir os objetivos básicos da WS. Assim, este artigo tem como objetivo descrever e apresentar especificas dimensões e respectivos problemas de qualidade no contexto da WS e dos dados ligados.**

***Palavras-chave*—Qualidade de Dados, Linked Data, Web Semântica**

***Abstract*— Since the Semantic Web was created in order to improve the current web user experience, the Linked Data is the primary means in which semantic web application is theoretically full, respecting appropriate criteria and requirements. Therefore, the quality of data and information stored on the linked data sets is essential to meet the basic semantic web objectives. Hence, this article aims to describe and present specific dimensions and their related quality issues.**

***Index Terms*—Data Quality, Linked Data, Semantic Web**

I. INTRODUÇÃO

VISTO que a forma como a informação é processada, gerenciada, organizada, armazenada e acessada é de grande interesse para a Ciência da Informação (CI), novas ferramentas e tecnologias foram desenvolvidas visando atender a implementação automatizada destes processos. O que torna imprescindível a adoção de padrões de descrição de informações na web para auxiliar a organização e posterior recuperação da informação, de forma que atenda a atual necessidade de beneficiar-se dos conteúdos de grandes centros de conhecimentos, porém de modo fácil, rápido e acessível.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456".

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

O surgimento de novos conceitos para organização da informação de modo que o usuário final tire total proveito tem revolucionado o ponto de vista sobre ferramentas para tal. A Web Semântica, por exemplo, foi fundamentada a partir do princípio de fazer o usuário se beneficiar ao máximo do grande volume de informações web, na qual por meio de ferramentas de descrição, organização e recuperação da informação (metadados, ontologias, frameworks) é possível definir uma estrutura de representação dos documentos inseridos na web de modo que o computador e o usuário possam explorar efetivamente os dados. Deste modo, o Linked Data auxilia esse processo para organizar dados relacionados.

O Linked Data consiste num conjunto de práticas que auxiliam a publicar e conectar dados estruturados na web. Berners-Lee [7] define quatro regras que expõe expectativas de comportamento para publicação, quais seguidas promovem a interconexão dos dados. Caso contrário limita a reutilização destes dados. Dentre as regras estão a (1) utilização de URI (*Uniform Resource Identifier*) para nomear 'coisas', (2) utilizar HTTP (*Hypertext Transfer Protocol*) como URI de modo que tais dados possam ser encontrados, (3) prover informações úteis utilizando os padrões RDF (*Resource Description Framework*), e SPARQL (*Protocol and RDF Query Language*) e por fim (4) incluir links que guiem a outras URIs de modo que o usuário possa encontrar mais informações relacionadas.

Assim uma vez que os dados estão de acordo com as quatro regras, são armazenados em conjuntos de dados em formato aberto. Esses dados são ligados a outros conjuntos de dados criando uma relação semântica entre eles. Tais conjuntos podem ser mantidos por uma ou duas organizações em diferentes localizações geográficas [11].

As quatro regras podem ser consideradas os primeiros requisitos de qualidade para a publicação de informação no *Linked Data*, visto que a qualidade pode ser considerada um conjunto de requisitos necessários para um dado atender as exigências de acordo com domínios específicos.

Para promover a interoperabilidade e proporcionar uma eficaz utilização, a qualidade de tais dados publicados é de extrema importância, visto que problemas de qualidade não somente nos dados, porém na estrutura provida para publicação pode dificultar seu acesso, até mesmo inviabilizar seu uso.

A literatura fornece abrangente conteúdo visando orientar o processo de construção das informações a serem publicadas em bases de dados linkados abertas (LOD – abreviação em inglês para *Linked Open Data*), visando mitigar erros de qualidade como: formatos de dados errados, links quebrados, criação de URIs, guias de quais padrões de metadados e

ontologias utilizar para descrição do conteúdo, etc., Porém, depois de 8 anos 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.

Problemas de qualidade são classificados de acordo com dimensões e as definições de dimensões de qualidade diferem de acordo com seu domínio de aplicação, não havendo um campo comum para tal. Os requisitos são divididos em dimensões ou métodos de aplicação os quais são totalmente dependentes do domínio. Este trabalho tem como objetivo abordar dimensões especificas de qualidade para aplicação no *Linked Data*. E assim, mediante de tais dimensões realizar a identificação de problemas existentes de acordo com cada dimensão.

II. LINKED DATA

O exemplo mais claro da materialização e utilização das tecnologias propostas para o funcionamento da Web Semântica é o Linked Data, o qual utiliza das tecnologias para estabelecer link de relacionamento entre os recursos na Web. O conjunto de tecnologias e conceitos estabelecidos pela Web Semântica, de modo a auxiliar agentes computacionais compreender a semântica de documentos e dados auxilia este processo de link entre os recursos.

Dentre as fundamentais tecnologias estabelecidas na web semântica que são utilizadas no *Linked Data* constam URI que consiste em uma cadeia de caracteres que identifica o nome de um recurso, não necessariamente web. Pode ser dividido em URN (*Uniform Resource* Name) e URL (*Uniform Resource Locator*), no qual URN é uma URI utilizada para nomear um recurso estando este na Web ou não e URL é uma URI de especificação para a localização de tal recurso. Tais tecnologias utilizam o protocolo HTTP que provê um mecanismo universal para recuperação de recursos que podem ser dispostos como um fluxo de bytes (como uma fotografia, por exemplo) ou descrições de entidades [13].

Outra tecnologia utilizada é o RDF um modelo padrão para troca de dados na web. De acordo com W3C (World Wide Web Consortium) o RDF facilita a fusão de dados independentemente do esquema no qual estes se encontram [24]. Posteriormente, o RDF Schema agregou uma vertente semântica aos dados RDF possibilitando a modelagem de dados para vocabulários. A sintaxe modelo RDF é representado por triplas, onde uma tripla é composta por recurso, propriedade e valor conforme apresentado na Figura 1, na qual o recurso Guerra nas Estrelas possui uma propriedade "trabalho notável" do cineasta George Lucas, qual valor é também um recurso que pode levar a mais informações sobre este recurso. Assim, por meio do RDF é possível declarar relações entre os recursos, que por sua vez são representados por URIs que permite o acesso a tal recurso relacionado, e tal recurso relacionado pode possuir n recursos outros relacionados.

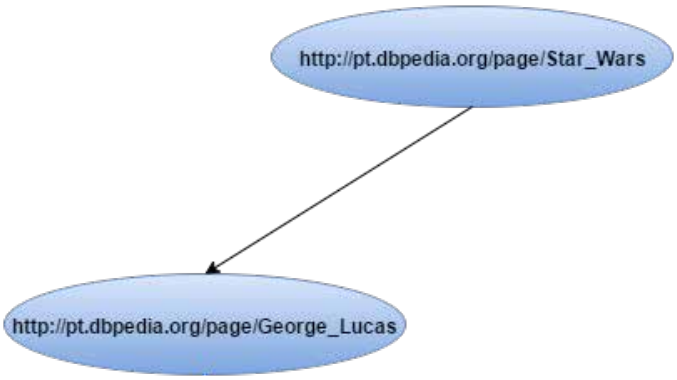

Fig. 1. Tripla apresentando recurso propriedade e valor

Para realizar a recuperação das informações em conjuntos de dados RDF, utiliza-se o SPARQL uma linguagem capaz de manipular os dados armazenados no formato RDF. Uma query SPARQL consiste em três partes o (1) modelo de correspondência, que inclui as configurações de interesse quanto ao grafo a ser pesquisado, tais como união dos dados, filtrar ou restringir valores possíveis; (2) modificadores de solução no qual uma vez que o resultado foi computado, permite a modificação dos valores por meio de operadores como ordem, distinção, limite; e por fim o (3) resultado da query SPARQL que pode ser de tipos diferentes: query de sim/não, seleção de valores que relacionam com o modelo, construção de novas triplas provenientes de tais valores e descrição de recursos [20].

O Linked Open Data é um exemplo da adoção destas tecnologias, consiste num projeto que viabiliza identificar conjuntos de dados disponibilizados sob licença aberta, republicá-los em RDF na web e interliga-los [12].

O processo de publicação de dados linkados na web consiste em três passos principais, primeiramente compreender o conjunto de dados em questão (quais são as entidades principais, suas propriedades, como tais entidades se relacionam com outras), publicar como RDF e linkar com outras bases de dados. Existem dois requisitos básicos para o conjunto de dados ser incluído no diagrama do LOD, são eles: (1) os itens devem estar acessíveis via URIs referenciáveis. Oferecendo apenas o SPARQL endpoint sem URIs referenciáveis não torna o conjunto de dados apto para inclusão; (2) o 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 [34]. Passos para constituição do processo:

- Selecionar os vocabulários: a utilização de vocabulários existentes auxilia a interoperabilidade dos dados, facilita o desenvolvimento das aplicações, dentre os vocabulários mais utilizados estão o Dublin Core (DC), Friend of a Friend (FOAF), Simple Knowledge Organization System (SKOS), Core Ontology Specification (SIOC). Pode acontecer de haver a necessidade de utilizar mais de um vocabulário para representação dos dados, ou que vocabulários existentes não atendam às necessidades do contexto, conduzindo à criação de vocabulários

específicos.

- Particionar grafos RDF em páginas de dados: quanto a grandes conjuntos de dados seria apropriado dividi-los em páginas de dados interligadas visto que apenas uma forma de apresentação grande, extensa e centralizada não seria prática para utilização.
- Atribuir uma URI para cada dado compartilhado: que consiste em colocar cada página de dados online como RDF
- Adicionar metadados e links para a pagina
- Adicionar um mapa semântico para o site [12]

Atualmente, os dados publicados são armazenados no Datahub, que consiste numa plataforma livre para o armazenamento de dados abertos. De acordo com as recomendações do W3C, o processo de publicação do conjunto de dados no Datahub pode ser feito do seguinte modo:

1) Primeiramente criar uma conta no datahub.io antes de editar ou adicionar pacotes;
2) Verificar se o conjunto de dados não existe no datahub.io antes de adicioná-lo.
3) Adicionar ou editar o conjunto de dados e descrevê-lo de acordo com o mínimo de informações requeridas: nome (identificador único), título, número de triplas e links para outros conjuntos de dados
4) Atribuir a tag lod ao novos conjuntos de dados inseridos
5) Se não souber de nenhum inlink ou outlink deve-se atribuir a tag lodcloud.nolinks
6) Prover a maior quantidade de informação adicional possível como SPARQL endpoint, descrição voID, licenças [34].

Sendo possível disponibilizar conjuntos dados em mais de 40 formatos diferentes, dentre eles RDF, OWL (Web Ontology Language), XML (eXtensible Markup Language), SPARQL endpoint (para fazer consulta no conjunto de dados). Pode se dizer que grande parte dos dados do projeto LOD estão armazenados no DataHub, visto que ele é indicado pelo W3C.

Biezer et al. [10] descrevem um processo de avaliação de qualidade dos dados que visa verificar se as informações estão sendo acessadas de modo correto. Quanto a URIs, o processo de verificação pode ser feito por meio de um validador chamado Vapour que valida se dados semânticos estão publicados corretamente de acordo com as melhores práticas definidas por Berners-Lee [6] e Berrueta et al. [8]. Visando verificar se a informação é mostrada corretamente em diferentes navegadores de dados linkados e se os navegadores seguem links RDF de acordo com os dados, os navegadores Marbles, OpenLink RDF Browser, Disco e Tabulator podem ser utilizados.

No Tabulator, caso leve muito tempo para mostrar a informação é um sinal de os grafos RDF do conjunto de dados são muito grande e devem ser divididos. Tal navegador faz inferências web sem a verificação de consistência, assim caso o navegador comporte de modo inesperado, é um indicativo de problemas com declarações RDF (rdfs:subClassOf e rdfs:subPropertyOf) e esquemas OWL.

De acordo com as extensas recomendações de melhores práticas fortemente sugeridas pelo W3C, avaliação da qualidade dos dados apresentados em tais ambientes semânticos pode ser considerada de extrema importância, não somente no processo de construção das bases e dos ambientes semânticos, mas também após a publicação dos conjuntos de dados em plataformas de acesso aberto. A seguir abordaremos uma introdução quanto à qualidade de dados e dimensões de classificação e também alguns problemas relacionados a estas.

III. QUALIDADE DE DADOS

O processo de avaliação de qualidade, assim como suas dimensões, é definido de acordo com o contexto considerando as dimensões definidas. Porém, pode-se afirmar que o processo acontece da seguinte maneira: (1) definição dos problemas de qualidade, (2) definição das dimensões de acordo com os problemas, (3) definição de atributos e métricas para avaliação da qualidade.

De acordo com a literatura, não há um padrão definido quanto a qualidade da informação em sistemas de tomada de decisão, os requisitos são divididos em dimensões ou métricas e a aplicação destas são altamente dependentes de domínio, considerando que a aplicação define seus respectivos significados de acordo com objetivos, tarefas e decisões associadas. As abordagens e descrições de diferentes perspectivas serão descritas a seguir.

O'Brien [42] define as dimensões de qualidade necessárias para sistemas de informação em três dimensões principais divididas em: Conteúdo, Tempo e Forma. Dentre os atributos de qualidade constam:

- Tempo: Prontidão, Aceitação, Frequência, Período
- Conteúdo: Precisão, Relevância, Integridade, Concisão, Amplitude, Desempenho
- Forma: Clareza, Detalhe, Ordem, Apresentação, Mídia

Wang e Strong [43] categorizam os atributos das dimensões da qualidade em quatro classes principais conforme a hierarquia apresentada na Figura 1, descritas a seguir:

Qualidade Intrínseca de dados implica a garantia da credibilidade e reputação dos dados, dentre os atributos constam a própria credibilidade e reputação, como precisão e objetividade.

Qualidade de dados contextual é formada por atributos que devem ser considerados e avaliados de acordo com o contexto da tarefa a ser realizada, tendo como atributos: relevância, tempo, completude, etc.

Quanto a qualidade de representação, os atributos são definidos de acordo com aspectos relacionados ao formato do dado (como a concisão e representação) e o significado em relação a compreensão e interpretação de tais dados. Por fim, os autores classificam individualmente os atributos relacionados à acessibilidade.

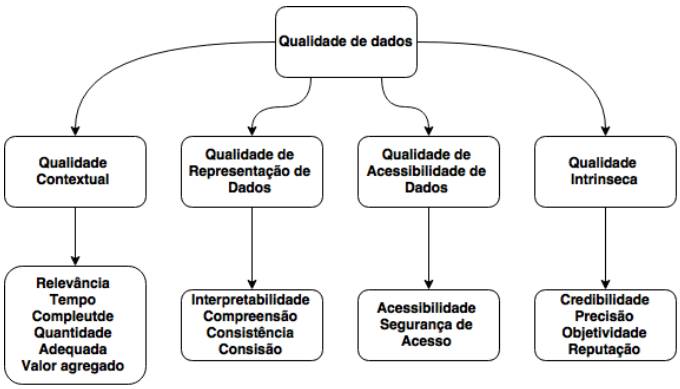

Fig. 4. Dimensões e atributos de qualidade [43]

Carlo Batini [4] afirmam existir muitas discrepâncias quanto a definição da maioria das dimensões devido à dependência contextual da qualidade. Confirmam que não há uma definição global de um conjunto específico de dimensões que definem a qualidade dos dados. Os autores [4] abordam quatro dimensões, sendo estas: Precisão, Completude, Consistência e aspecto temporal do dado (Volatilidade, Atualização). Adiante seguem devidas descrições:

A precisão de um dado é medida por meio de sua equivalência com o dado em questão, assim, determinada proximidade de um valor x em relação a outro valor x pode ou não ser considerado preciso mediante o contexto no qual o dado está sendo testado.

Completude pode ser definida como o grau no qual o dado em questão abrange sua correspondente situação no ambiente real. Conforme abordado na área de pesquisas de bancos de dados relacionais completude relaciona-se com o significado e a representação de valores nulos.

A consistência aborda a violação de regras semânticas para um determinado conjunto de dados. Destaca-se que a consistência pode ser mensurada e avaliada estatisticamente e de acordo com a teoria relacional [4]. É importante notar que grande parte da literatura considera a consistência para tratar problemas de qualidade em bancos de dados levando em consideração constantes pré-definidas para encontrar e solucionar tais problema [14, 25].

Scannapieco et al. [22] exemplificam inconsistência de um dado em uma resposta provida para um conjunto de dados no qual o estado marital é casado e a idade do sujeito é cinco anos de idade. Deste modo, métricas para avaliação são definidas segundo um conjunto de regras predefinidas semânticas para um conjunto de dados. Ainda conforme abordado por Batini et al [3] adotam duas diferentes métricas para avaliação, sendo a primeira com base em técnicas de ligação entre dados usada para identificar regras de consistência de chaves estrangeiras na presença de dados inconsistentes [19]; e a segunda métrica é utilizada para verificar regras de negócio.

A dimensão que abrange a relevância consiste no grau o qual determinado conjunto de informação atende as necessidades do usuário. Também é definida como a extensão na qual do dado é aplicável e útil para a tarefa a ser realizada [2]. De acordo com o levantamento de avaliações de métricas realizado por Batini et al [3] a relevância pode ser medida por meio de métodos subjetivos, como avaliações aplicadas com usuários experientes no domínio.

IV. QUALIDADE DE DADOS NO LINKED DATA

A literatura possui uma gama de informações quanto ao processo de construção de conjuntos de dados aptos para web semântica com base nas tecnologias e aptos a serem disponibilizados como LOD, bem como princípios e pilares base para validação dos dados, porém é provável a existência de dados incorretos, imprecisos ou desatualizados em tais bases e conjuntos de dados.

Visto que as métricas para avaliação da qualidade das informações são altamente dependentes de domínio, Heat et al., [16] e Biezer [13] proveem heurísticas classificadas em três categorias para realizar a avaliação de tais informações de acordo com o tipo da informação utilizada como indicador de qualidade:

- Conteúdo: que usa a própria informação a ser avaliada como um indicador de qualidade, as métricas analisam ou comparam o conteúdo da informação com informações relacionadas;
- Contexto: empregam meta-informações sobre o conteúdo da informação e as circunstâncias na qual a informação foi criada;
- Classificação: depende de classificações explicitas sobre o dado, provedores ou fontes de informações.

A seguintes medidas podem ser tomadas para lidar com dados de baixa qualidade: (1) classificar itens do dado de acordo com seu índice de qualidade, (2) filtrar dados de modo a evitar usuários com dados de baixa qualidade, algumas aplicações podem decidir apresentar apenas dados bem sucedidos na avaliação de qualidade e (3) realizar a fusão dos dados visto que aplicações de dados linkados podem basear-se neste processo por escolher heurísticas de resolução de conflitos apropriadas [16].

Ainda assim, conforme reportado por Kontokostas et al., [17] é possível encontrar uma considerável quantidade de erros de acordo com os testes de qualidade aplicados na ontologia do DBpedia, sendo:

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

O DBpedia é apenas um conjunto de dados dos que estão disponíveis para acesso aberto, deste modo podemos concluir que existem uma relevante quantidade de problemas de qualidade nos conjuntos de dados abertos, podendo estes

serem classificados de acordo com diversas dimensões de qualidade de dados. A seguir abordaremos a definição, o problema e um exemplo de problema de qualidade relacionado com as seguintes dimensões: timeliness, completude e verificabilidade.

A. Timeliness

As dimensões relacionadas ao tempo, comumente nomeadas como timeliness ou atualidade, podem relacionar-se ao atraso na atualização dos dados entre seu estado no ambiente real e no sistema de informação ou ambiente digital que pregue representar seu conteúdo real. Também pode referir-se à idade média do dado em sua fonte. Diferentes definições quanto a timeliness de acordo com diferentes autores é apresentada na Tabela 1.

Segundo Bouzeghoub e Peralta [27] esta dimensão pode ser subdividida duas categorias de acordo com o ponto de vista dos usuários. A primeira é denominada atualidade, captura a frequência da mudança ou atualização dos dados ou com que frequência novos dados são criados em uma fonte. Como por exemplo, a timeliness indica o com qual frequência o preço dos produtos mudam em uma loja ou novos livros são adicionados numa biblioteca. E progressão, que captura o intervalo entre a extração do dado de sua fonte and a entrega aos usuários. Indica o quão velho é o saldo do balanço da conta apresentada ao usuário em relação ao real balanço no banco.

A aplicação da timeliness pode ser guiada pelo fato de ser ou não possível ter dados atuais que são inúteis por estarem atrasados para um uso específico. Um calendário de cursos universitários, por exemplo, pode estar atual, conter dados mais recentes e ainda não ser útil se for disponibilizado somente após o início das aulas [39].

Quanto a dimensões temporais no Linked Data, Li et al. [44] afirmam que técnicas para fazer a relação dos dados ligados são falhas visto que ignoram informações temporais e podem fracassar quanto a representação de tais informações. Isso pode acontecer porque o processo é realizado da seguinte maneira, primeiro realiza-se a comparação da similaridade entre cada par de registros decidindo se estes pares são equivalentes ou não, o segundo passo consiste em agrupar os dados de acordo com objetivo que os dados no mesmo agrupamento se referem a esta entidade no ambiente real; e registros em diferentes conjuntos referem-se a diferentes entidades.

Os problemas de qualidade relacionados a esta dimensão podem acontecer ao tentar identificar a projeção das informações no decorrer do tempo. Um exemplo é abordado por Rula [21] que aborda um caso no qual um determinado autor x que pertencia a universidade "The Open University" no ano y1; e então se mudou para universidade "Univerisity of Milan Bicocca" no ano y2. Autores mudarem de instituição é algo comum de acontecer, porém a projeção de tal informação aconteceria apenas manualmente, por meio de uma consulta na qual seria possível identificar que em tal ano y1 o autor pertenceu à universidade 1 e no ano y2 à universidade 2 (Figura 2).

TABELA I
DEFINIÇÕES DE TIMELINESS DE ACORDO COM A LITERATURA [4]

Quantity	Conversion from Gaussian and CGS EMU to SI [a]
Wand e Wang (1996)	Timeliness refere-se apenas ao atraso entre uma mudança do estado do mundo real e da modificação resultante no estado do sistema de informação
Wang e Wand (1996)	Timeliness é a extensão na qual a idade do dado é apropriado para a tarefa a ser efetuada
Redman (1996)	É o grau no qual um dado está atualizado. O valor do dado está atualizado se estiver certo considerando possíveis discrepâncias causadas por mudanças temporais no valor correto
Jarke et al. (1995)	Descreve quando a informação foi registrada em fontes de dados. Descreve o período de tempo que a informação é válida no mundo real.
Bovee et al. (2001)	Possui dois componentes idade e volatilidade. Idade é a medida do quão velha é a informação com base em quanto tempo foi registrada. Volatilidade é a medida de instabilidade da informação, da frequência de mudança do valor para o atributo de uma entidade
Naumann (2002)	É a idade média de um dado em uma fonte
Liu e Chi (2002)	Grau no qual o dado é suficientemente atualizado para uma tarefa

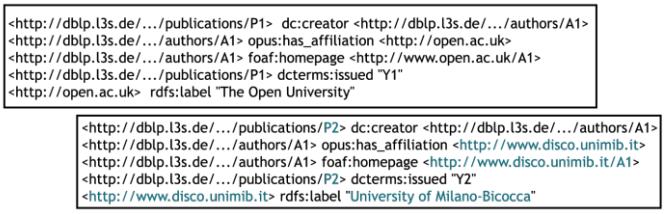

Fig. 5. Divergências temporais das informações quanto a filiação do autor [21]

Considerando um sistema de dados ligados, um exemplo quanto a timeliness consiste no seguinte evento, levando em consideração um cenário de agendamentos de voo no qual o usuário deseja agendar um voo de Milão para Boston e o sistema de busca retorna diferentes voos de conexão. O dado no conjunto de dados mostra uma companhia disponível de acordo com os requisitos do usuário. Quanto a timeliness, a informação relacionada ao voo está salva e é disponibilizada ao usuário a cada dois minutos os quais preenchem os requisitos decididos pelo mecanismo de busca que correspondem a volatilidade da informação do voo. Apesar dos valores do voo estarem atualizados, a informação recebida pelo usuário sobre a disponibilidade do voo não está atual. Assim, o usuário não percebe que a disponibilidade do voo se esgotou porque a mudança foi disponibilizada ao sistema logo após o usuário efetuar sua busca [26].

B. Completude

Esta dimensão é comumente definida e amplamente citada na literatura como: o grau no qual uma coleção de dados inclui dados, descrevendo o seu conjunto de objetos no mundo real [4, 43]. A aplicação desta dimensão pode acontecer de formas diferentes de acordo com o contexto na qual ela é estabelecida. Por exemplo, Souza et al., [23] definem um modelo de informações necessárias para atender denúncias de roubo. Assim, uma denúncia pode ser considerada completa se todos os objetos e atributos estiverem presente na denúncia. Amicis e Batini [25] descrevem um dado completo caso aconteça uma frequência exata dos atributos financeiros estiverem presente.

No contexto da web semântica, consequentemente dados linkados, a completude divide-se em três categorias completude de esquema, população e propriedade. A completude de esquema refere-se ao grau que os elementos da ontologia são representados, quanto a completitude de população indica se todos os objetos referentes a uma instância do mundo real estão sendo representados. Quanto a completude de propriedade, consiste em valores ausentes de acordo com uma propriedade ou coluna específica [15]. Um exemplo é apresentado na Tripla 1 (Figura 6).

Tripla 1:

dbpedia:Firewingdbpprop:isbn "978"^^xsd:integer

Fig. 6. Problema de qualidade encontrado numa tripla descrevendo um ISBN incompleto

Consiste num recurso do DBPedia sobre um livro infantil chamado “Firewing” com um valor incompleto e incorreto sobre o ISBN [1]. É comum encontrar problemas de qualidade nos quais buscas resultam em valores incorretos, como apresentados na tripla acima.

C. Verificabilidade

A literatura descreve a verificabilidade grau e facilidade na qual a informação pode ser verificada para correção [26, 38, 39]. Outro exemplo de verificabilidade é apresentado na Tripla 2, na qual os autores definem tal problema de qualidade como tipo de dado extraído incorretamente, onde aborda triplas com tipo de dados incorretos, visto que na ontologia do DBpedia, o intervalo da propriedade activeYearsStartYear é definido com xsd:gYear, apesar da declaração estar correta na Tripla 2 formatado como xsd:dateTime, o valor esperado era "1981"^^xsd:gYear [15].

Tripla 2:

dbpedia:Stephen_Frydbpediaowl:activeYearsStartYear"1981-01-01T00:00:00+02:00"^^xsd:gYear

Fig. 6. Problema de qualidade encontrado numa tripla de acordo com a dimensão da verificabilidade

Outro problema de qualidade abordado por Furber [15] que se enquadra na dimensão da verificabilidade acontece quando links para páginas web ou fontes de dados externas estão incorretos ou não mostram nenhuma informação relacionada ao conteúdo da fonte. Tais problemas podem ser encontrados no Datahub, e este pode ser um fator pelo qual uma base de dados pode ser considerada inapta para fazer parte do LOD.

Dimensões	Métodos
Timeliness	Atualidade do dado: tempo limite < tempo atual [15] Definir recência e frequência de validação dos dados: verificar se o dado publicado foi validado há não mais que um mês atrás [41]. Exclusão de dados desatualizados: número de triplas atualizadas no conjunto de dados / pelo total de número de triplas em um conjunto de dados [41]
Completude	Identificar a obsolescência do dado: checar nas instancias na base de dados da web semântica se há data de modificação mais velhas que a última data de modificação na instancia da fonte [15]. Quão completo o conjunto de dados está: quantidade de atributos presente / quantidade total de atributos Grau no qual objetos do mundo real estão presentes [18] Completude do conjunto de informações considerando presença e ausência de atributos normais e prioritários [23] Grau no qual interlinks não estão ausentes [40]
Verificabilidade	Verificar a exatidão do conjunto de dados [9] Verificar a autenticidade do conjunto de dados [41]

V. CONSIDERAÇÕES FINAIS

Descrevemos aqui três dimensões de qualidade e seus respectivos problemas, sendo estas timeliness, completude e verificabilidade. Timeliness consiste numa dimensão que descreve problemas temporais, abordando questões como a atualidade das informações, o atraso de seu estado no ambiente digital em relação ao seu estado no mundo real. Tais tipos de problemas podem acontecer principalmente em conjuntos de dados que descrevem recursos relacionados a pessoas, em informações como estado marital, local de trabalho, ou seja, informações que estão sujeitas a mudanças através dos anos.

A dimensão que aborda a completude é muito importante para a veracidade das informações dos conjuntos de dados. Um problema identificado foi a presenta de registros incompletos, como o ISBN incompleto de um livro. E por fim a verificabilidade que consiste na facilidade que uma informação pode ser verificada para correção. Um exemplo nesta dimensão consiste em links relacionados aos recursos que não levam a nenhum conteúdo.

Conjuntos de dados que possuem informações de baixa qualidade podem ser considerados inaptos para ser disponibilizados no LOD. Ainda existem recomendações expressas para evitar tais tipos de problemas para auxiliar no progresso do projeto dos dados abertos.

Por meio da análise bibliográfica pode-se afirmar que as

dimensões de qualidade, bem como a definição de seu contexto de aplicação é altamente depende do domínio podendo uma dimensão possuir diferentes significados e levar em consideração diferentes atributos como prioridade tanto para aplicação quanto para posterior avaliação.

Além dos métodos citados para avaliação dos problemas de qualidade relacionados a tais dimensões, é possível encontrar uma gama de métodos que consideram atributos diferentes, consecutivamente resultando em processos de avaliação distintos dos descritos. Realizar testes em conjuntos de dados pode resultar no levantamento de diferentes problemas de qualidade.

Desconsiderando o contexto de dados linkados existe uma vasta literatura relacionada a dimensões e aplicações especificas de métricas para definição e avaliação de problemas de qualidade. Acredita-se que ainda existem diversos problemas de qualidade relacionados às dimensões de qualidade aqui discutidas. Porém a identificação de tais problemas deve ser feita mediante uma análise em conjuntos de dados específicos. Futuramente objetiva-se definir um conjunto de dados para realizar testes e consultas a fim de definir a aplicação de dimensões e métodos de qualidade específicos para conjuntos de dados específicos do LOD.

REFERENCES

[1] Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., e Lehmann, J. Crowdsourcing linked data quality assessment. In:The Semantic Web–ISWC 2013. Springer Berlin Heidelberg, 2013. p. 260-276.

[2] Agre, J., Vassiliou, M. S., & Kramer, C. (2011). Science and Technology Issues Relating to Data Quality in C2 Systems. Institute for Defense Analyses ALEXANDRIA VA

[3] Batini, C., Barone, D., Mastrella, M., Maurino, A., & Ruffini, C. (2007). A Framework And A Methodology For Data Quality Assessment And Monitoring. In ICIQ (pp. 333-346).

[4] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR), 41(3), 16.

[5] BATINI, Carlo et al. Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR), v. 41, n. 3, p. 16, 2009.

[6] Berners-Lee, T. (2009). Cool URIs don't change, 1998.

[7] BERNERS-LEE, Tim. Linked data-design issues (2006). URL http://www.w3.org/DesignIssues/LinkedData. html, 2011.

[8] BERRUETA, Diego et al. Best practice recipes for publishing RDF vocabularies. Working draft, W3C, 2008.

[9] BIZER, C. Quality-Driven Information Filtering in the Context of Web-Based Information Systems. PhD thesis, Freie Universität Berlin, March 2007

[10] Bizer, C., Cyganiak, R., & Heath, T. (2007). How to publish linked data on the web.

[11] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, 205-227.

[12] *Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008, April). Linked data on the web (LDOW2008). In Proceedings of the 17th international conference on World Wide Web (pp. 1265-1266). ACM.*

[13] BIZER, Christian; CYGANIAK, Richard. Quality-driven information filtering using the WIQA policy framework. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 7, n. 1, p. 1-10, 2009.

[14] Cong, G., Fan, W., Geerts, F., Jia, X., & Ma, S. (2007, September). Improving data quality: Consistency and accuracy. In Proceedings of the 33rd international conference on Very large data bases (pp. 315-326). VLDB Endowment.

[15] FÜRBER, Christian; HEPP, Martin. Swiqa-a semantic web information quality assessment framework. In: **ECIS**. 2011. p. 19.

[16] HEATH, Tom; BIZER, Christian. Linked data: Evolving the web into a global data space. **Synthesis lectures on the semantic web: theory and technology**, v. 1, n. 1, p. 1-136, 2011.

[17] KONTOKOSTAS, Dimitris et al. Test-driven evaluation of linked data quality. In: **Proceedings of the 23rd international conference on World Wide Web**. ACM, 2014. p. 747-758.

[18] MENDES, P., M ÜHLEISEN, H., AND BIZER, C. Sieve: Linked data quality assessment and fusion. In LWDM (March 2012)

[19] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In SIGMOD Conference, pages 802–803, 2006

[20] PÉREZ, Jorge; ARENAS, Marcelo; GUTIERREZ, Claudio. Semantics and Complexity of SPARQL. In: **International semantic web conference**. 2006. p. 30-43.

[21] RULA, Anisa. DC proposal: Towards linked data assessment and linking temporal facts. In: **The Semantic Web–ISWC 2011**. Springer Berlin Heidelberg, 2011. p. 341-348.

[22] Scannapieco, M., Missier, P., & Batini, C. (2005). Data Quality at a Glance. Datenbank-Spektrum, 14, 6-14.

[23] Souza, J., Botega, L. C., Segundo, J. E. S., Berti, C. B., de Campos, M. R., & de Araújo, R. B. (2015). Conceptual framework to enrich situation awareness of emergency dispatchers. In *Human Interface and the Management of Information. Information and Knowledge in Context* (pp. 33-44). Springer International Publishing.

[24] *W3C Semantic Web. RDF – Resource Description Framework. < http://www.w3.org/RDF/>. Setembro 2014.*

[25] Winkler, W. E. (2004). Methods for evaluating and creating data quality. Information Systems, 29(7), 531-550.

[26] ZAVERI, Amrapali et al. Quality assessment methodologies for linked open data. **Submitted to Semantic Web Journal**, 2013.

[27] WAND,Y.AND WANG, R. 1996. Anchoring data quality dimensions in ontological foundations. Comm. ACM 39, 11.

[28] REDMAN, T. 1996. Data Quality for the Information Age. Artech House.

[29] JARKE, M., LENZERINI, M., VASSILIOU,Y.,ANDVASSILIADIS, P., Eds. 1995. Fundamentals of Data Warehouses. Springer Verlag

[30] BOVEE, M., SRIVASTAVA, R., ANDMAK, B. September 2001. A conceptual framework and belief-function approach to assessing overall information quality. In Proceedings of the 6th International Conference on Information Quality

[31] NAUMANN, F. 2002. Quality-driven query answering for integrated information systems. Lecture Notes in Computer Science, vol. 2261.

[32] LIU,L.ANDCHI, L. Evolutionary data quality. In Proceedings of the 7th International Conference on Information Quality. 2002.

[33] SANTAREM SEGUNDO, J. E. Web Semântica: Introdução a Recuperação de Dados Usando Sparql. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO: ALÉM DAS NUVENS, EXPANDINDO AS FRONTEIRAS DA CIÊNCIA DA INFORMAÇÃO, 15.1: 3863-3882, Belo Horizonte, MG. Anais.... 2014.

[34] W3C. Guidelines for Collecting Metadata on Linked Datasets in the datahub.io Data Catalog. Disponível em:

<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation>. Acesso em: 14 set. 2015

[35] HEATH, Tom; HAUSENBLAS, Michael; BIZER, Christian; CYGANIAK, Richard; HARTIG, Olaf. How to publish linked data on the web. In: Tutorial in the 7th International Semantic Web Conference, Karlsruhe, Germany. 2008

[36] BATINI, Carlo; SCANNAPIECA, Monica. Data Quality Dimensions. **Data Quality: Concepts, Methodologies and Techniques**, p. 19-49, 2006.

[37] BOUZEGHOUB, Mokrane. A framework for analysis of data freshness. In: Proceedings of the 2004 international workshop on Information quality in information systems. ACM, 2004. p. 59-67.

[38] Rivard, S., Poirier, G., Raymond, L., & Bergeron, F. Development of a measure to assess the quality of user-developed applications. ACM SIGMIS Database, v. 28, n. 3, p. 44-58, 1997.

[39] Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. A framework for information quality assessment. Journal of the American Society for Information Science and Technology, v. 58, n. 12, p. 1720-1733, 2007.

[40] GUÉRET, C., G ROTH, P., S TADLER, C., AND LEHMANN, J. Assessing linked data mappings using network measures. In ESWC (2012).

[41] FLEMMING, A. Quality characteristics of linked data publishing datasources. Master's thesis, Humboldt-Universität zu Berlin, 2010

[42] O'BRIEN, J. Sistemas de Informação e as Decisões Gerenciais na Era da Internet. 2. ed. São Paulo: Saraiva, 2004. v. 2, p. 431.

[43] WANG, R., & STRONG, D. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems. p. 5–33.

[44] LI, Pei et al. Linking temporal records. **Proceedings of the VLDB Endowment**, v. 4, n. 11, p. 956-967, 2011.