UNIVERSIDADE ESTADUAL PAULISTA "*JÚLIO MESQUITA FILHO*"
FACULDADE DE FILOSOFIA E CIÊNCIAS – CAMPUS DE MARÍLIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

JESSICA OLIVEIRA DE SOUZA FERREIRA MELO

METODOLOGIA DE AVALIAÇÃO DE QUALIDADE DE DADOS NO CONTEXTO DO LINKED DATA

MARÍLIA
2017

JESSICA OLIVEIRA DE SOUZA FERREIRA MELO

METODOLOGIA DE AVALIAÇÃO DE QUALIDADE DE DADOS NO CONTEXTO DO LINKED DATA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP) – Faculdade de Filosofia e Ciências de Marília, como requisito para obtenção do título de Mestre em Ciência da Informação.

Linha de Pesquisa: Informação e Tecnologia

Orientador: Dr. José Eduardo Santarém Segundo

MARÍLIA
2017

Melo, Jessica Oliveira de Souza Ferreira
M528m Metodologia de avaliação de qualidade de dados no contexto do linked data / Jessica Oliveira de Souza Ferreira Melo. – Marilia, 2017.
111 f. ; 30 cm.

Orientador: José Eduardo Santarém Segundo.
Dssertação (Mestrado em Ciência da Informação) – Universidade Estadual Paulista (UNESP), Faculdade de Filosofia e Ciências, 2017.
Bibliografia: f. 107-111.

1. Dados ligados. 2. Web semântica. 3. Avaliação. 4. Metodologia. I Título.

CDD 004.67

JESSICA OLIVEIRA DE SOUZA FERREIRA MELO

METODOLOGIA DE AVALIAÇÃO DE QUALIDADE DE DADOS NO CONTEXTO DO LINKED DATA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio Mesquita Filho (UNESP) – Faculdade de Filosofia e Ciências de Marília, como requisito para obtenção do título de Mestre em Ciência da Informação.

Área de Concentração: Informação, Tecnologia e Conhecimento
Linha de Pesquisa: Informação e Tecnologia
Data da Defesa: 09 de maio de 2017.

BANCA EXAMINADORA

José Eduardo Santarém Segundo (Orientador)
Docente do Programa de Pós-Graduação em Ciência da Informação da UNESP/FFC

Silvana Aparecida Borsetti Gregório Vidotti
Docente do Programa de Pós-Graduação em Ciência da Informação da UNESP/FFC

Leonardo Castro Botega
Docente do Departamento de Ciência da Computação da UNIVEM/Marília

AGRADECIMENTOS

Agradeço à Jeová Deus, por tudo de bom em minha vida.

Agradeço ao meu marido Danilo, que sempre esteve ao meu lado e fez o necessário para que eu pudesse alcançar mais esta conquista em minha vida.

Ao meu orientador Prof. Eduardo pela oportunidade, pelos conselhos, paciência e orientações que foram essenciais para a conclusão do mestrado.

Aos Prof. Leonardo por ter estado sempre à disposição para me auxiliar e aconselhar em diversos momentos no decorrer desta pesquisa.

À Profa. Silvana por ter disponibilizado um ambiente de estudo onde pude compartilhar experiências e aprender dos colegas de laboratório, agradeço também pelas suas contribuições e por ter participado da minha banca de defesa.

À Ana Maria, Sandra, Edgar, Fernanda e Caio, companheiros do PPGCI que me ajudaram em diversos momentos e contribuíram também para minha formação acadêmica.

À Maria Lucia Balestriero, pela correção ortográfica.

Aos professores, ao programa de PPGCI e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo fomento cedido para o desenvolvimento desta pesquisa.

Aos meus pais, irmãs e sobrinhos que sempre estiveram ao meu lado e compartilharam comigo a alegria de momentos importantes como este, em minha vida.

An extraordinary amount of arrogance is present in any claim of having been the first in inventing something
Benoit Mandelbrot

MELO, Jessica Oliveira de Souza Ferreira. **Metodologia de Avaliação de Qualidade de Dados no Contexto do Linked Data**. 2017. 111f. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Filosofia e Ciências, Marília/SP, 2017.

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 *Linked Open Data* (LOD), que consiste em uma iniciativa para promover a publicação de dados linkados (*Linked Data*). Com o evidente crescimento dos dados publicados como *Linked Data*, a qualidade tornou-se essencial para que tais conjuntos de dados (*datasets*) atendam os objetivos básicos da Web Semântica. Isso porque problemas de qualidade nos *datasets* 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 disponibilizados como *Linked Data* possibilitam um ambiente favorável para aplicações inteligentes, problemas de qualidade podem também dificultar ou impedir a integração dos dados provenientes de diferentes *datasets*. A literatura aplica diversas dimensões de qualidade no contexto do *Linked Data*, porém indaga-se a aplicabilidade de tais dimensões para avaliação de qualidade de dados linkados. Deste modo, esta pesquisa tem como objetivo propor uma metodologia para avaliação de qualidade nos *datasets* de *Linked Data*, bem como estabelecer um modelo do que pode ser considerado qualidade de dados no contexto da Web Semântica e do *Linked Data*. Para isso adotou-se uma abordagem exploratória e descritiva a fim de estabelecer problemas, dimensões e 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 sete dimensões de qualidade aplicáveis ao domínio do *Linked Da*ta e 14 fórmulas diferentes para a quantificação da qualidade de *datasets* sobre publicações científicas. Por fim realizou-se uma prova de conceito na qual a metodologia de avaliação de qualidade proposta foi aplicada em um *dataset* promovido pelo LOD. Conclui-se, a partir dos resultados da prova de conceito, que a metodologia proposta consiste em um meio viável para quantificação dos problemas de qualidade em *datasets* de *Linked Data*, e que apesar dos diversos requisitos para a publicação deste tipo de dados podem existir outros *datasets* que não atendam determinados requisitos de qualidade, e por sua vez, não deveriam estar inclusos no diagrama do projeto LOD.

Palavras-Chave: *Linked Data*, Qualidade de dados, Metodologia de avaliação, Web Semântica.

MELO, Jessica Oliveira de Souza Ferreira. **Data Quality Assessment Methodology in Linked Data Context**. 2017. 111f. Dissertation (Master in Information Science) – São Paulo State University "Júlio de Mesquita Filho", Faculty of Philosophy and Sciences, Marília/SP, 2017.

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 Linked Open Data (LOD) project was created, which consists of an initiative to promote the publication of Linked Data. With the evident growth of data published as Linked Data, quality has become essential for such datasets to meet the basic goals of the Semantic Web. This is because quality problems in published datasets are a hindrance not only to their use but also to applications that make use of such data. Considering that data made available as Linked Data enables a favorable environment for intelligent applications, quality problems can also hinder or prevent the integration of data coming from different datasets. The literature applies several quality dimensions in the context of Linked Data, however, the applicability of such dimensions for quality evaluation of linked data is investigated. Thus, this research aims to propose a methodology for quality evaluation in Linked Data datasets, as well as to establish a model of what can be considered data quality in the Semantic Web and Linked Data context. 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 indexes. This work resulted in the definition of seven quality dimensions applicable to the Linked Data domain and 14 different formulas for the quantification of the quality of datasets on scientific publications. Finally, a proof of concept was developed in which the proposed quality assessment methodology was applied in a dataset promoted by the LOD. It is concluded from the proof of concept results that the proposed methodology consists of a viable means for quantification of quality problems in Linked Data datasets and that despite the diverse requirements for the publication of this type of data there may be other datasets that do not meet certain quality requirements, and in turn, should not be included in the LOD project diagram..

Keywords: Linked Data, Data Quality, Assessment Methodology, Semantic Web

LISTA DE ILUSTRAÇÕES

Figura 1 – Estrutura da Web Semântica 24
Figura 2 – Representação conceitual por meio de URIs 27
Figura 3 – Sintaxe do modelo de representação RDF/XML 28
Figura 4 – Sintaxe do modelo de representação RDFa 28
Figura 5 – Sintaxe do modelo de representação *Turtle* 28
Figura 6 – Conjunto de triplas descrevendo o recurso Vincent Van Gogh 31
Figura 7 – Exemplo de uma *query* SPARQL para buscar obras relacionadas do pintor Vincent Van Gogh 32
Figura 8 – Resultado da busca realizada no exemplo da Figura 9 32
Figura 9 – Arquitetura geral da web 34
Figura 10 – Modelo conceitual do funcionamento da web semântica 35
Figura 11 – Nuvem de *Linked Data* publicados como LOD em 2011 37
Figura 12 – Nuvem de *Linked Data* publicados como LOD em 2014 38
Figura 13 – Diagrama atualizado LOD atualizado em 2017 39
Figura 14 – Crescimento de *datasets* no LOD 40
Figura 15 – Classificação das dimensões de qualidade 50
Figura 16 – Modelo de dimensões e requisitos de qualidade 51
Figura 17 – Declaração RDF com problemas de completude 52
Figura 18 – Relevância no contexto da recuperação da informação 57
Figura 19 – RDF link com problema de *interlinking* 63
Figura 20 – Problema de qualidade encontrado quanto ao ISBN incompleto de um recurso 65
Figura 21 – Modelo proposto para qualidade de dados no contexto do *Linked Data* 68
Figura 22 – Metodologia de avaliação de qualidade de dados para *Linked Data* 70
Figura 23 – Classificação das dimensões de qualidade para a metodologia de avaliação 71
Figura 24 – Localização da licença, quando disponibilizada pelos datasets 77
Figura 25 – Padrão de dados descritivos disponibilizados sobre os recursos analisados na amostragem 92
Figura 26 – Suposto site da documentação do vocabulário de descrição dos dados no *dataset* avaliado 95
Figura 27 – Classe não definida na documentação sendo utilizada para descrição do recurso 97

LISTA DE QUADROS

Quadro 1 – Regras de qualidade para avaliação no *Linked Data* do SWIQA.....16
Quadro 2 – Diferentes metodologias de avaliação de qualidade no *Linked Data*.....19
Quadro 3 – Tecnologias e protocolos da web de documentos e semântica.....35
Quadro 4 – Informações básicas, de nível 1 necessárias para publicar *datasets*.....45
Quadro 5 – Informações de nível 2 necessárias para publicação de *datasets* 45
Quadro 6 – Dados necessários para o nível 3 de completude do *dataset*.....45
Quadro 7 – Definições de completude de acordo com diferentes autores.....53
Quadro 8 – Diferentes definições de precisão.....55
Quadro 9 – Requisitos de qualidade para *datasets* de *Linked Data* 69
Quadro 11 – Metadados para descrição de artigos publicados em eventos científicos.....85
Quadro 12 – Metadados para descrição de teses e dissertações.....86
Quadro 13 – Metadados para descrição de artigos publicados em revistas científicas 87
Quadro 14 – Verificação da disponibilização dos dados descritivos de nível 1.....99
Quadro 15 – Verificação da disponibilização dos dados descritivos de nível 2.....99
Quadro 16 – Verificação da disponibilização dos dados descritivos de nível 3.....99
Quadro 17 – Quadro 10 preenchido com os índices de qualidade de cada dimensão.....103

SUMÁRIO

1. INTRODUÇÃO 12

2 WEB SEMÂNTICA, *LINKED DATA* E SUAS TECNOLOGIAS 22

2.1 Linked Data 36

2.2 Processo de publicação de datasets no LOD 41

3 QUALIDADE DE DADOS 47

3.1 Dimensões de Qualidade de Dados 51

3.2 Metodologias para avaliação de qualidade 59

3.3 Qualidade de dados no Linked Data 61

3.4 Dimensões de Qualidade para Linked Data 62

3.5 Metodologias para avaliação de qualidade no Linked Data 66

4 MODELO E METODOLOGIA DE AVALIAÇÃO DE QUALIDADE DE DADOS NO CONTEXTO DO *LINKED DATA* 68

4.1 Interlinking 73

4.2 Licenciamento 76

4.3 Consistência 78

4.4 Precisão Sintática 80

4.5 Precisão Semântica 81

4.6 Completude 82

4.7 Avaliação Temporal (Timeliness e Volatilidade) 88

5 PROVA DE CONCEITO 90

5.1 Interlinking 91

5.2 Licenciamento 93

5.3 Consistência 94

5.4 Precisão Sintática 96

5.5 Precisão Semântica 96

5.6 Completude 98

5.7 Avaliação Temporal (Timeliness e Volatilidade) 102

5.8 Índice de qualidade geral 102

6 CONSIDERAÇÕES FINAIS 104

REFERÊNCIAS 107

1. INTRODUÇÃO

Bizer et al (2009) descrevem *Linked Data* como a prática de utilizar a Web para criar *links* entre dados de diferentes fontes utilizando tecnologias e conceitos da Web Semântica. No entanto, os conjuntos de dados publicados podem variar na qualidade, podendo conter estruturas precisas, ou seja, dados bem formatados e livres de erros ou conjuntos dotados de diversos problemas de qualidade, como incompletude, imprecisão, *links* quebrados, etc. Qualidade pode ser considerada um conjunto de requisitos necessários para um dado atender às exigências, de acordo com domínios específicos.

Uma das medidas mais conhecidas para evitar problemas de qualidade no *Linked Data*, 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. Caso contrário limita a reutilização desses 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 (*Resource Description Framework*), 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.

Além das regras quatro regras definidas por Berners-Lee (2006), 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 dados linkados abertos (LOD – abreviação em inglês para *Linked Open Data*), 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.

1.1 Definição do problema

Conforme explicitado por Marcondes e Sayão (2009), resultados e informações provenientes de atividades de pesquisa na forma de publicações devem necessariamente também ser públicos, de modo que sejam usados amplamente. No contexto do *Linked Data*, os problemas de qualidade que têm afetado os conjuntos de dados publicados na categoria de Publicações podem ser considerados obstáculos não apenas para a utilização, mas também para aplicações que fazem uso de tais dados. Definições quanto às dimensões de qualidade para o *Linked Data* e para os métodos de avaliação são disponibilizadas, porém indaga-se como tais problemas e dimensões acontecem nos conjuntos de dados sobre publicações científicas. Indaga-se também, de que modo os problemas de qualidade podem ir contra os princípios de utilização dos dados linkados e qual é a proporção de tais problemas de acordo com tal categoria específica de *datasets*.

Desse modo, pretende-se por meio desta pesquisa responder aos seguintes questionamentos quanto à qualidade dos dados em *datasets* de publicações:

- Quais os requisitos de qualidade para *Linked Data*?
- Quais dimensões de qualidade podem ser aplicadas no domínio do *Linked Data*?
- Como verificar se o *dataset* realmente cumpre com requisitos de qualidade?
- Há fórmula para julgar quantitativamente os *datasets*?

1.2 Motivação e Justificativa

Conforme cunhado por Borko (1968), investigar propriedades, comportamento, meios de processamento de informações para acessibilidade, usabilidade, recuperação, interpretação, transmissão, armazenamento e utilização são temas de grande interesse para a comunidade da Ciência da Informação.

Considerando que os dados linkados contribuem de forma relevante para aplicações computacionais, a detecção de problemas de qualidade nesse tipo de dados, bem como meios de mensurá-los ou eliminá-los representa a principal motivação desta pesquisa.

O estudo justifica-se pela evidente presença de problemas de qualidade que afetam os *datasets* e a necessidade de uma descrição específica de como tais problemas acontecem na categoria de publicações científicas do diagrama LOD (MENDES et al., 2012; BIZER; CYGANIAK, 2008; ZAVERI et al., 2012; KONTOKOSTAS et al., 2013; RULA; ZAVERI, 2014;).

1.3 Objetivos

Esta pesquisa tem como objetivo geral definir uma metodologia de avaliação de qualidade para dados publicados em um segmento específico do *Linked Data*, o de publicações. Quanto aos objetivos específicos:

- Descrever os problemas de qualidade de modo geral;
- Definir um modelo do que é qualidade de dados no contexto da Web Semântica e do *Linked Data*
- Definir dimensões de problemas de dados para o *Linked Data*;
- Estabelecer funções de avaliação para detecção de problemas;
- Realizar um estudo de caso para validação da metodologia proposta.

1.4 Metodologia

No desenvolvimento da pesquisa foram utilizados diferentes métodos para a concretização do objetivo. A metodologia, as dimensões de qualidade, as métricas e os requisitos da avaliação foram definidos por meio de um estudo exploratório e descritivo da bibliografia.

A primeira etapa consistiu em realizar o levantamento bibliográfico, buscando por publicações nacionais e internacionais a fim de obter embasamento teórico sobre Web Semântica, qualidade de dados, qualidade de dados no *Linked Data* e metodologias de avaliação de qualidade no *Linked Data*.

Na segunda etapa foi realizada a leitura, a interpretação e a análise do material selecionado, que resultou em uma definição, a partir do conhecimento agregado, das dimensões, métricas e métodos a serem empregados para a etapa seguinte.

A terceira etapa foi composta pela definição do processo de avaliação e fórmulas de acordo com as métricas e as dimensões de qualidade aplicáveis para o contexto do *Linked Data*.

Foi constituído um modelo do que significa qualidade de dados no contexto do *Linked Data,* baseado em princípios de qualidade obtidos nas especificações e recomendações para Web Semântica e *Linked Data* do (1) W3C, (2) do *Linked Open Data* e da (3) literatura. Então, a partir dos resultados provenientes do modelo, foi realizada uma análise aplicada nos dados de publicações publicados como *Linked Data,* para a definição da metodologia de avaliação de qualidade.

Na quarta etapa, obtiveram-se informações sobre quais as dimensões que, de fato, estão presentes no conjunto de *datasets* selecionados para realizar a análise, quais são as dimensões

mais comuns, as que menos ocorrem, e, caso existam, as dimensões que não foram descritas na literatura analisada.

Na quinta etapa foi elaborada e proposta uma metodologia para avaliação de qualidade em *datasets* da categoria de Publicações, visando detectar e quantificar a ocorrência de tais problemas.

A metodologia foi definida visando auxiliar usuários que pretendem avaliar *datasets* já publicados e usuários que pretendem publicar dados que atendam aos princípios e requisitos de qualidade para *datasets* de *Linked Data*.

No final, foi realizada uma prova de conceito na qual foi realizada a avaliação de um *dataset* do LOD, onde foram aplicadas as sete dimensões definidas, as métricas e as fórmulas para avaliação de cada dimensão.

1.5 Trabalhos Relacionados

Existem diferentes formas de avaliar a qualidade de dados no *Linked Data;* verifica-se que elas vêm sendo implementadas principalmente na área da Ciência da Computação e serão descritas a seguir. Na sequência serão descritos alguns trabalhos que descrevem metodologias ou modelos de avaliação de qualidade para *Linked Data*, os quais foram selecionados em vista da similaridade com o tema e dos objetivos desta pesquisa, por terem uma grande quantidade de citações e pela metodologia utilizada para a definição das dimensões de qualidade específicas para *Linked Data*.

Fürber e Hepp (2011) descrevem um *framework* para avaliação da qualidade das informações na Web Semântica (SWIQA). A avaliação proposta pelos autores é aplicada de acordo com as seguintes dimensões: precisão sintática, precisão semântica, completude, *timeliness* e singularidade. De acordo com os autores, o processo de avaliação de qualidade de dados consiste em um processo de atribuição de valores numéricos e categóricos às dimensões avaliadas. O *framework* tem como objetivo aumentar o nível de objetividade da avaliação por meio de cálculos específicos para cada dimensão.

A fim de atender a objetividade da avaliação proposta, os autores definiram nove regras genéricas as quais foram aplicadas, verificadas e quantificadas de acordo com as cinco dimensões. As regras, que foram definidas com base em tipologias de pesquisas orientada a banco de dados, são apresentadas no Quadro 1.

Quadro 1 – Regras de qualidade para avaliação no *Linked Data* do SWIQA

Regra de Qualidade	Definição	Exemplo
Propriedade mandatória e regras literais	Propriedades e suas literais se tornam mandatórias, quando o dado é requisitado para a tarefa em mãos	As propriedades indicando coordenadas geográficas devem existir e ter valores para todas as instâncias de classes *foo:Location* possibilitando navegar para cada localização
Regras sintáticas	Regras sintáticas definem o tipo de caracteres e/ou o modelo de valores literais	Literais valores da propriedade *foo:country-name* deve conter apenas letras
Regras de dependência funcional	Dependências funcionais são dependências entre valores de duas ou mais propriedades diferentes	O valor literal para *foo:city* é sempre dependente do valor literal para *foo:country*, visto que certos nomes de cidades existem em apenas alguns países
Regras de valores legais	Valores legais são a definição explícita dos valores permitidos para uma propriedade específica	A propriedade *foo:gender* deve conter apenas valores "masculino" e "feminino"
Regras de coleção de valores legais	São a definição explícita de uma coleção de valores para propriedades numéricas. Contém valor máximo e mínimo.	A propriedade *foo:population* deve conter apenas valores maiores que zero.
Regras de valores ilegais	Regras que explicitam a definição de valores que podem ou não ser atribuídos a determinada propriedade	A propriedade *foo:gender* não poderia nunca conter o valor de uma propriedade de e-mail.
Regras de coleção de valores ilegais	Conjuntos de valores proibidos para propriedades de valores numéricos. Um conjunto de valores contém limite máximo e mínimo.	A propriedade foo:population não pode conter valor menor que um.
Regras de valores únicos	Definem propriedades que talvez contenham cada valor literal, não mais do que uma vez dentro de uma coleção de valores	Cada valor para a propriedade foo:ISBN em instâncias da classe *foo:Book* não podem ocorrer mais de uma vez
Regras de valores desatualizados	Identificam instâncias que representam um estado desatualizado da entidade correspondente no mundo real	Instâncias da classe *foo:Offer* estão desatualizadas se o valor para *foo:validThrough* é mais velho do que a data e horário atual.

Fonte: Fürber e Hepp (2011)

As métricas para o índice de qualidade dos dados realizado no SWIQA basearam-se no cálculo de média simples, aplicada por subtrair a média entre o número total de instâncias que violaram as regras de qualidade (DQRV) do número total de instâncias relevantes (T). Outra fórmula foi utilizada pelos autores no caso da atribuição de peso às propriedades, que possibilita atribuir a importância a propriedades específicas. Os autores propõem quatro fórmulas, sendo:

- IQ-Score = (1- (QDRV / T)); para o cálculo de índices individuais de qualidade que pode ser aplicado em cada propriedade do *dataset*;
- IQ-Score = (1- (DQRV / T)) x ω; multiplica a média pelo peso, ω, que representa

a importância da propriedade para a tarefa a ser realizada;

- Agregado-IQ-Dim-Score$_w$ = (IQScore$_w$) / W; corresponde a um índice de qualidade agregado para cada dimensão que leva em consideração diferentes níveis de importância do dado para a tarefa a ser realizada. Calcula-se o peso para cada propriedade e então divide-se a soma do resultado por todos os fatores de peso das propriedades consideradas (W);

- Agregado-IQ-Dim-Score = (IQScore) / P; a ser utilizada quando não é possível definir os valores de importância, esta fórmula realiza a agregação dos valores sem o peso, onde o resultado da soma dos valores é dividido pelo número de propriedades testadas (P);

Mendes et al (2011) descrevem um *framework* de integração de *Linked Data* chamado Sieve, responsável por lidar com conflitos provenientes do processo de integração de dados, que causa a existência de múltiplos valores para o mesmo atributo que admite apenas um valor.

No Sieve, o usuário avaliador define as características que indicam se um dado é de alta qualidade, como a qualidade é definida e como deve ser armazenada no sistema. As métricas foram definidas de acordo com indicadores de qualidade e a partir dos indicadores foi realizado o cálculo do índice de qualidade. Conforme estabelecido pelos autores, os indicadores de qualidade utilizados no Sieve podem ser metadados sobre as circunstâncias na qual os dados foram criados, informações sobre o provedor dos dados ou as classificações providas pelos consumidores dos dados ou experientes no domínio. São propostas duas dimensões de avaliação de qualidade:

- Proximidade do tempo: calcula a distância entre a data de entrada do grafo de proveniência e a data atual; os dados mais recentes recebem um valor próximo a 1.

- Preferência: atribui pontuação decrescente para cada grafo URI provido na configuração.

Rula e Zaveri (2014) descrevem uma metodologia para avaliação de qualidade no *Linked Data* partindo de um conceito popular na literatura, o qual relaciona a qualidade dos dados com a aptidão para o uso. Para conduzir a análise proposta na metodologia, os autores sugerem a comparação com os valores da fonte original de dados, ou com um *dataset* do mesmo domínio. Esta metodologia tem o usuário avaliador como fonte de requisitos, identificação de problemas e definição de dimensões para a análise. As fases e suas etapas são explicadas a seguir:

- Fase 1: Análise de requisitos: realizada pelo usuário avaliador.
    - Passo 1 - Análise do caso de uso: o usuário identifica detalhes sobre a utilização do *dataset*, provendo assim requisitos para a aptidão da utilização do conjunto de dados;
- Fase 2: Avaliação de qualidade de dados
    - Passo 2 - Identificação de problemas de qualidade: o usuário identifica o conjunto mais relevante de problemas de qualidade por meio de uma lista de verificação.
    - Passo 3 - Análise estatística e de baixo-nível: executa análise de estatísticas genéricas que podem ser calculadas automaticamente;
    - Passo 4 - Análise avançada: as métricas de índice são baseadas no cálculo de média simples. A média é medida por subtrair a média entre o número total de instâncias que violam uma regra de qualidade (V) e o número total de instâncias relevantes (T); o resultado é chamado de índice DQscore. Também é possível calcular a qualidade das propriedades e atributos como um todo; neste cálculo, o DQscore é multiplicado pelo peso wi, que representa a importância da tarefa a ser realizada para cada propriedade no *dataset*, e em seguida o peso do DQscore é dividido pela soma de todos os fatores de ponderação das propriedades consideradas (W). Ao final desta fase, os índices dos passos 2 a 5 são agregados e providos como resultado do indicador de qualidade do *dataset*.
- Fase 3: Aperfeiçoamento da qualidade
    - Passo 5 - Análise da causa raiz: este passo consiste em analisar se o problema acontece no *dataset* original e se o *dataset* original não estiver disponível, analisar o *dataset* e detectar a causa.
    - Passo 6 - Consertar problemas de qualidade: sugere-se a utilização de abordagens automáticas ou semiautomáticas e mecanismos de *crowdsourcing*.

O Quadro 2 apresenta metodologias para avaliação de qualidade de dados no *Linked Data,* que foram desenvolvidas de acordo com diferentes dimensões, ferramentas e objetivo.

Quadro 2 – Diferentes metodologias de avaliação de qualidade no *Linked Data*

Definição	Dimensões	Metodologia
Metodologia de avaliação para *Linked Data* por meio de uma ferramenta de *Crowdsourcing*	Precisão, Relevância, Consistência representacional e *Interlinking*	1 - Seleção de recurso: o *dataset* é selecionado para a avaliação, processo que pode ser feito de três maneiras diferentes: por classe, em que são selecionados recursos de uma classe específica; aleatoriamente, em que um recurso é selecionado de modo aleatório no *dataset* e manualmente, em que os recursos a serem avaliados são selecionados de forma manual. 2 - Seleção do modo de avaliação: pode ser realizada de três maneiras diferentes: manualmente, em que os recursos são atribuídos a um avaliador humano, que irá avaliar os recursos manual e individualmente; semiautomático: os recursos são inseridos em uma ferramenta de avaliação semiautomática, que retorna um feedback ao usuário e automática, em que os recursos são inseridos na ferramenta de avaliação automática sem qualquer interferência do humano. 3 - Avaliação dos recursos: potenciais problemas de qualidade são identificados. 4 - Melhoria da qualidade dos dados: pode ser realizada de forma direta ou indireta, sendo que na forma direta ocorre a edição da tripla em que ocorreu o problema pelo valor correto. E na forma indireta os autores incentivam a utilização de um *framework* que permite ao usuário obter feedbacks sobre triplas incorretas (KONTOKOSTAS et al., 2013).
Metodologia para avaliação de qualidade que utiliza *crowdsourcing* como um meio para lidar com os problemas encontrados em *datasets* do DBpedia	Interlinking, completude e precisão	1 - Seleção dos recursos: pode ser realizada de forma manual ou por classes específicas. 2 - Avaliação das triplas em que foram detectadas as triplas incorretas. Os autores lançaram um debate alvejando um público de pesquisadores e entusiastas especialistas em *Linked Data* a fim de encontrar e classificar triplas RDF incorretas. 3 - Seleção dos problemas de qualidade foi realizada a fim de verificar três tipos de problemas nos *datasets* do DBpedia: (1) valores incorretos de objetos ou extraídos de modo incompleto; (2) tipo de dados extraídos incorretamente e (3) link incorreto entre entidades do DBpedia e fontes relacionadas na Web. 4 - Lista das triplas incorretas classificadas de acordo com o problema de qualidade 5 - Avaliação das triplas: A avaliação foi conduzida a fim de investigar as seguintes questões, que foram encaradas como requisitos: (1) é possível detectar problemas de qualidade no *dataset* do *Linked Data* por meio de mecanismos de *crowdsourcing*? (2) que tipo de público é o mais adequado para cada tipo de problema de qualidade? (3) que tipos de erros são cometidos por leigos e especialistas? (ACOSTA et al., 2013)

Definição	Dimensões	Metodologia
Metodologia para avaliar recursos do *Linked Data* utilizando o conceito de testes dirigidos, provenientes do desenvolvimento de *software*	Consistência, completude e precisão	1 - São definidos os modelos de testes de qualidade de dados por meio do feedback da comunidade de usuários do DBpedia e de geradores automático de testes (TAG) que utilizam inferências e axiomas OWL/RDF. 2 - Os testes são instanciados de acordo com diferentes requisitos disponibilizados pelos usuários e pelas TAGs 3 - Os testes são utilizados na avaliação de qualidade

Ao analisar a forma como as metodologias foram estabelecidas podem-se notar diferentes pontos de vistas e objetivos para cada uma delas. A maioria das metodologias descritas na literatura segue um princípio base da qualidade de dados: um dado é de qualidade quando se adequa às tarefas que o usuário executará. Por esse motivo, não somente as citadas, mas grande parte das metodologias levam em conta o usuário, em todo ou em alguma parte do processo de avaliação.

Nota-se também que não há um padrão das dimensões avaliadas em cada metodologia, visto que cada uma foi desenvolvida para avaliar um conjunto de dados, ou realizar alguma tarefa específica. Pode-se afirmar também que nem todas adotam dimensões exclusivas para o domínio do *Linked Data*, assim como não houve também um processo de definição de quais seriam as prioridades e dimensões aplicáveis ao contexto.

Outro fator importante é que os testes, métodos e requisitos utilizados nas metodologias derivam de domínios diferentes do avaliado, como por exemplo, o *framework* descrito por Fürber e Hepp (2011) define as regras para avaliação com base em tipologias de pesquisas orientadas a banco de dados. Verificou-se também que todas as metodologias para avaliação do *Linked Data* vêm sendo implementadas na área da Ciência da Computação.

1.6 Organização do trabalho

Excluindo este capítulo inicial, que introduziu o contexto, abordou trabalhos relacionados, objetivos e meios para a concretização desta pesquisa, o trabalho está organizado do seguinte modo:

O Capítulo 2 – WEB SEMÂNTICA, *LINKED DATA* E SUAS TECNOLOGIAS – fornece embasamento teórico sobre a Web Semântica, o que é, como funciona, quais são seus

requisitos. Também descreve o que é o *Linked Data*, quais são os requisitos para publicação e inserção no diagrama LOD.

O Capítulo 3 – QUALIDADE DE DADOS – apresenta uma introdução sobre qualidade de dados no geral e metodologias para avaliação de qualidade de dados. Define dimensões e metodologias de avaliação de qualidade aplicadas no *Linked Data* de acordo com a literatura.

O Capítulo 4 – METODOLOGIA DE AVALIAÇÃO DE QUALIDADE DE DADOS NO CONTEXTO DO *LINKED DATA* – aborda o modelo de qualidade para *Linked Data* proposto, bem como a metodologia de avaliação de acordo com as dimensões definidas, métricas, processos e cálculos de índices.

O Capítulo 5 – ESTUDO DE CASO – aborda um estudo de caso no qual aplica a metodologia de avaliação de dados proposta em um *dataset* de *Linked Data* inserido no LOD.

Por fim, o Capítulo 6 – CONSIDERAÇÕES FINAIS – discorre sobre os obstáculos experienciados no decorrer da pesquisa, conclusões e possíveis trabalhos futuros.

2 WEB SEMÂNTICA, *LINKED DATA* E SUAS TECNOLOGIAS

Este capítulo tem como objetivo descrever os conceitos de Web Semântica, *Linked Data*, bem como o processo de publicação de dados como *Linked Data*.

A Web de documentos proporcionou um avanço significativo para o modo como as informações eram disponibilizadas e as tarefas eram realizadas na década de 90. O nome Web de documentos dá-se pelo fato desta conectar recursos (páginas Web) que podem ser relacionados com outros recursos por meio URIs. Diferentemente da Web de dados (ou Web Semântica), na qual não apenas os recursos, mas também os dados contidos nas páginas possuem elementos descritivos de modo que as máquinas possam realizar um processamento otimizado dos dados (ISOTANI; BITTENCOURT, 2015).

Uma das primeiras menções de uma Web de documentos foi feita por Berners-Lee (1989) que propunha um ambiente no qual as informações sobre projetos, códigos, documentos, etc. seriam acessadas por meio de um sistema de ligação entre esses itens. Isto, por sua vez, auxiliaria na organização e facilitaria o acesso das informações em rede de grandes ambientes computacionais. Em sua proposta, Berners-Lee (1989) apresentou a necessidade de um sistema que auxiliasse o gerenciamento e a utilização da informação no CERN (*Conseil Européen pour la Recherche Nucleáire*), ressaltando as características dessa organização, que envolvia milhares de pessoas e apresentava dificuldades para que informações sobre projetos e códigos já existentes fossem armazenadas de forma que facilitasse o posterior acesso e guiasse a informações relacionadas.

Posteriormente, já na década de 90, esse conceito, visando primeiramente atender às necessidades do CERN, impulsionou a criação da *World Wide Web*, descrita como um universo global de informações no qual é possível navegar por meio dos documentos inseridos nele, bem como pelas referências por meio de *hiperlinks*. E para que isso acontecesse, seria necessário um esquema de endereçamento, um protocolo de comunicação e um formato para os documentos (BERNERS-LLE et al., 1992).

Diferentemente da proposta inicial, a web de documentos atual utiliza padrões e tecnologias consolidadas para realizar o endereçamento, a comunicação e a apresentação de documentos, sendo eles: URI, HTTP e HTML (*HyperText Markup Language*). Tais tecnologias possibilitam a identificação do recurso (URI), seja este um documento, uma empresa, seres humanos, livros, imagens, serviços etc., que são apresentados por meio de uma linguagem de marcação (HTML) e acessados por meio de um protocolo de comunicação global (HTTP).

Desde o seu propósito inicial de criação, que visava proporcionar uma melhor

organização dos documentos em uma grande organização, a sua finalidade expandiu de modo a conectar e disponibilizar documentos acessíveis em qualquer lugar do mundo. Assim, os questionamentos atuais relacionam-se com a utilização do grande volume de dados disponibilizados nos documentos web.

Uma das principais características da web de documentos é que os dados estão disponibilizados de forma não estruturada visto que o HTML se encontra na estruturação de documentos textuais, não dos dados. Isso significa que os dados não possuem uma organização que atribua um modelo definido.

A existência de um modelo para estruturação dos dados auxilia agentes web a conectar dados distribuídos em diferentes documentos disponibilizados na web. Atualmente, existem as Web APIs (*Application Programming Interface*) que proveem um meio de acesso a dados estruturados utilizando o protocolo HTTP. Por meio dessas aplicações grandes organizações disponibilizam os dados gerados em seus ambientes web, porém pode-se dizer que ainda não podem ser consideradas uma solução, visto que muitas utilizam identificadores de escopo local, ou seja, os dados não têm nenhum sentido quando tirados do contexto de tal API específica (HEATH; BIZER, 2011).

Tais fatores impulsionaram a conceitualização da Web Semântica. Esta foi idealizada também por Berners-Lee et al (2001), que ilustraram um cenário onde os dados estão disponibilizados na web utilizando padrões que aderem a uma estrutura, de modo que agentes web possam trocar informações entre si por meio de aplicações inteligentes. Neste caso, agentes consistem em sistemas de software que são programados para realizar tarefas sem o controle ou supervisão de humanos; tais programas realizam tarefas de filtro, coleta e processamento de informações (BERNERS-LEE et al, 2001).

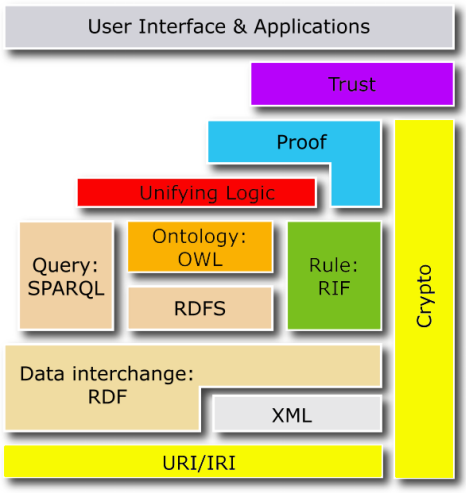
No exemplo proposto por Berners-Lee et al (2011) Pete e Lucy são irmãos que precisam realizar uma tarefa em conjunto: levar a mãe para consultar um especialista e realizar uma série de sessões de fisioterapia. Neste caso, ambos os agentes trabalhariam em conjunto para localizar informações como clínicas que realizassem o tratamento que foi prescrito à mãe deles, que fosse coberto pelo plano de saúde da mãe, em um raio de 32 km da casa dela, que fosse adequado aos horários de ambos os irmãos. Tudo aconteceria com o mínimo de esforço dos usuários, visto que os agentes percorreriam a Web Semântica de acordo com os pré-requisitos de ambos em busca da clínica, horários, melhores rotas etc., podendo se comunicar com diversos agentes, como o da clínica, o do médico e assim por diante.

Atualmente, os dados encontram-se em documentos HTML, que não possuem uma estrutura de significado para que agentes web possam realizar este tipo de tarefa. Aplicando o

exemplo de Pete e Lucy no cenário da web atual, os irmãos realizariam as seguintes tarefas: (1) ambos precisariam relacionar os dias em que estão disponíveis para levar a mãe no consultório e entrar em consenso quanto à disponibilidade de cada um, (2) acessar o *website* do plano de saúde a fim de verificar quais são os consultórios cobertos e quais são perto da casa de Lucy. (3) acessar o *website* dos consultórios visando encontrar as informações sobre o tratamento, o médico e o telefone, (4) entrar em contato com a clínica visando agendar o tratamento de acordo com a disponibilidade de ambos os irmãos. Por fim, (5) caso os horários disponíveis não coincidirem com os horários dos irmãos, realizar novamente o processo de busca por outro consultório ou adaptar-se aos horários disponíveis.

Para que a Web Semântica possa existir, é necessário que as informações estejam organizadas em conjuntos estruturados de acordo com os padrões, para que o processamento automatizado sobre os dados possa ser realizado (BERNERS-LEE et al 2011). Comparando com a web de documentos, esse processo de aderir significado e estrutura aos dados disponíveis é realizado pelo humano que consome o conteúdo. Para que isso aconteça na Web Semântica, é necessário que estejam publicados de acordo com padrões de representação de conhecimento. A Figura 1 apresenta a estrutura na qual a Web Semântica está fundamentada e seus respectivos componentes, sendo: URI, XML (*eXtensible Markup Language*), RDF, SPARQL, ontologias e agentes computacionais que realizam o processamento dos dados nos formatos definidos (BERNERS-LEE et al, 2011).

Figura 1 – Estrutura da Web Semântica



Fonte: Layer Cake (2007)

Os conceitos e tecnologias da Web Semântica, conforme apresentado na Figura 1, são

descritos a seguir:

Uma URI consiste em uma cadeia de caracteres que identifica recursos de interesse, que podem ser classificados como recursos físicos ou abstratos. De acordo com Masinter et al (2005) as URIs são:

- Uniformes: por permitirem que sejam utilizadas no mesmo contexto, independentemente do mecanismo de acesso; permite uma interpretação semântica em diferentes tipos de identificadores de recursos, permite que sejam utilizados em diferentes contextos, etc.;
- Recursos: visto que tudo pode ser identificado por uma URI, quer sejam documentos eletrônicos, imagens, fontes de informações, serviços, seres humanos, empresas, conceitos abstratos, tipos de relacionamentos.
- Identificadores: por distinguirem a informação dentro de um escopo de identificação e de outros recursos.

Tal identificação descreve e aborda categorias do recurso em questão. Podem ser divididas em URN (*Uniform Resource Name*) e URL (*Uniform Resource Locator*), sendo que URN é uma URI utilizada para nomear um recurso, estando ele na Web ou não, e URL é uma URI de especificação para a localização de tal recurso. As URLs são utilizadas para localização dos documentos ou recursos em um servidor; assim, por exemplo, a URL http://sitederoupas.com/Main_Page consiste em um recurso no qual sitederoupas.com consiste no domínio e dentro desse domínio o documento HTML Main_Page é o que se pretende localizar.

URN é um tipo de URI que estabelece um identificador único, ou seja, um nome em um domínio específico sem explicitar a sua localização ou como acessá-lo. Para exemplificar, no sistema do ISBN (*International Standard Book Number*), ISBN 978-85-8057-226-1 identifica uma versão do livro *A culpa é das estrelas;* a URN para este livro seria urn:isbn:978-85-8057-226-1. Então, para acessar esse conteúdo sua localização seria necessária; consequentemente, uma URL poderia ser especificada.

Heat e Biezer (2011) ressaltam a importância de não confundir URIs que identificam os objetos em si com os documentos web que os descrevem, objetos do mundo real não são transmissíveis utilizando o protocolo HTTP; desse modo é importante poder distinguir entre a data de criação de um documento que descreve um recurso do mundo real e a data de nascimento do recurso em si. O mesmo conceito se aplica para diferenciar URIs de URLs, uma URI descreve um recurso e não necessariamente possui uma localização (URL) para tal recurso em sua descrição.

O XML é uma linguagem de marcação de texto para representar informações

estruturadas. É aplicado no contexto da Web Semântica em razão de possibilitar a utilização de *tags* para aderir estrutura aos dados. Para realizar uma leitura do conteúdo é necessário saber para que cada *tag* é utilizada. O XML simplesmente atribui uma estrutura ao documento, não adere significado. A utilização do XML possibilita a serialização, ou seja, o processamento da estrutura a fim de armazenar os dados, de diferentes linguagens de marcação; pode ser utilizado para marcação de páginas web para processar diferentes elementos; e por fim para transferir objetos de dados entre duas aplicações (DECKER et al., 2000). O RDF consiste em um modelo padrão para a descrição dos dados que é realizada por meio de uma estrutura de triplas contendo recurso, propriedade e valor. Por meio do RDF é possível realçar as relações entre os recursos e objetos; será analisado detalhadamente nas seções a seguir (WORLD WIDE WEB CONSORTIUM, 2014).

O RDF consiste em um modelo para representação de dados na web, sendo que tal representação acontece por meio de grafos dirigidos. Por meio do RDF informações provenientes de diferentes fontes de dados são integradas e relacionadas. De acordo com W3C (*World Wide Web Consortium*) o RDF facilita a fusão de dados independentemente do esquema no qual eles se encontram (WORLD WIDE WEB CONSORTIUM, 2014).

Uma descrição em RDF é feita por meio de triplas, que são compostas por sujeito, predicado e objeto. Nela o sujeito consiste no recurso descrito, correspondendo então a uma URI identificando o recurso descrito. O objeto pode corresponder a qualquer tipo de valor, podendo ser uma cadeia de caracteres, uma data, números ou até mesmo uma URI correspondente a outro recurso relacionado. O predicado representa qual é o tipo de relação existente entre o sujeito e o objeto, e também é descrito por meio de uma URI proveniente de vocabulários, para descrever informações de domínios específicos.

A Figura 2 apresenta uma representação conceitual do RDF que aplica URIs para representação do conteúdo; observa-se uma representação teórica de um grafo RDF, na qual apresenta-se uma pessoa, chamada Jéssica, que tem interesse artístico por uma obra chamada A Noite Estrelada, do pintor Vincent Van Gogh. Pode-se notar como Jessica e Vincent Van Gogh estão relacionados; isso ocorre por meio da utilização de vocabulários que atribuem significado nas relações. Para inferir que Jéssica se interessa por uma obra chamada A Noite Estrelada utilizou-se um elemento do vocabulário FOAF (*Friend of a friend*) chamado *topic_interest,* que possibilita declarar elementos que são do interesse de uma pessoa. Para declarar quem é o autor dessa obra foi utilizado outro elemento do vocabulário *DC Terms* chamado *creator*, que é utilizado para declarar uma entidade responsável pela criação do recurso em questão. Assim, as relações, recursos e objetos são descritos por meio de URIs e vocabulários.

Figura 2 – Representação conceitual por meio de URIs

Fonte: Elaborada pela autora.

Ao modelar um conjunto de dados de acordo com os padrões da Web Semântica é vital inserir URIs relacionadas, visto que conforme apresentado na Figura 2, as URIs auxiliam a expansão do conhecimento sobre os tópicos descritos. Isto é o que acontece no *Linked Data*, onde milhares de dados são interligados por meio do RDF, possibilitando assim uma estrutura semântica dos dados armazenados nesse ambiente.

O RDF possui diferentes tipos de apresentação do conteúdo apresentado, ou seja, diferentes sintaxes para serem utilizadas na web. Existem diferentes tipos de sintaxe RDF que adequam o formato de grafos para uma linguagem passível de processamento; são elas:

- RDF/XML: sintaxe amplamente utilizada para publicar *Linked Data*, porém é considerada difícil para humanos lerem e escreverem a partir dela; sendo assim, aconselha-se a utilização de outra sintaxe, como o *Turtle*;
- RDFa: incorpora triplas RDF em documentos HTML possibilitando que os mecanismos de busca agreguem os dados descritos na busca web para aprimorar os resultados (WORLD WIDE WEB CONSORTIUM, 2014).
- *Turtle*: considerado um formato de texto simples devido ao fato de suportar formas abreviadas de *namespaces* e outros componentes de uma descrição RDF; é indicado para leitura e escrita à mão das triplas RDF (HEAT; BIEZER, 2011).

As diferentes sintaxes utilizadas para descrição de recursos são apresentadas nas Figuras 3, 4 e 5, as quais descrevem as formas como as triplas apresentadas na Figura 2 são descritas para processamento computacional.

Figura 3 – Sintaxe do modelo de representação RDF/XML

```
<rdf:RDF

  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/">

  <rdf:Description rdf:about="http://dbpedia.org/page/Jessica_Alba">
  <foaf:topic_interest rdf:resource="http://dbpedia.org/page/The_Starry_Night"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/page/The_Starry_Night">
  <dc:creator rdf:resource="http://dbpedia.org/page/Vincent_van_Gogh"/>
  <dc:title> A Noite Estrelada </dc:title>
  </rdf:Description>

</rdf:RDF>
```

Fonte: Elaborada pela autora.

Figura 4 – Sintaxe do modelo de representação RDFa

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc11: <http://purl.org/dc/elements/1.1/> .

<http://dbpedia.org/page/Jessica_Alba> foaf:topic_interest
<http://dbpedia.org/page/The_Starry_Night> .
<http://dbpedia.org/page/The_Starry_Night>
 dc11:creator <http://dbpedia.org/page/Vincent_van_Gogh> ;
 dc11:title " A Noite Estrelada " .
```

Fonte: Elaborada pela autora.

Figura 5 – Sintaxe do modelo de representação *Turtle*

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" "http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<head>
  <meta http-equiv="Content-Type" content="application/xhtml+xml; charset=UTF-8"/>
  <title>Título da Página</title>
</head>
<div xmlns="http://www.w3.org/1999/xhtml"
 prefix="
  foaf: http://xmlns.com/foaf/0.1/
  rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
  dc: http://purl.org/dc/elements/1.1/
  rdfs: http://www.w3.org/2000/01/rdf-schema#">
 <div typeof="rdfs:Resource" about="http://dbpedia.org/page/Jessica_Alba">
  <div rel="foaf:topic_interest">
   <div typeof="rdfs:Resource" about="http://dbpedia.org/page/The_Starry_Night">
    <div property="dc:title" content=" A Noite Estrelada "></div>
    <div rel="dc:creator" resource="http://dbpedia.org/page/Vincent_van_Gogh"></div>
   </div>
  </div>
 </div>
</div>
```

Fonte: Elaborada pela autora.

Em síntese, o RDF consiste em um modelo para a descrição de recursos, ou seja, permite

modelar a descrição dos recursos no formato de triplas: recurso ou objeto – propriedade – valor, em que um objeto consiste em um recurso e o valor de tal objeto pode também ser um recurso, conforme apresentado na Figura 2, onde a atriz Jessica Alba (recurso) se interessa (propriedade) pela obra de arte chamada "*The Starry Night*" (valor que é também um recurso). Assim, o modelo RDF proporciona um ambiente interativo no qual novas informações podem ser descobertas. Ou seja, todo recurso possui um valor, e o relacionamento entre valor e o recurso é expresso por meio das propriedades; na prática os vocabulários são utilizados para definir o tipo do relacionamento/propriedade. O RDF permite também que recurso e valor troquem de papel, ou seja, o valor de um recurso pode ser também um recurso. Como exemplo, para realizar a descrição dos recursos em RDF, na linha 7 da Figura 2 o elemento chamado '*about',* em rdf:about, faz uma explicitação sobre de que trata o recurso, indicando uma URI que fornece informações sobre a atriz Jessica Alba. Neste caso, o recurso é a atriz, representado pela URI que a representa, o valor consiste em outro recurso, uma obra de arte chamada *The Starry Night*, e a relação entre os dois recursos é realizada pelo termo disponível no vocabulário FOAF para declarar o tópico de interesse do recurso 1 (Jessica Alba) no recurso 2 (obra de arte). Assim, o RDF, além de permitir tal modelo de descrição e relacionamento de recursos, descreve-os com auxílio dos vocabulários.

A fim de prover o suporte para vocabulários no próprio RDF, há o RDFS (RDF *Schema*), que permite definir características semânticas aos dados. Apesar de ser comumente utilizado, o RDFS possui apenas 15 classes e 16 propriedades, o que pode ser considerado uma limitação para descrever recursos de diferentes áreas. Sendo assim, faz-se uso de diferentes vocabulários existentes.

Os vocabulários são utilizados para definir conceitos e relacionamentos por meio de termos para representar diferentes áreas. Sua utilização auxilia no tratamento de ambiguidades, bem como na organização do conhecimento. Dentre os vocabulários conhecidos estão: *DC Terms*, FOAF (*Friend of a Friend*), SKOS (*Simple Knowledge Organization System*), no qual cada vocabulário é utilizado para descrever diferentes tipos de informações, conforme descrito a seguir:

- *DC Terms*: vocabulário desenvolvido sob o princípio de ser amplo e genérico a fim de ser utilizado para descrever uma variedade ampla de recursos. Possui 15 elementos básicos, um exemplo de sua utilização é apresentado nas linhas 11 e 12 da Figura 5, onde foram utilizados dois elementos: '*creator'* e '*title',* para descrição do criador/pintor e do nome da obra de arte descritos na tripla RDF;
- FOAF: utilizável para descrever pessoas de acordo com três tipos de redes: (1) redes de informação, que utilizam ligações baseadas na web com objetivo de

compartilhar descrições publicadas de forma independente de seu mundo interconectado; (2) redes sociais, de colaboração humana, amizades e associações; e (3) redes representacionais, que descrevem uma visão simplificada do universo de desenhos animados em termos reais. Os termos são agrupados de acordo com as seguintes categorias: **Core,** que descreve pessoas e grupos sociais, bem como suas informações históricas, herança histórica; **Web Social** descreve informações relacionadas a atividades realizadas na web e **Utilidades para *Linked Data,*** para documentos RDF ligados no contexto do *Linked Data* (BRICKLEY; MILLER, 2014);

- SKOS: visa disponibilizar uma forma de descrever sistemas de organização de conhecimento como tesauros, esquemas de classificação, taxonomias, etc. Possui 32 elementos descritivos (MILES; BECHHOFER, 2009).

Quanto ao SPARQL, inicialmente foi denominado linguagem e posteriormente, em 2008, tornou-se um padrão para realizar consultas em dados RDF.

Para realizar a recuperação das informações em conjuntos de dados RDF, utiliza-se o SPARQL, um protocolo capaz de manipular os dados armazenados no formato RDF. Uma *query* SPARQL consiste em três partes: (1) o 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 os quais, uma vez que o resultado foi computado, permitem 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 (PÉREZ et al., 2006).

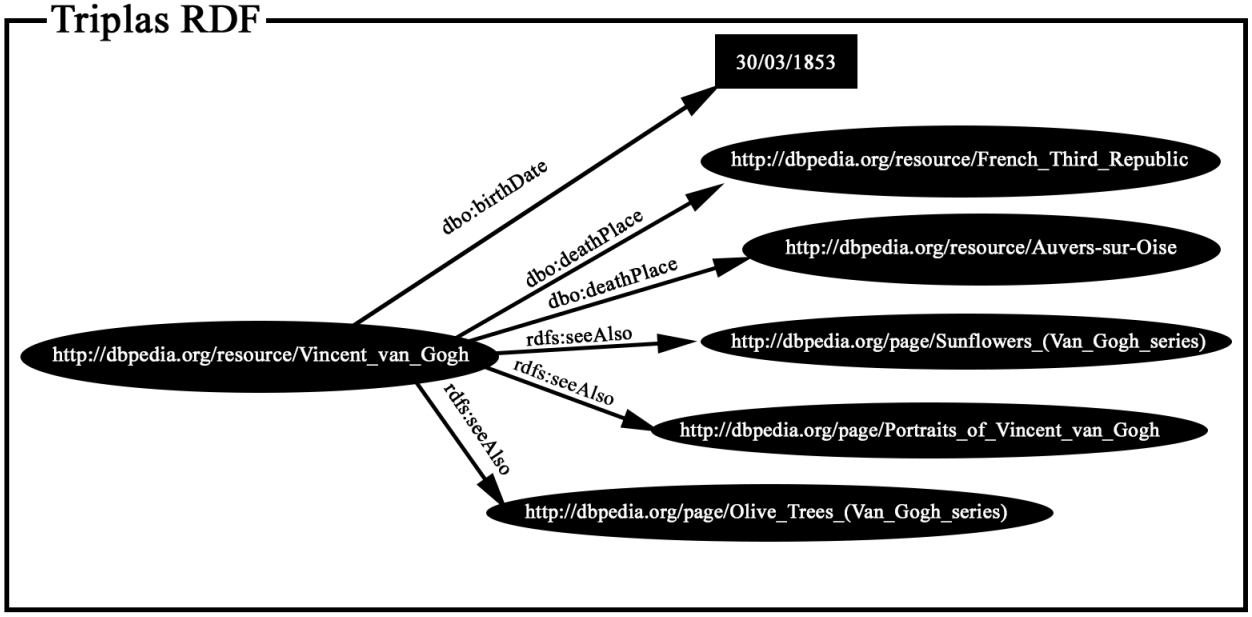
As buscas utilizando SPARQL diferem de buscas realizadas em buscadores web. Para recuperar dados de um *dataset* são utilizadas linhas de comando que representam: o que está sendo buscado, onde deve ser selecionado e, finalmente, como será representado. Para realizar a seleção do que se deseja buscar, utiliza-se um comando chamado SELECT e, então, atribui-se um rótulo para a informação a ser buscada; o comando WHERE é utilizado para definir onde será realizada a busca. E, por fim, é possível agrupar os resultados de acordo com diferentes operadores.

Os comandos podem ser executados em SPARQL *endpoints*, quando disponibilizados, que consistem em ambientes abertos para a realização de buscas. A sintaxe em que o resultado é apresentado pode variar de acordo com as diferentes sintaxes RDF citadas anteriormente, como RDF/XML, *Turtle*, etc. Algumas organizações disponibilizam conjuntos de dados em arquivos chamados *dump,* onde é possível realizar a consulta por meio de programas e recursos computacionais, como o Apache Jena.

Assim, o primeiro passo para realizar uma *query* SPARQL é (1) declarar um prefixo, que possibilita a utilização da URI por meio de uma chave identificadora, excluindo a necessidade de inserir a URI toda vez que for referenciá-la. O segundo passo (2) é definir em qual *dataset* a busca será realizada e, para isso, utiliza-se o comando *FROM*; quando a busca está sendo realizada no *endpoint* do próprio *dataset* não se faz necessária a declaração do *WHERE*. Em seguida, (3) deve-se declarar que tipo de informação se espera que a busca retorne, utilizando o comando *SELECT*, (4) o que pretende ser buscado no *dataset* é definido por meio do comando *WHERE* e assim (5) os resultados podem ser ordenados, divididos, agrupados por meio de diferentes operadores, como o *ORDER BY* que ordena os resultados de acordo com determinado valor dos resultados. Conforme ressaltado por Santarém Segundo (2014), é necessário conhecer a estrutura semântica, os vocabulários associados, e o contexto da construção dos dados relacionados em um *dataset* para realizar a recuperação por meio do SPARQL.

Para exemplificar, a Figura 6 apresenta um conjunto de triplas sobre o pintor Vincent Van Gogh, que é um recurso, neste caso contendo 6 informações relacionadas, das quais 5 são também recursos, que estão relacionados a ele por meio das seguintes propriedades: data de nascimento, lugar de nascimento, composto por dois recursos (cidade e país), e veja também que relaciona as obras realizadas por tal pintor.

Figura 6 – Conjunto de triplas descrevendo o recurso Vincent Van Gogh



Fonte: Elaborada pela autora.

A Figura 7 apresenta uma *query* executada no *endpoint* do DBpedia, que tem como objetivo apresentar as obras criadas pelo autor, onde PREFIX consiste em meio de tornar o vocabulário conhecido pelo mecanismo de busca, ?Pinturas_Van_Gogh será o rótulo utilizado para nomear os resultados, a cláusula WHERE é composta pelo recurso, a propriedade e o valor, em que o recurso é representado pela URI de identificação do autor, a propriedade é indicada pelo indicador definido no PREFIX, rdfs, seguido da propriedade, sendo rdfs:seeAlso, e por fim o valor será indicado de acordo com o rótulo definido.

Figura 7 – Exemplo de uma *query* SPARQL para buscar obras relacionadas do pintor Vincent Van Gogh

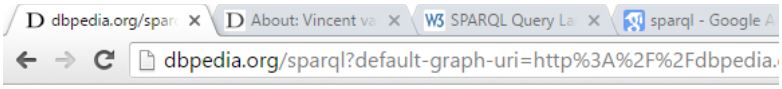
SPARQL

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?Pinturas_Van_Gogh
 WHERE {
    <http://dbpedia.org/resource/Vincent_van_Gogh>  rdfs:seeAlso ?Pinturas_Van_Gogh

}
```

Fonte: Elaborada pela autora.

O resultado da *query* da Figura 7 é apresentado na Figura 8 no formato de tabela com as obras criadas pelo autor.

Figura 8 – Resultado da busca realizada no exemplo da Figura 7



Pinturas_Van_Googh
http://dbpedia.org/resource/Sunflowers_(Van_Gogh_series)
http://dbpedia.org/resource/Portraits_of_Vincent_van_Gogh
http://dbpedia.org/resource/Flowering_Orchards
http://dbpedia.org/resource/Still_life_paintings_by_Vincent_van_Gogh_(Netherlands)
http://dbpedia.org/resource/Portraits_by_Vincent_van_Gogh
http://dbpedia.org/resource/Double-square_painting
http://dbpedia.org/resource/Hospital
http://dbpedia.org/resource/Japonaiserie_(Van_Gogh)
http://dbpedia.org/resource/Square
http://dbpedia.org/resource/Langlois_Bridge_at_Arles
http://dbpedia.org/resource/Wheat_Fields_(Van_Gogh_series)
http://dbpedia.org/resource/The_Letters_of_Vincent_van_Gogh
http://dbpedia.org/resource/Early_works_of_Vincent_van_Gogh
http://dbpedia.org/resource/Olive_Trees_(Van_Gogh_series)
http://dbpedia.org/resource/Arles_(Van_Gogh_series)
http://dbpedia.org/resource/His_art
http://dbpedia.org/resource/Van_Gogh's_family

Fonte: Elaborada pela autora.

Uma forma de realizar uma descrição completa dos recursos e suas relações, utilizando vocabulários, é por meio da utilização de ontologias.

A palavra vem do grego "*ontos*" (ente) e *logos* (saber) e relaciona-se com a ciência do saber ou estudo do ser. Tal estudo foi realizado por filósofos com o objetivo de tratar a natureza e a estrutura da realidade dos seres. No contexto da Web Semântica, as ontologias correspondem a estruturas que possibilitam realizar uma descrição completa de determinada área ou conceito, atribuindo então semântica aos dados, de forma que agentes computacionais possam realizar um processamento inteligente de tais dados.

De acordo com Gruber (1993), uma ontologia consiste em uma especificação explícita de uma conceitualização, realizada por meio de um vocabulário do domínio, possibilitando que consultas e afirmações possam ser realizadas entre agentes computacionais.

Para Santarém Segundo e Coneglian (2015, p 227), "[...] do ponto de vista de aplicação, ontologias são: artefatos computacionais que descrevem um domínio do conhecimento de forma estruturada, através de: classes, propriedades, relações, restrições, axiomas e instâncias".

De acordo com Roussey et al (2011), as ontologias podem ser classificadas em diferentes categorias: ontologia de **domínio**, aplicáveis apenas em domínios específicos, de acordo com uma visão específica de como determinado grupo conceitua e visualiza fenômenos específicos. Ontologias **locais** ou de aplicações, consideradas uma especialização da ontologia de domínio, com objetivo de realizar propósitos específicos em aplicações específicas. Ontologias de **referência central** consistem em ontologias padrão utilizadas por diferentes grupos de usuário, integrando diferentes panoramas relacionados a grupos específicos de usuários. Ontologias **gerais** são as designadas para conter conhecimentos gerais de uma grande área e por fim ontologias de **alto nível**, que são genéricas e aplicáveis a diversos domínios.

De um ponto de vista técnico, pode-se dizer que a unificação de um modelo de descrição de recursos (RDF), juntamente com vocabulários descritivos e a atribuição de uma estrutura de organização de conhecimento por meio de classes, propriedades, relacionamentos e restrições, consiste em uma ontologia.

Assim, nota-se que o estado da web de documentos não favorece a Web Semântica, tornando necessário um ambiente que suporte a web de dados, que é considerada uma evolução da web de documentos. Para que isso aconteça, as informações devem possuir uma ligação entre elas, de modo que uma informação esteja interligada com informações relacionadas. Uma das aplicações mais conhecida da Web Semântica é o *Linked Data*, que consiste em um conjunto de práticas para publicar e relacionar dados estruturados na web, sendo elas:

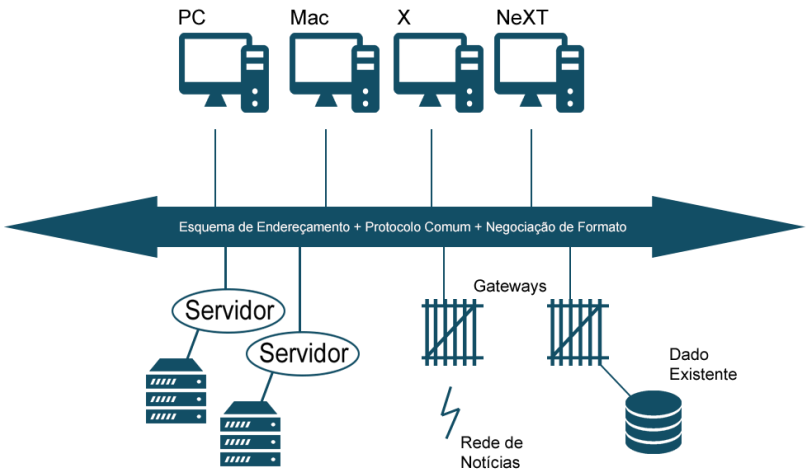
- Utilização de URI para nomear recursos

- Utilizar HTTP como URI de modo que tais dados possam ser encontrados
- Prover informações úteis utilizando os padrões RDF e SPARQL
- Incluir *links* que guiem a outras URIs de modo que o usuário possa encontrar mais informações relacionadas (BERNERS-LEE, 2006; HEATH; BIZER, 2011).

Na Web Semântica, ainda se espera que haja uma comunicação independentemente da localização física, objetiva-se não somente uma apresentação, mas uma descrição dos recursos na web, conceitos abstratos, animais, objetos físicos, pessoas, empresas, etc.

A Figura 9 apresenta o esboço da arquitetura da web de documentos, que necessitava de um canal aberto de comunicação, para possibilitar o intercâmbio de documentos. Visando atender à necessidade de troca de documentos foi, inicialmente, proposta a utilização do FTP (*File Transfer Protocol*). Posteriormente, foi proposta a utilização do protocolo HTTP, visto que se mostrou mais rápido que o FTP na recuperação de documentos e possibilitando busca por índices (BERNERS-LLE et al., 1992). Então, para que os documentos fossem identificados seria necessário a existência de um esquema de nomeação. Quando primeiramente idealizada por Berners-Lee et al (1992) os autores mencionaram a importância de um esquema de nomeação, visto que um nome provê uma forma para que o servidor encontre os documentos na web. Eventualmente, as URIs foram definidas como um esquema de identificação dos recursos, não somente de documentos, visto que por meio delas é possível identificar recursos tanto físicos como abstratos. E, por fim, para que os documentos pudessem ser visualizados, o HTML foi definido como um meio de apresentação de documentos de texto.

Figura 9 – Arquitetura geral da web



Fonte: Berners-Lee et al. (1992)

Visto que a Web Semântica é considerada uma extensão da web atual, ela faz uso dos

protocolos de comunicação e identificação dos recursos. Porém, diferentemente da web de documentos, utiliza novas tecnologias com o objetivo de estabelecer uma estrutura semântica para a descrição de recursos, de forma que agentes computacionais possam realizar a identificação dos dados. Espera-se que, ao aplicar tal estrutura semântica, as aplicações da web de dados possam compreender e manipular os dados de forma significativa a favor dos usuários.

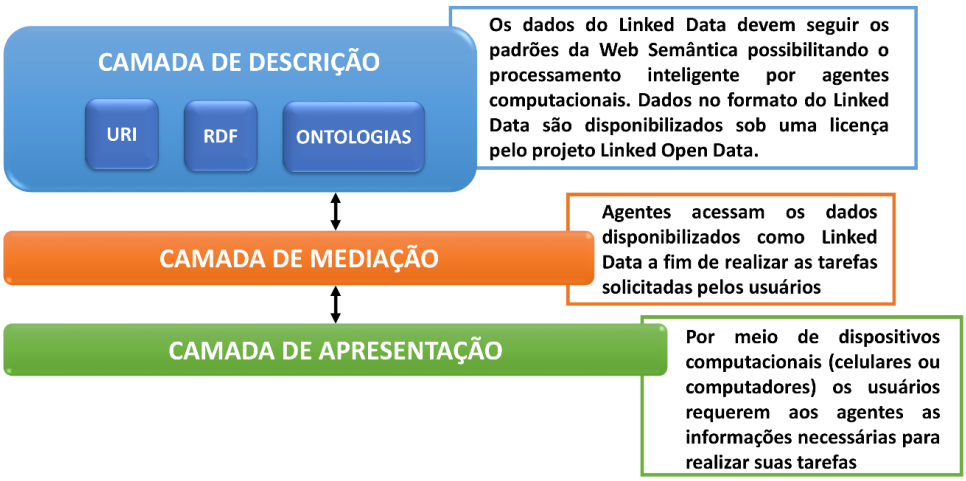
O Quadro 3 relaciona as tecnologias e protocolos que compõem a web semântica e de documentos, e o objetivo da utilização de cada um deles. Ainda se faz uso das URIs, que são utilizadas como um meio de identificar e nomear os recursos; do XML, como uma forma de integrar as definições da Web Semântica; porém é utilizado o RDF para realizar declarações sobre os recursos por meio de URIs e definir vocabulários que podem ser referenciados por elas; por meio das ontologias, relações entre diferentes conceitos são definidas (KOIVUNEN; MILLER, 2001). A forma de interação entre agentes computacionais e os dados da Web Semântica é ilustrada na Figura 10.

Quadro 3 – Tecnologias e protocolos da web de documentos e semântica

Padrões	Objetivo	Compõe
HTTP	Protocolo utilizado para realizar a comunicação/conexão de itens em diferentes servidores	Web de documentos e de dados
URI	Identificador de recursos na web	Web de documentos e de dados
HTML	Representar conteúdo textual na web	Web de documentos
RDF	Descrever recursos web e estabelecer relações entre eles	Web de dados
SPARQL	Linguagem de busca e manipulação de conteúdo no formato RDF	Web de dados

Fonte: Elaborado pela autora.

Figura 10 – Modelo conceitual do funcionamento da web semântica



Fonte: Elaborada pela autora.

Conforme abordado, para que ferramentas da Web Semântica existam, como citado no

exemplo de Pete e Lucy (BERNERS-LEE et al., 2001), seriam necessários dados publicados de acordo com os padrões da Web Semântica; desse modo agentes computacionais poderiam realizar o processamento inteligente por meio da estrutura semântica dos dados. Tendo em vista a maneira como os dados são estruturados na web de documentos, pode-se afirmar que não é possível a existência desse tipo de aplicação.

No entanto, o conjunto de melhores práticas para *Linked Data* possibilita a utilização dos dados publicados para construção de aplicações inteligentes, nas quais os agentes possam atuar e realizar tarefas utilizando os dados linkados disponibilizados no formato necessário, como no exemplo de Pete e Lucy (Berners-Lee et al., 2001). Para esse fim, diferentemente da web de documentos, os dados publicados como *Linked Data* descrevem todo tipo de conteúdo como artigos científicos, que por sua vez podem se relacionar com diversos outros recursos. Todos os conteúdos publicados de acordo com os padrões da Web Semântica são agrupados em *datasets* (conjuntos de dados), classificados em diferentes categorias. Para realizar o acesso aos conjuntos armazenados em tais *datasets* utiliza-se um identificador de espaço controlado pelo provedor dos dados chamado *namespace*. Por meio do *namespace* é possível atribuir uma chave às URIs, e por sua vez utilizá-las para referenciá-las, sendo assim um meio de tornar as URIs menos extensas.

2.1 Linked Data

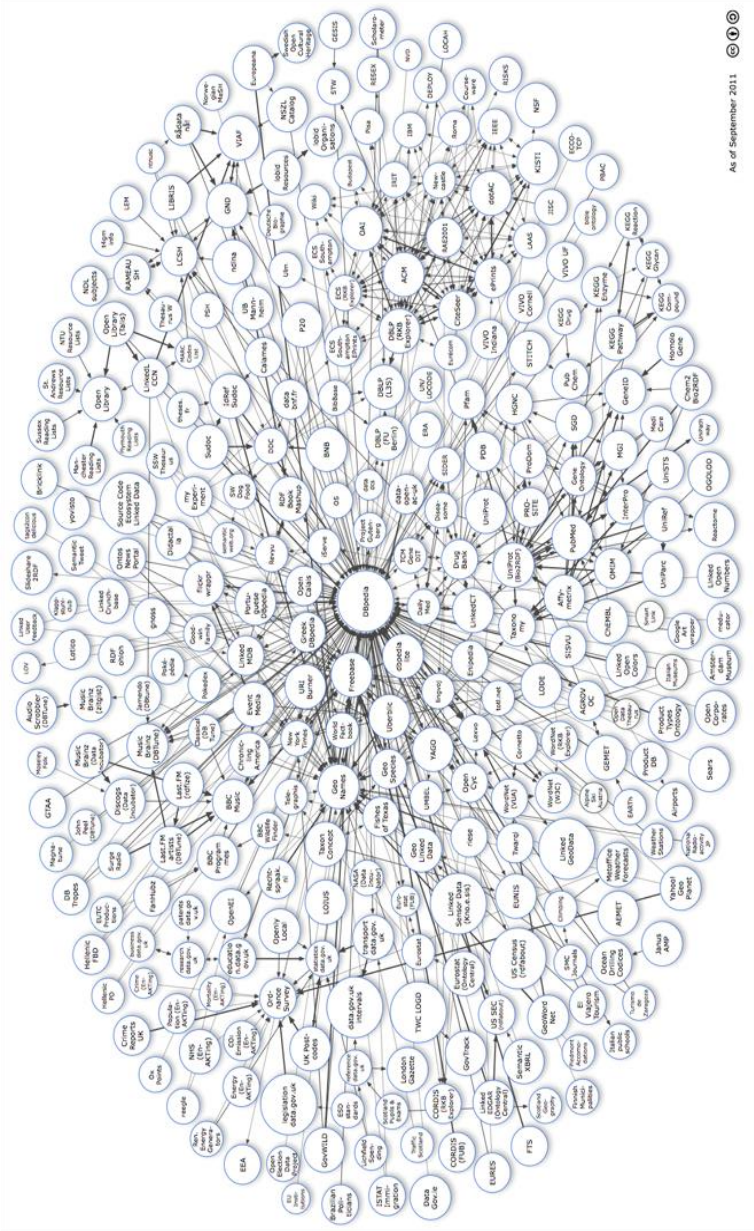
O maior exemplo de adoção e aplicação do *Linked Data* é o *Linked Open Data*, um projeto fundado em 2007 que teve como objetivo iniciar a web de dados identificando *datasets* disponíveis sob licenças abertas, convertê-los de acordo com os princípios do *Linked Data* e publicá-los na Web (BIZER et al, 2009).

Os dados são publicados no formato de *datasets*, e podem ser classificados de acordo com diferentes categorias, sendo elas: publicações, ciências da vida, domínios cruzados, redes sociais, geográficas, governamentais, mídia, conteúdos gerados por usuários e logística.

O primeiro relatório das estatísticas de publicação de dados no formato de *Linked Data* no LOD foi publicado em setembro 2011 (Figura 11), no qual se observou que havia 295 *datasets* publicados, sendo a maioria dos *datasets* na categoria de publicações, 87; em seguida dados governamentais, com 49 *datasets* publicados; dados de domínios cruzados e ciências da vida tiveram a mesma quantidade de *datasets*, 41, seguidos de dados geográficos, que somaram 31 *datasets*, mídia com 25 e, por fim, conteúdos gerados por usuários com 20 *datasets*. Nesse ano, a quantidade de triplas RDF somaram mais de 31bilhões (BIZER et al, 2011).

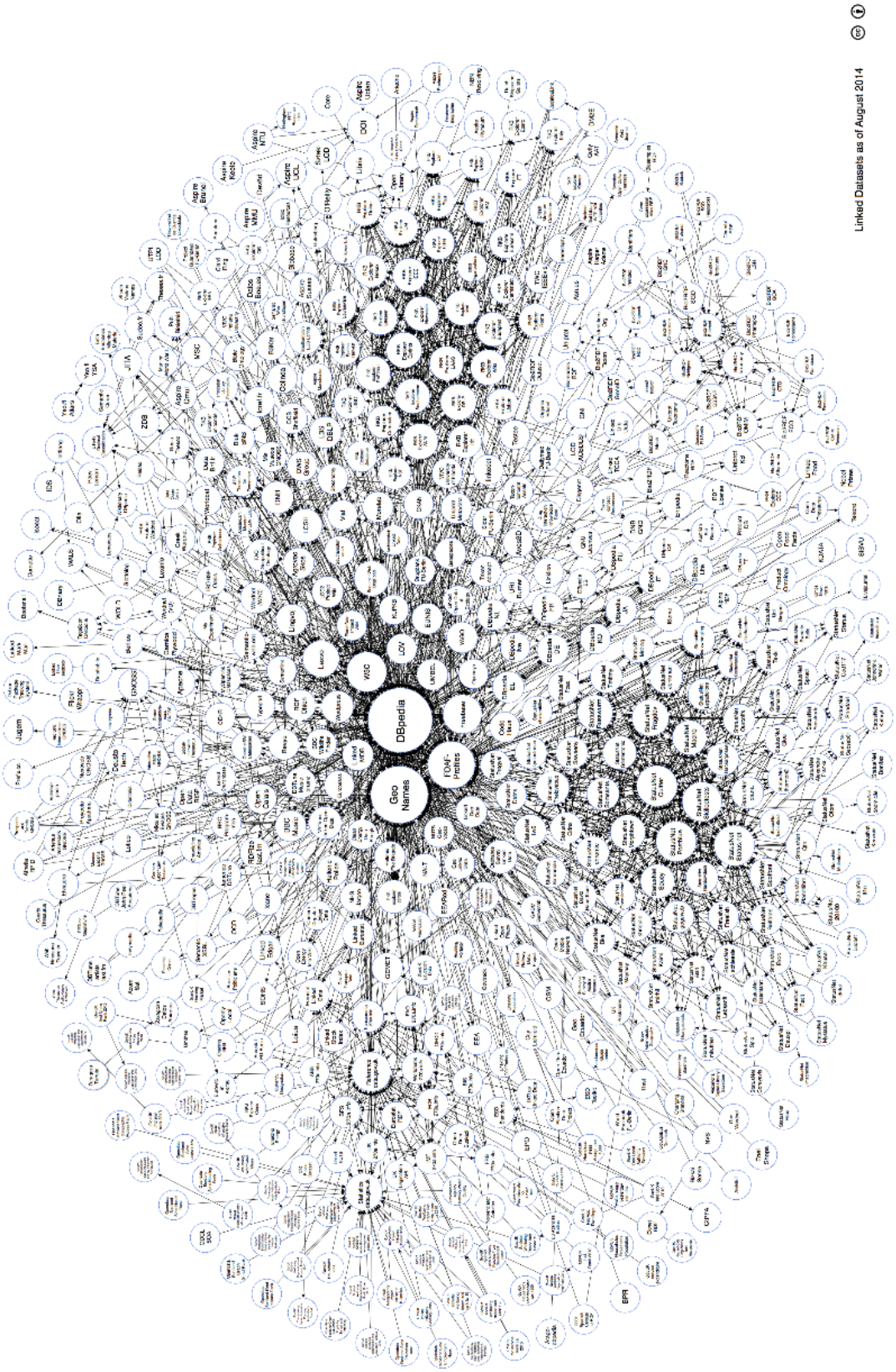
Apesar do diagrama ter sido atualizado em fevereiro de 2017 (Figura 13), o último relatório sobre as estatísticas do LOD foi publicado em agosto de 2014 (Figura 12), e permite notar um crescimento considerável em relação ao ano anterior. Foram publicados 1014 *datasets*, sendo 520 *datasets* de dados de redes sociais, 183 no domínio governamental, 96 de publicações, 83 no domínio de ciências da vida, 48 de conteúdos gerados pelos usuários, 41 de domínios cruzados, 22 de mídia e por fim 21 de dados geográficos (SCHMACHTENBERG et al, 2014).

Figura 11 – Nuvem de *Linked Data* publicados como LOD em 2011



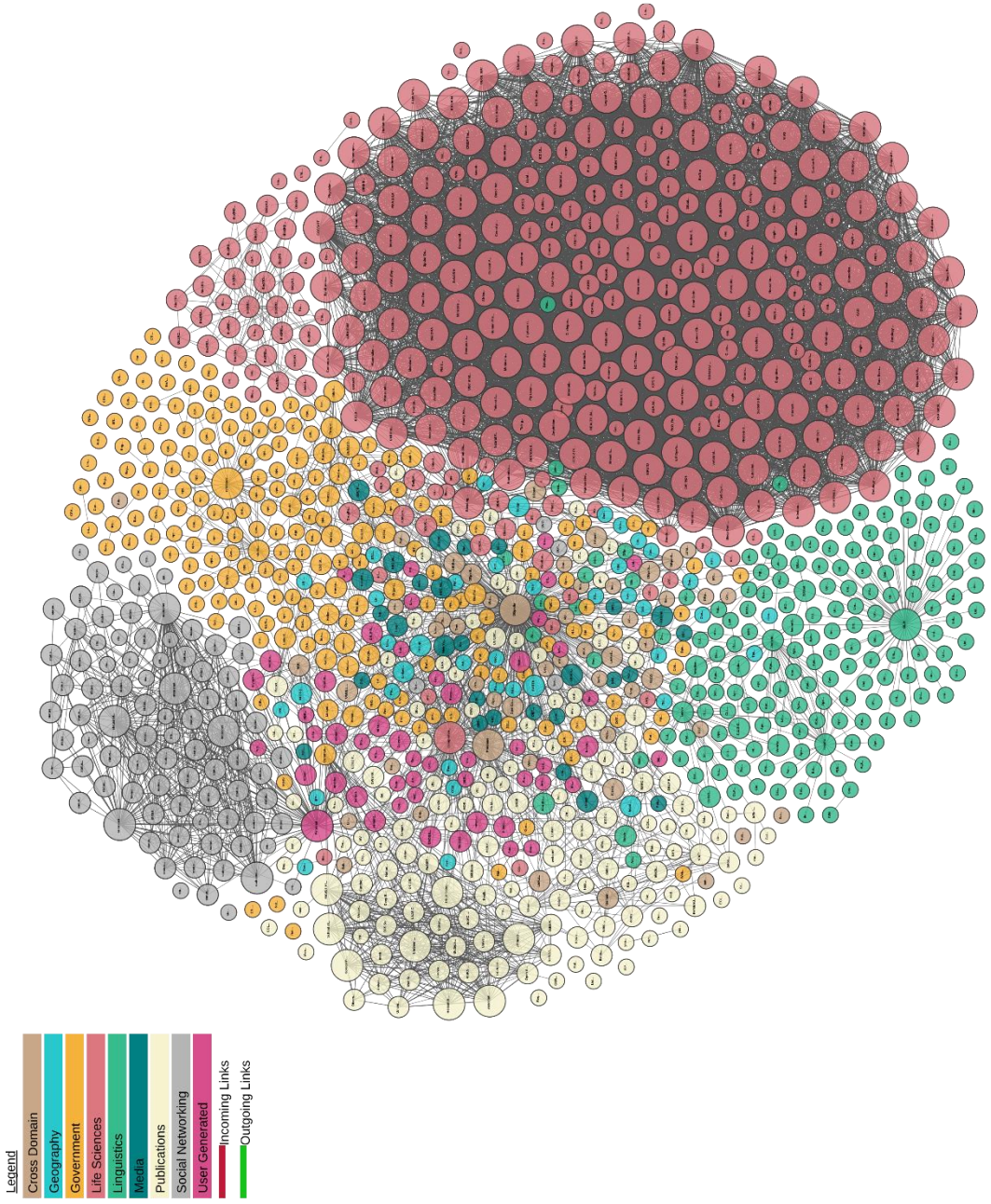
Fonte: Bizer et al (2011)

Figura 12 – Nuvem de *Linked Data* publicados como LOD em 2014



Fonte: Bizer et al (2014)

Figura 13 – Diagrama LOD atualizado em 2017

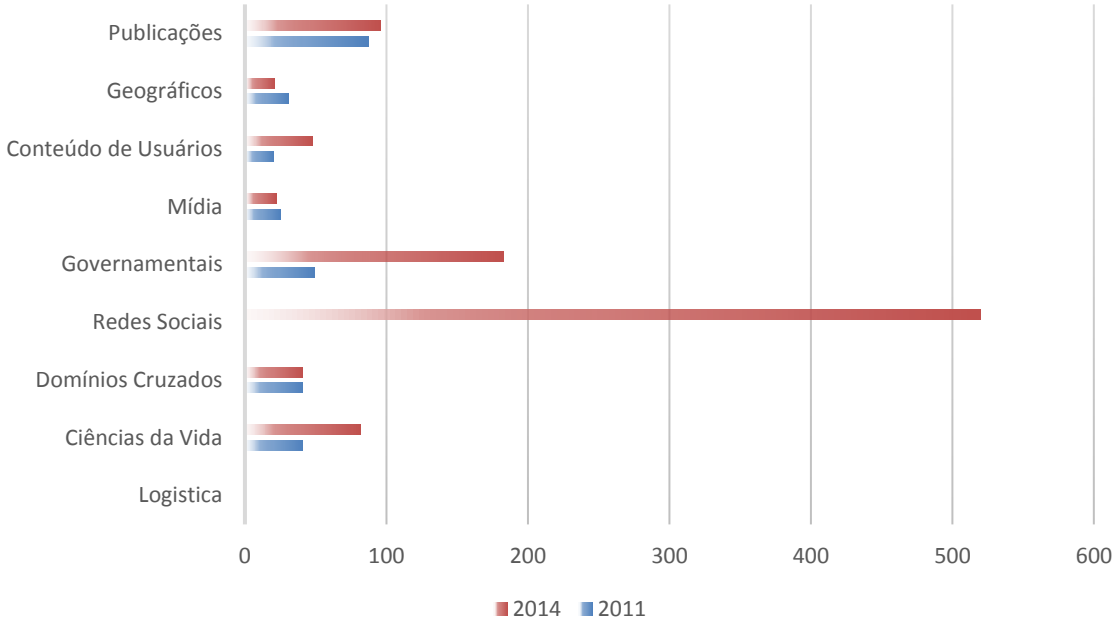


Fonte: Andrejs Abele et al (2017)

O gráfico apresentado na Figura 14 foi gerado a partir dos relatórios disponibilizados pelo projeto LOD em relação aos anos de 2011 e 2014; ressalta-se que, apesar do diagrama ter sido atualizado no ano de 2017, nenhum relatório sobre o diagrama atualizado foi publicado durante o desenvolvimento desta pesquisa. O gráfico apresenta uma comparação da publicação de *datasets;* ao analisarmos o gráfico notamos que a categoria com mais dados publicados é a de Redes Sociais com mais de 500 *datasets*, um grande aumento comparado com o primeiro levantamento, visto que no ano de 2011 não foi publicado nenhum *dataset* nessa categoria. As

categorias com a maior quantidade de *datasets* são de dados governamentais, com quase 200, e de rede sociais, com um pouco mais de 500. Porém o domínio de logística não registrou a publicação de nenhum *dataset* desde que a primeira análise foi realizada. Em algumas categorias houve pouco ou nenhum crescimento, como a categoria de domínios cruzados, que permaneceu sem mudanças nos dois relatórios e a categoria de mídia, na qual foram publicados apenas 3 *datasets* no intervalo entre 2011 e 2014. A categoria de dados geográficos apresentou uma mudança surpreendente, uma vez que houve uma diminuição de *datasets* publicados, sendo 31 em 2011 e apenas 21 no ano de 2014. Por fim, a categoria de publicações não apresentou um crescimento notável entre 2011 e 2014, porém, embora ainda não houvesse um relatório referente ao ano de 2017, ao observar o diagrama atualizado (Figura 13) e o gráfico, pode-se concluir que houve um aumento, visto que há mais de 100 *datasets* nessa categoria, que representa o objeto de estudo desta pesquisa.

Figura 14 – Crescimento de *datasets* no LOD



Fonte: Elaborada pela autora.

Os dados publicados de forma aberta no LOD têm um papel importante para a web de dados. Isso porque, na web de documentos, os conteúdos publicados podem ser buscados por meio de mecanismos de busca, que por sua vez raramente disponibilizam *links* para redirecionar a mais informações do que foi mencionado no documento. Esse seria um processo que seria realizado individualmente, em que o usuário realizaria novamente o processo de busca, desta vez sobre um conteúdo mencionado no documento visualizado previamente. Na web de dados,

são utilizados buscadores de *Linked Data* que buscam os dados em *datasets* publicados, e os resultados possuem *links* RDF que possibilitam navegar em diversas outras fontes, viabilizando uma navegação interminável em fontes de dados web conectados por *links* RDF (BIZER et al., 2008). Assim, o LOD pode ser considerado um servidor de informações, visto que os dados publicados podem ser acessados por diferentes aplicações da Web Semântica, como os navegadores de *Linked Data*, neste caso.

Os dados no LOD podem ser publicados por organizações ou usuários individuais. Para que isto aconteça é necessário que os *datasets* atendam a alguns requisitos específicos. Ressalta-se que, quando os conjuntos de dados são publicados, não são incluídos no diagrama LOD imediatamente. Para que os *datasets* sejam incluídos no diagrama, eles devem atender a requisitos de qualidade, avaliados, e somente após considerados aptos são incluídos no diagrama. Requisitos e a descrição do processo de publicação de dados no LOD são abordados a seguir.

2.2 Processo de publicação de datasets no LOD

O crescimento dos dados publicados no LOD torna-se evidente, quando se comparam as Figuras 11 e 12 (p. 39 e p.40), dos anos 2011 e 2014, respectivamente. Na representação utilizada pelos autores, cada circunferência representa um *dataset* e quanto maior a circunferência, maior o volume de dados publicados no conjunto; as flechas dirigidas representam as relações entre os *datasets*. Cada *dataset* no diagrama possui uma página de referência no Datahub, o qual fornece dados descritivos sobre o conjunto de dados publicados como *Linked Data*.

O Datahub consiste em uma plataforma de gerenciamento de dados desenvolvida com base no sistema de gerenciamento e catalogação de dados chamado CKAN (*Comprehensive Knowledge Archive Network*) desenvolvido pela *Open Knowledge Foundation*. O CKAN tem como objetivo tornar dados acessíveis e, para isso, provê ferramentas para aprimorar a publicação, o compartilhamento e a descoberta de dados. Assim, o Datahub utiliza o sistema do CKAN para gerenciamento de *datasets* e permite buscas nos dados, registrar *datasets* já publicados, criar e gerenciar grupos de *datasets* e disponibilizar alertas quanto a atualizações de *datasets* de interesse.

Deste modo, o Datahub armazena um catálogo dos *datasets*, os quais podem estar armazenados em diferentes domínios, como dados de universidades, governamentais, geográficos, etc. Todos os *datasets* contidos no diagrama LOD possuem um registro de

identificação no Datahub. Atualmente, existem mais de 10 mil *datasets* no Datahub e, segundo o último relatório do LOD, publicado em 2014, havia apenas 1014 *datasets*, o que evidencia que a grande maioria não atendia aos requisitos mínimos de qualidade para serem inseridos no diagrama.

Três princípios devem ser levados em consideração ao publicar *Linked Data*: (1) primeiramente compreender o conjunto de dados em questão (quais são as entidades principais, suas propriedades, como tais entidades se relacionam com outras), (2) publicar como RDF e (3) conectar com outras bases de dados.

Um fator relevante é que nem todos os conjuntos de dados publicados no Datahub estão no formato de *Linked Data* por meio do RDF, visto que a plataforma aceita a publicação dos dados em diversos formatos, como CSV (*Comma-Separated Values*), planilhas Excel, arquivos XML, PDF, arquivos de imagem, o que reforça a necessidade de uma verificação para seleção de quais *datasets* publicar no diagrama LOD.

Assim, será abordado a seguir o modo como é realizado o processo de criação e publicação de *datasets* e quais são os requisitos para que o *dataset* seja inserido no diagrama LOD.

2.2.1 Publicação de *datasets*

Conforme mencionado anteriormente, o Datahub consiste em uma plataforma que permite a publicação e criação de *datasets* e o gerenciamento de grupos e comunidades. Visto que os *datasets* que atendem aos requisitos de qualidade devem estar inseridos no Datahub, o primeiro passo para realizar a publicação de um *dataset* é criar uma conta de usuário nessa plataforma.

Para realizar o cadastro são necessárias as seguintes informações: nome de usuário, nome completo, e-mail e a senha. Para realizar a publicação do *dataset* é necessário ser membro de uma organização no Datahub, isso porque todos os conjuntos publicados são relacionados a uma organização e não ao usuário cadastrado.

A criação de organizações é gerenciada pela *Open Knowledge Foundation* – organização sem fins lucrativos que promove o conhecimento livre – por meio de uma rede de fóruns de discussão. Desse modo, é necessário criar também uma conta de usuário para fazer uso do fórum, visto que, para criar uma organização é necessário escolher um título e *slug* para a URL da organização, utilizar os dados da conta do Datahub para se tornar o administrador da organização e, então, realizar a requisição por meio de uma postagem no fórum com os dados

necessários. A organização é criada, assim, por meio de um dos administradores do fórum, e somente após a criação de uma organização é possível criar *datasets* no Datahub.

Finalmente, o processo de publicação do *dataset* é composto pelos seguintes passos:

- Criar uma conta de usuário no Datahub.
- Cadastrar a organização pela qual o usuário disponibilizará o catálogo do *dataset*.
- Verificar se o conjunto de dados não existe no datahub.io antes de adicioná-lo.
- Adicionar ou editar o conjunto de dados e descrevê-lo de acordo com o mínimo de informações requeridas: nome (identificador único), título, tipo da licença, organização, a visibilidade e a URL da fonte indicando de onde o *dataset* provém.
- Atribuir *tags* úteis relacionadas com o tema do *dataset.*
- Prover o arquivo com os dados a serem publicados (*WORLD WIDE WEB CONSORTIUM*, 2014)

Os passos listados anteriormente consistem nos requisitos para a publicação de *datasets* contendo qualquer tipo de dado. Os requisitos necessários para publicação diferem, quando se considera uma posterior inserção do *dataset* no diagrama LOD. Tais requisitos são abordados a seguir.

2.2.2 Inserção de *datasets* no LOD

Existem dois requisitos básicos para o conjunto de dados ser incluído no diagrama do LOD, que são: (1) os itens devem estar acessíveis via URIs referenciáveis, uma vez que o simples oferecimento do SPARQL *endpoint* (ambientes abertos para a realização de buscas), sem URIs referenciáveis, não torna o conjunto de dados apto para inclusão; e o (2) 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 (*WORLD WIDE WEB CONSORTIUM*, 2014). Passos para constituição do processo:

- Selecionar os vocabulários: a utilização de vocabulários existentes auxilia a interoperabilidade dos dados e facilita o desenvolvimento das aplicações; dentre os vocabulários mais utilizados estão o *DC Terms*, FOAF, SKOS, SIOC (*Semantical Interlinked Online Comunities*). 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 on-line como RDF;
- Adicionar metadados e *links* para a página.
- Adicionar um mapa semântico para o site (BIEZER et al., 2008).

Considerando uma possível inserção no diagrama LOD, os passos para publicação diferem nesse contexto, visto que se espera que campos específicos para publicação de *Linked Data* sejam preenchidos. Desse modo, o processo de publicação deve ser composto pelos seguintes passos:

- 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.
- Atribuir a *tag lod* aos novos conjuntos de dados inseridos. Se não souber de nenhum *inlink* ou *outlink* deve-se atribuir a *tag* lodcloud.*nolinks*.
- Prover a maior quantidade de informação adicional possível como SPARQL *endpoint*, descrição voID, licenças (*WORLD WIDE WEB CONSORTIUM*, 2011).

Ainda assim, o Instituto Hassu Plattner, incentivado pelo *World Wide Web Consortium*, disponibiliza um guia de completude para a publicação dos *datasets*, ou seja, um manual descrevendo quais são as informações necessárias a serem informadas sobre os *datasets* ao publicá-los, de modo que possam, quando avaliados, ser inseridos no diagrama LOD. As informações necessárias são divididas em três níveis de completude, sendo: nível 1 (básico), nível 2 (mínimo) e nível 3 (completo). Os dados necessários para cada nível são apresentados nos Quadros 4, 5 e 6.

Quadro 4 – Informações básicas, de nível 1, necessárias para publicar *datasets*

Campo	Descrição
Nome	Utilizar um ID (identificador) único para o *dataset*
Título	Nome completo do *dataset*
URL	*Link* para página do *dataset*
Autor	Nome da organização e/ou pessoa
E-mail	Contato do autor ou do mantedor
Tag	Utilizar a *tag* lod para que o *dataset* seja identificado como *Linked Data*

Fonte: Instituto Hasso Plattner (2016)

Os requisitos para avaliação disponibilizados nos Quadros 2, 3 e 4 são utilizados para verificar os *datasets* já publicados no Datahub, a fim de que, caso bem-sucedido nos testes realizados, o *dataset* seja incluído no diagrama LOD. Esses podem ser considerados como métodos de verificação para uma avaliação de qualidade, uma vez que conduzidos antes da inserção do *dataset* poderiam evitar que problemas de qualidade se propagassem no Datahub.

Quadro 5 – Informações de nível 2 necessárias para publicação de *datasets*

Campo	Descrição
Tag	Uma das 9 categorias, visto que essa informação é utilizada para colorir o diagrama.
Link para um exemplo RDF	Podendo ser nos formatos: rdf+xml, *turtle*, ntriples, x+quads, rdfa, x-trig
URL para SPARQL *endpoint*	API para SPARQL *endpoint*
URL de *download* para cada arquivo RDF	De acordo com cada formato disponibilizado (rdf+xml, *turtle*, x+ntriples, x-nquads, x-trig)
URL para uma página com a lista de *downloads*	Disponibilizar o *download* para múltiplos arquivos
Triplas (Informação adicional)	Valor aproximado do tamanho do *dataset* em triplas RDF
Links (Informação adicional)	Quantidade de *links* RDF apontando ao *dataset*
SPARQL *endpoint* (Informação adicional)	*Link* de acesso

Fonte: Instituto Hasso Plattner (2016)

Quadro 6 – Dados necessários para o nível 3 de completude do *dataset*

Campo	Descrição
Versão	Data da modificação ou da versão do *dataset*
Notas	Descrição do *dataset*
Licença	Licença padrão
Abreviação (Informação adicional)	Para inserir na circunferência do diagrama LOD
Link da licença (Informação adicional)	*Link* customizado
Namespace (Informação adicional)	Declaração da instância da *namespace*

Campo	Descrição
Arquivo voiD	*Link* para download
XML Sitemap	*Link* para download
RDF Schema	*Link* para download
Vocabulário	*Link* para download
Inserir uma das *tags*: no-proprietary-vocab, deref-vocab, no-deref-vocab	A primeira *tag* indica que o vocabulário utilizado não é proprietário e as seguintes que o *dataset* utiliza vocabulário proprietário e se são dereferenciáveis ou não, de acordo com as boas práticas para publicação de vocabulários RDF (BERRUETA et al., 2008).
Inserir as *tags*: vocab-mappings, no-vocab-mappings	Para informar se mapeamento para os termos de vocabulário proprietário são disponibilizados
Inserir as *tags*: provenance-metadata, no-provenance-metadata	Para indicar se o *dataset* provê a proveniência das meta-informações
Inserir as *tags*: *license-metadata*, *no-license-metadata*	Indica se o *dataset* provê meta-informações de licenciamento
Inserir as *tags*: *published-by-producer*, *published-by-third-party*	Informa se o *dataset* foi publicado por um produtor de dados ou por terceiros
Inserir a *tag*: *limited-sparql-endpoint*	Para informar se o SPARQL *endpoint* realiza buscas em todo o *dataset*
Inserir a *tag*: format-<prefix>	Indicar o vocabulário utilizado
Inserir *tag*: lodcloud.nolinks	Quando o *dataset* não possuir *links* RDF externos para ou de outros datasets
Inserir a *tag*: lodcloud.unconnected	Quando o *dataset* não possuir *links* RDF externos para ou de outros *datasets*
Utilizar a *tag*: lodcloud.needinfo	Quando o provedor ou o a página inicial do *dataset* não provê informações mínimas
Utilizar a *tag:* lodcloud.needsfixing	Quando o *dataset* não está funcionando e fornecer detalhes nas notas.

Fonte: Instituto Hasso Plattner (2016)

Mesmo realizando esse processo de verificação antes da inserção no diagrama, é possível que os *datasets* sejam afetados por problemas de qualidade, visto que podem ser atualizados, bem como *datasets* podem ser tirados do ar ocasionando *links* quebrados.

Em síntese, este capítulo abordou definições de conceitos e tecnologias que são essenciais para a Web Semântica, de modo que agentes e outras aplicações computacionais possam beneficiar-se dos avanços que a Web Semântica provê. É essencial que os dados publicados, além de seguir as boas práticas para publicação, estejam livres de problemas de qualidade. Conforme abordado neste capítulo, medidas de qualidade são definidas para que os dados possam ser considerados aptos para serem inseridos no diagrama LOD. Assim, a qualidade dos dados desempenha um papel importante para que as aplicações atendam a necessidades específicas. Portanto, o capítulo a seguir realiza uma análise de quais problemas e dimensões de qualidade são detectados em *datasets* do *Linked Data.*

3 QUALIDADE DE DADOS

Qualidade pode ser encarada como um conceito subjetivo sobre a percepção de um indivíduo em relação a um serviço, produto, dado, informação, etc. De modo geral, 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. Nota-se 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.

Quando se fala de qualidade aplicada a dados, significa 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.

Juran et al (1974) realizaram a definição comumente adotada por autores que trabalham com qualidade (WANG; STRONG, 1996; STRONG et al., 1997; NAUMANN, 2002; BATINI et al., 2008), em que qualidade de dados consiste na aptidão para o uso, visando realizar determinadas tarefas. Considerando qualidade como sendo adequação para o uso, Bizer e Cyganiak (2009) abordam dois aspectos relevantes: (1) a definição de um dado, objeto, produto ou serviço de qualidade consiste em uma tarefa dependente, visto que um usuário pode considerar a qualidade de tal dado útil para realizar determinadas tarefas, porém a mesma qualidade do dado pode ser considerada insuficiente em uma tarefa diferente; (2) consiste também em uma tarefa subjetiva, uma vez que um terceiro usuário poderia ter uma perspectiva diferente da qualidade do mesmo conjunto de dados e considerá-lo útil para realizar ambas as tarefas.

Quanto à qualidade de dados, outro fator relevante identificado consiste na forma como as dimensões são definidas, ou seja, são estabelecidas, visando atender às necessidades específicas do domínio ao qual estão sendo aplicadas. Em razão disso, as dimensões variam de acordo com cada metodologia e domínio; assim, algumas dimensões mostraram-se presentes em diversas metodologias, como completude, precisão, *timeliness*. Ressalte-se o fato de que a maioria das metodologias utiliza a avaliação subjetiva para a definição do que pode ser considerado de qualidade ou não.

O domínio do sistema de informação desempenha um papel muito importante para medir a qualidade dos dados, visto que é por meio dele que os requisitos são estabelecidos. Tais requisitos podem ser definidos como itens aos quais o dado deve atender para ser considerado de qualidade. Os problemas de qualidade são classificados em diferentes categorias, chamadas de dimensões, nas quais problemas do mesmo tipo são classificados de acordo com uma categoria específica. No geral, as dimensões possuem descrições e definições similares, porém a aplicação e como funciona cada dimensão possuem características distintas, de acordo com o domínio no qual estão sendo aplicadas.

Problemas de qualidade podem resultar em grandes desastres. Algumas falhas conhecidas relacionadas a problemas de qualidade resultaram não somente em danos financeiros, mas afetaram também a vida das pessoas envolvidas. Fisher e Kingma (2001) descrevem dois desastres envolvendo qualidade, os quais resultaram em grandes perdas financeiras e mais de 200 pessoas perderam suas vidas. O primeiro acidente ocorreu no dia 28 de janeiro de 1986, quando, segundos após o lançamento, o ônibus espacial *Challenger* explodiu matando todos os tripulantes a bordo (7 pessoas) e teve o prejuízo financeiro de mais de um bilhão de dólares. Isto aconteceu quando o *O-ring* do foguete de combustível sólido falhou, rompendo o selamento e conduzindo à explosão da nave. De acordo com a revisão realizada pelos autores, o sistema de informações gerenciais da Nasa possuía graves problemas de qualidade de dados, como inconsistências e erros, violações de relatórios, falta de modelagem de tendências e uma pobre integração de componentes e testes. Foram constatados erros, de acordo com diferentes dimensões de qualidade (consistência, precisão e completude), tanto no banco de dados como nos relatórios.

Outro acidente catastrófico aconteceu no dia 3 de julho de 1988, quando uma embarcação da Marinha dos Estados Unidos disparou dois mísseis em um avião, acreditando ser um caça militar em procedimento de ataque. Os dispositivos da embarcação identificaram de modo errado a aeronave, o que resultou na destruição do Airbus Iraniano, voo 655, ocasionando a morte de 290 pessoas a bordo. Problemas de qualidade de dados contribuíram

para esse acidente, no qual foram detectados a utilização de indicadores errados de alvos, informações incompletas e conflitantes, problemas de comunicação por voz e sobrecarga de informações (FISHER; KINGMA, 2001).

Os exemplos citados representam grandes falhas que podem ser causadas por problemas de qualidade. Um ponto importante é que nos dois acidentes os sistemas, tanto para gerenciar informações, quanto para tomada de decisões, possuíam problemas de qualidade. Sob o panorama de que um sistema de informação utiliza dados para transformá-lo em informações úteis para tomada de decisão e elaboração de planos estratégicos, as grandes falhas ocasionadas por problemas de qualidade comprovam que os dados são afetados de modo direto. Isso porque, ao inserir dados livres de problemas em um sistema falho quanto à qualidade, consequentemente as informações geradas por tal sistema também serão afetadas. Destaca-se, assim, a importância de que os dados estejam livres de problemas de qualidade para auxiliarem o processamento eficaz, pelas aplicações que os integram.

Visto que cada domínio possui diferentes requisitos de qualidade, é possível encontrar diversos pontos de vista quanto à aplicação das dimensões de qualidade na literatura (WANG; STRONG, 1996; VEREGIN, 1999; LEE et al., 2002; PIPINO et al., 2002; BATINI et al., 2008; LAUDON, 1986; FISHER; KINGMA, 2001). Por meio da análise realizada nota-se que existem algumas dimensões que são mais utilizadas, porém não existe um padrão estabelecido de conjunto de dimensões; cada domínio utiliza dimensões de avaliação que atendam aos seus requisitos. A seguir serão analisados diferentes conjuntos de dimensões e suas definições, de acordo com diferentes autores.

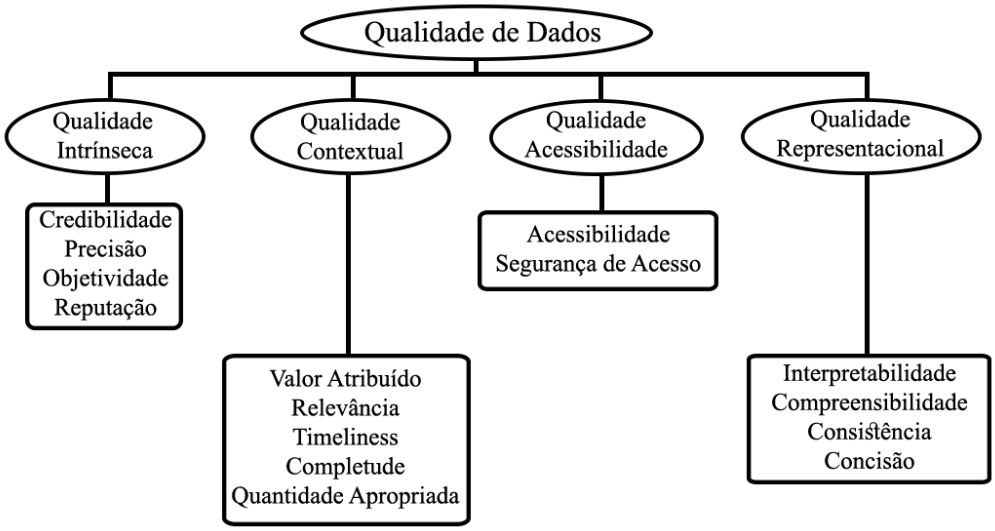
Wang et al (1995) classificam os problemas de qualidade em quatro dimensões: acessibilidade, interpretabilidade, utilidade e credibilidade, no sentido de que o dado tem que estar acessível, o usuário deve ter permissão para acessá-lo, compreender a sintaxe e a semântica, deve ser útil no processo de tomada de decisão e, por fim, deve ser possível utilizá-lo como insumo ao tomar decisões. Conforme descrito pelos autores, cada dimensão abrange diferentes requisitos de qualidade, sendo eles:

- Acessível: disponível;
- Interpretável: sintaxe, semântica;
- Útil: relevante, conveniência (atual e não volátil);
- Acreditável: completa, consistente, fonte confiável, precisa;

O estudo relatado por Wang e Strong (1996) possui a descrição comumente utilizada em muitos trabalhos sobre qualidade de dados. Esses autores categorizam os atributos das

dimensões da qualidade em quatro classes principais, conforme a hierarquia apresentada na Figura 15.

Figura 15 – Classificação das dimensões de qualidade



Fonte: Wang e Strong (1996)

Qualidade intrínseca de dados implica a garantia da credibilidade e reputação dos dados, sendo que dentre os atributos constam a própria credibilidade e a reputação, assim como a precisão e a 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, sendo tais atributos: relevância, tempo, completude, etc.

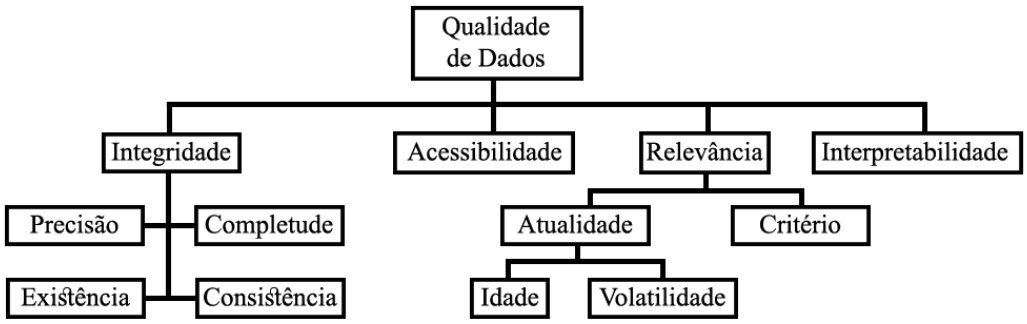
Qualidade representacional aborda aspectos relacionados ao formato dos dados (representação concisa e consistente) e o significado do dado (interpretabilidade e facilidade de compreensão). Nessa classe, para a informação estar bem apresentada, além de concisa e consistente, ela deve ser interpretável e de fácil compreensão.

E, por fim, qualidade de acessibilidade é considerada um aspecto muito importante para a qualidade de dados. Sugere-se que, dependendo do contexto, a qualidade quanto acessibilidade pode ser tratada como um aspecto individual ou incorporada a outras categorias.

Com base nos modelos definidos por Wang et al (1995), Wang e Strong (1996), Boove et al (2003) realizam uma diferente abordagem quanto à classificação das dimensões de qualidade. A metodologia adotada para a definição do modelo foi composta pelos seguintes passos: (1) juntar informações que possam ser úteis (acessibilidade), (2) compreender e aplicar o significado (interpretabilidade), (3) aplicá-la a determinado domínio e propósito de interesse dentro de um contexto (relevância) e por fim (4) acreditar que tal dado está livre de erros

(integridade). E, então, de acordo com os quatro princípios, as dimensões são definidas conforme a hierarquia apresentada na Figura 16.

Figura 16 – Modelo de dimensões e requisitos de qualidade



Fonte: Boove et al (2005)

Nota-se que a classificação de dimensões de qualidade de acordo com as abordagens de Wang et al (1995) e Wang e Strong (1996) são realizadas conforme princípios aos quais os dados precisam atender, e de acordo com os princípios são descritas as dimensões que os compõem, o que é diferente do modelo definido por Boove et al (2005), o qual é composto por dimensões principais, que por sua vez podem ou não serem divididas em subdimensões, como no caso da acessibilidade e interpretabilidade.

Os modelos diferem quanto às principais dimensões, mas nota-se que existem dimensões similares, com algumas diferenças entre as principais dimensões de qualidade; todos abordam similares problemas de qualidade, como relevância, completude, consistência e precisão. Desse modo, não há uma concordância unânime quanto às dimensões de qualidade, uma vez que os modelos existentes são definidos de acordo com requisitos, princípios, metodologias e domínios diferentes. Porém, as dimensões presentes nos modelos analisados estão presentes também nos diversos estudos documentados quanto a problemas e dimensões de qualidade. Desse modo, os próximos tópicos abordarão a definição das dimensões mais utilizadas, ao discutirem qualidade de dados e diferentes metodologias para avaliação e detecção de problemas.

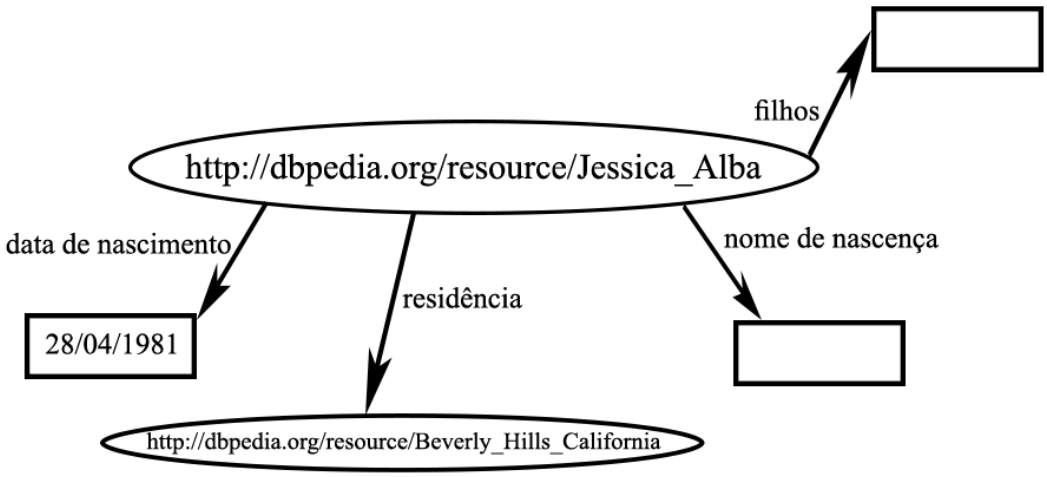
3.1 Dimensões de Qualidade de Dados

Considerando a presença unânime de tais dimensões na literatura, uma descrição dos diferentes conceitos e princípios quanto a sua descrição é apresentada a seguir.

3.4.1 Completude

No geral, relaciona-se com quanto o dado pode ser considerado completo de acordo com o que se espera dele; diferentes descrições de completude são apresentadas no Quadro 4. Aplicando essa dimensão na área de pesquisa de banco de dados relacionais, Batini et al. (2009) relacionam completude com valores nulos, significando que eles existem no mundo real, mas não constam no conjunto de dados. Assim, é importante analisar porque tal valor não está disponível. Para exemplificar, a Figura 17 apresenta uma declaração RDF na qual a atriz Jessica Alba é representada como um recurso pela URI, correspondente aos dados da atriz no DBPédia. Considerando que os itens consistem em todos os dados existentes no DBPédia quanto a esse recurso, a ausência de dados quanto ao nome e à quantidade de filhos classifica essa declaração como incompleta.

Figura 17 – Declaração RDF com problemas de completude



Fonte: Elaborada pela autora.

De acordo com Fox et al (1994) os valores nulos podem significar três hipóteses: (1) o dado existe no mundo real, mas é desconhecido por sua entidade digital, (2) o valor não existe no mundo real e (3) não se sabe se o valor existe ou não no mundo real. Considerando as definições de completude dispostas por Batini et al. (2009) e Fox et al. (1994) a única hipótese na qual uma informação poderia ser considerada sem problemas de completude é quando o valor não existe no mundo real.

Em uma análise em registros criminais dos Estados Unidos da América (LAUDON, 1986), a definição de completude foi realizada com base em registros criminais que indicavam uma prisão, mas não havia qualquer disposição formal da corte no período de um ano da data

de prisão. Nessa análise, outro requisito que influenciou a consideração de que o registro estaria incompleto foram formulários que demandavam prisão, sendo que nenhum crime estava relacionado; assim também, registros que possuíam as sentenças, mas não apresentavam informações convincentes.

Souza et al. (2015) descrevem um modelo de informações esperadas para atender chamadas de roubo e a completude relaciona-se com a obtenção da quantidade de informações esperadas. Nesse domínio, caso a quantidade de informações esperada seja 10, e por meio da denúncia recebida as 10 informações esperadas forem obtidas a informação pode ser considerada completa. Diferentes definições de completude são apresentadas no Quadro 7.

Quadro 7 – Definições de completude de acordo com diferentes autores

Referência	Descrição
Wand e Wang (1996)	Capacidade de um sistema de informação para representar cada estado significativo de um sistema do mundo real
Wang e Strong (1996)	Extensão em que os dados são de amplitude, profundidade e escopo suficiente para a tarefa em mãos.
Jarke et al. (1999)	Nível de cobertura do contexto, podendo se relacionar com o número de valores e entidades ausentes.
Lee et al. (2002)	Ausência de qualquer valor
Pipino et al. (2002)	Grau em que os dados não estão faltando e são de amplitude e profundidade suficientes para a tarefa em mãos
Amicis e Batini (2004)	Grau da presença e da ausência de valores de dados em uma variável

Fonte: Adaptado de Batini et al (2009)

Cada domínio de aplicação possui suas próprias peculiaridades, características e requisitos. Realizar uma avaliação de qualidade pode ser um processo delicado e que exige esforço, mas pode impactar de forma positiva os sistemas de informação.

Para um domínio de banco de dados relacionais, é muito importante refletir sobre questões como: quais medidas serão tomadas para distinguir se a informação existe no mundo real e é desconhecida em sua entidade digital? Como distinguir se existe um valor no mundo real e esse valor é desconhecido por sua entidade digital? É possível criar um método de controle para prever problemas de completude nesses casos?

Tais questionamentos podem ser importantes não somente para bancos de dados, mas também para qualquer tecnologia que possua uma estrutura de organização de informações

inseridas por humano.

No caso da aplicação de completude descrita por Laudon (1986) e Souza et al. (2015), pode-se dizer que os ambientes são controlados, no sentido de que já é de conhecimento prévio quais tipos e que quantidade de informações se espera obter; caso tais informações não estejam presentes significa que há um problema de qualidade quanto à completude. Outro fator a ser levado em consideração é que ambos necessitam do papel do humano no processo de avaliação de qualidade.

Definiu-se para este trabalho de pesquisa que completude se relaciona com a inteireza de um conjunto de informações, de acordo com as definições de determinado domínio. Nesse sentido, pode-se classificar completude em dois tipos: global e local. Completude global acontece quando se considera o domínio como um todo, ao passo que se aplica o termo completude local para conjuntos de informações necessárias dentro de um domínio, no qual a falta de alguma informação impossibilita a realização de tarefas específicas.

3.4.2 Precisão

Assim como a completude, a precisão também possui diferentes descrições, de acordo com diferentes autores. Wand e Wang (1996) descrevem precisão como uma descrição autêntica do mundo real; assim, uma informação imprecisa implica que o sistema de informação apresenta um estado no mundo real diferente do que deveria ter sido apresentado. De acordo com Wang e Strong (1996), precisão consiste no grau em que os dados estão corretos, confiáveis e certificados como livres de erros. Também é descrita como valor, já inserido no sistema, para exemplificar, o registro de um usuário que nasceu no ano de 1992; porém, se a idade do funcionário for diferente de 24, os campos estão com problemas de precisão. Problemas de precisão também acontecem com dados que não podem ser exclusivamente definidos, como nomes estrangeiros, por exemplo, que possuem grafias alternativas (FOX et al., 1994).

Para Batini et al. (2006) a precisão consiste na proximidade de dois valores: A e A', sendo que o valor A' representa o estado real de um dado, e o valor A consiste no dado correspondente ao A' em um ambiente digital. Assim, para exemplificar, caso A' = Jessica e A = Jssica, o valor de A está impreciso em relação ao valor de A’. Outro exemplo, considerando o nome Cides; nota-se que a partir de uma conversa por telefone podem ser gerados diferentes problemas de ortografia, tais como “Alcides”, “Euclides”, “Cide”, “Sidis”, que consistem em versões imprecisas do nome original. Diferentes definições de precisão são apresentadas no Quadro 8.

Quadro 8 – Diferentes definições de precisão

Escala de exatidão do conteúdo do dado (requer uma fonte de referência para verificação) (MCGILVRAY, 2008)
Dados certificados corretos, sem falhas, de confiança, livre de erros ou que podem ser facilmente identificados, relaciona-se com a integridade e exatidão dos dados. (WANG; STRONG, 1996)
Consiste em uma informação verdadeira ou livre de erros em relação a um valor conhecido, medido ou mensurado (BOOVE et al., 2003).
Quando o valor registrado está em conformidade com o real valor (WANG et al., 1995)

Fonte: Elaborado pela autora.

Podem-se considerar dois tipos de precisão: precisão sintática e semântica (BATINI et al, 2009). Precisão sintática relaciona-se com palavras individuais e a real representação delas em um determinado domínio. Desse modo, considerando um conjunto de dados com nomes de animais, mesmo que a palavra Higuana possua um erro de ortografia, pode ser considerada sem problemas de precisão, se existir essa exata palavra no conjunto de dados em questão. Outro exemplo: considerando que a = Rebeca e A = Rebecca, A pode ser considerado preciso, caso exista no domínio avaliado; ambos podem ser considerados precisos se existirem no conjunto de dados de tal domínio. Assim, o domínio pode ser um fator predeterminante nesse caso, mesmo existindo problemas de ortografia, porque no domínio em que o dado está inserido, ele é considerado um dado preciso.

Quanto à precisão semântica, consiste em quão correta está uma informação, de acordo com o que ela representa no mundo real. Esse tipo de precisão relaciona-se com o significado da informação representada. Considerando a Figura 17 apresentada anteriormente, um problema quanto à precisão semântica é apresentada na descrição RDF da atriz Jessica Alba, que possui os seguintes dados: ano de nascimento, residência, quantidade de filhos e o nome de nascença. De acordo com a descrição, a atriz reside na cidade de Beverly Hills, Los Angeles; tal informação poderia estar com problema caso esta informação fosse verdadeira e o recurso correspondente a outro local estivesse disponível.

Assim, a precisão pode ser definida de acordo com três pontos de vista diferentes. Como precisão no geral, precisão sintática e precisão semântica. Nomeia-se precisão geral, como o fato de um dado estar livre de erros e correto para a sua utilização. Problemas quanto a esse tipo de precisão podem ocorrer ao se lidar com nomes estrangeiros, erros ortográficos ou de digitação. Precisão sintática consiste em um dado livre de erros dentro de um determinado domínio, considerando como domínio um conjunto de dados predeterminados, no qual mesmo existindo erros ortográficos ou de digitação o dado pode estar preciso, caso exista dentro de tal

domínio. E por fim precisão semântica, na qual informações devem descrever a sua exata instância no mundo real. Nesse caso, os problemas acontecem quando sua entidade representativa não descreve seu correspondente.

3.4.3 Relevância

De acordo com a literatura sobre o tema, a relevância consiste no grau segundo o qual determinado conjunto de informação atende às necessidades do usuário, bem como em que nível o dado é aplicável e útil para a tarefa a ser realizada (AGRE et al, 2011; WANG; STRONG, 1996; BATINI; SCANNAPIECO, 2006). De acordo com o levantamento de avaliações de métricas realizado por Batini et al (2007), a relevância pode ser medida por meio de métodos subjetivos, como avaliações aplicadas por usuários experientes no domínio.

Quando a relevância é dividida em subcategorias, passa a ser composta por diversas dimensões que são tratadas de modo individual na literatura, como, por exemplo, atualidade e volatilidade (WANG et al, 1995; WANG; STRONG, 1996), que consistem em dimensões que abordam problemas temporais da informação. Nesses casos, tais aspectos são tratados como requisitos para a relevância, visto que uma informação desatualizada pode não ser relevante para realizar determinadas tarefas.

No contexto da recuperação da informação, a relevância relaciona-se com recuperar conteúdos que estejam de acordo com as necessidades do usuário. Em vista da subjetividade de definir o que é relevante de acordo com o ponto de vista, diversos modelos têm sido utilizados e aplicados visando atingir de modo razoável tais necessidades. O modelo booleano, por exemplo, tem como base a teoria dos conjuntos e a álgebra de Boole. As consultas combinam expressões de busca com operadores booleanos e, ou e não (utilizados em inglês, sendo AND, OR e NOT). O processo é realizado com base no processamento sintático das palavras, por meio de uma comparação extada dos termos da busca. Nas buscas realizadas a partir desse modelo, o usuário define palavras-chave que possam estar contidas no documento e serem relevantes de acordo com sua necessidade. Nesse modelo, os resultados são considerados relevantes quando o documento possui palavras similares às chaves de busca inseridas pelo usuário, o que pode ser considerada uma tática falha, visto que mesmo contendo a palavra no título do documento ou até mesmo no assunto, o resultado recuperado pode não ser relevante do ponto de vista do usuário.

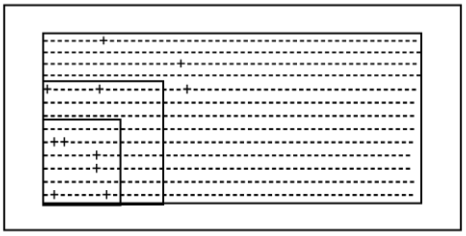
O modelo vetorial determina a relevância de um documento por meio de um vetor no qual pesos são atribuídos aos elementos do termo de indexação, sendo que cada vetor

corresponde à posição do documento dentro de um espaço multidimensional e os termos de indexação consistem nos eixos. Então um peso entre zero e um é atribuído a cada elemento do vetor e os mais próximos de um são considerados de maior relevância para o documento. Assim, os termos de busca e de indexação mais pontuados são considerados os mais relevantes nesse modelo (FERNEDA, 2003).

Há também o modelo realizado pelo Google, no seu processo de recuperação da informação. Uma pesquisa realizada no Google funciona da seguinte maneira: (1) primeiro acontece um rastreamento a fim de descobrir páginas novas e atualizadas a serem incluídas no índice do Google; nesse processo é levado em consideração a frequência e quantas páginas devem ser buscadas em cada site; (2) em seguida as páginas são processadas visando compilar um índice com as palavras encontradas e a localização delas na página (atributos, *tags* de conteúdo, como de título e atributos alternativos são levados em consideração); (3) o próprio Google define a relevância dos resultados de acordo com mais de 200 fatores diferentes, dentre eles o *PageRank* (BRIN; PAGE, 2012). O *PageRank* utiliza como medida a qualidade, a quantidade e o contexto dos *links* que a página recebe, para definir sua relevância. Esses três princípios são considerados para definir o valor de cada *link*, reforçando o valor de *PageRank* de cada página (GOOGLE, 2015). Assim, quem define a relevância é o mecanismo de busca e não o usuário, o que muitas vezes exclui resultados que são realmente relevantes.

A Figura 18 ilustra como problemas de qualidade quanto à relevância podem afetar esse processo de recuperação, no qual o retângulo maior representa uma base de dados e os itens armazenados nela. Os sinais de adição "+" representam dados que poderiam hipoteticamente atender às necessidades do usuário. Os arquivos representados por "-" consistem em dados não úteis. Para cada necessidade de informação haverá mais itens – do que +. Assim "o problema está em recuperar tantos itens *relevantes* quanto for possível, e o menor número possível de itens *irrelevantes*" (LANCASTER, 2004).

Figura 18 – Relevância no contexto da recuperação da informação



Fonte: Lancaster, 2004

De acordo com a análise realizada, nota-se que a relevância pode ser encarada como

uma dimensão de caráter subjetivo, visto que definir se uma informação é relevante ou não para realizar tarefas específicas depende do usuário responsável por executá-las. Diferentemente das dimensões citadas anteriormente, a relevância é uma dimensão que requer uma avaliação de usuários, podendo, às vezes, ser identificada por complexos e custosos sistemas computacionais, mas tendo a possibilidade de ser detectada por processos automatizados quando em ambientes pequenos e controlados, nos quais se sabe ao certo que tipo de dado se espera obter, mediante requisitos pré-definidos.

3.4.4 Consistência

Consistência pode relacionar-se com firmeza, coerência, e quando qualificando uma pessoa, transmitir o quadro mental de um sujeito que é regular quanto ao seu modo de ser, agir e se comportar. Aplicando o sentido literal para o contexto dos dados, este, quando considerado consistente, equivale a um conjunto de dados verdadeiros, regulares e aptos para realizar as tarefas necessárias.

A literatura define consistência como violação de regras semânticas para um determinado conjunto de dados. Grande parte de literatura relaciona consistência com a área de bancos de dados e estatística, que utiliza constantes pré-definidas para detectar e solucionar os problemas (FOX et al, 1994; WINKLER, 2004; CONG et al, 2007). Batini e Scannapieco (2006) descrevem consistência como coerência do mesmo dado, representado em múltiplas cópias, ou de dados diferentes, quanto a constantes e regras de integridade.

Scannapieco et al (2005) exemplificam diferentes regras semânticas para detectar problemas de inconsistência, como: se o estado marital é casado, a idade do indivíduo não deve ser menor do que 18 anos; já a idade de um funcionário deve ser entre 18 e 55; se os anos de trabalho forem em número menor que 3, o salário não pode ser mais que 25.000 por ano; o número correspondente ao ano de refilmagem de um filme não pode ser menor do que o correspondente ao ano do primeiro lançamento. No domínio de banco de dados, erros durante a modelagem da estrutura podem resultar em resultados duplicados, os quais, nesse caso, são encarados como inconsistentes.

Desse modo, métricas para avaliação são definidas segundo um conjunto de regras predefinidas semânticas para um conjunto de dados. Batini et al (2007) 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 (KOUDAS et al, 2006); e a segunda métrica é utilizada para verificar regras de

negócio.

Entende-se então que um conjunto de informações está consistente quando não há contradições inseridas nele. Considerando o contexto do *Linked Data*, os *datasets* possuem regras e instruções específicas de modo que sejam publicados e atendam aos requisitos mínimos de qualidade. Tais requisitos podem ser considerados como regras semânticas, que, nesse caso, quando quebradas, resultam em um *dataset* inconsistente.

Em síntese, este capítulo abordou diferentes definições quanto ao que é considerado qualidade. Devido aos diversos pontos de vista, notamos que a qualidade se relaciona com o que é necessário para que um dado seja utilizado de acordo com o pretendido, ou para que um produto seja confeccionado de forma a atender as necessidades de seu consumidor, ou, então, medidas necessárias para que uma organização gerencie de forma positiva seus recursos. Independentemente do domínio, nota-se que todos se utilizam de sistemas para auxiliar em seus processos, o que por sua vez promove um ambiente para que os dados possam ser afetados de maneira direta por problemas de qualidade. Os problemas causados pela qualidade ressaltam a importância de medidas para assegurar que tais problemas não aconteçam, a fim de proporcionar o sucesso das tarefas desenvolvidas.

Desse modo, considerando o impacto social e o desenvolvimento que o *Linked Data* proporciona ao cenário atual, será abordado a seguir: como os problemas de qualidade podem afetá-lo de acordo com suas respectivas dimensões, quais tipos de problemas podem ser encontrados e quais são os requisitos de qualidade seguidos pelos *datasets* publicados.

Os padrões comumente utilizados para projetar ontologias (OWL, RDFs) auxiliam a definir a semântica ou o significado do dado RDF, por meio da definição de classes e propriedades, o que permite interpretar tais recursos e suas relações e, em função disso, considerar que problemas de qualidade quanto à consistência são detectáveis somente por meio da análise do avaliador.

Considerando que problemas de qualidade podem afetar qualquer ambiente que lide com dados, é possível encontrar diversas metodologias para avaliação de qualidade, na literatura. Serão abordadas, a seguir, metodologias para avaliação de qualidade de dados em diferentes domínios, visando identificar como acontece o processo de avaliação, o que é levado em consideração, quais são os requisitos e as dimensões avaliados.

3.2 Metodologias para avaliação de qualidade

Em vista do grande índice de citação, a metodologia chamada *Total Data Quality*

Management (TDQM) é a mais descrita para avaliação de qualidade. A TDQM possui quatro fases: (1) definição na qual são relacionados os dados, como produtos e suas características, dimensões de qualidade e requisitos, (2) medição, onde métricas de qualidade de dados para dimensões, definidas na fase 1, são identificadas. Na fase (3) análise, as dimensões que carecem de melhorias são analisadas, bem como a causa dos possíveis problemas de qualidade. Por fim, na fase (4) melhoria, são aplicadas as ferramentas e *frameworks* para solucionar os problemas de qualidade identificados (WANG, 1998). Os autores classificam os problemas de qualidade de acordo com as seguintes dimensões: precisão, objetividade, credibilidade, reputação, acesso, segurança, relevância, valor agregado, *timeliness*, quantidade de dados, interpretabilidade, concisão representacional e consistência representacional.

Quanto à qualidade de dados, Amicis e Batini (2004) abordam uma metodologia para avaliação de dados financeiros. Nesse contexto, dados financeiros correspondem a registros de dados usados para descrever instrumentos financeiros, dados referentes a preços e taxas de câmbio, provenientes de teorias temporais e de modelos financeiros. A metodologia tem como objetivo avaliar a qualidade desses dados armazenados em bancos de dados, a fim de aprimorar o desempenho dos negócios. A avaliação de qualidade acontece sobre dados de registros fornecidos por fontes externas, utilizados por institutos financeiros, que são considerados cruciais para a operacionalidade dos negócios. Nesse contexto, os problemas de qualidade podem acontecer no carregamento de dados descontrolados e não estruturados e no processo de atualização. Os problemas são classificados de acordo com diferentes categorias e/ou dimensões, sendo elas: precisão sintática e semântica, consistência interna e externa, completude, *currency*, *timeliness* e singularidade.

A metodologia proposta por Amicis e Batini (2004) é composta por cinco fases:

- Fase 1 (seleção de variáveis): são selecionadas de acordo com o grau de importância para conduzir análises posteriores e avaliar os resultados finais;
- Fase 2 (análise): avalia dimensões de qualidade por meio de métricas e regras de negócio para identificação de erros;
- Fase 3 (análise objetiva): o grau de erros detectados já fornece uma avaliação objetiva;
- Fase 4 (análise subjetiva): profissionais experientes no domínio financeiro e em qualidade de dados realizam uma avaliação subjetiva qualitativa;
- Fase 5 (comparação): é por fim realizada uma análise comparativa entre a avaliação subjetiva e objetiva, provendo assim sugestões para melhorar a qualidade dos dados.

Nota-se que, no domínio de dados financeiros, a metodologia adotou dois tipos de

avaliação, uma avaliação objetiva, por meio de cálculos matemáticos, considerando diversos fatores como: regras de negócio e métricas, segundo os autores Amicis e Batini (2004), apropriadas para a avaliação. A segunda consistiu em uma avaliação subjetiva, na qual profissionais experientes foram utilizados como um recurso para conduzir a avaliação.

3.3 Qualidade de dados no Linked Data

Por meio do *Linked Open Data* (LOD), um grande volume de dados linkados têm sido disponibilizado na Web de modo público por meio de licença livre. A publicação de tais dados possibilita agregar informações de diferentes fontes e domínios para a construção de poderosas aplicações capazes de descobrir informações novas e relacionadas. No entanto, estudos recentes mostram que a maioria dos *datasets* publicados de acordo com diferentes categorias apresentam problemas de qualidade como inconsistência, problemas representacionais, interoperabilidade, completude, etc. (HOGAN et al., 2012; RULA; ZAVERI, 2014).

Na análise realizada por Acosta et al (2013), inúmeros problemas de qualidade foram identificados no *dataset* do DBpédia, que é um dos maiores conjuntos de dados 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, mas sem que conste 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

Conforme abordado anteriormente, os *datasets* inseridos no Datahub passam por um processo de avaliação que verifica se os conjuntos atendem aos seguintes requisitos: (1) os itens de dados devem ser acessíveis por meio de URIs dereferenciáveis e (2) o *dataset* declara ao menos 50 *links* RDF que apontam para outros *datasets* ou deve existir pelo menos um *dataset* que possua 50 RDF *links* apontando para o *dataset* em questão. Assim, o processo de avaliação

considera apenas os componentes estruturais dos dados, não o que eles significam e/ou se possuem problemas de qualidade. Por esse motivo, o DBpédia possui problemas de qualidade quanto aos dados, e conforme os dados disponibilizados pelo Wikipédia aumentam, é possível que os problemas de qualidade se propaguem através de tal *dataset*.

Entrementes, o DBpédia consiste em apenas um *dataset* que compõe uma das nove categorias de *datasets* que podem ser publicadas no LOD, categorias que possuem diversos *datasets* que, por sua vez, podem possuir também problemas de qualidade. Desse modo, este trabalho tem como objetivo propor uma metodologia para avaliação de *Linked Data* dos dados segmentados na categoria publicações, visando, por meio de testes, verificar quais são os problemas de qualidade que afetam esse tipo de dados. Objetiva-se, com o resultado da análise, determinar medidas de avaliação que auxiliem na detecção de dados com problemas de qualidade, antes do processo de publicação no LOD.

Considerando o fato de que antes que um *dataset* seja inserido no diagrama LOD ele passa por um processo de validação de acordo com os requisitos estabelecidos, a análise será realizada nos *datasets* já inseridos, em busca de problemas de qualidade. Visto que o último relatório foi publicado em 2014, é provável que muitos *datasets* possam estar desatualizados, ou com problemas temporais, assim como *links* podem estar desativados ou inexistentes. Desse modo, todos os *datasets* analisados já estão inseridos no diagrama LOD.

3.4 Dimensões de Qualidade para Linked Data

Visto que o *Linked Data* possui características singulares, algumas dimensões são únicas e outras são adaptadas de acordo com o contexto de dados linkados. A seguir descrevem-se as dimensões para *Linked Data* mais utilizadas:

3.4.1 Interlinking

Interlinking é o termo aplicado aos RDF *links* responsáveis por relacionar as entidades identificadas pelo sujeito com as entidades identificadas pelo objeto. Como dimensão consiste no grau de relacionamento das entidades entre um ou mais *datasets* (ZAVERI et al, 2015; HOGAN et al, 2010; HOGAN et al, 2012).

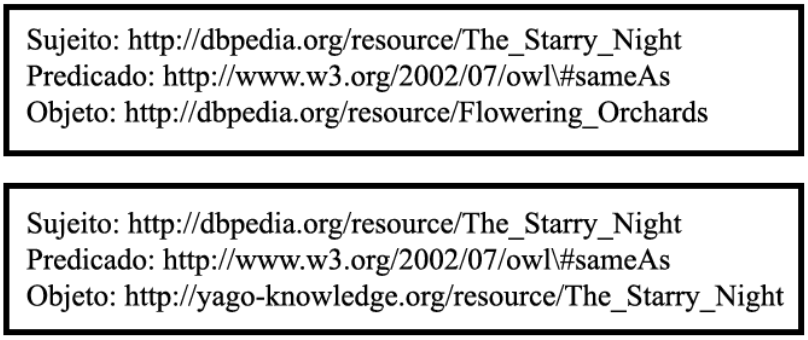
Visto que possuir ao menos 50 *links* para outro *dataset* ou de um *dataset* apontando para o em questão consiste em um dos requisitos para que o conjunto de dados seja publicado no diagrama LOD, esta dimensão avalia se esta prática está sendo realizada pelos *datasets*, se os

links funcionam, se possui *links* externos, bem como se as URIs atendem os critérios de boas práticas estabelecidos.

Conforme abordado por Zaveri et al (2013), problemas de *interlinking* acontecem quando os *links* para *websites* ou *datasets* externos estão incorretos, não apresentam nenhuma informação, ou estão expirados.

A Figura 19 exemplifica um problema de *interlinking* no qual o RDF *link* superior, o recurso "*The Starry Night*", uma obra do pintor Vincent Van Gogh do DBpedia, é interligado por meio da propriedade 'mesmo que', com o recurso chamado "*Flowering Orchards*". Nesse caso, o problema de qualidade acontece em razão do *link* que foi declarado ser o mesmo recurso não condiz com o que propõe, visto que *Flowering Orchards* consiste em uma série de obras realizadas por Van Gogh. O RDF *link* inferior ilustra a forma correta, sem problemas de qualidade quanto a *interlinking.*

Figura 19 – RDF *link* com problema de *interlinking*



Fonte: Elaborada pela autora.

Dependendo da forma como o problema de *interlinking* ocorra, este poderá associar-se com outras dimensões; com completude, por exemplo, quando o *link* estiver incompleto, com consistência, como no caso ilustrado na Figura 19, etc.

3.4.2 Licenciamento

Uma característica singular para dados linkados consiste no fato de que os documentos RDF devem conter uma licença, de modo que o conteúdo possa ser usado e reutilizado, permitindo aos consumidores das informações utilizar os dados sob termos legais (HOGAN et al, 2012; BIZER; CYGANIAK, 2009).

Conforme abordado por Bizer et al (2008), o LOD é um projeto que tem como objetivo identificar *datasets* disponíveis sob licença e publicá-los em RDF, interligando-os. Atualmente,

a comunidade do W3C disponibiliza guias e princípios a serem seguidos, para que os dados sejam publicados no LOD, e possuir licença consiste em um deles.

Essa dimensão permitirá avaliar as questões, quanto à licença para utilização do conteúdo.

3.4.3 Consistência

No contexto do *Linked Data*, a consistência tem um significado similar ao aplicado em diferentes contextos. Uma base de conhecimento livre de erros lógicos e formais, ou seja, sem contradições ao que alega representar é considerada uma base consistente (ZAVERI et al, 2012).

É também definida como a ausência de contradições nos dados de um *dataset* (HOGAN et al, 2010) e livre de informações ou dados conflitantes (MENDES et al, 2012).

Zaveri et al (2013) definem consistência como o grau em que o formato e a estrutura de dados correspondem à informação devolvida anteriormente; assim, os autores aplicam a ocorrência em um caso real, no qual a quantidade de assentos do Estádio Drei Flüsse, de acordo com o Wikipédia, é de 20 mil, ao passo que em seu registro do DBpedia é de 20 somente.

3.4.4 Precisão Sintática e Semântica

As definições de precisão sintática e semântica para dados linkados consiste na mesma aplicada nos demais domínios. Quanto à precisão sintática, ocorre quando, mesmo que o valor atribuído não consista em seu equivalente ao do mundo real, ele consiste em um valor presente num conjunto de valores aceitáveis dentro de tal domínio. Ou seja, caso a instância represente um carro com o valor da cor sendo vermelho, mas no mundo real o carro é azul, o valor vermelho pode ser encarado como sintaticamente correto se a cor vermelha fizer parte do conjunto de cores aceitáveis para esse domínio (ZAVERI et al, 2012; FÜRBER; HEPP, 2011).

Fleming (2011) aplica precisão sintática para a utilização correta de vocabulários e sintaxe para descrição de documentos.

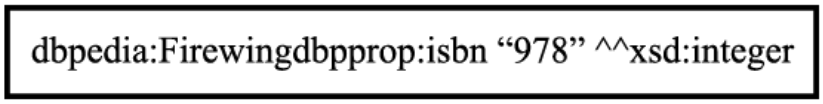
No caso da precisão semântica, consiste no quanto os conjuntos de dados representam sua entidade no mundo real. Ainda utilizando o exemplo do carro, considerando que o carro é azul, caso este seja representado na entidade como um carro da cor azul, e não vermelho, este consiste em um dado semanticamente preciso (FÜRBER; HEPP, 2011). Problemas de precisão semântica podem relacionar-se com etiquetas e classificações imprecisas, utilização errada de propriedades e valores (LEI et al, 2007).

3.4.5 Completude

De modo geral, a completude consiste no grau em que uma coleção os dados, ao se descrever seu conjunto de objetos do mundo real (BATINI et al, 2009; LI et al, 2011). No contexto do *Linked Data*, a completude divide-se em três categorias: completude de esquema, de população e de propriedade (ZAVERI et al, 2012; FÜRBER; HEPP, 2011). A completude de esquema refere-se ao grau em que os elementos da ontologia são representados, enquanto que a completude de população indica se todos os objetos referentes a uma instância do mundo real estão sendo representados. Já a completude de propriedade consiste em valores ausentes ou não, de acordo com uma propriedade ou coluna específica.

O exemplo da Figura 20 consiste num recurso do DBPedia sobre um livro infantil chamado *Firewing*, com um valor incompleto e incorreto sobre o ISBN (ACOSTA et al, 2013). É comum encontrar problemas de qualidade, a partir dos quais buscas resultam em valores incorretos, como apresentados na tripla acima.

Figura 20 – Problema de qualidade encontrado quanto ao ISBN incompleto de um recurso



Fonte: Acosta et al (2013)

Considerando os princípios para a publicação do conjunto de dados no diagrama LOD, a completude será avaliada também sob o panorama dos campos de descrição necessários ao publicar um *dataset* no Datahub, de modo que este possa ser posteriormente inserido no diagrama.

Assim, a completude será avaliada de duas maneiras: em nível de dados, conforme ilustrado pela Figura 18, na qual será realizada uma busca por dados incompletos, e em nível global, em que será verificada a disponibilidade das informações como um todo, do *dataset.*

3.4.6 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 pretenda representar seu conteúdo real. Também pode referir-se à idade média do dado em sua fonte.

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 (STVILIA et al., 2007).

Quanto a dimensões temporais no *Linked Data*, Li et al. (2011) afirmam que técnicas para fazer a relação dos dados ligados são falhas, visto que ignoram informações temporais e podem fracassar quanto à 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 esses pares são equivalentes ou não, o segundo passo consiste em agrupar os dados de acordo com o objetivo a que os dados no mesmo agrupamento se referem, a essa entidade no ambiente real; e registros em diferentes conjuntos referem-se a diferentes entidades.

Os problemas de qualidade relacionados a essa dimensão podem acontecer ao se tentar identificar a projeção das informações no decorrer do tempo. Um exemplo é abordado por Rula (2011), que aponta um caso no qual um determinado autor x pertencia à universidade "*The Open University*" no ano y1, e então se mudou para universidade "*University 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 a univesidade1 e no ano y2 à universidade2.

3.5 Metodologias para avaliação de qualidade no Linked Data

Conforme abordado nos trabalhos relacionados, não há um padrão das dimensões avaliadas em cada metodologia, cada uma foi desenvolvida para avaliar um conjunto de dados ou realizar alguma tarefa específica. Pode-se afirmar também que nem todas adotam dimensões exclusivas para o domínio do *Linked Data*, assim como não houve também um processo de definição de quais seriam as prioridades e dimensões aplicáveis ao seu contexto.

No contexto do *Linked Data* 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.,

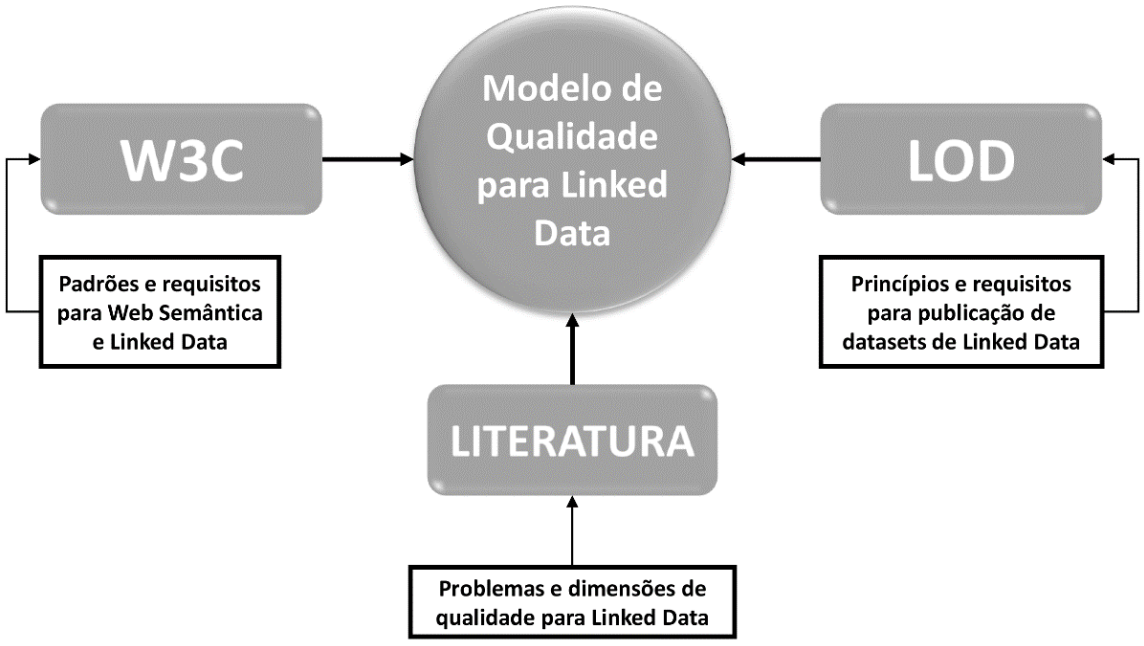
2012; GUÉRET et al., 2012), evidenciando, assim, a ausência de metodologias para realizar a avaliação de qualidade nos conjuntos de dados do *Linked Data*.

Desse modo, o capítulo a seguir abordará a metodologia para avaliação de qualidade proposta neste trabalho para o *Linked Data*.

4 MODELO E METODOLOGIA DE AVALIAÇÃO DE QUALIDADE DE DADOS NO CONTEXTO DO *LINKED DATA*

Este capítulo tem como objetivo apresentar a metodologia proposta e suas etapas, bem como apresentar cada dimensão de qualidade a ser avaliada, suas métricas e o cálculo de atribuição dos índices de qualidade. Para realizar a definição da metodologia foi necessário realizar primeiramente uma definição do que é qualidade no contexto do *Linked Data*. Tal processo resultou na definição de um modelo composto por 3 pilares, os quais forneceram subsídios para o estabelecimento dos requisitos, das dimensões e das métricas de avaliação, sendo eles: (1) literatura por meio da qual obtiveram-se informações sobre problemas e dimensões de qualidade para *Linked Data,* (2) W3C, que estabelece os padrões para o funcionamento tanto da Web Semântica, como do *Linked Data* e o (3) projeto *Linked Open Data*, o qual estabelece princípios de qualidade, reúne *datasets*, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios. A Figura 21 ilustra o modelo definido e seus componentes.

Figura 21 – Modelo proposto para qualidade de dados no contexto do *Linked Data*



Fonte: Elaborada pela autora.

Notou-se durante a análise e descrição bibliográfica das metodologias para avaliação da qualidade de dados em diferentes domínios, que cada dimensão e requisito de qualidade foram definidos de acordo com as necessidades do contexto no qual seriam aplicadas. Assim, o modelo foi estabelecido a fim de auxiliar a identificação dos requisitos e necessidades no domínio do

Linked Data. 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 Linked Data, deste modo foi realizada uma análise de como funciona o processo de inclusão de *datasets* no diagrama a fim de obter requisitos de qualidade.

O Quadro 9 apresenta os insumos e requisitos extraídos desses dois pilares, para a definição do modelo, dimensões e métricas do processo de avaliação. A descrição de cada metadado é realizada detalhadamente nos Quadros 4, 5 e 6 (p.47 e p.48).

Quadro 9 – Requisitos de qualidade para *datasets* de *Linked Data*

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 no-*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 *dataset*, como a combinação do nome e sobrenome do recurso.	

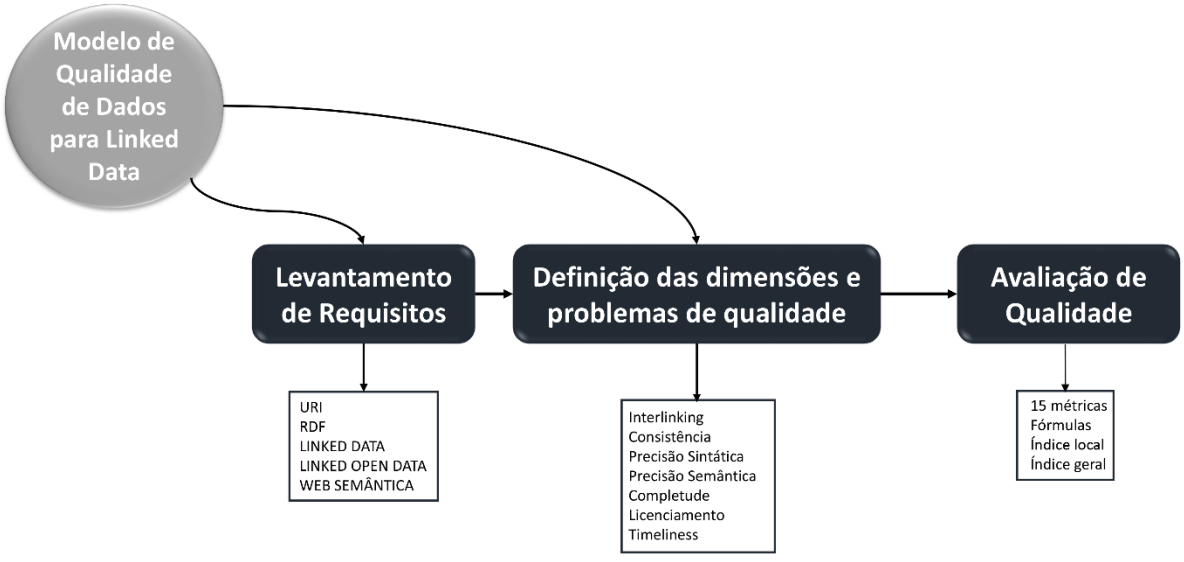
Fonte: Elaborado pela autora.

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 aplica-se a mesma dimensão para diferentes problemas de qualidade, dois fatores foram levados em consideração na definição das dimensões para a metodologia proposta: (1)

quais das dimensões abordadas relacionavam-se 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; BOUZEGHOUB, 2004; BATINI; SCANNAPIECO, 2016).

Diferente das metodologias descritas na literatura, a metodologia proposta visa avaliar, além das dimensões interdisciplinares, as que são aplicáveis especificamente no domínio do *Linked Data*. A partir do modelo obteve-se subsídios para cada etapa da metodologia proposta, dividida em três passos, sendo esses (1) levantamento de requisitos de qualidade para o *Linked Data*, (2) definição das dimensões e métricas e, por fim, (3) avaliação de qualidade, conforme apresentado na Figura 22. O modelo proposto complementa a metodologia, visto que os subsídios para as três etapas foram obtidos por meio dele, que auxiliou também a definir quais seriam os requisitos para a avaliação, os quais são descritos no Quadro 9.

Figura 22 – Metodologia de avaliação de qualidade de dados para *Linked Data*



Fonte: Elaborada pela autora.

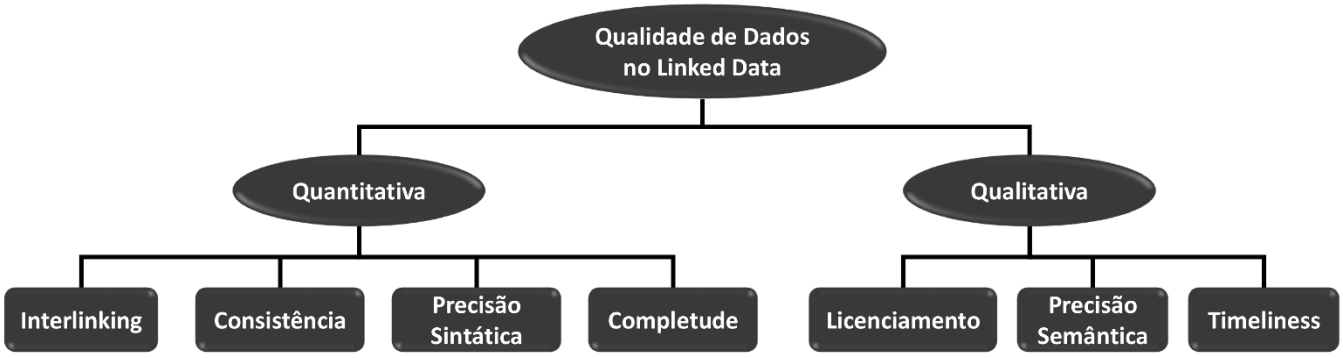
Para realizar o levantamento de requisitos foi realizada pesquisa bibliográfica quanto às diferentes metodologias de avaliação de dados no *Linked Data*, as dimensões aplicáveis para esse domínio, as diferentes métricas e os objetivos. Conforme abordado na seção de trabalhos

relacionados, pode-se concluir o seguinte, sobre as metodologias descritas na literatura: (1) não há um padrão quanto às dimensões específicas para o domínio de *Linked Data*; (2) não é possível identificar as prioridades de cada dimensão no processo de avaliação; (3) as regras e requisitos para atender qualidade no *Linked Data* foram elaboradas a partir de metodologias para avaliação de qualidade de dados dedicadas a diferentes domínios de aplicação.

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.

O processo de avaliação, que consiste no terceiro passo da metodologia, será realizado a partir de diferentes métricas descritas na literatura, onde os autores identificaram 69 métricas para avaliação de qualidade no *Linked Data,* que são aplicáveis em 18 dimensões diferentes. A Figura 23 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 23 – Classificação das dimensões de qualidade para a metodologia de avaliação



Fonte: Elaborada pela autora.

Acredita-se que as avaliações propostas podem ser conduzidas tanto no volume total dos dados, quanto em pequenas amostragens. Para realizar avaliação de grandes conjuntos de dados, bem como no volume total dos dados, no caso dos *datasets* de *Linked Data*, sugere-se o desenvolvimento de programas, visto que possibilitam que as avaliações sejam feitas de modo automático.

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, a fim de classificá-los como de baixa qualidade ou não, diferentemente de outras dimensões nas quais a qualidade pode ser quantificada objetiva e automaticamente (NAUMANN; ROLKER, 2000; PIPINO et al., 2002; LEE et al., 2002; AMICIS; BATINI,

2004; 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.

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.

Propõe-se 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 possua importância maior para o avaliador, sendo, pois, 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 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, propõe-se lidar com esse problema 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 aplica-se 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:

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

A seguir serão descritos os requisitos e os passos necessários para realizar a avaliação, de acordo com as dimensões propostas. Quando as métricas forem avaliadas por meio de propriedades ou classes dos conjuntos de dados, o OWL será adotado como exemplo da aplicação dos conceitos definidos nas métricas, visto que faz parte do conjunto de tecnologias recomendadas para Web Semântica, de acordo com o W3C.

4.1 Interlinking

Esta dimensão tem como objetivo avaliar problemas que possam afetar *links* RDF, conforme apresentados nas métricas 1 e 2. Ao final do cálculo de cada métrica a Fórmula 3 será utilizada para obter o resultado do índice local de *interlinking*. Propõe-se avaliação quantitativa para definir um índice de porcentagem de quantos *links* RDF não foram afetados com problemas de qualidade.

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

4.1.1 Métrica 1: Detectar *links* de boa qualidade

Essa avaliação será conduzida levando em consideração os princípios para identificar *links* de boa qualidade, sendo esses os que atendem aos seguintes requisitos:

(1) Não utilizar *namespaces* que fazem uso de conjuntos ou identificadores numéricos para os recursos. Quando utilizados, as URIs não podem ser referenciáveis, visto que apenas o domínio que as criou conseguiria identificá-las. Uma URI é referenciável quando é possível recuperar o recurso identificado por ela. O exemplo abaixo apresenta duas URIs, ambas sobre uma série de TV americana chamada *Suits*; a primeira é proveniente do *Internet Movie Database* (IMDb), uma base on-line de dados sobre filmes, séries, atores, etc. E a segunda provém de uma rede social chamada Banco de Séries, na qual os usuários podem avaliar filmes, séries, curta metragem, etc. Ambas utilizam identificadores aleatórios (tt1632701 no IMDb e 4355 no banco de séries).
   a. http://www.imdb.com/title/tt1632701/
   b. http://bancodeseries.com.br/index.php?serieid=4355&action=ss

(2) Abstrair detalhes de implementação: não utilizar nomes dos *hosts*, extensão de páginas, detalhes sobre o desenvolvimento na URI que não apresentem informações sobre o recurso. Tendo em vista as duas URIs consideradas no requisito anterior, este requisito é atendido na URI do IMDb, mas isso não acontece no Banco de Séries, o qual utilizou a extensão da página ". php", indicando que a página foi desenvolvida utilizando a linguagem de programação chamada PHP.

(3) Utilizar chaves naturais nas URIs: utilizar identificadores significativos para o domínio do *dataset*, como a combinação do nome e sobrenome do recurso (no caso de uma pequena organização), ISBN, etc. (HEATH; BIZER, 2011 SAUERMANN et al 2008; BERNERS-LEE, 1998). A URI abaixo mostra o cumprimento desse requisito, que apresenta informações sobre o pintor Vincent Van Gogh:
   a. http://dbpedia.org/resource/Vincent_van_Gogh

As seguintes verificações serão conduzidas a fim de cumprir com os requisitos listados: primeiramente, verificar as URIs a fim de analisar se números ou termos conhecidos do domínio foram utilizados, de acordo com os requisitos 1 e 3. A segunda consiste em verificar se não foram utilizados termos de desenvolvimento de páginas web para compor a URI, conforme o requisito 2. Tais verificações podem ser realizadas tanto de forma manual, como automática. A avaliação dessa dimensão acontecerá do seguinte modo: verificar as propriedades que têm como valor URIs, que podem ser identificadas no arquivo por meio da propriedade *rdf:about*, por exemplo, visto que ela é utilizada para descrever um recurso, que consequentemente é descrito por meio de uma URI. Em seguida, verificar se foi utilizado algum identificador numérico.

Então, no próximo passo, verificar se termos de implementação foram utilizados. A última verificação identifica se os termos utilizados na URI consistem em termos significativos para o domínio de recurso analisado.

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.

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

4.1.2. Métrica 2: Detectar a existência de *links* para fontes externas.

Considerando que a quantidade de *links* que apontam para *datasets* externos é encarada como uma informação adicional, é possível que o registro de alguns *datasets* no Datahub não informem esse valor. De acordo com a proposta do *Linked Data* o ideal é que *datasets* tenham pelo menos 50 *links* para fontes externas e essa métrica será disposta em duas etapas: (1) será realizada uma verificação na página do *dataset* no Datahub, se tais informações não forem disponibilizadas, (2) uma busca no arquivo dos dados será realizada visando encontrar o conteúdo sob a propriedade owl:sameAs.

Essa propriedade será utilizada porque cria um relacionamento entre indivíduos, visto que indica que duas referências URI dizem respeito ao mesmo indivíduo, ou seja, que eles têm a mesma identidade. É utilizada para declarar que o indivíduo do *dataset* X tem a mesma identidade que um indivíduo no *dataset* Y.

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 *dataset* 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.

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

Propõe-se a aplicação da fórmula, 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 *dataset* 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.

4.2 Licenciamento

Esta dimensão tem como objetivo verificar os requisitos e princípios de qualidade quanto à disponibilização e ao tipo da licença, bem como efetuar uma verificação para conferir se a licença correta para o domínio foi disponibilizada. Propõe-se avaliação qualitativa objetiva (métrica 1 e 3) e subjetiva (métrica 2).

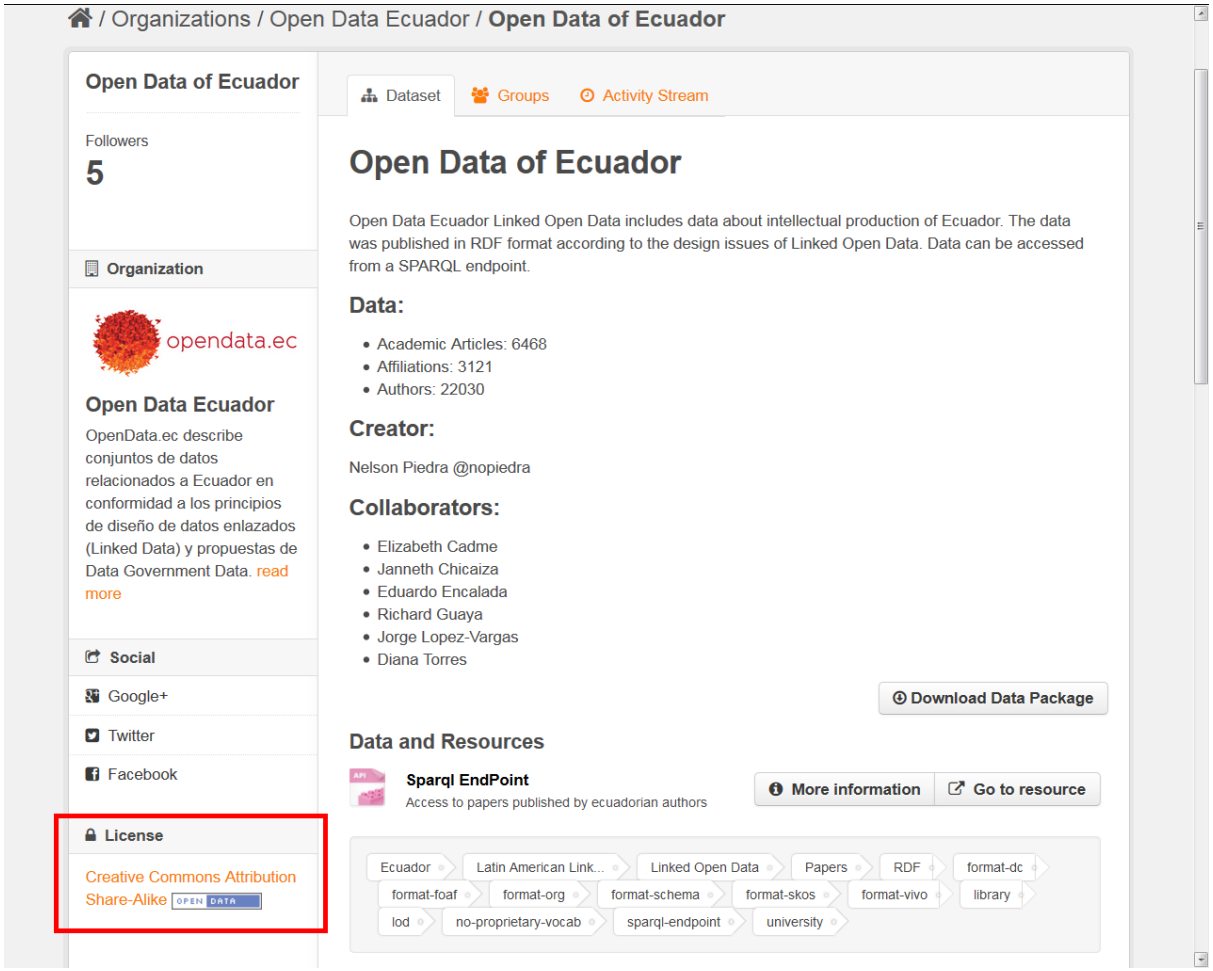
4.2.1. Métrica 1 - Detectar a existência de uma licença na documentação do *dataset*.

Um requisito para os dados abertos e para que o *dataset* seja inserido no diagrama do projeto LOD é a disponibilização da licença na página do Datahub, 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.

Essa verificação será realizada na página do *dataset* no Datahub, a qual está localizada no lado esquerdo da página, logo abaixo das redes sociais, conforme apresentado na Figura 22.

O *dataset* do exemplo apresentado na Figura 24 disponibiliza uma licença, mas é comum se deparar com *datasets* que não disponibilizaram ainda suas licenças.

Figura 24 – Localização da licença, quando disponibilizada pelos *datasets*



Embora seja comum encontrar *datasets* que não disponibilizaram ainda suas licenças, observa-se que o *dataset* do exemplo apresentado na Figura 24 disponibiliza uma licença '*Creative Commons Attribution Share-Alike*'. A avaliação do tipo da licença conduz a outra métrica proposta a seguir.

4.2.2. 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 *dataset*.

Essa verificação também é realizada na página do *dataset*; 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.

4.2.3. Métrica 3 – Disponibilizar uma licença aberta para o *dataset*

Constatou-se, na análise realizada para a definição do modelo de qualidade para o *Linked Data,* que a prática ideal para a publicação, de acordo com os princípios de qualidade do LOD, é disponibilizar uma licença aberta para os *datasets.* Assim, esta verificação terá como objetivo analisar se a licença disponibilizada é do tipo aberta.

4.3 Consistência

A consistência tem como objetivo verificar inconsistências nos conjuntos de dados analisados, ou seja, se o valor não apresenta contradições. Propõe-se avaliação quantitativa objetiva nas métricas, onde será gerado um índice da porcentagem dos dados analisados que não apresentam inconsistências.

4.3.1. 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:

- 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: refere-se a diferentes indivíduos; será analisado se esta propriedade foi utilizada junto com owl:sameAs, sobrepondo assim tal propriedade;

Ressalta-se que 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.

A avaliação será conduzida sob o arquivo do conjunto de dados; assim, é fundamental para a análise que esse arquivo tenha sido disponibilizado na página do *dataset* no Datahub. Nesse sentido, é importante que o avaliador tenha um pouco de conhecimento dos termos utilizados para a descrição de conteúdo, a fim de listar quais propriedades são passíveis de gerar inconsistências e contradições quando utilizadas juntas. Após buscar pelos termos, o avaliador irá analisar se as classes e propriedades foram utilizadas de modo satisfatório.

Será atribuído um índice por meio do cálculo apresentado na Fórmula 6, onde: TTi consiste na somatória de classes/propriedades da amostragem, da qual é subtraída a somatória de Ti, que representa a quantidade de itens inconsistentes. O resultado é então dividido pelo total de itens TTi, que ao final é multiplicado por 100 para encontrar a porcentagem final.

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

Conforme ressaltado no início do capítulo, a utilização de propriedades e classes OWL auxilia a aplicação da avaliação da dimensão. No caso da utilização de diferentes e/ou vocabulários próprios, a avaliação consiste em detectar propriedades e classes do vocabulário do *dataset* que se contradizem e não poderiam ser utilizadas na mesma declaração.

4.3.2. Métrica 2 - Verificar o tipo de dados permitido para a propriedade.

Cada campo permite um tipo diferente de dados; esta análise consiste em verificar se o tipo do dado inserido condiz com o tipo permitido para o campo. Exemplo: no caso de uma propriedade datadeNascimento ser do tipo data, e uma *string* ter sido utilizada para preencher o campo, isso implicará em problemas de consistência. Para a avaliação seria necessário:

(1) Uma lista com o tipo de dados de cada propriedade, tais informações são disponibilizadas na documentação do vocabulário;
(2) Uma lista das propriedades utilizadas para a descrição dos dados;
(3) A comparação dos campos e do tipo de dados.

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.

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

4.4 Precisão Sintática

Esta dimensão tem como objetivo verificar se o valor da propriedade analisada condiz com o seu equivalente no mundo real. Propõe-se uma avaliação quantitativa objetiva a qual resultará na porcentagem do quanto das propriedades analisadas se encontram livres de problemas de consistência.

4.4.1. Métrica 1 - Detectar o uso de regras sintáticas.

Essa verificação é conduzida para analisar o tipo de caracteres ou modelo de valores permitidos. Assim como no processo de avaliação de consistência, é necessário que o avaliador tenha um conhecimento das propriedades que possuem valores de quantidade e formatos pré-definidos. 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 ISSN que consiste em um código internacional para individualizar o título de publicações periódicas, ISBN que identifica numericamente livros de acordo com título, autor, país, editora e edição etc.

Assim, essa verificação será realizada do seguinte modo:

(1) Será necessário um conjunto com as máscaras dos dados referentes ao domínio analisado;

(2) Conjunto das propriedades que utilizam tais tipos de dados;

(3) Será realizada, então, a comparação dos dados utilizados com a máscara referente ao campo.

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.

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

4.5 Precisão Semântica

Esta dimensão tem como objetivo identificar problemas semânticos como etiquetas, atribuições e classificações imprecisas, bem como utilização errada de propriedades e valores. Propõe-se uma avaliação qualitativa subjetiva para as métricas desta dimensão.

4.5.1. Métrica 1 - Detectar a utilização de propriedades inexistentes.

Um problema possível de acontecer consiste na declaração de classes e propriedades não existentes no conjunto do vocabulário/linguagem utilizada para descrição dos dados, com exceção, ressalte-se, da criação de classes e propriedades específicas de um vocabulário criado exclusivamente para determinados *datasets*, o que implicaria na disponibilização e especificação na web.

Um meio de conduzir uma avaliação genérica seria por meio de uma lista das propriedades e das classes OWL, e comparar com as classes e propriedades de fato utilizadas. Mas, visto que existem inúmeros vocabulários em uso nos *datasets* publicados, pode ser conduzida do seguinte modo:

(1) Uma lista com as propriedades utilizadas para a descrição dos dados, o que implica em uma busca da documentação dos termos do vocabulário usado;
(2) Uma busca no arquivo, realizando a exclusão dos termos utilizados;
(3) Uma análise dos termos restantes a fim de identificar se não foram referenciados ou se são propriedades inexistentes que foram utilizadas incorretamente.

4.5.2. 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.

Os passos para avaliar este problema são os mesmos realizados na métrica 1, porém visando encontrar a utilização de atributos como relacionamentos e de relacionamentos como valores literais:

(1) Verificação será conduzida no arquivo dos dados, na página do *dataset* no Datahub;
(2) Verificar e excluir da análise as vezes em que os atributos e os relacionamentos foram utilizados de modo correto;
(3) Analisar os atributos e os relacionamentos restantes e identificar se foram utilizados incorretamente.

Conforme abordado nos tópicos anteriores, quando não utilizado OWL, será necessário aplicar o mesmo conceito (utilização de atributos em relacionamentos, e relacionamentos como valores literais) nos vocabulários utilizados no *dataset* em questão.

4.6 Completude

A completude consiste em uma dimensão que avalia o quão completo um conjunto de dados, esquema, informação etc., está. Propõe-se aqui uma 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 *dataset* no Datahub, 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 *dataset* é de 0%.

4.6.1. 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. Para realizar a verificação será necessário o arquivo da ontologia, que é disponibilizado na página do *dataset*. Será então verificado quantas classes e propriedades foram representadas.

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.

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

4.6.2. 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.

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

4.6.3. 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.

$$COMm2 = \left(\frac{\sum Tcp}{TC}\right) * 100 \quad (12)$$

4.6.4. Métrica 4 - Quantidade de metadados necessários a serem disponibilizados no Datahub

Conforme abordado no Capítulo 2, o Datahub é uma plataforma que disponibiliza diversas funcionalidades do CKAN, uma ferramenta de gerenciamento e publicação de grandes coleções de dados. No Datahub é possível criar um registro de *datasets* publicados, criar e gerenciar grupos de *datasets* e obter atualizações. Assim, todos os *datasets* inseridos no diagrama LOD possuem um registro no Datahub, e, quando inseridos no diagrama, subentende-se que foram disponibilizadas informações descritivas sobre eles, divididas em três níveis, conforme os Quadros 4, 5 e 6 (p. 47 e p. 48).

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.

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

Uma exceção acontece no caso das *tags* *lodcloud.nolinks*, *lodcloud.unconnected*, *lodcloud.needinfo*, *lodcloud.needsfixing*, que são obrigatórias, conforme especificado no Quadro 6 (p. 48), porém não serão utilizadas, caso o problema retratado por elas não for aplicável ao *dataset* em avaliação. Desse modo, tais propriedades não serão incluídas no cálculo, quando essa exceção acontecer.

4.6.5. Métrica 5 - Completude considerando atributos prioritários

Diferente da verificação conduzida na métrica 4, essa avaliação tem como objetivo analisar o quão completa está a tripla em relação à descrição do recurso analisado. Para conduzir a verificação são necessárias as seguintes informações: o que pode ser considerado como um registro completo? Quais dados descritivos compõem tal registro? 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).

No contexto de dados exclusivamente de publicações constatou-se que existem apenas algumas diferenças no modo como cada instituição descreve a mesma categoria de publicação. Tais diferenças podem ser observadas nos Quadros 11, 12 e 13, que apresentam os metadados utilizados por cada instituição para descrever as publicações.

Quadro 10 – Metadados para descrição de artigos publicados em eventos científicos

Artigo de Evento				
UNESP	USP	IEEE	BIELEFELD	SOUTHAMPTON
Autor	Autor	Autor	Autor	Autor
URI	URI	URI	URI	URI
Resumo	Resumo	Resumo	Resumo	Resumo
Título	Título	Título	Título	Título
Tipo	Tipo	Tipo	Tipo	Tipo
Assunto	Assunto	Assunto	Assunto	Fonte
Fonte	Fonte	Data	Fonte	Editor
Idioma	Idioma	Nome do compêndio	Idioma	Data
Relacionamento	Relacionamento	Relacionamento	Data	
Direitos	Direitos	Proceedings	Direitos	
Contribuidor	Editor			
Data de acesso	Contribuidor			
Data de disponibilização	Data de acesso			
Data de publicação	Data de disponibilização			
Identificador (DOI)	Data de publicação			
Citação	Detentor dos direitos			
Extensão	Patrocínio			
Licença	Remissiva do patrocinador			
Instituição	Cidade de publicação			
Afiliação	País de publicação			
Identificador WOS				
Identificador Scopus				

Fonte: Elaborado pela autora.

Quadro 11 – Metadados para descrição de teses e dissertações

Teses e dissertações			
UNESP	USP	BIELEFELD	SOUTHAMPTON
Autor	Autor	Autor	Autor
URI	URI	URI	URI
Assunto	Assunto	Assunto	Assunto
Título	Título	Título	Título
Tipo	Tipo	Tipo	Tipo
Resumo	Resumo	Data	Resumo
Data de acesso	Data	Editor	Data
Data de disponibilização	Editor	Idioma	
Data de publicação		Direitos	
Citação			
Extensão (páginas)			
Idioma ISO			
Fonte			
Direitos de acesso			
Editor			
Patrocínio			
Instituição contribuidora			
Programa de graduação			
Área de conhecimento			
Área de pesquisa			
Campus			

Fonte: Elaborado pela autora.

Quadro 12 – Metadados para descrição de artigos publicados em revistas científicas

Artigos de revista			
UNESP	USP	BIELEFELD	SOUTHAMPTON
Autor	Autor	Autor	Autor
Título	Título	Título	Título
URI	URI	URI	URI
Data de acesso	Data de acesso	Idioma	Resumo
Data de disponibilização	Data de disponibilização	Parte de	Tipo
Data de publicação	Data de publicação	Fonte	Data
Citação	Resumo	Tipo	Editor
Resumo	Idioma	Diretos	ISSN
Extensão	Parte de	Data	
Idioma	Fonte	Editor	
Parte de	Assunto		
Fonte	Tipo		
Assunto	Contribuidor		
Instituição	Direitos		
Afiliação	Editor		
Direitos	Cidade de publicação		
ISSN	País de publicação		
	Remissiva do patrocinador		
	ISSN		
	Volume		
	Número		
	Página inicial		
	Última página		

Fonte: Elaborado pela autora.

Ao analisar a forma como cada instituição descreve o conhecimento científico produzido observa-se que ela varia; nota-se 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, propõe-se 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.

Visto que o modelo de avaliação proposto tem como objetivo avaliar os *datasets,* tanto antes como depois de publicados, pode-se afirmar que, quando aplicada, a avaliação será realizada por dois tipos de usuários diferentes: o produtor e o consumidor dos dados. Assim, pode-se dizer que ambos possuem experiência, contexto e objetivos diferentes, o que influencia a importância dos dados para o contexto no qual serão utilizados.

Em vista disso, propõe-se a utilização da fórmula para completude descrita por Botega (2016), na qual a prioridade dos atributos descritivos pode ser atribuída pelo avaliador, visto que tanto o produtor quanto o consumidor do *dataset* 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.

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

A análise será realizada do seguinte modo: o avaliador deve definir a amostragem de dados para aplicar o cálculo, que deve ser aplicado nas descrições de recursos, ou seja, em triplas individuais. Quando aplicado em mais de uma tripla, deve ser calculada a média dos resultados que será o resultado final do valor da métrica. Será necessário o arquivo dos dados, disponibilizado na página do Datahub, sendo que a não disponibilização desse arquivo inviabiliza a aplicação dessa métrica.

4.7 Avaliação Temporal (Timeliness e Volatilidade)

Essa dimensão tem como objetivo avaliar dados temporais sobre o *dataset*; propõe-se uma avaliação qualitativa para as métricas definidas a seguir:

4.7.1 Métrica 1 - Verificar quão atuais os dados são

Essa métrica tem como objetivo obter duas informações temporais sobre o *dataset*, sendo que tais informações são classificadas em duas categorias:

- *Timeliness*: verifica se o dado é atual para o contexto no qual será utilizado;
- 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 *dataset*; 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 *dataset* 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 *dataset* esteja desatualizado. Consequentemente, quando o valor for maior que 0, significa que o *dataset* já foi atualizado.

Este capítulo considerou as duas contribuições deste trabalho: (1) o modelo proposto composto por três pilares (literatura, LOD, W3C) o qual auxilia a definição do que pode ser considerado como qualidade de dados no contexto do *Linked Data*; e (2) a metodologia de avaliação de qualidade para o *Linked Data* composta por sete dimensões, 16 métricas e 14 fórmulas. . A aplicação da avaliação proposta será descrita no capítulo a seguir.

5 PROVA DE CONCEITO

Neste capítulo será abordada a aplicação dos métodos de avaliação propostos. A avaliação proposta tem como objetivo detectar problemas no contexto do *Linked Data,* tanto em *datasets* a serem publicados, como em avaliações posteriores, nos que já foram disponibilizados. Visto que o *Linked Data* consiste em um conjunto de práticas para expor, compartilhar e conectar dados, informação e conhecimento na Web Semântica, uma avaliação de qualidade pode ser considerada um fator de extrema relevância para os objetivos da Web Semântica e a validade dos dados e informações compartilhados.

Desse modo, a avaliação foi conduzida em um *dataset* de *Linked Data*, que, por estar inserido no diagrama LOD, acredita-se atender aos requisitos de qualidade descritos no capítulo 4 (Quadro 9, p. 72). Todos os requisitos e princípios da Web Semântica e *Linked Data* foram levados em consideração para realizar a avaliação de qualidade, nas dimensões propostas.

Visto que OWL faz parte das tecnologias recomendadas para Web Semântica pelo W3C, suas classes e propriedades serão utilizadas, quando aplicáveis, nos exemplos para a avaliação proposta, nas dimensões de qualidade.

Para conduzir a avaliação foi utilizado um *dataset* no domínio proposto (publicações), chamado IEEE Papers (RKBExplorer), criado por Hugh Glaser, que contém dados sobre artigos do IEEE (*Institute of Electrical and Electronics Engineers*), provenientes do RBKExplorer, um repositório semântico que contém e publica dados de acordo com as práticas do *Linked Data*. Pode-se dizer que o *dataset* analisado possui um grande volume de dados: mais de 11.894 *links* para fontes externas e mais de 91.564 triplas.

Um item fundamental para realizar a avaliação foi a disponibilização do conjunto dos dados na página do *dataset* no Datahub. Visto que a avaliação pode ser aplicada tanto no conjunto completo dos dados como em pequenas amostragens individuais, o *dataset* analisado possui um grande volume de dados e a avaliação foi conduzida manualmente; foi selecionado um conjunto controlado, composto de 20 triplas, para a aplicação dos métodos de avaliação de acordo com as sete dimensões de qualidade: *interlinking*, licenciamento, consistência, precisão sintática, precisão semântica, completude e avaliação temporal.

O índice de qualidade será dividido entre local, quando os dados analisados estão de acordo com a dimensão, sendo esta de 0 (não cumpriu) e 100 (cumpriu totalmente), e global, que dispõe um valor relativo ao quanto as dimensões mensuráveis, quantitativamente juntas, atenderam aos requisitos de qualidade. Cada dimensão quantitativa corresponde a 25% do índice global. Desse modo, será calculado o valor do índice local, cujo resultado deverá se situar

no intervalo de 0 a 25; em seguida será efetuada a soma dos valores de cada índice local resultando, então, no índice global.

5.1 Interlinking

5.1.1 Aplicação da métrica 1 - *Links* de boa qualidade

Foi verificado se as URIs da amostragem analisada cumpriam com os seguintes requisitos: (1) não utilizar *namespaces* impossíveis de controlar, ou seja, conjuntos ou identificadores numéricos para os recursos. Quando utilizadas, as URIs não podem ser referenciáveis, visto que apenas o domínio que as criou conseguiria identificá-las. (2) não utilizar detalhes de implementação para compor as URIs, tais como: nome dos *hosts*, extensão das páginas, detalhes sobre o desenvolvimento. E (3) utilizar identificadores significativos para o domínio do *dataset*, como, por exemplo, a combinação do nome e sobrenome do recurso.

A amostragem analisada possuía 76 recursos, dentre os quais 20 descreviam diferentes artigos publicados, 20 remetiam a conferências a partir das quais cada artigo foi publicado e 36 referiam-se aos autores dos 20 artigos descritos.

Visto que os três requisitos compõem a métrica, quando falho em ao menos um deles, a URI é considerada com problema de *Interlinking*. Quanto aos requisitos, todas as URIs analisadas mostraram-se falhas, ou seja, cumpriram o requisito (2) mas não cumpriram os requisitos (1) e (3). Assim, essa métrica atingiu 0% de seu percentual.

No decorrer da análise notou-se que foi utilizado o mesmo padrão na criação das URIs, conforme apresentado na Figura 25. Acredita-se, então, que esse padrão deve acontecer em todo o conteúdo dos dados disponibilizados.

Figura 25 – Padrão de dados descritivos disponibilizados sobre os recursos analisados na amostragem


```
<rdf:Description rdf:about="http://ieee.rkbexplorer.com/id/publication-00008092">
  <akt:has-date><support:Calendar-Date rdf:about='http://www.aktors.org/ontology/date#1988'>
    <support:year-of>1988</support:year-of>
    <support:has-pretty-name>1988</support:has-pretty-name>
  </support:Calendar-Date></akt:has-date>
  <akt:has-title>The algebra of security</akt:has-title>
  <rdf:type rdf:resource="&akt;Proceedings-Paper-Reference" />
  <akt:has-author><akt:Person rdf:about="http://ieee.rkbexplorer.com/id/person-1822
6fd4028f6d8db2e752e12b69545a-56de0b49bcec41c8ec27dd3c20939a03">
    <akt:full-name>J. McLean</akt:full-name>
  </akt:Person></akt:has-author>
  <extension:has-abstract>A general framework is developed in which various mandatory access control security
models that allow changes in security levels can be formalized. These models form a Boolean algebra.
The framework is expanded to include models that allow &lt;e1&gt;n&lt;/e1&gt;-person rules necessary
for discretionary access controls in an industrial security setting.
The resulting framework is a distributive lattice</extension:has-abstract>
 <akt:paper-in-proceedings><akt:Conference-Proceedings-Reference rdf:about="http://ieee.rkbexplorer.com/id/proceedings-427" >
    <akt:has-title>IEEE Symposium on Security and Privacy, 1988</akt:has-title>
    <akt:has-web-address>http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=427</akt:has-web-address>
    <akt:has-date>
      <support:Calendar-Date rdf:about='http://www.aktors.org/ontology/date#1988'>
        <support:year-of>1988</support:year-of>
        <support:has-pretty-name>1988</support:has-pretty-name>
      </support:Calendar-Date>
    </akt:has-date>
  </akt:Conference-Proceedings-Reference></akt:paper-in-proceedings>
  <akt:has-web-address>http://ieeexplore.ieee.org/iel5/201/427/00008092.pdf</akt:has-web-address>
  <iai:has-ieee-keyword>Boolean algebra</iai:has-ieee-keyword>
  <iai:has-ieee-keyword>security of data</iai:has-ieee-keyword>
  <iai:has-ieee-keyword>systems analysis</iai:has-ieee-keyword>
 </rdf:Description>
```


Fonte: Elaborada pela autora.

Desse modo, das 76 URIs presentes, nenhuma cumpriu com os requisitos para criação/utilização de *links* de boa qualidade, resultando em um índice de 0%. O resultado aplicado na fórmula proposta (Fórmula 5) é apresentado na Fórmula 15.

$$m1 = \frac{(76 - 76)}{76} = 0 \quad (15)$$

5.1.2 Aplicação da métrica 2 - Utilizar *links* para fontes externas

Um dos requisitos envolvidos para a avaliação de qualidade dos *datasets* consiste em inserir no mínimo 50 *links* para fontes externas. A verificação pode ser realizada de dois modos: o primeiro, ao conferir a página do *dataset* no Datahub; porém, visto que consistiria em uma informação adicional, é possível que não seja informada.

Nesse caso, outro modo de identificar, é por meio da propriedade owl:sameAs. Porém, a verificação da utilização da propriedade owl:sameAs tem de ser realizada no arquivo RDF disponibilizado no Datahub.

Foi constatado que esse *dataset* possui *links* para 20 conjuntos diferentes, todos

incluídos no diagrama LOD, sendo eles: ACM, Citeseer, DBLP, ePrints, KISTI (Korean Institute of Science Technology and Information), OAI, LAAS, Newcastle, Dotac, IBM, Pisa, Rae2001, Wiki, Southampton, Resex, ULM, Roma, RISKS, Curriculum e NSF.

A quantidade de *links* para cada uma das fontes foi disponibilizada na Datahub, onde 2.949 *links* apontam para ACM, 1.182 para o Citeseer, 5.857 para o DBLP, 643 para o ePrints, 516 para o KISTI, 417 para o OAI, 97 para o LAAS, 73 para o Newcastle, 50 para o Dotac, 29 para o *dataset* da IBM, 18 para Pisa, 17 para Rae2001, 9 para Wik, 7 para Southampton, 6 para Resex, 5 para ULM, 3 para Roma, 3 para RISKIN, 2 para Curriculum e 1 para NSF.

Assim, esse *dataset* totalizou 11.894 links para fontes externas, superando por uma grande diferença a quantidade de *links* requeridos (50), não sendo assim necessário realizar a aplicação da Fórmula proposta, conforme definido no capítulo 4. Desse modo, a M2 totalizou os 50%, o índice local de *interlinking* foi então de 50%, conforme apresentado na aplicação da Fórmula 2 na Fórmula 16.

$$I = \frac{0 + 100}{2} = 50 \quad (16)$$

5.2 Licenciamento

5.2.1 Aplicação da métrica 1 - Existência de uma licença

A indicação de uma licença é realizada de duas maneiras: por declarar o tipo da licença, informação que fica do lado esquerdo da página, abaixo da indicação das redes sociais, e a segunda maneira é por verificar a disponibilização da *tag licence-metadata*, a qual indica se foram disponibilizados metadados sobre a licença, assim como *no-licence-metadata* quando não disponibilizados. Durante a verificação constatou-se que não foi disponibilizada a licença, uma vez que o campo da indicação estava em branco, e a *tag no-licence-metadata* foi inserida no conjunto de *tags* do *dataset*.

5.2.2 Aplicação da métrica 2 - Tipo certo da licença

Essa métrica tem como objetivo disponibilizar a licença para que o avaliador distinga se ela é a licença correta para os dados disponibilizados. Visto que nenhuma licença foi disponibilizada, a avaliação dessa métrica é inviabilizada.

5.2.3 Métrica 3 - Disponibilizar uma licença aberta para o *dataset*

Foi verificado se a licença disponibilizada pelo *dataset* é aberta. Visto que nenhuma licença foi relacionada na página do *dataset*, consequentemente, essa métrica não foi aplicável.

5.3 Consistência

5.3.1 Aplicação da métrica 1 - Verificar se os dados estão de acordo com a especificação da ontologia.

O *dataset* do IEEE utiliza um vocabulário chamado *AKT Reference Ontology*. Considerando que os exemplos de validação foram sugeridos na linguagem OWL, foi necessário analisar os termos do vocabulário utilizado visando definir quais são as classes e propriedades que não poderiam ser utilizadas juntas, visto que iriam se contradizer, apresentando assim inconsistências.

Um obstáculo encontrado durante a busca da documentação da ontologia foi que esse vocabulário não é mais suportado pelos seus criadores, apesar de ainda ser utilizado em 24 *datasets*, conforme revela o resultado de uma busca por vocabulário no Datahub.

A avaliação foi conduzida do seguinte modo: primeiro foi baixado o arquivo da ontologia no Datahub, em seguida o arquivo foi examinado, no cabeçalho do arquivo RDF, onde os vocabulários são inseridos, a fim de verificar o nome e o endereço para verificação. E então foi realizada uma busca na Web pela documentação do vocabulário utilizado.

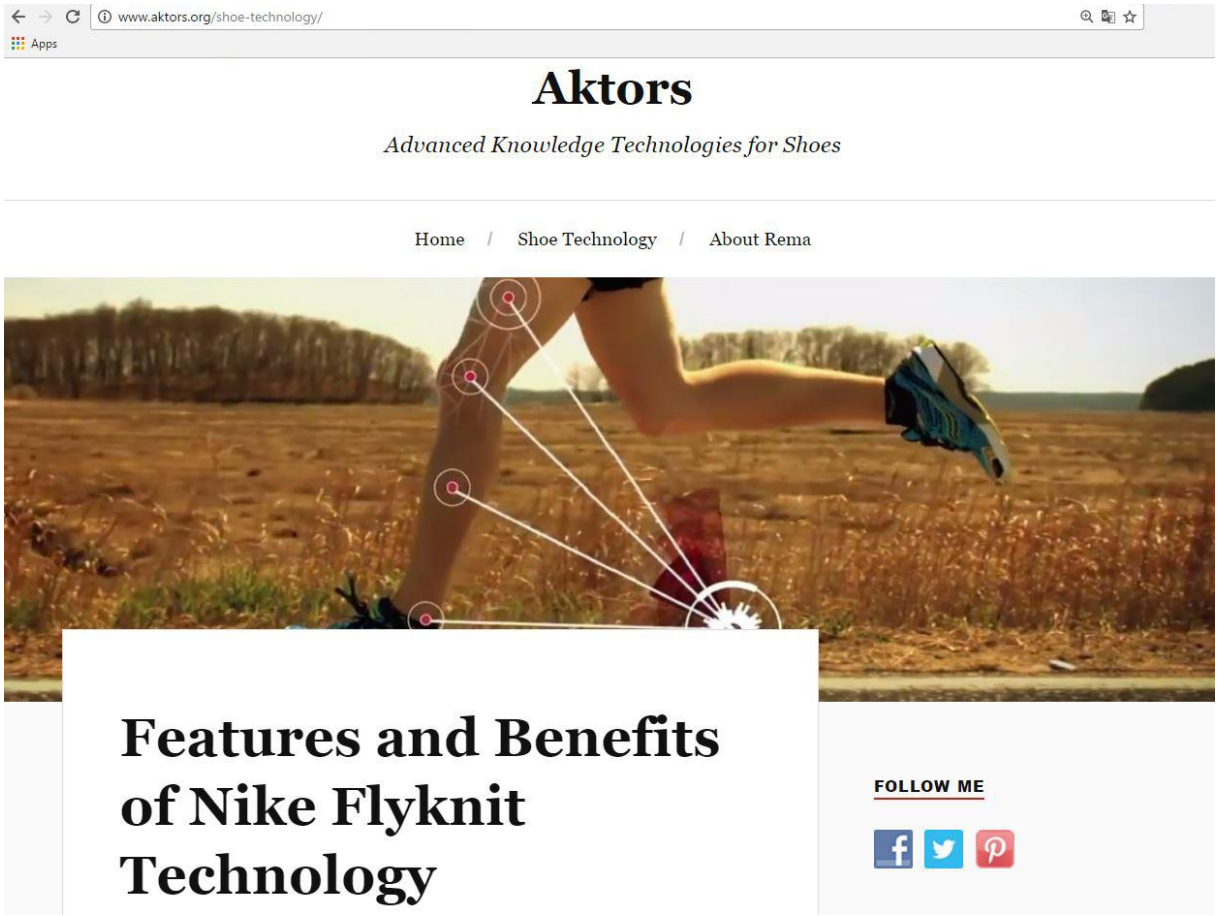
Constatou-se que foram utilizados os vocabulários chamados AKT, IAI, Support e Extension, todos com o mesmo endereço, o que indicava ser do mesmo mantedor. Foi possível encontrar algumas informações sobre o vocabulário AKT no *Linked Open Vocabularies* (LOV), um repositório que disponibiliza informações sobre classes e propriedades, diversos vocabulários, bem como informações sobre os criadores, mantedores etc., porém, não foram encontradas informações sobre os outros vocabulários (IAI, Support e Extension).

Outro obstáculo correspondeu ao fato de que o endereço que supostamente direcionava para informações sobre tais vocabulários é, atualmente, um blog intitulado Aktors - *Advanced Knowledge Technologies for Shoes*, o site de uma blogueira sobre tecnologia para calçados, conforme apresentado na Figura 26.

Desse modo, a avaliação foi conduzida por meio dos termos obtidos no LOV, apenas sob os termos do vocabulário AKT, que totalizam 274, sendo 152 classes e 122 propriedades.

Foi conduzida uma análise sobre cada termo do vocabulário, tanto nas propriedades quanto nas classes, a fim de verificar prováveis inconsistências em relação às classes e propriedades que foram de fato utilizadas no *dataset*. Considerando que foram disponibilizados exemplos de inconsistências por meio do OWL, uma aplicação para o vocabulário akt seria a utilização da classe akt:Journal, visto que a classe akt:Conference-Proceedings-Reference foi utilizada para descrever os artigos publicados. Porém, tal inconsistência não foi detectada durante a avaliação de qualidade.

Figura 26 – Suposto site da documentação do vocabulário de descrição dos dados no *dataset* avaliado



Desse modo, essa métrica atingiu 100% de seu índice de qualidade, porém, visto que representa metade do valor local, seu valor inteiro corresponde a 50%, conforme apresentado na aplicação da Fórmula 7, na Fórmula 17.

$$CONm1 = \left(\frac{(274) - (0)}{274}\right) * 100 \quad = 100 \ (17)$$

5.3.2 Aplicação da métrica 2 - Tipo de dados permitido

Essa verificação teve como objetivo conferir se o tipo de dado inserido condizia com o que foi especificado para a propriedade, na documentação do vocabulário. Em vista dos obstáculos citados acima, não foram encontradas as informações necessárias para conduzir essa verificação. Assim, essa métrica foi invalidada nesse caso excepcional, visto que não houve as informações necessárias, e avaliar negativamente poderia afetar de modo impreciso o índice de qualidade.

Assim, nesse caso em especial, o índice da métrica 1 passa a valer 100%, visto que foi a única avaliada de fato.

5.4 Precisão Sintática

5.4.1 Aplicação da métrica 1 - Detectar o uso de regras sintáticas.

Após uma verificação das propriedades do vocabulário, constatou-se a presença de propriedades que têm um modelo específico, como o ISSN, mas essa propriedade não foi utilizada nas triplas analisadas, visto que a dimensão não pôde ser avaliada em razão da não disponibilização dos atributos. Desse modo, nesse caso específico, tal dimensão foi excluída do índice geral de qualidade.

5.5 Precisão Semântica

5.5.1 Aplicação da métrica 1 - Propriedades não definidas

Ao analisar a amostragem selecionada, constatou-se que a propriedade akt:paper-in-proceedings, que apareceu em todas as triplas, não consta na documentação do vocabulário disponibilizada no LOV (Figura 27).

Figura 27 – Classe não definida na documentação utilizada para descrição do recurso

```
<rdf:Description rdf:about="http://ieee.rkbexplorer.com/id/publication-00008092">
  <akt:has-date><support:Calendar-Date rdf:about='http://www.aktors.org/ontology/date#1988'>
     <support:year-of>1988</support:year-of>
     <support:has-pretty-name>1988</support:has-pretty-name>
   </support:Calendar-Date></akt:has-date>
  <akt:has-title>The algebra of security</akt:has-title>
  <rdf:type rdf:resource="&akt;Proceedings-Paper-Reference" />
  <akt:has-author><akt:Person rdf:about="http://ieee.rkbexplorer.com/id/person-1822
6fd4028f6d8db2e752e12b69545a-56de0b49bcec41c8ec27dd3c20939a03">
     <akt:full-name>J. McLean</akt:full-name>
   </akt:Person></akt:has-author>
  <extension:has-abstract>A general framework is developed in which various mandatory access control security
models that allow changes in security levels can be formalized. These models form a Boolean algebra.
The framework is expanded to include models that allow &lt;e1&gt;n&lt;/e1&gt;-person rules necessary
for discretionary access controls in an industrial security setting.
The resulting framework is a distributive lattice</extension:has-abstract>
  <akt:paper-in-proceedings><akt:Conference-Proceedings-Reference rdf:about="http://ieee.rkbexplorer.com/id/proceedings-427" >
     <akt:has-title>IEEE Symposium on Security and Privacy, 1988</akt:has-title>
     <akt:has-web-address>http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=427</akt:has-web-address>
     <akt:has-date>
       <support:Calendar-Date rdf:about='http://www.aktors.org/ontology/date#1988'>
         <support:year-of>1988</support:year-of>
         <support:has-pretty-name>1988</support:has-pretty-name>
       </support:Calendar-Date>
     </akt:has-date>
   </akt:Conference-Proceedings-Reference></akt:paper-in-proceedings>
  <akt:has-web-address>http://ieeexplore.ieee.org/iel5/201/427/00008092.pdf</akt:has-web-address>
  <iai:has-ieee-keyword>Boolean algebra</iai:has-ieee-keyword>
  <iai:has-ieee-keyword>security of data</iai:has-ieee-keyword>
  <iai:has-ieee-keyword>systems analysis</iai:has-ieee-keyword>
</rdf:Description>
```

Fonte: Elaborada pela autora.

A análise foi realizada por meio de uma verificação de cada termo utilizado na amostragem, sendo que para cada termo uma busca nos termos da documentação foi realizada, tanto nas classes quanto nas propriedades, e esta foi a única propriedade não definida. Não foi detectada nenhuma classe indefinida na amostragem.

5.5.2 Aplicação da métrica 2 - Uso incorreto das propriedades

Essa verificação objetiva avaliar se alguma classe foi utilizada como uma propriedade. Foi verificado cada propriedade e classe das triplas analisadas e constatou-se que não ocorreu este tipo de problema.

Assim, todas as classes utilizadas eram de fato classes, bem como todas as propriedades utilizadas, excluindo a propriedade akt:paper-in-proceedings, conforme apontado na M1, foram empregadas corretamente.

5.6 Completude

Cada métrica equivale a 20% do índice de completude local. Os cálculos das métricas 1, 2 e 3 foram realizados do seguinte modo: primeiro foi realizado uma verificação em cada tripla das classes e propriedades utilizadas, excluindo as repetições.

Constatou-se que dos 274 termos do vocabulário, sendo 152 classes e 122 propriedades, foram utilizadas 7 propriedades e 2 classes. A propriedade akt:paper-in-proceedings não foi considerada nas avaliações, visto que não consta na documentação do vocabulário.

E assim o resultado, de acordo com cada métrica, foi o seguinte:

- Aplicação da métrica 1: completude de esquema verificou quantas classes e propriedades foram utilizadas e o percentual representado; como foram 9 os termos utilizados, e considerando que o AKT possui 274 termos, o resultado foi de 3,28%. Ou seja, apenas 3,28% dos termos foram utilizados, conforme a aplicação da Fórmula 10, na Fórmula 18;

$$COMm1 = \left(\frac{9}{274}\right) * 100 = 3{,}28 \quad (18)$$

- Aplicação da métrica 2: completude de propriedade verificou o quanto das propriedades foram utilizadas, que foram 7 das 122, assim 5,73% dos termos foram utilizados, conforme a aplicação da Fórmula 11, na Fórmula 19:

$$COMm2 = \left(\frac{7}{122}\right) * 100 = 5{,}73 \quad (19)$$

- Aplicação da métrica 3: completude de população verifica o quanto das classes foram utilizadas; durante a análise constatou-se que foram utilizadas 2 classes (akt:Person e akt:Conference-Proceedings-Reference) das 152, resultando em 1,31%, conforme a aplicação da Fórmula 12, na Fórmula 20:

$$COMm3 = \left(\frac{2}{152}\right) * 100 = 1{,}31 \quad (20)$$

5.6.1 Aplicação da métrica 4 - Uso incorreto das propriedades

Essa métrica tem como objetivo verificar se os atributos descritivos necessários no registro do *dataset* no Datahub foram disponibilizados pelo criador. O resultado da verificação é apresentado nos Quadros 13, 14 e 15, que apresentam as informações e os recursos que foram declarados.

Quadro 13 – Verificação da disponibilização dos dados descritivos de nível 1

Campo	**Foi utilizado?**
Nome	Sim
Título	Sim
URL	*Sim*
Autor	Sim
E-mail	Não
Tag	Sim

Fonte: Elaborado pela autora.

Quadro 14 – Verificação da disponibilização dos dados descritivos de nível 2

Campo	**Foi informado?**
Tag	Sim
Link para um exemplo RDF	Sim
URL para SPARQL *endpoint*	Não
URL de *download* para cada arquivo RDF	Sim
URL para uma página com a lista de *downloads*	Sim
Triplas (Informação adicional)	Sim
Links (Informação adicional)	Sim
SPARQL *endpoint* (Informação adicional)	*Sim*

Fonte: Elaborado pela autora.

Quadro 15 – Verificação da disponibilização dos dados descritivos de nível 3

Campo	**Foi informado?**
Versão	Sim
Notas	Não
Licença	Não
Abreviação (Informação adicional)	Sim
Link da licença (Informação adicional)	*Não*
Arquivo voiD	Sim
XML Sitemap	Sim

Campo	**Foi informado?**
Namespace (Informação adicional)	Sim
RDF Schema	Sim
Vocabulário	Não
Inserir uma das *tags*: no-proprietary-vocab, deref-vocab, no-deref-vocab	Sim
Inserir as *tags*: vocab-mappings, no-vocab-mappings	Sim
Inserir as *tags*: provenance-metadata, no-provenance-metadata	Sim
Iserir as *tags*: license-metadata, no-license-metadata	Sim
Inserir as *tags*: published-by-producer, published-by-third-party	Sim
Inserir a *tag*: limited-sparql-endpoint	Não aplicável
Inserir a *tag*: format-<prefix>	Não
Inserir *tag*: lodcloud.nolinks	Não aplicável
Inserir a *tag*: lodcloud.unconnected	Não aplicável
Utilizar a *tag*: lodcloud.needinfo	Não aplicável
Utilizar a tag: lodcloud.needsfixing	Não aplicável

Fonte: Elaborado pela autora.

Dos 35 atributos descritivos, 5 foram excluídos do cálculo por não serem aplicáveis ao *dataset* em avaliação; do restante, 6 consistem em atributos adicionais; e dos 24 atributos restantes, 6 não foram informados.

Assim, a primeira etapa da avaliação revelou que o *dataset* avaliado utilizou 18 dos elementos descritivos requeridos, incluindo os dois atributos prioritários (título e assunto). O cálculo foi conduzido a fim de identificar a porcentagem equivalente aos 18 atributos, sendo 24 o valor máximo. Desse modo, o *dataset* mostrou-se 66,66% completo, de acordo com a métrica 4, conforme a aplicação da Fórmula 13, na Fórmula 21:

$$COMm4 = \left(\frac{(18) - (2)}{(30) - 6}\right) * 100 = 66{,}66 \quad (13)$$

5.6.2 Aplicação da métrica 5 - Completude considerando atributos prioritários

O primeiro passo para avaliar essa métrica consistiu no fato de que o avaliador deve

estabelecer qual seria o modelo de registro descritivo completo no contexto de sua avaliação. Pode-se dizer que, do ponto de vista do criador do *dataset*, o modelo de registro completo de um recurso corresponde ao que foi disponibilizado; porém, para um avaliador externo, que pretende utilizar os dados, pode ser que o registro não esteja completo.

Em vista dessa subjetividade, incentiva-se, aqui, que o avaliador defina o que, para ele, significa um registro completo e quais atributos tal registro deve conter.

De acordo com a avaliação proposta, na qual o avaliador define quais são os atributos descritivos necessários para a amostragem avaliada, para esta avaliação, o registro considerado completo com base nos atributos definido pelo avaliador, deve conter: título, autor, resumo, assunto, data, nome do evento, ISSN, volume e direitos de acesso.

Nesse contexto de prova de conceito, a prioridade dos atributos descritivos foi atribuída com base no estudo Delphi, conduzido por Santos (2013) para identificar quais são os principais elementos para a descrição de um recurso informacional. Nesse estudo, Santos (2013) aplicou um questionário com 3 bacharéis em Biblioteconomia que não atuavam em bibliotecas; 8 bibliotecários atuando em bibliotecas universitárias, mas que não trabalhavam diretamente com catalogação; e 15 bibliotecários catalogadores, 2 atuando em bibliotecas escolares e 13 em bibliotecas universitárias.

De acordo com os resultados do estudo, os elementos descritivos mais indicados por todos os participantes da pesquisa foram: título e assunto da publicação. Desse modo, esses dois atributos foram definidos como prioritários, entre os 9 atributos descritivos definidos para a avaliação das triplas.

A avaliação foi aplicada nas 20 triplas da amostragem selecionada e, para o cálculo, foi utilizada a Fórmula 14 (p. 91). O cálculo foi realizado do seguinte modo: primeiro acontece a multiplicação do peso (2 = atributo prioritário e 1 = atributos não-prioritários).

Conforme explicitado, são 9 atributos (título, autor, resumo, assunto, data, nome do evento, ISSN, volume, direitos de acesso), sendo dois deles prioritários, de peso igual a 2 (título e assunto), e sete não-prioritários, de peso igual a 1. Desses 9, 3 de peso 1 estão ausentes, sendo eles: ISSN, volume e direitos de acesso.

A somatória do peso multiplicado pela presença foi igual a 8 e a somatória do peso, conforme indicado na fórmula, foi igual a 11. Em seguida, o resultado da divisão de 8 por 11 é multiplicado por 0.9, e então somado 0.1. O resultado é então multiplicado por 100 para chegar no resultado final de completude do registro do recurso, que foi de 75,45%, conforme a aplicação da Fórmula 14, na Fórmula 22:

$$C = 1\left[\left(\frac{8}{11} * 0{,}9\right) + 0{,}1\right] = 75{,}45\ (22)$$

O mesmo processo foi realizado em cada uma das 20 triplas da amostragem, o que resultou sempre em 75,45%, visto que os mesmos atributos descritivos foram utilizados para descrever cada artigo. A média dos resultados dessa métrica, que corresponde então ao seu valor local, foi de 75,45%.

Para o resultado final, local, de completude, M1 = 3,28, M2 = 5,73, M3 = 1,31, M4= 75% e M5= 75,45, a média dos valores, correspondente a completude local, foi de 32,152%.

5.7 Avaliação Temporal (Timeliness e Volatilidade)

Essa métrica é avaliada de acordo com duas categorias: *timeliness*, resulta na idade do *dataset* a partir de sua última atualização, e volatilidade, que resulta na quantidade de vezes que o *dataset* foi atualizado.

5.7.1. Aplicação da métrica 1 - Verificar o quão atual os dados são

Quanto à *timeliness*, a última atualização do *dataset* ocorreu no dia 30 de julho de 2016, e a avaliação foi realizada no dia 26 de fevereiro de 2017, resultando então em 241 dias corridos.

Quanto à volatilidade, observou-se uma única informação disponibilizada sobre atualização, deduzindo-se, assim, que o *dataset* foi atualizado uma vez.

5.8 Índice de qualidade geral

Conforme abordado no capítulo 4, o índice global é composto pelo índice local das dimensões quantitativas, sendo: precisão sintática, completude, consistência e interlinking. Visto que o índice global é composto por quatro dimensões, cada uma é responsável por 25% do valor total. Desse modo, o índice local, que consiste em um valor de 0 a 100%, é reajustado de modo a prover seu valor equivalente dentro do limite onde 25% corresponde a 100%.

A partir dos resultados, foi aplicado o cálculo da fórmula 1, para o índice geral, conforme apresentado na fórmula 23, após a aplicação da fórmula 2 para obter o índice de qualidade local de completude (apresentado na fórmula 24) e da fórmula 3, para obtenção do índice de *interlinking* (apresentado na fórmula 25).

$$Ig = \frac{50 + 100 + 30{,}486}{3} \quad (23)$$

$$Com = \frac{3{,}28 + 5{,}73 + 1{,}31 + 75{,}45}{5} = 30{,}486 (24)$$

$$I = \frac{0 + 100}{2} = 50 \ (25)$$

Assim, o resultado do índice local de cada dimensão, sem a equivalência, foi o seguinte:

- Precisão sintática = 0.
- Consistência = 100%, visto que a M1 atingiu seu valor total e a M2 não foi aplicável;
- Interlinking = 50%
- Completude = 30,486%.

O Quadro 17 apresenta um cartão proposto para auxiliar o processo de avaliação de acordo com cada dimensão, onde QT corresponde à avaliação quantitativa e QL à qualitativa. Propõe-se a inserção dos índices de locais de qualidade, para as dimensões quantitativas, para o cálculo do índice de qualidade geral.

Quadro 16 – Quadro 10 preenchido com os índices de qualidade de cada dimensão

Dimensão	Métricas	Tipo da avaliação	Pontuação	Índice Local
Interlinking	M1	QT	0%	50%
	M2	QT	50%	
Licenciamento	M1	QL	Não se aplica	X
	M2	QL	Não se aplica	
	M3	QL	Não se aplica	
Consistência	M1	QT	100%	100%
	M2	QT	Invalidada	
Precisão Sintática	M1	QT	Invalidada	Invalidada
Precisão Semântica	M1	QL	Não se aplica	X
	M2	QL	Não se aplica	
Completude	M1	QT	3,28%	30,486%
	M2	QT	5,73%	
	M3	QT	1,31%	
	M4	QT	66,66%	
	M5	QT	75,45%	
Timeliness	M1	QL	Não se aplica	X
	M2	QL	Não se aplica	
Índice de qualidade geral				60

Fonte: Elaborado pela autora.

6 CONSIDERAÇÕES FINAIS

A Web Semântica foi idealizada com a finalidade de possibilitar uma descrição para promover a identificação de dados, de modo que agentes computacionais possam manipular tais dados de modo significativo a favor dos usuários. Assim como todos ambientes que utilizam dados estão sujeitos a ser afetados por problemas de qualidade, a literatura aponta uma gama de problemas de qualidade tratados por diversas metodologias para avaliação de qualidade no contexto do *Linked Data*.

Pode-se dizer que um fato consensual na comunidade de qualidade de dados é o de que a definição e a avaliação de qualidade estão altamente relacionadas com um domínio específico, seus requisitos e os dos usuários que realizarão suas tarefas a partir de tais dados. Assim, constatou-se que, das diferentes metodologias de avaliação de qualidade para o *Linked Data*, não se encontrou em nenhuma delas embasamento e justificativa, com base nos requisitos do domínio, sobre o porquê as dimensões abrangidas por tais metodologias foram definidas. Identificou-se, também, que, assim como na aplicação de qualidade de dados de modo geral, não há um conjunto pré-estabelecido sobre quais dimensões utilizar.

Assim, este trabalho propôs-se a definir um modelo do que é realmente qualidade de dados para o domínio de *Linked Data*, modelo o qual foi constituído sobre três pilares, sendo eles: (1) literatura, por meio da qual obtiveram-se informações sobre problemas e dimensões de qualidade para *Linked Data*, (2) W3C, que estabelece os padrões para o funcionamento tanto da Web Semântica, como do *Linked Data* e o (3) projeto *Linked Open Data* (LOD), o qual estabelece princípios de qualidade, reúne *datasets*, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios.

Por meio do modelo definido obtiveram-se insumos para construir uma metodologia composta por três etapas da metodologia de avaliação proposta: (1) levantamento de requisitos de qualidade para o *Linked Data*, (2) definição das dimensões e métricas e, por fim, (3) avaliação de qualidade. Quanto aos requisitos, definiu-se 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 *datasets*. Na segunda etapa da metodologia foram definidas as seguintes dimensões de qualidade: *interlinking*, consistência, completude, licenciamento, avaliação temporal, precisão sintática e semântica. E, 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 *dataset.*

A metodologia foi desenvolvida a fim de avaliar não somente *datasets* já publicados, mas também para auxiliar a verificação de problemas de qualidade antes da publicação de um *dataset*, especificamente no segmento de dados sobre publicações no *Linked Data*. No caso de uma avaliação prévia do *dataset*, ao invés de verificar no Datahub, o avaliador pode utilizar os Quadros 4, 5 e 6 para a verificar os metadados necessários durante o processo de inserção do *dataset* na plataforma.

Ao realizar o processo de avaliação podem-se constatar 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 problema experienciado durante o processo de avaliação do *dataset* avaliado, foi a utilização de um vocabulário não mais suportado pelos criadores, o que dificultou e inviabilizou a obtenção de informações para conduzir a análise. O que pode, também, fora do contexto da análise, dificultar a interoperabilidade e integração dos dados do *dataset,* o que iria contra os propósitos da Web Semântica;
- Um dos meios de obtenção de requisitos e princípios de qualidade foi LOD, projeto que estabelece princípios de qualidade, reúne *datasets*, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios. A visibilidade dos *datasets* que atendem aos requisitos de qualidade é feita por meio da disponibilização de um diagrama composto pelos *datasets* que atendem a tais requisitos, o qual é dividido em 9 categorias diferentes. Porém, ao conduzir a avaliação de qualidade constatou-se que o *dataset* analisado, o qual está inserido no diagrama não cumpre com um dos requisitos de qualidade do LOD: não disponibiliza uma licença. Desse modo, pode-se concluir que podem existir outros *datasets* que não atendem a determinados requisitos de qualidade, e, por sua vez, não deveriam estar inclusos no diagrama.

A qualidade tem sido muito estudada na área da Ciência da Computação, entretanto, entende-se que hoje em dia a publicação de dados é algo assumido pela Ciência da Informação; sendo assim, é imprescindível que a publicação de dados seja realizada com boa qualidade e a metodologia proposta contribui, nesse contexto, para melhorar a qualidade de dados, principalmente na Ciência da Informação. Acredita-se que a metodologia pode ser utilizada

para detectar problemas de qualidade nos processos de armazenamento, recuperação e manipulação dos dados.

Como trabalhos futuros propõe-se a investigação de problemas de qualidade em outros *datasets* 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 *datasets* 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.

REFERÊNCIAS

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

AGRE, J.; VASSILIOU, M. S.; KRAMER, C. **Science and technology issues relating to data quality in C2 systems**. Alexandria, VA: Institute for Defense Analyses, 2011.

AMICIS, F.; BATINI, C. A methodology for data quality assessment on financial data. **Studies in Communication Sciences**, v. 4, n. 2, p. 115-137, 2004.

BATINI, C. et al. A framework and a methodology for data quality assessment and monitoring. In: INTERNATIONAL CONFERENCE ON INFORMATION QUALITY, 12., 2007, Cambridge, MA. **Proceedings...** Cambridge: MIT, 2007. p. 333-346.

BATINI, C. et al. A comprehensive data quality methodology for web and structured data. **International Journal of Innovative Computing and Applications**, v. 1, n. 3, p. 205-218, 2008.

BATINI, C. et al. Methodologies for data quality assessment and improvement. **ACM Computing Surveys (CSUR)**, v. 41, n. 3, 2009. 52 p.

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

BATINI, C.; SCANNAPIECO, M. **Data quality dimensions**: data and information quality. Switzerland: Springer, 2016.

BERNERS-LEE, T. Cool URIs don't change. 1998. Disponível em: <https://www.w3.org/Provider/Style/URI>. Acesso em: 29 jun. 2016.

BERNERS-LEE, T. Information management: a proposal. CERN, 1989. Disponível em: <https://www.w3.org/History/1989/proposal.html>. Acesso em: 29 jun. 2016.

BERNERS-LEE, T. Linked Data: design issues. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 29 jun. 2016.

BERNERS-LEE, T. et al. World-wide web: the information universe. **Electronic Networking: Research, Applications and Policy**, v. 2, n. 1, p. 52-58, 1992.

BERNERS-LEE, T.; FIELDING, R.; MASINTER, L. **Uniform resource identifier (URI): generic syntax**. [S.l.]: RFC Editor, 2005.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 28-37, 2001.

BERRUETA, D.; PHIPPS, J. (Ed.). Best practice recipes for publishing RDF vocabularies. W3C Working draft. 2008. Disponível em: <https://www.w3.org/TR/swbp-vocab-pub/>. Acesso em: 5 jun. 2016.

BIZER, C.; CYGANIAK, R. 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.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: the story so far. **International Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1-22, 2009.

BIZER, C. et al. Linked data on the web (LDOW2008). In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 17., 2008, Beijing. **Proceedings...** New York: ACM, 2008. p. 1265-1266.

BIZER, C.; JENTZSCH, A.; CYGANIAK, R. State of the LOD Cloud 2011. 2011. Disponível em: <http://lod-cloud.net/state/state_2011/>. Acesso em: 5 jun. 2016.

BORKO, H. Information science: what is it? **American Documentation**, v. 19, n. 1, p. 3-5, 1968.

BOTEGA, L. C. **Modelo de fusão dirigido por humanos e ciente de qualidade de informação**. 2016. 247 f. Tese (Doutorado em Ciência da Computação)-Centro de Ciências Exatas de Tecnologia, Universidade Federal de São Carlos, São Carlos, 2016.

BOTEGA, L. C. et al. Methodology for data and information quality assessment in the context of emergency situational awareness. **Universal Access in the Information Society**, p. 1-14, 2016.

BOUZEGHOUB, M. A framework for analysis of data freshness. In: INTERNATIONAL WORKSHOP ON INFORMATION QUALITY IN INFORMATION SYSTEMS, 2004, Paris. **Proceedings...** New York: ACM, 2004. p. 59-67.

BOVEE, M.; SRIVASTAVA, R. P.; MAK, B. A conceptual framework and belief-function approach to assessing overall information quality. **International Journal of Intelligent Systems**, v. 18, n. 1, p. 51-74, 2003.

BRICKLEY, D.; MILLER, L. **FOAF vocabulary specification 0.99**. Namespace Document 14 January 2014 - Paddington Edition. 2014. Disponível em: <http://xmlns.com/foaf/spec/>. Acesso em: 5 jun. 2016.

CONG, G. et al. Improving data quality: consistency and accuracy. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 33., Vienna. **Proceedings...** New York: ACM, 2007. p. 315-326.

DECKER, S. et al. The semantic web: the roles of XML and RDF. **IEEE Internet Computing**, v. 4, n. 5, p. 63-73, 2000.

FERNEDA, E. **Recuperação de informação**: análise sobre a contribuição da Ciência de Computação para a Ciência da Informação. 2003. 137 f. Tese (Doutorado em Ciências da Comunicação)-Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2003.

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

FOX, C.; LEVITIN, A.; REDMAN, T. The notion of data and its quality dimensions. **Information Processing & Management**, v. 30, n. 1, p. 9-19, 1994.

FÜRBER, C.; HEPP, M. SWIQA: a semantic web information quality assessment framework. In: EUROPEAN CONFERENCE ON INFORMATION SYSTEMS, 18., 2011, Roksilde. **Proceedings...** [S.l.]: AISeL, 2011. p. 19.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, n. 2, p. 199-220, 1993.

GUÉRET, C. et al. Assessing linked data mappings using network measures. In: EXTENDED SEMANTIC WEB CONFERENCE, 9., 2012, Heraklion. **Proceedings...** Heidelberg: Springer, 2012. p. 87-102.

HASSO PLATTNER INSTITUT. Data Hub LOD validator. 2016. Disponível em: <http://validator.lod-cloud.net/validate.php>. Acesso em: 5 jun. 2016.

HEATH, T.; BIZER, C. **Linked Data**: evolving the web into a global data space. New York: Morgan & Claypool, 2011.

HOGAN, A. et al. Weaving the pedantic web. In: INTERNATIONAL WORKSHOP ON LINKED DATA ON THE WEB, 3., 2010, Raleigh. **Proceedings...** Aachen: CEUR-WS, 2010.

HOGAN, A. et al. An empirical survey of linked data conformance. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 14, p. 14-44, 2012.

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

JARKE, M. et al. Architecture and quality in data warehouses: an extended repository approach. **Information Systems**, v. 24, n. 3, p. 229-253, 1999.

JURAN, J. M.; GRYNA, F. M.; BINGHAM JR, R. S. (Ed.). **Quality control handbook**. [S.l.]: McGraw-Hill, 1974.

KOIVUNEN, M.; MILLER, E. W3C semantic web activity. In: HYVÖNEN, E. (Ed.) **Semantic web kick-off in Finland**: vision, technologies, research, and applications. Helsinki: HIIT Publications, 2001. p. 27-44.

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

KOUDAS, N.; SARAWAGI, S.; SRIVASTAVA, D. Record linkage: similarity measures and algorithms. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2006, Chicago. **Proceedings...** New York: ACM, 2006. p. 802-803.

LEE, Y. W. et al. AIMQ: a methodology for information quality assessment. **Information & Management**, v. 40, n. 2, p. 133-146, 2002.

LEI, Y.; UREN, V.; MOTTA, E. A framework for evaluating semantic metadata. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE CAPTURE, 4. 2007, Whistler. **Proceedings...** New York: ACM, 2007. p. 135-142.

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

LINKING Open Data cloud diagram. 2017. Disponível em: <http://lod-cloud.net/>. Acesso em: 5 fev. 2017.

MARCONDES, C. H.; SAYÃO, L. F. Introdução: repositórios institucionais e livre acesso. In: SAYÃO, L. F. et al. (Ed.). **Implantação e gestão de repositórios institucionais**: políticas, memória, livre acesso e preservação. Salvador: EDUFBA, 2009. p. 9-21.

MCGILVRAY, D. **Executing data quality projects**: ten steps to quality data and trusted information. Burlington, MA: Morgan Kaufmann, 2008.

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

MENDES, P. N. (Coord.). D2.1: Conceptual model and best practices for high-quality metadata publishing. Technical report. [S.l.]: PlanetData, 2012.

MILES, A.; BECHHOFER, S. SKOS Simple Knowledge Organization System Reference. W3C recommendation, v. 18, 2009.

NAUMANN, F. **Quality-driven query answering for integrated information systems.** Berlin: Springer Science & Business Media, 2002.

NAUMANN, F.; ROLKER, C. Assessment methods for information quality criteria. In: INTERNATIONAL CONFERENCE ON INFORMATION QUALITY, 5., 2000, Cambridge. **Proceedings...** Cambridge, MA: MIT, 2000. p. 148-162.

PÉREZ, J.; ARENAS, M.; GUTIERREZ, C. Semantics and complexity of SPARQL. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 5., 2006, Athens. **Proceedings...** Berlin: Springer, 2006.p. 30-43.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Communications of the ACM**, v. 45, n. 4, p. 211-218, 2002.

ROUSSEY, C. et al. An introduction to ontologies and ontology engineering. In: FALQUET, G. et al. **Ontologies in urban development projects.** London: Springer, 2011. p. 9-38.

RULA, A. DC proposal: towards linked data assessment and linking temporal facts. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 10., 2011, Bonn. **Proceedings...** Berlin: Springer, 2011. p. 341-348.

RULA, A.; PALMONARI, M.; MAURINO, A. Capturing the age of linked open data: Towards a dataset-independent framework. In: INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING, 6., Palermo. **Proceedings...** [S.l.]: IEEE, 2012. p. 218-225.

RULA, A.; ZAVERI, A. Methodology for Assessment of Linked Data Quality. In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 10., 2014, Leipzig. **Proceedings...** Leipzig: LDQ, 2014.

SANTAREM SEGUNDO, J. E. Web Semântica: introdução a recuperação de dados usando SPOARQL. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais...** Belo Horizonte: ECI/UFMG, 2014. p. 3863-3882.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Tecnologias da Web Semântica aplicadas a organização do conhecimento: padrão SKOS para construção e uso de vocabulários controlados descentralizados. In: José Augusto Chaves Guimarães; Vera Dodebei. (Org.). **Organização do Conhecimento e Diversidade Cultural**. 1.ed.Marília: Fundepe, v. 3, p. 224-233, 2015.

SANTOS, Plácida Leopoldina Ventura Amorim da Costa. Catalogação, formas de representação e construções mentais. **Tendências da Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 6, n. 1, jan./jun. 2013.

SAUERMANN, L.; CYGANIAK, R.; VÖLKEL, M. Cool URIs for the semantic web. Saarbrücken: Deutsches Forschungszentrum für Künstliche Intelligenz, 2007.

SCANNAPIECO, M.; MISSIER, P.; BATINI, C. Data quality at a glance. **Datenbank-Spektrum**, v. 14, p. 6-14, 2005.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. State of the LOD Cloud 2014. 2014. Disponível em: <http://lod-cloud.net/state/state_2014/>. Acesso em: 5 jun. 2016.

STRONG, D. M.; LEE, Y. W.; WANG, R. Y. Data quality in context. **Communications of the ACM**, v. 40, n. 5, p. 103-110, 1997.

STVILIA, B. et al. A framework for information quality assessment. **Journal of the American Society for Information Science and Technology**, v. 58, n. 12, p. 1720-1733, 2007.

VEREGIN, H. Data quality parameters. **Geographical Information Systems**, v. 1, p. 177-189, 1999.

WAND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological foundations. **Communications of the ACM**, v. 39, n. 11, p. 86-95, 1996.

WANG, R. Y. A product perspective on total data quality management. **Communications of the ACM**, v. 41, n. 2, p. 58-65, 1998.

WANG, R. Y.; REDDY, M. P.; KON, H. B. Toward quality data: An attribute-based approach. **Decision Support Systems**, v. 13, n. 3, p. 349-372, 1995.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumer. **Journal of Management Information Systems,** v. 12, n. 4, p. 5-33, 1996.

WINKLER, W. Methods for evaluating and creating data quality. **Information Systems,** v. 29, n. 7, p. 531-550, 2004.

WORLD WIDE WEB CONSORTIUM. Guidelines for collecting metadata on linked datasets in the datahub.io data catalog. 2011. Disponível em: <https://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation>. Acesso em: 25 jun. 2016.

WORLD WIDE WEB CONSORTIUM. Guidelines for collecting metadata on linked datasets in the datahub.io data catalog. 2014. Disponível em: <https://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation>. Acesso em: 25 jun. 2016.

WORLD WIDE WEB CONSORTIUM. Layer Cake. Disponível em: <http://www.w3.org/2007/03/layerCake.png>. Acesso em: 25 jun. 2016.

WORLD WIDE WEB CONSORTIUM. Ontologies: Vocabularies. 2015. Disponível em: <https://www.w3.org/standards/semanticweb/ontology>. Acesso em: 25 jun. 2016.

WORLD WIDE WEB CONSORTIUM. RDF 1.1 Primer. 2014. Disponível em: <http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/>. Acesso em: 25 jun. 2016.

WORLD WIDE WEB CONSORTIUM. RDF: Resource Description Framework. 2014. Disponível em: <http://www.w3.org/RDF/>. Acesso em: 25 jun. 2016.

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

ZAVERI, A. et al. Quality assessment for linked data: a survey. **Semantic Web**, v. 7, n. 1, p. 63-93, 2015.