UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DEFESA DE DISSERTAÇÃO DE MESTRADO

UMA ABORDAGEM SEMIAUTOMÁTICA DIRIGIDA A MÉTRICAS PARA AVALIAÇÃO DA QUALIDADE DE *datasets* CONECTADOS

MESTRANDA
DANILA FEITOSA DE CARVALHO OLIVEIRA

ORIENTADOR
IG IBERT BITTENCOURT SANTANA

MACEIÓ, AL
FEVEREIRO - 2017

DANILA FEITOSA DE CARVALHO OLIVEIRA

UMA ABORDAGEM SEMIAUTOMÁTICA DIRIGIDA A MÉTRICAS PARA AVALIAÇÃO DA QUALIDADE DE *datasets* CONECTADOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Prof. Dr. Ig Ibert Bittencourt Santana

MACEIÓ, AL
FEVEREIRO - 2017

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Bibliotecário Responsável: Valter dos Santos Andrade

S232e Oliveira, Danila Feitosa de Carvalho.
Uma abordagem semiautomática dirigida a métricas para avaliação da qualidade de datasets conectados / Danila Feitosa de Carvalho Oliveira. – 2017.
134 f.: il.

Orientador: Ig Ibert Bittencourt Santana.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Programa de Pós-Graduação em Informática. Maceió, 2017.

Bibliografia: f. 107-111.
Apêndice: f. 112-134.

1. Dados conectados. 2. Qualidade de dado - *Datasets*. 3. Métrica. I. Título.

CDU: 004.89

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Membros da Comissão Julgadora da Dissertação de Mestrado de Danila Feitosa de Carvalho Oliveira, intitulada: *"Uma Abordagem Semiautomática Dirigida a Métricas para Avaliação da Qualidade de Datasets Conectados"*, apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 07 de fevereiro de 2017, às 09h00min, na Sala de Reuniões do Instituto de Computação da UFAL.

COMISSÃO JULGADORA

Prof. Dr. Ig Ibert Bittencourt Santana Pinto
UFAL – Instituto de Computação
Orientador

Prof. Dr. Rafael de Amorim Silva
UFAL – Instituto de Computação
Examinador

Prof.ª Dr.ª Bernardette Farias Lóscio
UFPE – Universidade Federal de Pernambuco
Examinador

A Deus por ter aberto essa porta e ter me guiado nesta longa caminhada.

A minha família que, com grande apoio, sempre me manteve perseverante nesta caminhada, principalmente nos momentos mais difíceis.

Ao meu noivo, pelo companheirismo de ter estado ao meu lado, vivenciando todos os momentos e me orientando com sabedoria e paciência.

Ao meu Orientador Prof. Dr. Ig Ibert Bittencourt Santana por, além de ter trabalhado junto comigo neste mestrado, ter acreditado em mim durante os momentos de desafios.

Agradecimentos

Em primeiro lugar, agradeço a Deus por ter me dado forças e ter cuidado de mim fazendo com que esta jornada fosse cumprida.

Agradeço a minha família pelo apoio e confiança que a mim foram depositadas, especialmente ao meu Pai José Ribamar de Oliveira, minha Mãe Maria Carleusa Feitosa de Carvalho Oliveira e ao Meu Irmão Dárcio Feitosa de Carvalho Oliveira que sempre estiveram presentes em todos os momentos desta jornada e por terem dedicado um pouco de suas vidas para me proporcionar a oportunidade de estudar e ter um futuro melhor.

Ao meu Noivo Ítalo Marcos Marinho de Carvalho Beltrão que com paciência e sabedoria esteve comigo e soube me guiar em todos os momentos, principalmente nos mais difíceis. Agradeço imensamente a sua Família, em especial a sua mãe Nilma Marinho, ao seu Pai Antônio Marcos Beltrão, ao seu Irmão Iago Antônio Marinho, as suas tias Dilma Marinho e Wilma Marinho, e a sua avó Doralice Marinho por me ajudarem sempre quando precisei e por estarem juntos comigo nesta jornada tão importante.

A minha grande amiga Pâmela Carvalho dos Santos que em todos os momentos esteve presente, mesmo distante, me apoiando e ajudando nas diversas etapas da vida e do mestrado.

Ao meu Orientador Prof. Dr. Ig Ibert Bittencourt Santana por ter me orientado com paciência não só no programa de mestrado, mas também em momentos difíceis da minha vida pessoal. Pois seus ensinamentos foram de grande valia para o meu crescimento. Agradeço também por ter acreditado em meu potencial e ter me feito também acreditar quando nem mesma eu acreditava, fazendo concretizar meu objetivo, que antes de conhecê-lo era apenas um sonho.

A Thiago José Tavares Ávila pelos ensinamentos e por ter me ajudado a conseguir realizar um sonho, que era fazer parte de uma empresa, e melhor ainda, aplicar um pouco do conhecimento adquirido no Mestrado.

Aos meus amigos do Núcleo de Excelência em Tecnologias Sociais (NEES), que nunca hesitaram em me ajudar nos momentos em que precisei, em especial a Josmário de Albuquerque Silva, Jário José dos Santos Júnior, Williams Lourenço de Alcantara, André Vinicius Teixeira de Lima, Thyago Tenório Martins de Oliveira, Judson Bandeira, Sivaldo Joaquim de Santana, Daniel Borges F. da Silva, Denys Fellipe S. Rocha, Armando Barbosa Sobrinho, Ranilson Paiva, Diego D. Medeiros da Cunha Matos e Maria das Graças Cavalcante da Silva. Agradeço pelo apoio, colaboração e amizade.

Á Professora Patrícia Ospina, pela grande contribuição que tem dado neste trabalho, pela paciência e humildade que tem tido em trabalhar comigo.

Á equipe da Superintendência de Produção da Informação e do Conhecimento (SINC) da SEPLAG/AL pelo grande apoio e amizade que foram fundamentais na conclusão desta conquista, em especial a Roberson Leite Silva Junior, Maria Teonia Melo Amorim, Marcia

Nubia Barbosa Lopes, Teresa Marcia da Rocha Lima Emery e Salete Costa Cabral, Allisson Nascimento Goncalves da Silva e Gilvandro Freitas.

À Floripes Teixeira e Lúcia Maria de Souza, servidoras da UFAL, pela amizade, apoio, diálogo e conselhos durante esta trajetória que foram fundamentais para que este sonho fosse realizado.

Á Universidade Federal de Alagoas por proporcionar todo o ambiente e meios necessários para a formação de pesquisadores.

Á Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL) por fomentar o trabalho de pesquisadores impulsionando a formação de profissionais e o desenvolvimento econômico, científico e tecnológico de Alagoas.

Resumo

Dados Conectados tem contribuído com uma grande quantidade de informações na *Web*, representadas em formatos estruturados e conectados com outras informações. O principal objetivo das iniciativas de dados conectados é criar conhecimento pela conexão de dados dispersos e relacionados. A atual *Linked Open Data Cloud* (LOD *Cloud*) consiste em mais de 50 bilhões de fatos representados como triplas RDF. Essas informações pertencem a um grande número de *datasets* que cobrem diversos domínios, como ciência, dados geográficos, governamentais, etc. Entretanto, estudos recentes mostram que a maioria desses *datasets* sofrem de vários problemas de qualidade de dados, tais como, representacionais, inconsistências e questões de interoperabilidade. Esses problemas dificultam a interpretação dos dados e afetam a qualidade dos resultados. Desta forma, um desafio da área é analisar a qualidade de *datasets* conectados e deixá-la explícita. Com isso, este trabalho tem como objetivo criar uma solução computacional baseada em dimensões de qualidade e boas práticas de publicação que execute a verificação e validação semiautomática da qualidade de *datasets* conectados. Para isto, foram analisadas dimensões de qualidade e as mesmas foram correlacionadas com as melhores práticas de qualidade de dados contidas nos documentos "*Data on the Web Best Practices*" e "*Best Practices for Publishing Linked Data*". Para validação da proposta, foi executado um experimento com o objetivo de avaliar a solução desenvolvida, visando identificar se a mesma torna eficiente a avaliação da qualidade de *datasets* conectados, através da comparação da solução computacional semiautomática, proposta nesta dissertação, com a abordagem manual de avaliação da qualidade de *dataset* conectados. Como resultado, espera-se que a solução semiautomática seja um meio eficiente de executar a avaliação da qualidade de um *dataset* conectado, diminuindo o tempo de avaliação, bem como a carga de trabalho do avaliador. A contribuição dessa dissertação é disponibilizar um meio de avaliação voltado às melhores práticas do W3C, com base em dimensões de qualidade existentes na literatura.

Palavras-chave: Publicação de dados conectados. Dimensões de qualidade. Melhores práticas. Qualidade de dados.

Abstract

Linked Data has contributed to a lot of information on the Web represented in structured formats and linked to other information. The main purpose of linked data initiatives is to create knowledge by linking scattered and relational data. The current Linked Open Data Cloud (LOD Cloud) consists of more than 50 billion facts represented as RDF triples. This information belongs to a large number of covering various domains, such as science, geography, government, etc. However, recent studies show that most of these datasets suffer from various data quality problems, such as representational problems, inconsistency problems, and interoperability issues. These problems make data interpretation difficult and affect the quality of the results. In this way, a challenge in the area is to analyze the quality of linked datasets and make it explicit. This work aims to create a computational solution based on quality dimensions and best practices for publishing that performs the semiautomatic verification and validation of the quality of linked datasets. For this, quality dimensions were analyzed and correlated to the best practices of data quality contained in the documents, "Data on the Web Best Practices" and "Best Practices for Publishing Linked Data". To validate the proposal, an experiment was carried out to evaluate the developed solution and identify if it makes the evaluation of the quality of linked datasets more efficient by comparing the semiautomatic computational solution proposed in this dissertation to the manual approach of quality evaluation of linked datasets. As a result, a semiautomatic solution is expected to be an efficient way of performing the quality evaluation of a linked dataset and reduce the evaluation time as well as the workload of the user. The contribution of this dissertation is to provide an evaluation alternative focused on the best practices of the W3C, based on the quality dimensions existing in the literature.

Keywords: Publishing Linked Data. Quality dimensions. Best Practices. Data quality.

Lista de Figuras

1 Estágios do ciclo de vida de qualidade de dados 3
2 Metodologia GQM 10
3 Motivação para a publicação de dados conectados sobre áreas 22
4 Dimensões identificadas por Zaveri (2012) 35
5 Diagrama de casos de uso da dimensão Disponibilidade 51
6 Diagrama de casos de uso da dimensão Proveniência 52
7 Diagrama de casos de uso da dimensão Licenciamento 52
8 Diagrama de casos de uso da dimensão Compreensibilidade 53
9 Diagrama de casos de uso da dimensão Atualidade 54
10 Diagrama de casos de uso da dimensão Versatilidade 55
11 Diagrama de casos de uso da dimensão Consistência Representacional . . . 55
12 Ontologia Vocabulário de *Datasets* Conectados (VDC) 56
13 Uso da Classe Categoria 57
14 DataProperty name 57
15 Propriedades hasCategory e hasDimension 58
16 Propriedades hasCategory no Protégé 58
17 Propriedades hasMetric e hasDimension 59
18 Propriedades hasMetric e hasDimension no Protégé 59
19 Propriedades belongsToMetric e hasCriteria 60
20 Propriedades **belongsToMetric** e **hasCriteria** no Protégé 60
21 Propriedades **belongsToCriteriaEvaluation** e **hasCriteriaEvaluation** . . . 60
22 Propriedades **belongsToCriteriaEvaluation** e **hasCriteriaEvaluation** no Protégé. 61
23 Arquitetura de componentes da ferramenta semiautomática 64
24 Tela Inicial 65
25 Tela da Dimensão Disponibilidade 66
26 Tela de avaliação da disponibilidade de um *endPoint* SPARQL 67
27 Tela de avaliação da dimensão Proveniência 68
28 Tela de avaliação da dimensão Licença 69
29 Tela de avaliação da dimensão Compreensibilidade 69
30 Tela de avaliação da dimensão Atualidade 70
31 Tela de avaliação da dimensão Versatilidade 71
32 Relatório final da avaliação da qualidade de *datasets* conectados 71
33 Histogramas das variáveis: *Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency* e Score *Global* . 85

34 QQplots das variáveis: *Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency* e *Score* Global considerando todas as observações. 86
35 Histograma e QQplots da variável: Time, considerando todas as observações. 87
36 Boxplot de *Avaliability, representational consistency* e *score* global considerando os grupos: Nota base, Semiautomático e Manual 89
37 Boxplot de *Provenance, Licensing* e *Understandibility* considerando os grupos: Nota base, Semiautomático e Manual. 89
38 Boxplot de *Currency* e *Versatility*, considerando os grupos: Nota base, Semiautomático e Manual. 90
39 Boxplot de *Time* considerando os grupos: Nota base, Semiautomático e Manual. 90
40 Tela de avaliação da dimensão Disponibilidade 101
41 Tela de avaliação da disponibilidade de um *endPoint* SPARQL 102
42 Tela de avaliação da dimensão Proveniencia relacionada a página *Web* . . . 102
43 Tela de avaliação da dimensão Proveniencia relacionada a API 103
44 Tela de avaliação da dimensão Proveniencia relacionada ao RDF 103
45 Tela de avaliação da da dimensão Licença 104
46 Tela de avaliação da dimensão Atualidade relacionada a página *Web* 104
47 Tela de avaliação da dimensão Atualidade relacionada a API 104
48 Tela de avaliação da dimensão Atualidade relacionada ao RDF 105
49 Tela de avaliação da dimensão Compreensibilidadeq 105
50 Tela de avaliação da Dimensão Versatilidade 106
51 Gráfico de resíduos. Modelo 4 113
52 Gráfico de resíduos. Modelo 5 115
53 Gráfico de resíduos. Modelo 6 116
54 Gráfico de resíduos. Modelo 7 117

Conteúdo

1 Introdução **1**

1.1 Motivação e contextualização do trabalho 1
1.2 Problemática 3
1.3 Objetivos 4
1.4 Escopo 5
1.5 Contribuição do trabalho 5
1.6 Organização da dissertação 5

2 Fundamentação teórica **7**

2.1 Dados conectados 7
2.2 A abordagem Goal, Question, Metric (GQM) 8
2.3 Publicação de dados na *Web* 10
2.4 Dimensões de qualidade de *datasets* conectados 14
2.4.1 Dimensões contextuais 15
2.4.2 Dimensões de confiabilidade 15
2.4.3 Dimensões intrínsecas 16
2.4.4 Dimensões de Acessibilidade 16
2.4.5 Dimensões representacionais 17
2.4.6 Dimensões de dinamicidade 17

3 Trabalhos Relacionados **18**

3.1 Revisão da literatura sobre o uso de melhores práticas para a publicação de dados conectados 18
3.1.1 Resultados 19
3.2 Abordagem dirigida a métricas para avaliar a qualidade de dados abertos conectados (Behkamal et. al, 2013) (Behkamal et. al, 2013) 24
3.2.1 Limitações deste trabalho 27
3.3 Avaliação de qualidade dirigida à usuário do DBpedia (Zaveri, 2013) 27
3.3.1 Limitações deste trabalho 29
3.4 Luzzu – Um *framework* para avaliação da qualidade de dados conectados . 29
3.4.1 Limitações deste trabalho 31
3.5 QualityStamp – Avaliando a Qualidade de *Linked Datasets* para Aplicações de Domínio Específico (Travassos, 2014) 32
3.5.1 Limitações deste trabalho 32
3.6 Sumarização dos trabalhos relacionados 33

4 Uma abordagem semiautomática dirigida à métricas para avaliação da qualidade de *datasets* conectados **34**
4.1 Seleção de dimensões e métricas 34
4.1.1 Dimensões e métricas para verificação e validação da qualidade de *datasets* conectados 40
4.2 Implementação da abordagem semiautomática para avaliação da qualidade de *datasets* conectados 49
4.2.1 Cenários de funcionalidades da solução 50
4.2.2 Reuso e expansão da ontologia *Data Quality Vocabulary* (DQV) . . 56
4.2.2.1 Especificações do vocabulário 56
4.2.2.2 Exemplos de uso da ontologia Vocabulário de Dados Conectados 61
4.2.3 Arquitetura do *LinkedDatasetEvaluation* 63
4.2.4 Apresentação do sistema 65

5 Validação - Experimento da abordagem para avaliação da qualidade de *datasets* conectados **72**
5.1 Definição do problema 72
5.1.1 Contextualização 72
5.1.2 Relevância do problema 72
5.1.3 Problema técnico 73
5.2 Objetivos da investigação 73
5.3 Planejamento do experimento 74
5.3.1 Questões de pesquisa e hipóteses 74
5.3.1.1 Hipóteses 74
5.3.1.2 Fatores e variáveis de resposta 78
5.3.1.3 Definição formal das hipóteses 80
5.4 Unidades experimentais 80
5.5 Plano de execução 81
5.6 Instrumentação 83
5.7 Ameaças à validade 83
5.7.1 Ameaças à validade interna 83
5.7.2 Ameaças à validade de constructo 83
5.8 Análise estatística 84
5.9 Discussão 91

6 Conclusão **93**
6.1 Trabalhos futuros 94

7 REFERÊNCIAS **96**

A Apêndice **101**

B Apêndice **107**

B.1 Modelos de regressão 107

B.1.1 Modelos de regressão gerados 108

B.1.1.1 Análise de Diagnóstico 112

C Apêndice **118**

1 Introdução

Este capítulo fornece uma visão geral desta pesquisa e apresenta o contexto no qual este trabalho está inserido. Inicialmente, é mostrada a motivação do estudo e a definição do problema abordado. Depois são elencados os objetivos pretendidos, seguido por uma discussão sobre as contribuições esperadas deste trabalho. Por fim, é apresentada a descrição da estrutura desta dissertação.

1.1 Motivação e contextualização do trabalho

Dados estão sendo publicados na *Web* em diferentes formatos, por exemplo, PDF, imagens, CSV, planilhas eletrônicas, tabelas englobadas em documentos e muitas outras formas. Arquivos como os citados são normalmente disponibilizados em páginas *Web* e conectados com outros arquivos de conteúdos relacionados através de *urls*.

Desta forma, na *Web* atual (também conhecida por *Web* de documentos) podemos ver que o relacionamento entre documentos e a publicação de conteúdo é feita de maneira despadronizada, ou seja, cada usuário publica conteúdos na forma que for mais conveniente às suas necessidades.

No entanto, os formatos de dados utilizados, como PDF, imagens, CSV, planilhas eletrônicas, etc. possuem uma importante desvantagem: requerem frequentemente uma habilidade especializada para utilizá-los (processados por máquina), pois são formatos de dados para consumo humano. Por esta razão, não é uma tarefa fácil acessar, pesquisar ou reusar esses dados de uma maneira automática (Wood et al., 2009).

Com o objetivo de tornar os dados publicados mais fáceis de serem reusados, acessados e processados, um número crescente de provedores de dados no mundo estão optando por disponibilizar seus dados seguindo padrões Web de publicação. Tais padrões visam facilitar o reuso e processamento dos dados publicados, e os mesmos podem ser usados para muitos propósitos, por exemplo, para serem consumidos ou integrados em suas aplicações de maneira automática (Isotani & Bittencourt, 2015).

Esses padrões de publicação de dados conectados pertencem a um novo conceito da *Web*, chamado *Web* Semântica, a qual tem como finalidade produzir e disponibilizar dados que não só os usuários entendam as informações contidas nos mesmos, mas também agentes computacionais (Berners-Lee et. al, 2014). Com isso, o desenvolvimento e padronização de tecnologias *Web* semântica tem resultado em um grande volume de dados sendo publicados na *Web* como Dados Conectados (Zaveri et al., 2012). Dados Conectados tem contribuído com uma grande quantidade de informações na *Web*, representados em formatos estruturados e conectados com outras informações (Heath & Bizer, 2011).

O principal objetivo das iniciativas de dados conectados é criar conhecimento através da conexão de dados dispersos e relacionados. Por exemplo, a atual *Linked Open Data*

Cloud (LOD Cloud) consiste em mais de 50 bilhões de fatos representados como triplas RDF (Behkamal, 2013). Essas informações pertencem a um grande número de *datasets* que cobrem diversos domínios, como ciência, dados geográficos, governamentais, etc. (Batini & Scannapieco, 2006).

Dados conectados consistem em um conjunto de 5 (cinco) princípios definidos por Tim Berners-Lee (2009). Esses princípios, quando utilizados com tecnologias *Web* como HTTP (*Hypertext Transfer Protocol*) e URI (*Uniform Resource Identifier*) permitem a leitura dos dados conectados, de forma automática, por agentes de *software* (Isotani & Bittencourt, 2015).

A *Web* de Dados compartilha muitas características com a *Web* de Documentos. Da mesma forma em que temos avaliação de qualidade na *Web* de documentos, nós temos avaliação de qualidade na *Web* de Dados. Qualidade na *Web* de documentos é usualmente medida indiretamente usando ranqueamento de páginas.

Há uma grande variedade de dimensões e medidas de qualidade de dados (Zaveri, 2012) que podem ser computadas automaticamente. Segundo Debattista et. al (2016), a avaliação de qualidade, quando considerada isoladamente, não pode melhorar a qualidade de um *dataset*, mas quando considerada juntamente com outras fases do ciclo de vida de gerenciamento do dado, pode ser usada continuamente para monitorar e melhorar os *datasets* à medida que eles evoluem. Os estágios do ciclo de vida do dado podem ser vistos na Figura 1.

1. **Identificação e Definição de Métricas:** consiste em encontrar as métricas adequadas para um *dataset* com base em uma tarefa em mãos;

2. **Avaliação:** avaliação do *dataset* baseada nas métricas encontradas;

3. **Reparação e Limpeza do Dado:** garantir que, seguindo uma avaliação de qualidade, um *dataset* será reparado e limpado com o objetivo de melhorar sua qualidade;

4. **Armazenamento, Catalogação e Arquivamento:** Atualização do *dataset* melhorado na nuvem (*cloud*) ao passo que metadados de qualidade tornam-se disponíveis ao público.

5. **Exploração e Ranqueamento:** finalmente, consumidores de dados podem explorar os *datasets* com melhor qualidade de acordo com seus metadados de qualidade.

O presente trabalho foca na primeira e segunda fases do ciclo de vida do dado, que consistem na Identificação e Definição de Métricas e a Avaliação. A Solução para verificação e validação da qualidade de *datasets* conectados discutida no presente trabalho será tratada na seção 4.1.

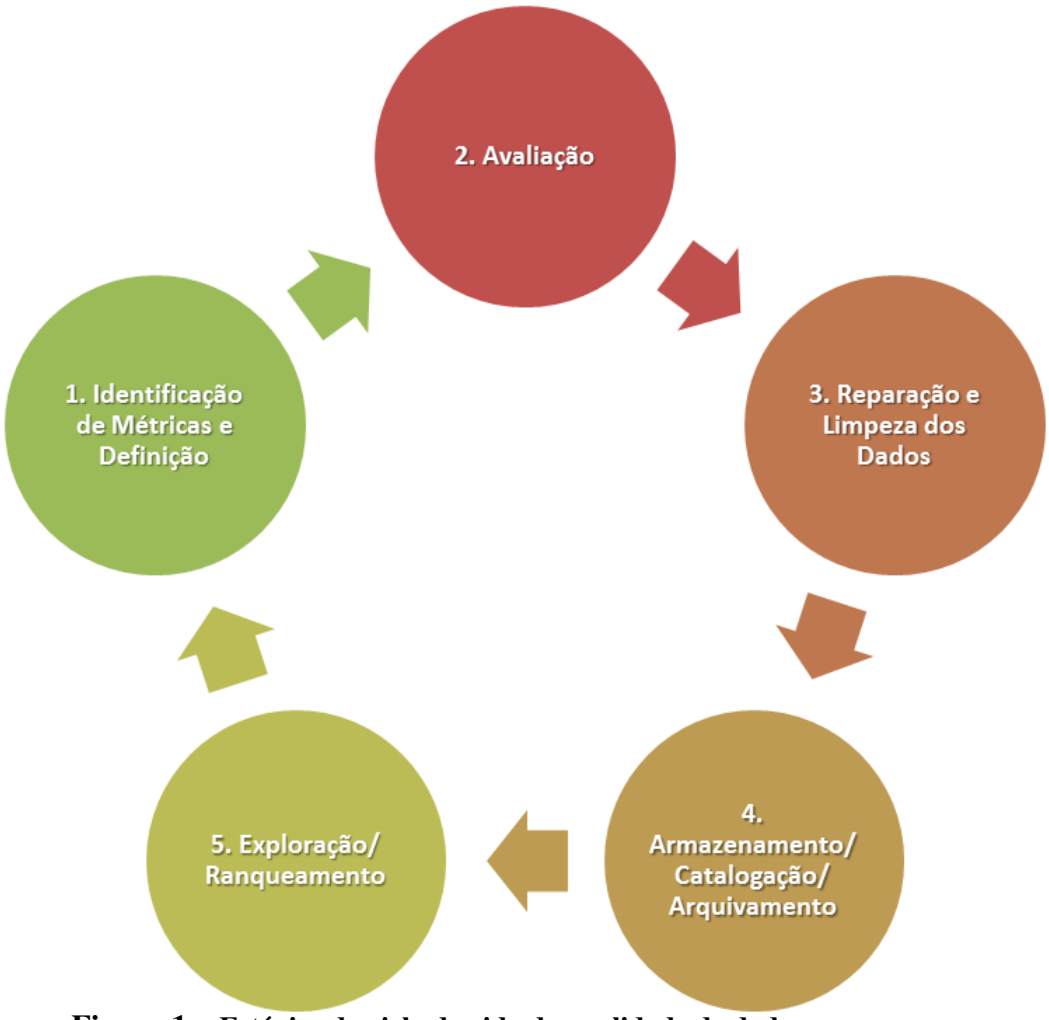

Figura 1 – Estágios do ciclo de vida de qualidade de dados
Fonte: Debattista et. al (2016)

1.2 Problemática

Estudos recentes mostram que muitos *datasets* sofrem de vários problemas de qualidade de dados, tais como, problemas representacionais, inconsistências e questões de interoperabilidade (Hogan et al, 2012). Esses problemas dificultam a interpretação dos dados em casos de uso particulares e afetam a qualidade dos resultados, se propagando nos *datasets* conectados.

De acordo com os editores do documento *Data on the Web Best Practices*, "qualidade de dados pode afetar a potencialidade de aplicações que usam dados, como isso, sua inclusão na publicação e consumo de dados é de fundamental importância"(Lóscio 2015 *apud* Debattista, 2016).

Com isso, um particular desafio da área é analisar a qualidade de *datasets* conectados e deixá-la explícita (Zaveri et al., 2013). A qualidade dos dados normalmente não pode ser descrita usando uma única medida, mas geralmente requer uma grande variedade de medidas de qualidade a serem calculadas.

Segundo O. Harting et al. (2009), qualidade da informação (QI) consiste em um agregado

de valores de múltiplos critérios, como por exemplo *accuracy, completeness, believability* e *timeliness*. Assim como o trabalho de O. Harting et.al. (2009), há outros trabalhos que propõem tais critérios, chamados também de dimensões, como Loshin (2011), Behkamal (2013) e Zaveri (2014).

Por outro lado, um número de melhores práticas e orientações para publicar dados com maior qualidade, são disponibilizadas para provedores de dados com o objetivo de ajudá-los na publicação de dados conectados com melhor qualidade (Debattista et. al, 2016). O W3C por exemplo, disponibiliza diversas orientações para publicação de dados conectados para a comunidade, com a finalidade de fomentar o reuso desses dados, bem como sua qualidade.

Com isso, podemos perceber que a comunidade não está se preocupando apenas em disponibilizar dados na *Web* do ponto de vista do publicador, mas também está se preocupando em publicar dados com melhor facilidade de interpretação e reuso do ponto de vista do consumidor.

1.3 Objetivos

Este trabalho tem como objetivo geral a concepção, implementação e avaliação de uma solução computacional baseada em dimensões de qualidade e boas práticas de publicação que execute a avaliação semiautomática da qualidade de *datasets* conectados. Para isto, dimensões de qualidade foram analisadas e correlacionadas com melhores práticas para publicação de dados, contidas nos documentos "*Data on the Web Best Practices*" e "*Best Practices for Publishing Linked Data*".

Os objetivos específicos do presente trabalho são:

1. Realizar um levantamento das melhores práticas para publicação de dados conectados e dimensões de qualidade utilizadas para avaliar a qualidade desses dados, existentes na literatura;

2. Relacionar práticas para publicação de *datasets* conectados com dimensões de qualidade de *datasets* conectados;

3. Estabelecer métricas para cada dimensão de qualidade selecionada;

4. Elaborar uma abordagem de avaliação de *datasets* conectados que englobe dimensões, práticas e métricas;

5. Especificar e desenvolver uma solução computacional que implemente a abordagem elaborada;

1.4 Escopo

O trabalho apresentado ao longo dessa dissertação trata sobre a concepção, desenvolvimento e validação de uma abordagem de avaliação de qualidade a ser aplicada no contexto de *datasets* conectados. Esta abordagem é baseada em dimensões de qualidade e melhores práticas de publicação de dados conectados.

Não está no escopo deste trabalho, dimensões de qualidade que não possuem relação com as melhores práticas, pois há uma grande variedade de dimensões de qualidade existentes na literatura, e abordá-las faria com que o processo de avaliação da qualidade de *datasets* conectados seja exaustivo.

Também, não está no escopo deste trabalho entrar na análise da qualidade do dado bruto, pois seria necessário conhecer o processo de geração do dado, bem como, o dado fonte (caso o dado que se deseja avaliar seja resultante de outros dados). Processos deste tipo podem ser visto no livro *The practitioner's guide to data quality improvement de Loshin* (2010)

Esta dissertação visa auxiliar publicadores e consumidores de dados a identificar a qualidade de um *dataset* conectado, evitando uma possível sobrecarga na avaliação e consequentemente uma melhor avaliação, além de possibilitar uma análise crítica do *dataset* ao final, através do relatório gerado pela abordagem semiautomática.

1.5 Contribuição do trabalho

Este trabalho visa contribuir com a comunidade de *Web* Semântica, apresentando uma abordagem de avaliação da qualidade de *datasets* conectados elaborada com base na metodologia GQM (*Goal, Questions and Metrics*) utilizando dimensões de qualidade.

Outra contribuição deste trabalho é a ferramenta semiautomática, que implementa a solução desenvolvida, diminuindo a carga de trabalho do avaliador, facilitando a avaliação do *dataset* conectado e realizando uma melhor análise ao final da avaliação, através de um relatório gerado que detalha a qualidade do *dataset*.

Além disso, outra contribuição do trabalho é a apresentação de uma revisão sistemática sobre quais melhores práticas estão sendo utilizadas pela literatura para a publicação de dados conectados, bem como, quais as vantagens em utilizar cada melhor prática, evidenciando que é benéfico utilizar as melhores práticas. Neste sentido, as contribuições científicas deste trabalho poderão servir como base para novas pesquisas em outras áreas, como publicação de dados no âmbito da Web Semântica.

1.6 Organização da dissertação

Esta dissertação está organizada da seguinte forma: O Capítulo 2 apresenta os principais conceitos que envolvem este trabalho, contribuindo para um melhor entendimento do mesmo; o

Capítulo 3 traz os principais trabalhos relacionados, identificados na literatura, que tratam da mesma problemática; o Capítulo 4 trata sobre a abordagem proposta com mais detalhes; o Capítulo 5 mostra como foi feita a validação da abordagem proposta e finalmente o Capítulo 6 apresenta as conclusões e trabalhos futuros.

2 Fundamentação teórica

Neste capítulo serão apresentados os principais conceitos teóricos que fundamentam esta dissertação para a análise e compreensão dos elementos que envolvem a abordagem apresentada, bem como o entendimento completo de sua implementação e processo de sua validação.

Neste sentido, as seções foram distribuídas da seguinte maneira: 2.1 Dados conectados; posteriormente será apresentada a abordagem Goal, Question, Metric (GQM) na seção 2.2; em seguida será abordado os conceitos das Melhores práticas na seção 2.3, seguida de uma Revisão da literatura sobre o uso de melhores práticas para a publicação de dados conectados na seção 2.4; e finalmente serão abordadas dimensões de qualidade de *datasets* conectados na seção 2.5.

2.1 Dados conectados

A área de Dados Conectados surgiu no ano de 2006, através da publicação *Design Issues* feita por Tim Berners-Lee, com uma subseção de *Web* Semântica exclusiva para Dados Conectados (Isotani & Bittencourt, 2015). A partir de então um número crescente de provedores de dados no mundo estão optando por publicar seus dados de maneira conectada na *Web* como (Wood, et al., 2014 *apud* Isotani & Bittencourt, 2015):

i) O *Google* anunciou a utilização do formato de serialização *JSON*-LD para o *Gmail*,

ii) IBM anunciou que o Banco de Dados DB2 se tornará um servidor de Dados Conectados;

iii) *Facebook* expôs os Dados Conectados via *Graph* API;

iv) BBC usou Dados Conectados para gerar páginas de três de seus produtos;

v) O Governo Britânico disponibiliza vários de suas fontes de dados em formato RDF (Data.gov, 2014).

Dados conectados referem-se a um conjunto de melhores práticas para publicação e conexão de dados estruturados na *Web* (Wood, 2014). Publicar dados utilizando o conceito de dados conectados permite a reutilização dos mesmos tanto por humanos, quanto por máquinas de forma automática.

> "Estas práticas são fundamentadas em tecnologias *Web*, como HTTP (*Hypertext Transfer Protocol*) e URI (Uniform Resource Identifier), com o objetivo de permitir a leitura dos dados conectados, de forma automática, por agentes de software. A *Web* de Dados cria inúmeras oportunidades para a integração semântica

dos próprios dados, motivando o desenvolvimento de novos tipos de aplicações e ferramentas, como navegadores e motores de busca" (Isotani & Bittencourt, 2015).

Dados conectados organizam as informações usando quatro regras básicas (Wood, 2014):

- Uso de URIs (*Uniform Resource Identifier*) como nome para recursos: O uso de URI permite que cada recurso tenha seu identificador único, não permitindo o problema de ambuiguidade de dados;
- Uso de HTTP (*Hypertext Transfer Protocol*) URIs para que as pessoas possam encontrar esses nomes: A partir do momento que cada dado tem seu identificador, torná-los disponíveis para busca.
- Quando as pessoas acessarem um URI, provê informações úteis usando padrões (RDF, SPARQL): RDF (*Resource Description Framework*), é um modelo de representação e estruturação dos dados que permite raciocínio por parte da máquina. SPARQL é uma linguagem de consulta sobre estes tipos de dados, semelhante ao SQL. A partir do momento que o dado é descoberto na *Web*, ele deve conter informações úteis e não apenas ser um dado comum.
- Incluir *links* para outros URIs, para que as pessoas possam descobrir mais recursos relacionados: Informações úteis contidas no dado possuem *links* que conectam para mais informações a respeito deste.

2.2 A abordagem Goal, Question, Metric (GQM)

Victor Basili, do Software Engineering Laboratory, Universidade de Maryland, Estados Unidos, propôs a abordagem GQM, cujo objetivo era servir como embasamento na elaboração e execução de programas de avaliação da qualidade de processos e produtos de software (BASILI, 1994a). Da mesma forma, a abordagem vem sendo utilizada por outros pesquisadores, centros de pesquisas e organizações (GRESSE, 1995, 1998).

O paradigma GQM (Briand et al., 1996; Caldiera et al., 1994) é uma abordagem de mensuração orientada a metas que apoia a definição e implementação top-down de metas operacionais e mensuráveis para melhoria de software e a interpretação botton-up dos dados coletados. É um enfoque de mensuração orientada a metas que ajuda na definição e implementação de um espectro amplo de metas de melhoria de softwares operacionais e mensuráveis, isto em todo o processo de desenvolvimento de software (Guenther et al., 1996; Bassman et al., 1994).

O processo GQM (BRIAND et al. 1996; GRESSE et al. 1995 *apud* VON WANGENHEIM et al., 1999) cobre o planejamento, execução de um programa de mensuração e a

captura de experiências para reutilização futura, por exemplo, em forma de modelos e diretrizes. A abordagem GQM foi utilizada em diversas empresas, como:

- NASA (EUA) (BASILI, 1992 *apud* Saraiva, 2006);
- Robert Bosch GmbH (Alemanha) (Bröckers, 1996 *apud* Saraiva, 2006);
- Kontinuierliche Qualitätsverbesserung inder Software Entwicklung - Erfahrungen bei der Allianz Lebensversicherungs-AG (H. Günther et al., 1994 *apud* Saraiva, 2006);
- Digital SPA (Itália) (Fuggetta, 1998 *apud* Saraiva, 2006);
- Motorola (Daskalantonakis, 1992 *apud* Saraiva, 2006);

E também vem sendo aderida pela comunidade científica, como podemos ver nos trabalhos abaixo:

- FOCA: Uma Metodologia que utiliza princípios da Representação do Conhecimento para Avaliação de Ontologias (Bandeira,2015);
- *A Metrics-Driven Approach for Quality Assessment of Linked Open Data* (BEHKAMAL, 2014)
- Usando GQM para gerenciar riscos em projetos de *software* (Fontoura et al., 2004)
- Uso do GQM para avaliar documentos de utilização de *framework* (de Souza, 2009)
- Utilização do GQM no Desenvolvimento de *Software* (WANGENHEIM, 2000)
- Aplicando a Abordagem GQM para Avaliar o Impacto da Adoção da Metodologia Ágil *Scrum* (Neves, 2012)

Essa abordagem tem várias vantagens: ela suporta a identificação das métricas úteis e relevantes, tanto quanto, suporta a análise e interpretação dos dados coletados (Wangenheim, 2000). Uma vantagem decorrente dessa abordagem é que ela sustenta a identificação das métricas apropriadas, de acordo com o contexto e os objetivos da avaliação, tanto quanto, sustenta a análise e legitimidade dos dados coletados, assim como a interpretação e armazenamento desses dados (Martins, 2011).

Segundo ABIB (1998) a abordagem GQM tem como principais componentes os objetivos, as questões e as métricas, conforme descritos a seguir:

- Objetivo: Sua definição envolve o objeto, o propósito, o foco de qualidade, o ponto de vista e o ambiente.

- Questão: corresponde a necessidade de se obter informações em uma linguagem natural, podendo-se formular uma ou mais indagações ; quanto à resposta, esta deve estar de acordo com o objetivo.

- Métrica: Sua função é especificar os dados ou as informações que se deseja obter durante as avaliações, em termos quantitativos e avaliáveis, podendo-se, utilizar uma ou mais métricas para cada questão.

É muito importante para o sucesso da aplicação do GQM que os objetivos estejam bem traçados, pois somente assim a escolha das métricas e posterior avaliação dos dados será bem sucedida (Gomes, 2001 *apud* de Souza, 2009). Com isso o GQM pode ser subdivido em três níveis de realização (Soligen, 1999), os mesmos estão ilustrados na Figura 2

- Conceitual: que consiste na definição do escopo da avaliação, ou seja, do objeto a ser medido, como por exemplo, Processos, Produtos ou Recursos.

- Operacional: definição de um conjunto de questões que auxilie na caracterização do objeto de estudo.

- Quantitativo: definição de um conjunto de dados a serem obtidos, relacionado a cada uma das questões definidas anteriormente, a fim de respondê-las de forma quantitativa, denominado por métricas.

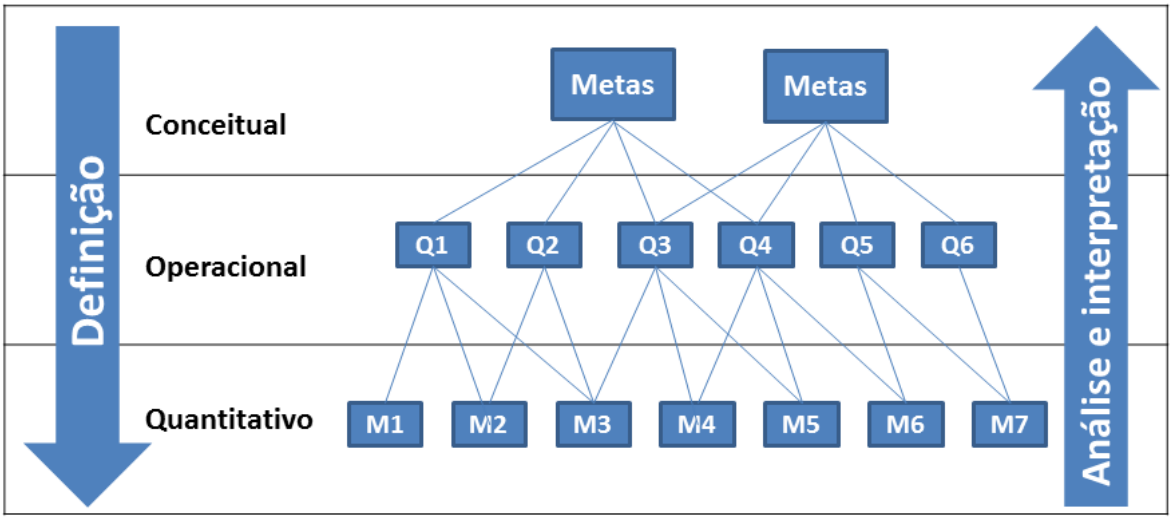

Figura 2 – Metodologia GQM
Fonte: Soligen (1999)

2.3 Publicação de dados na *Web*

Nas últimas décadas houve um "crescimento exponencial de dados gerados pela sociedade e a necessidade de minerar informações obtidas por meio da análise das conexões semânticas

entre conceitos e relações presentes nestes dados"(Isotani, et al., 2009; Bittencourt, et al., 2008).

Desta maneira, com a finalidade de orientar o processo de publicação de dados por parte publicadores, e também objetivando tornar os dados melhores compreensíveis para os consumidores de dados, o W3C disponibilizou diversas recomendações para publicação tanto de dados na *Web*, de uma forma geral, quanto para dados conectados.

Com isso, Nunes (2010) corrobora que práticas incluem processos, atividades, funções ou orientações. Com isso, práticas para publicação de dados na *Web* são um conjunto de orientações desenvolvidas para encorajar e possibilitar a expansão continuada da *Web* como um meio para o intercâmbio de dados (Lóscio e Calegari, 2016). Essas orientações podem ser utilizadas para representar, descrever e tornar os dados fáceis de ser encontrados, reusados e entendidos na *Web*.

Melhores práticas cobrem diferentes aspectos relacionados á publicação de dados, como formatos de dados, acesso, identificação e metadados (Lóscio and Calegari, 2016). Atualmente existem diversas melhores práticas, tanto para publicação de dados na *Web*, quanto para a publicação de dados conectados na *Web*, consideradas padrões *Web* internacionais para publicação de dados (Wood, 2014).

O W3C disponibiliza orientações para publicação de dados na *Web* através dois documentos, os quais são "*Data on the Web Best Practices*" que compreende um conjunto de melhores práticas para a publicação de dados mais generalizadas e "*Best Practices for Publishing Linked Data*" que compreende melhores práticas para a publicação de dados conectados.

Best Practices for Publishing Linked Data é um documento que provê melhores práticas relacionadas à publicação e uso dos dados na *Web*, neste são propostas 32 melhores práticas:

Tabela 1 – Melhores práticas do documento "*Best Practices for Publishing Linked Data*"

#	***Prática***	***Descrição***
MP 1	Prover metadados	Disponibilizar metadados legíveis por humanos e máquinas
MP 2	Prover metadados descritivos	As características gerais de conjuntos de dados e as distribuições devem ser descritos por metadados
MP 3	Prover metadados de parâmetros de localidade	Informações sobre parâmetros (data, tempo, número de formatos, linguagem) devem ser descritas no metadado
MP 4	Prover metadados sobre estrutura do *dataset*	Informações sobre o esquema e a estrutura interna de uma distribuição devem ser descrita pelo metadado
MP 5	Provê informações sobre a licença dos dados	Informações sobre a licença do dado devem estar disponível
MP 6	Provê informações de Proveniência do dado	Informações de proveniência de dados devem estar disponível
MP 7	Provê informações sobre a qualidade do dado	Informações sobre a qualidade do dado devem estar disponível
MP 8	Provê informações de versionamento	Informações sobre o versionamento do *dataset* devem estar disponíveis
MP 9	Provê histórico de versão	O histórico de versão sobre o *dataset* deve estar disponível
MP 10	Evitar grandes alterações para a API, Comunicar mudanças para desenvolvedores	Evitar alterações da API que quebram o código do cliente, e informar qualquer mudança na API para os desenvolvedores, quando evoluções acontecerem.
MP 11	Usar URIs persistentes como identificadores dos *datasets*	*Datasets* devem ser identificados com URI persistente, ou seja, que não muda
MP 12	Usar URIs persistentes como identificadores dos *datasets*	*Datasets* devem usar e reusar URIs de outras pessoas como identificadores quando for possível
MP 13	Atribuir URIs para versões do *dataset* e para a séries	URIs devem ser atribuídos para versões individuais do *dataset* assim como na série global
MP 14	Usar formato de dados padronizados legíveis por máquina	Dados devem estar disponíveis em um formato padronizado legível por máquina que é adequado para o potencial pretendido ou uso

(Continuação)

#	***Prática***	***Descrição***
MP 15	Prover dados em múltiplos formatos	Dados devem ser disponíveis em múltiplos formatos
MP 16	Usar termos padronizados	Termos padronizados devem ser usados para prover dados e metadados
MP 17	Reusar vocabulário	Vocabulários compartilhados devem ser utilizados para fornecer metadados
MP 18	Escolher nível de formalização adequado	Ao reusar um vocabulário, o publicador de dados deve optar pelo nível de semântica formal que encaixa o dado e a aplicação
MP 19	Prover referência a indisponibilidade do dado	Fornecer explicação sobre as diferentes restrições para dados que não estão abertos, ou indisponíveis e como o referido dado pode ser acessado e quem pode acessá-lo
MP 20	Provê *download* em massa	Dados devem estar disponíveis para *download* em massa
MP 21	Usar interfaces *Web* padronizadas	É recomendado o uso de URIs, verbos HTTP, códigos de resposta HTTP, tipos de MIME, tipos de *links* HTTP e conteúdo de negociação
MP 22	Servindo dados e recursos em diferentes formatos	É recomendado o uso de negociação de conteúdo para servir dados disponíveis em múltiplos formatos
MP 23	Provê acesso em tempo real	Quando dados são produzidos em tempo real, os mesmos devem ser disponibilizados na *Web* da mesma maneira
MP 24	Fornecer dados atualizados	Dados devem ser disponíveis de maneira atualizada e a frequência de atualização deve estar explícita
MP 25	Documentar sua API	Prover informações completas sobre como utilizar uma API
MP 26	Usar uma API	Oferecer uma API para servir de dados
MP 27	Cobertura de acesso a *datasets*	A cobertura de um *dataset* deve ser acessada antes de sua preservação
MP 28	Usar um formato de serialização confiável para preservar dumps de dados	Provedores de dados dispostos a enviar dumps de dados para preservação de termos longos devem usar uma serialização bem estabelecida
MP 29	Atualizar o status dos identificadores	Recursos preservados devem ser conectados com seus homólogos (relacionados)
MP 30	Capturar *feedback* de consumidores de dados	Publicadores de dados deve prover uma maneira para consumidores de dados oferecerem *feedback*
MP 31	Prover informações sobre o *feedback*	Informações sobre o *feedback* devem estar disponíveis
MP 32	Enriquecer os dados pela geração de novos metadados	Os dados devem ser enriquecidos sempre que possível através da geração de metadados mais ricos que os representem e descreva-os.

Fonte: Elaboração própria

Outro documento que disponibiliza melhores práticas é o "*Data on the Web Best Practices*", este trata de orientações que objetivam facilitar o desenvolvimento e entrega de dados abertos governamentais como dados abertos conectados. O mesmo provê um conjunto de 10 melhores práticas as quais são:

Tabela 2 – Melhores práticas do documento "*Data on the Web Best Practices*"

#	***Prática***	***Descrição***
MP 1	Preparar Partes Interessadas	Informar as partes interessadas sobre o processo de criação e manutenção dos dados abertos conectados.
MP 2	Selecionar Conjuntos de Dados	Selecionar um conjunto de dados que provê benefícios para reuso.
MP 3	Modelar os Dados	Representar objetos de dados e como eles se relacionam com a uma aplicação de maneira independente.
MP 4	Especificar uma Licença Apropriada	Explicitar uma licença apropriada para dados abertos.
MP 5	Estabelecer bons identificadores universais (URIs)	Utilizar bons URIs baseados em HTTP URIs.
MP 6	Utilizar Vocabulários Padrão	Descrever objetos com vocabulários definidos anteriormente.
MP 7	Converter Dados	Transformar os dados para uma representação de dados conectados.
MP 8	Prover Acesso Automatizado aos Dados	Fornecer várias maneiras para processos automatizados pesquisarem e acessarem os dados utilizando mecanismos *Web* padronizados.
MP 9	Anunciar Conjuntos de Dados	Inserindo novos conjuntos de dados em um domínio.
MP 10	Estabelecer um contrato social para os dados publicados	Reconhecer sua responsabilidade na manutenção de dados, uma vez que é publicado.

Fonte: Elaboração própria

2.4 Dimensões de qualidade de *datasets* conectados

O conceito de dimensões evoca pensamentos de medição, e é exatamente o que significa quando o termo é usado no contexto de qualidade de dado. Diferentes dimensões tem o propósito de representar diferentes aspectos de medição de qualidade de dados e são usadas na caracterização relevante através de um conjunto de domínios de aplicações para monitoramento de acordo com um padrão organizacional específico de qualidade de dados (Loshin, 2010).

Zaveri (2012) em um survey realiza uma revisão da literatura com a finalidade de elencar dimensões de qualidade de *datasets* conectados. Nesse estudo a autora identifica e descreve um conjunto de 26 dimensões de qualidade de dados e as define no contexto de dados conectados. Tais dimensões foram classificadas em seis categorias: (1) Contextual (2) Confiabilidade (3) Intrínseca (4) Acessibilidade (5) Representacional e (6) Dinamicidade do *dataset*.

As dimensões identificadas por Zaveri (2012), bem como as categorias serão descritas a seguir.

2.4.1 Dimensões contextuais

As dimensões desta categoria são altamente dependentes do contexto da tarefa em mãos, assim como, da preferência subjetiva do consumidor de dados.Sendo elas: Completude, Quantidade de dados e Relevância.

i. **Completude**: o grau em que as informações não estão faltando (Zaveri, 2012), ou seja, é o grau em que todas as informações requeridas estão presentes em um particular *dataset*.

ii. **Quantidade de dados**: refere-se medida em que o volume de dados é apropriado para a tarefa em mãos, ou seja, o grau em que a quantidade e volume de dados são apropriados para uma tarefa particular. A quantidade de dados fornecidos por uma fonte de dados influencia sua usabilidade e deve ser suficiente para aproximar-se do verdadeiro cenário.

iii. **Relevância**: a extensão em que a informação é aplicável e útil para a tarefa em mãos, ou seja, refere-se ao provimento de informações que estão em conformidade com a tarefa e consultas importantes para o usuário.

2.4.2 Dimensões de confiabilidade

As dimensões desta categoria focam na confiabilidade do *dataset*. Cinco são as dimensões que fazem parte desse grupo: Proveniência, Verificabilidade, Credibilidade, Reputação e Licenciamento.

iv. **Proveniência:** refere-se à metadados contextuais que focam em como representar, gerenciar e usar informações sobre a origem da fonte.

v. **Verificabilidade:** refere-se ao grau e a facilidade com que a informação pode ser checada para correção, ou seja, refere-se a medida em queum consumidor pode avaliar dados da exatidão de um conjunto de dados e, consequentemente, a sua confiabilidade.

vi. **Credibilidade:** refere-se a medida em que a informação é considerada verdadeira e credível, ou seja, ao grau em que a informação é aceito para ser correto, verdadeiro, real e credível.

vii. **Reputação:** "reputação de uma entidade" resultado da experiência direta ou recomendações de outras pessoas. Com isso, reputação é um julgamento feito por um usuário para determinar a integridade de uma fonte. É principalmente associado com um publicador de dados, uma pessoa, organização, grupo de pessoas, etc.

viii. **Licenciamento:** com o objetivo de possibilitar que consumidores usem dados sobre termos legais, cada documento RDF deve conter uma licença sobre a qual o conteúdo pode ser usado. A dimensão licenciamento é definida como a concessão de permissão para um consumidor reusar um *dataset* sobre condições definidas.

2.4.3 Dimensões intrínsecas

As dimensões pertencentes a esta categoria independem do contexto do usuário. As mesmas focam em verificar a o grau de correção da representação do *dataset* em relação ao mundo real e se a informação está logicamente consistente com o mesmo. As dimensões pertencentes a esta categoria são: Precisão, Objetividade, Validade de documentos, Interconexão, Consistência e Concisão.

ix. **Precisão:** refere-se ao grau de correção e precisão com que as informações em um sistema de informações representam estados do mundo real, ou seja, é a maneira em que os dados representam corretamente os fatos do mundo real e também estão livres de erro.

x. **Objetividade:** está relacionada ao grau em que a interpretação e utilização de dados é sem preconceito e imparcial.

xi. **Validade de documentos:** consistem em dois aspectos que influenciam a usabilidade do documento: o uso válido dos vocabulários subjacentes e a sintaxe válida dos documentos.

xii . **Interconexão:** refere-se ao grau em que entidades que representam o mesmo conceito estão conectadas.

xiii. Consistência: está relacionada a dois ou mais valores que não conflitam entre si, ou seja, não contradições de dados.

xiv. **Concisão:** refere-se à redundância de entidades, estando elas no nível do esquema ou no nível de dados.

2.4.4 Dimensões de Acessibilidade

As dimensões pertencentes a esta categoria envolvem aspectos relacionados à maneira com que os dados podem ser acessados e recuperados. As dimensões desta categoria são: Disponibilidade, Desempenho, Segurança e Tempo de resposta.

xv. **Disponibilidade:** está relacionada com o grau em que a informação está presente, obtenível e pronta para usar.

xvi. **Desempenho:** refere-se à eficiência de um sistema que se liga a um grande *dataset*, ou seja, o maior desempenho é de uma fonte de dados e a maior eficiência é de um sistema que pode processar os dados.

xvii. **Segurança:** pode ser definida como sendo à medida que o acesso a dados pode ser restringido e protegido contra sua alteração ilegal e uso indevido.

xviii. **Tempo de resposta:** é definido como o tempo de atraso entre a submissão de uma requisição feita pelo usuário e a recepção da resposta dada pelo sistema.

2.4.5 Dimensões representacionais

As dimensões pertencentes a esta categoria estão relacionadas com o design do dado. As dimensões desta categoria são: Consistência representacional, Consistência representacional, Compreensibilidade, Versatilidade e Interpretabilidade do dado.

xix. **Concisão representacional:** refere-se à representação do dado que por um lado é compacta e bem formatada, mas também clara e completa.

xx. **Consistência representacional:** é definida como a medida em que a informação é representada no mesmo formato (Zaveri, 2012), ou seja, é o grau em que o formato e a estrutura da informação está em conformidade com informação anteriormente retornada.

xxi. **Compreensibilidade:** pode ser definida como "o grau em que os dados são facilmente compreendidos pelos consumidores de informações".

xxii. **Versatilidade:** é definida como "representações alternativas do dado e o manuseio dos mesmos".

xxiii. **Interpretabilidade do dado:** refere-se a aspectos técnicos do dado, ou seja, verificar se a informação é representada usando uma notação apropriada e se a mesma está em conformidade com habilidades técnicas de consumidores.

2.4.6 Dimensões de dinamicidade

As dimensões pertencentes a esta categoria estão relacionadas com a dinamicidade de um *dataset*. As dimensões desta categoria são: Atualidade, Volatilidade, e Atualidade.

xxiv. **Atualidade:** refere-se a idade do dado, que é a diferença entre a data atual e a data de última modificação do dado [50], ou seja, refere-se a velocidade em que a informação é atualizada após a alteração de informações do mundo real.

xxv. **Volatilidade:** é definida como o período de tempo em que os dados são válidos.

xxvi. **Tempo oportuno:** refere-se ao ponto de tempo em que os dados são efetivamente utilizados, ou seja, se a informação está disponível no tempo em que é útil.

3 Trabalhos Relacionados

Nesta seção serão apresentados os trabalhos identificados na literatura que, semelhante a esta pesquisa, desenvolveram abordagens para avaliação da qualidade dados conectados. Cada um dos trabalhos apresentados buscam prover soluções para diferentes problemas de qualidade. Com isso, estes trabalhos foram utilizados como base para o desenvolvimento da solução proposta.

Desta forma, na subseção 3.1 é apresentado uma abordagem dirigida a métricas para avaliar a qualidade de dados abertos conectados. Na seção 3.2 é detalhado uma avaliação de qualidade dirigida à usuário do DBpedia. Ademais, as subseção 3.3 é falado sobre a ferramenta Luzzu, que consiste em um *framework* para avaliação da qualidade de dados conectados. Na subseção 3.4 é apresentado uma metodologia para avaliação da qualidade de dados conectados. Na subseção 3.5 é apresentado a ferramenta QualityStamp que busca avaliar a qualidade de *datasets* conectados para aplicações de domínio específico.

3.1 Revisão da literatura sobre o uso de melhores práticas para a publicação de dados conectados

Essa revisão sistemática foi conduzida com o objetivo de identificar as melhores práticas na literatura utilizadas para publicar dados conectados, bem como quais são os benefícios ao utilizá-las. Com isso, pretendeu-se responder à principal questão de pesquisa:

O que tem sido pesquisado/recomendado sobre as melhores práticas para a publicação de dados conectados?

Com base na principal questão de pesquisa, foram elencadas questões específicas de acordo com os aspectos que nos interessavam.

Tabela 3 – Perguntas de pesquisa.

Questões	*Descrição*
RQ 1: Quais as motivações para recomendação de melhores práticas para publicação de dados conectados?	A questão pretende identificar as vantagens do uso das práticas para publicação de dados conectados.
RQ 1.1: Há evidencias empíricas que justificam a recomendação de melhores práticas para publicação de dados conectados?	A sub-pergunta tem como objetivo identificar os resultados obtidos com o uso das melhores práticas para a publicação de dados conectados que foram validados empiricamente.
RQ 2: Quais são os formatos de dados abordados nos estudos?	Esta pergunta pretende identificar quais são os formatos de dados usados para publicar dados conectados.
RQ 3: Quais são as áreas abordadas pelos estudos ao publicar dados conectados?	Esta pergunta pretende identificar quais as áreas tem publicado dados de forma conectada.
RQ4: Quais as instituições que tem publicado dados conectados?	Esta pergunta busca identificar quais organizações (por exemplo, governos, cientistas e instituições) publicaram seus dados conectados na *Web*.

Fonte: Elaboração própria

3.1.1 Resultados

RQ 1: Quais as motivações para recomendação de melhores práticas para publicação de dados conectados?

O objetivo desta pergunta de pesquisa foi identificar as motivações para recomendação de melhores práticas para a publicação de dados conectados na *Web*. Classificamos essas motivações de acordo com oito benefícios significativos. Os benefícios são baseados em Auer (2011) para as categorias de Uniformidade, De-Referência, Coerência, Integrabilidade e *timeliness* e com base em Mendonça et al. (2013) para a categoria de Proveniência. Além disso, as categorias "Práticas padrão"e "Não identificadas"foram criadas pelos autores. A Tabela 4 apresenta essas categorias, juntamente com a descrição de cada uma, bem como a distribuição dos estudos nelas contidas.

Tabela 4 – Motivações para recomendação de melhores práticas para publicação de dados conectados

Motivação	*Descrição*	*%*
Uniformidade	Todos os conjuntos de dados publicados de forma conectada compartilham um modelo de dados uniforme, como por exemplo, o modelo de dados RDF. Em RDF a informação é representada em fatos organizados em formato de triplas consistindo em um sujeito, um predicado e um objeto. Os componentes utilizados para nomeação das triplas são chamados de IRIs (Internationalized Resource Identifier). Na posição do objeto também podem ser utilizados literais, isto é, valores de dados que possem tipo, como por exemplo, inteiro, string, double, etc. (Auer, 2011).	13,158%
De-Referênciamento	Os IRIs também podem ser usados na forma de URLs, permitindo a localização e recuperação de recursos, que descrevem e representam entidades (Auer, 2011)	7,895%
Coerência	Quando uma tripla RDF contém IRIs a partir de diferentes namespaces no sujeito e no objeto, o objeto estabelece uma ligação com a entidade identificada pelo sujeito. Através desses *links* RDF, os itens de dados são efetivamente conectados (Auer, 2011).	2,632%
Integrabilidade	Como todas as fontes de dados conectados compartilham o modelo de dados RDF, é fácil obter uma integração semântica e sintática de diferentes conjuntos de dados.	15,789%
timeliness	Uma vez que uma fonte de dados conectados é atualizada, acessa esta fonte e utiliza-la é mais fácil, porque algumas tarefas tornam-se desnecessárias, como a extração, transformação e carregamento de erros (Auer, 2011).	5,263%
Proveniência	Compreende todas as informações associadas a Linked Open Data, descrevendo "quem", "como", "quando", e "por que" o dado foi publicado, enriquecendo o contexto em torno dos dados e apoiando a avaliação da confiabilidade e qualidade dos dados (Mendonça et al., 2013)	2,632%
Práticas padrão	As práticas são consideradas boas diretrizes pelo W3C para a publicação de dados conectados porque informam as principais atividades, formatos e componentes que os dados publicados na *web* devem ter.	36,842%
Não identificado	Motivação ou benefícios não informados no artigo.	15,789%

Fonte: Elaboração própria

Com isso, considerando os artigos analisados, podemos perceber que a categoria de motivação predominante foi a "Práticas padrão" (36,842%), seguida de "Integrabilidade" e "Não informado" (15,789%), "Uniformidade" (13,158%), "De-Referênciamento" (7,895%), *Timeliness* (5,263%), Coerência e Proveniência (2,632%). Considerando que um estudo pode estar agrupado em mais de uma categoria de motivação.

RQ 1.1: Há evidencias empíricas que justificam a recomendação de melhores práticas para publicação de dados conectados?

Especialistas em desenvolvimento e produção consideram que evidência empírica é definida como a visão desenvolvida através da experiência, enquanto os especialistas em design instrucional definem evidências empíricas como dados derivados da experimentação estruturada (Schindelka, 2000).

Tomando esta definição em consideração, infelizmente, não foi possível identificar qualquer estudo que tenha realizado uma avaliação empírica que explora os benefícios do uso de melhores práticas para a publicação de dados conectados. Desta forma, esta pode ser uma questão que poderia ser investigada por pesquisadores da comunidade de dados conectados.

RQ 2: Quais são os formatos de dados abordados nos estudos?

O objetivo desta pergunta de pesquisa foi identificar os formatos de dados que os provedores de dados estão usando para a publicação de conjuntos de dados. Como mostrado na tabela, mais de 80% dos trabalhos utilizaram o formato de dados RDF. Por outro lado, apenas 11% não informaram qual formato de dados foi utilizado no processo de publicação.

RQ 3: Quais são as áreas abordadas pelos estudos ao publicar dados conectados?

O objetivo desta pergunta de pesquisa foi identificar os domínios abordados pelos artigos. Conforme apresentado na Tabela 5, a maioria dos artigos trata da publicação de dados relacionados a domínio Governamental, correspondendo a 23,33% dos trabalhos. Outro resultado interessante identificado, é que há um grande número de trabalhos no domínio de registro eletrônico de saúde, correspondendo a cinco artigos. Além disso, quatro artigos estão usando as melhores práticas no domínio de dados geoespaciais (13,33%). Outros domínios identificados são de dados estatísticos e artefatos de patrimônio cultural, com dois trabalhos cada. Por fim, foram identificados outros domínios, como a base de dados de currículos (3,33%), dados de transporte (3,33%), dados meteorológicos (3,33%), dados de jornalismo de viagem (3,33%) e dados linguísticos (3,33%). Cada uma dessas áreas é abordada por apenas um artigo. Não foi possível categorizar o domínio de quatro trabalhos (13,33%).

Tabela 5 – Áreas de conhecimento que estão publicando dados conectados

Áreas	%
Geoespacial	13,33 %
Base de dados de currículo	3,33%
Dados Governamentais	23,33%
Artefatos de património cultural	6,67%
Registro eletrônico de saúde	16,67%
Dados de transporte	3,33%
Dados estatísticos	6,67%
Dados meteorológicos	3,33%
Dados jornalístico de viagem	3,33%
Dados linguísticos	3,33%
Dados científicos	3,33%
Não informado	13,33%

Fonte: Elaboração própria

A Figura 3 mostra o número de estudos considerando as motivações para a publicação de dados conectados sobre esses domínios. Nota-se que a soma dos números de estudos sobre motivações específicas excede o número total de estudos dentro de uma categoria específica porque um estudo pode estar inserido em mais de um domínio.

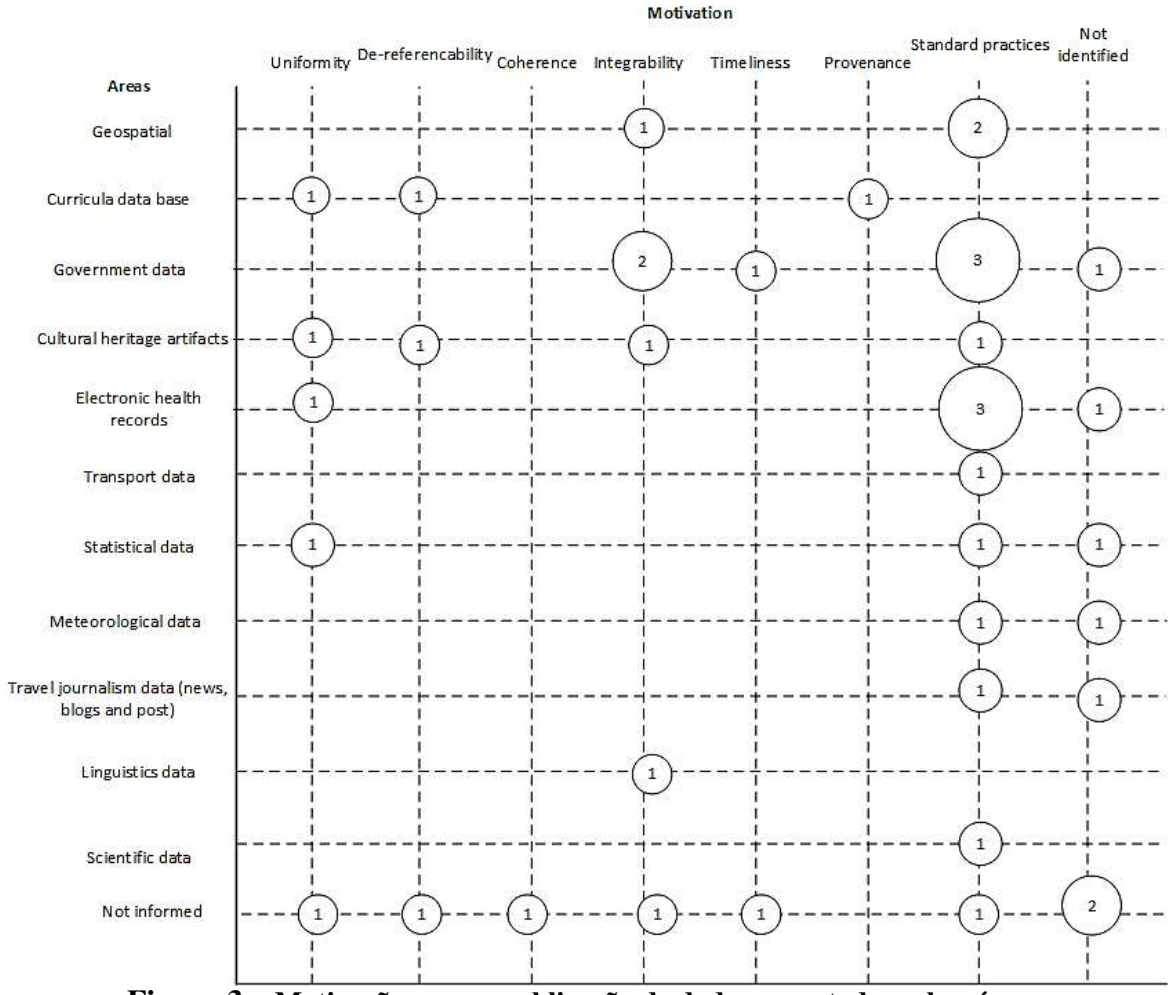

Figura 3 – Motivação para a publicação de dados conectados sobre áreas
Fonte: Elaboração própria

Na área do Governo, Villazon-Terrazas et al. (2011) sustentam que as iniciativas de Open *Government Data* em todo o mundo estão disponibilizando grandes quantidades de dados governamentais brutos ao público na *Web*. A abertura desses dados aos cidadãos possibilita a transparência e encoraja um maior uso público e comercial, e reutilização de informações governamentais. Esta área corresponde a 23,33% dos trabalhos. As motivações desta área (Figura 3) na publicação de dados utilizando as melhores práticas são Integrabilidade, *timeliness* e Práticas Padrão.

Na área de registro eletrônico de saúde, vários conjuntos de dados foram publicados na *Web* usando o RDF e estendendo-o estabelecendo *links* RDF entre itens de dados de diferentes fontes de dados contendo informações sobre genes, proteínas, vias, doenças e drogas (Pathak et al. 2012). Esta informação oferece um grande potencial para a realização de muitas combinações de dados de pacientes reais de registros de saúde eletrônicos, levando à criação de uma poderosa plataforma para consulta de dados heterogêneos. Esta área corresponde a 16,67% do total de trabalhos. Nesta área, as motivações para o uso de melhores práticas para a publicação de dados são Uniformidade e Práticas Padrão.

Na área geoespacial, Consoli et al. (2014) sustentam que dados geoespaciais ou informações geográficas são os dados que identificam uma localização geográfica de características naturais ou construídas, e limites na Terra (por exemplo, oceanos, edifícios, países, rios, etc.). *Linked Open Data* oferece a possibilidade de usar dados entre domínios ou organizações para diversos fins, como estatísticas, análises, mapas e publicações. Ao vincular esse conhecimento, inter-relações e correlações podem ser compreendidas por seres humanos e máquinas, e novas conclusões podem ser obtidas. Esta área corresponde a 13,33% do total de trabalhos e suas motivações para o uso das melhores práticas são Integrabilidade e Práticas Padrão.

Na área de artefatos do patrimônio cultural, de acordo com Marden et al. (2013), embora seja uma prática recente, muitas bibliotecas, arquivos e museus estão começando a adotar dados abertos conectados como uma forma de organizar e disseminar seus catálogos de explorações. Esta área corresponde a 6,67% do total de trabalhos. As motivações para o uso das melhores práticas são Uniformidade, De-Referênciamento, Integrabilidade e Práticas Padrão.

Na área de dados estatísticos, os censos são um dos mais relevantes tipos de dados estatísticos, permitindo análises da população em termos de demografia, economia, sociologia e cultura. Atualmente 77 países ou áreas disseminam microdados de censos e a América Latina é a região com maior percentual de países que disseminam microdados (Pablon et al., 2013). Esta área corresponde a 6,67% do total de trabalhos. As motivações, nesta área, são as práticas de Uniformidade e Padrão.

Além disso, como visto na Tabela 5, outras áreas vêm publicando dados, por exemplo, Banco de Dados de Currículos, Transporte, Meteorologia, Jornalismo de Viagem e Linguística

RQ4: Quais as instituições que tem publicado dados

A Tabela 6 apresenta a distribuição de artigos. 37% artigos tratam de publicação de dados governamentais. Os outros três artigos não estão apresentando claramente quais dados de formato eles estão usando. Dez artigos estão relacionados a instituições acadêmicas, seguidas por instituições de patrimônio cultural (um artigo) e instituições de saúde (três artigos). Em três artigos, não foi possível identificar o formato de dados usado para publicar dados conectados.

Tabela 6 – Instituições que estão publicando dados conectados

Instituições	%
Instituições governamentais	37%
Instituições acadêmicas	37%
Instituições do património cultural	4%
Instituições de saúde	11%
Não informado	11%

Fonte: Elaboração própria

Em muitos países, governos, autoridades locais e organizações não-governamentais, uma variedade de conjuntos de dados está disponível ao público. Esses conjuntos de dados abrangem diversos tópicos, como dados geoespaciais, dados de currículo dos pesquisadores, dados de transporte, dados de jornalismo de viagem (notícias, blogs e postagens) e muitos outros tipos de informações. Conforme mostrado na Tabela 6, as instituições governamentais e acadêmicas estão muito mais preocupadas com o uso de melhores práticas para a publicação de dados conectados. Esses resultados podem sugerir que eles dependem de tais práticas para aumentar a qualidade dos dados editoriais e, portanto, facilitar a utilização desses dados pelos consumidores.

3.2 Abordagem dirigida a métricas para avaliar a qualidade de dados abertos conectados (Behkamal et. al, 2013) (Behkamal et. al, 2013)

Este trabalho tem como objetivo propor e validar um conjunto de métricas para avaliar características de qualidade inerentes a um *dataset* antes da sua liberação na *Linked Open Data Cloud*. Segundo Behkamal (2013) qualidade inerente é: (i) a habilidade do *dataset* em representar efetivamente e refletir informações do mundo real nos dados; e (ii) a coerência interna dos dados que são representados como uma parte do mesmo *dataset*.

Com isso, em seu trabalho, Behkamal (2013) reúne três dimensões relevantes que são aspectos de qualidade inerentes de *Linked Open Data* (LOD), as quais são: *Accuracy, Completeness* e *Consistency*. Com o objetivo de melhor estudar a dimensão *accuracy*, o mesmo propõe três dimensões que são chamadas de *semantic accuracy, syntatic accuracy* e *uniqueness*.

Existem outras dimensões de qualidade, como *Credibility* e *Timeliness*, no entanto as mesmas não foram incluídas, porque elas não são consideradas dimensões de qualidade inerentes no contexto de LOD. Por exemplo, Credibility é relacionado à confiabilidade do *dataset* assim como , Provenance, Verifiability, Believability, e Licensing que são outras dimensões de qualidade. A dimensão *timeliness* é um importante aspecto de LOD que está relacionado à dinamicidade do *dataset* e refere-se ao tempo de duração em que o dado atualizado é usado.

Behkamal (2013) em seu trabalho utiliza as seguintes dimensões:

- *Syntatic Accuracy:* é definida como sendo "a proximidade dos valores de dados para um conjunto de valores definidos em um domínio considerado sintaticamente correto".
- *Semantic Accuracy:* refere-se à correção de um valor de dados em comparação com o valor real do mundo real, ou seja, todos os recursos ou entidade descrita em um *dataset* deve ter uma correspondência no mundo real.
- *Uniqueness:* é o grau em que uma ontologia é livre de redundâncias levando em consideração classes e propriedades.
- *Consistency:* é definida como o grau para o qual os atributos dos dados estão livres de contradições e são coerentes com outros dados em um contexto de uso específico.
- *Completeness:* refere-se ao grau em que todas as informações requeridas para descrever um conceito estão presentes em um *dataset*. O nível de detalhe é definido como a quantidade de dados disponíveis para a descrição de entidades do mundo real em um *dataset*.

Behkamal (2013) em seu trabalho emprega a abordagem *Goal-Question-Metric* (GQM) com o objetivo de levantar métricas para avaliação de cada dimensão de qualidade. Dessa forma, o mesmo define um obetivo principal (Goal), que é "avaliar a qualidade inerente de um *dataset* a partir do ponto de vista do usuário no contexto de LOD". Para avaliar cada dimensão de qualidade o mesmo define cinco sub-objetivos (SG – Sub-Goal), onde cada um é voltado para avaliar uma dimensão de qualidade. Nas tabelas abaixo estão descritas as questões e as métricas.

Tabela 7 – Avaliação da *Semantic Accuracy* de um *dataset* a partir pa perspectiva do usuário no contexto de LOD

Questão	***Métrica***	***Definição***
As entidades são descritas com valores corretos?	M1. Razão das triplas que contém objetos faltantes.	O número de triplas contendo propriedades com os valores em falta no que diz respeito às propriedades definidas no esquema, dividido pelo número total de triplas num conjunto de dados, subtraído a partir de 1.
	M2. Razão das triplas com objetos fora do alcance.	O número de triplas que contém propriedades com valores discrepantes em relação ao alcance dos valores aceitáveis definidos no esquema, dividido pelo número total de triplas no *dataset* subtraídas por 1.
	M3. A razão de triplas que contém valores de dados com erros de ortografia.	O número de triplas que contém propriedades com valores com erros de ortografia, dividido pelo número total triplas no *dataset*, subtraído por 1.
As entidades representam precisamente o mundo real?	M4. Razão das entidades sem correspondente no mundo real.	O número de entidades sem objetos correspondente no mundo real, dividido pelo número total de entidades descritas no *dataset*, subtraído por 1.

Fonte: Fonte: Behkamal (2013)

Tabela 8 – Avaliação de *Syntactic Accuracy* de um *dataset* a partir da perspectiva do usuário no contexto de LOD.

Questão	***Métrica***	***Definição***
A sintaxe do documento RDF é válida?	M5. Razão das triplas sintaticamente incorretas.	O número de triplas que contém propriedades com erros de sintaxe, dividido pelo número total de trilhas do *dataset*, subtraídas por 1.
Os recursos estão descritos com propriedades adequadas?	M6. Razão de triplas com atribuições impróprias de tipos de dados literais.	O número de triplas que contém propriedades com atribuições inadequadas de tipos de dados literais, divididos pelo número total de triplas do *dataset*, sutraído por 1.
	M7. Razão de instâncias usando classes/propriedades indefinidas.	Número de instâncias que usam classes/propriedades indefinidas, divididas pelo número total de entidades do *dataset*, subtraída por 1.
	M8. Razão de instâncias sendo membros de classes disjuntas.	Número de instâncias sendo membros de classes disjuntas, dividido pelo número total de instâncias do *dataset*, subtraídas por 1.
	M9. Razão de triplas que contém uso indevido de vocabulário.	O número de triplas que contém propriedades com uso indevido de vocabulário, dividido pelo número total de triplas do *dataset*, subtraídas por 1.

Fonte: Behkamal (2013)

Tabela 9 – Avaliação de *Uniqueness* de um *dataset* a partir da perspectiva do usuário no contexto de LOD

Questão	***Métrica***	***Definição***
Qual é o grau de redundância no contexto de classes?	M10. Razão de redundância das classes.	O número de classes com a mesma propriedade, dividido pelo número de classes definido no esquema do *dataset*, subtraído por 1.
Qual é o grau de redundância no contexto de propriedades?	M11. Razão de propriedades similares.	O número de propriedades similares, dividido pelo número total de propriedades definidas no esquema do *dataset*, subtraído por 1.
O *dataset* contém múltiplas representações para a mesma entidade?	M12. Razão de instâncias redundantes.	O número de instâncias com URI diferentes, mas a mesma propriedade dividido pelo número total de instâncias de um *dataset*.
O *dataset* contém valores redundantes para propriedades?	M13. Razão de propriedades funcionais com valores diferentes.	Número de propriedades funcionais que contém valores diferentes, divididos pelo número total de propriedades definidas no esquema de um *dataset*, subtraído por 1.

Fonte: Behkamal (2013)

Tabela 10 – Avaliação de *Consistency* de um *dataset* a partir da perspectiva do usuário no contexto de LOD

Questão	*Métrica*	*Definição*
Existe alguma inconsistência no esquema do *dataset*?	M14. Quantidade de membros (membership) de classes disjuntas.	O número de entidades que são membros de classes disjuntas.
	M15. Uso inválido de propriedades funcionais inversas.	O número de triplas que contém uso inválido de propriedades funcionais inversas
	M16. Razão das triplas que usam propriedades similares.	O número de triplas que usam propriedades similares, dividido pelo número total de triplas em um *dataset*, subtraído por 1.
	M17. Tipos de dados heterogêneos.	O número de tipos de dados heterogêneos no esquema.
Qual é o grau de conflitos no contexto do valor do dado?	M18. Valores de propriedades inconsistentes.	O número de propriedades com valores inconsistentes.

Fonte: Behkamal (2013)

Tabela 11 – Avaliação de *completeness* de um *dataset* a partir da perspectiva do usuário no contexto de LOD

Questão	*Métrica*	*Definição*
Todos os recursos foram descritos com o número de propriedade adequado?	M19. Razão de propriedades para classes.	O número de propriedades, dividido pelo número de classes descritos em um *dataset*.
Todas as informações necessárias para cada entidade foram apresentadas?	M20. Falta de propriedades por instância.	Soma do número de propriedades apresentado por instância, dividido pelo número total de instâncias em um *dataset*.

Fonte: Behkamal (2013)

3.2.1 Limitações deste trabalho

Uma das limitações consiste em avaliar as dimensões de forma manual, ou seja, a avaliação proposta pelo autor não é automática e nem semiautomática, consumindo assim, um tempo considerável da mesma e aumento da carga de trabalho. Não foi desenvolvida ferramenta para auxiliar no processo de avaliação, dificultando assim, a avaliação por parte do usuário. Com isso, não há geração de relatório, e consequentemente, não fica transparente a qualidade do *dataset*.

3.3 Avaliação de qualidade dirigida à usuário do DBpedia (Zaveri, 2013)

Zaveri (2013) propõe um processo para avaliar a qualidade de recursos de dados conectados, o qual compreende um processo manual e semiautomático. O processo consiste dos seguintes passos:

Passo I Seleção de recurso: Os recursos pertencentes a um *dataset* particular são selecionados. Esta seleção pode ser feita de três maneiras:

- **Por Classe:** seleciona recursos pertencentes a uma classe particular (ex.: animal, esporte, lugar). Esta forma possibilita ao usuário a flexibilidade para escolher o recurso pertencente a uma classe que lhe é familiar.
- **Completamente randomizada:** o recurso do *dataset* é selecionado de forma aleatória. Esta opção garante cobertura imparcial e uniforme do *dataset*.
- **Manual:** o recurso é selecionado manualmente. Nesta opção o usuário é livre para selecionar recursos com problemas possivelmente identificados anteriormente.

Passo II Seleção do modo de avaliação: A atribuição dos recursos para uma pessoa ou máquina, selecionado no Passo I pode ser realizada de três maneiras:

- Manual: a seleção dos recursos é realizada por uma pessoa (ou grupo de indivíduos) que irão proceder manualmente a avaliação de recursos individuais.
- Semiautomática: os recursos selecionados são atribuídos a uma ferramenta semiautomática que realizará a avaliação de qualidade de dados empregando alguma forma de *feedback* do usuário.
- Automática: o recurso selecionado é dado como entrada para uma ferramenta automática que realiza a avaliação de qualidade sem intervenção do usuário.

Passo III Avaliação do recurso:

- No caso da atribuição manual do recurso, a pessoa avalia cada recurso individualmente para detectar potenciais problemas de qualidade de dados.
- Se o recurso for atribuído a uma ferramenta semiautomática, a mesma irá apontar para triplas que provavelmente estão erradas.

Passo IV melhoria da qualidade do dado: há duas maneiras de realizar a melhoria:

- Direta: editando as triplas identificadas com problemas, com o valor correto.
- Indireta: usando o Patch Request Ontology, que permite capturar *feedbacks* de usuários sobre triplas erradas.

Para execução da avaliação foi utilizado a dimensão:

- *accuracy*: o grau em que os dados estão corretos, ou seja, o grau de corretude da representação do mundo real. A mesma é subdividida em:

    i. Objeto extraído incorretamente: refere-se a problemas que surgem quando o valor de um objeto de uma tripla está incorreto.

ii. Problemas de tipo de dados: refere-se as triplas que são extraídas com um tipo de dado incorreto para um literal tipado.

iii. Relações implícitas entre atributos: esta categoria de problemas pode surgir devido:

* Representação de um fato em vários atributos,
* Vários fatos codificados em um atributo, ou
* Um valor de um atributo computado a partir de outro atributo

3.3.1 Limitações deste trabalho

Uma das limitações deste trabalho consiste nas avaliações manual e automática. Na manual o usuário deve executar todos os passos da avaliação de forma manual, necessitando de um tempo considerável para fazê-la, levando em consideração a execução da verificação das dimensões e métricas, com isso, pode ocorrer do usuário ficar cansado. É trabalhoso o usuário analisar uma ontologia (partindo do princípio que o mesmo tenha conhecimento sobre ontologia).

A avaliação automática também é considerada uma limitação, devido existirem dimensões e métricas que são melhor avaliadas utilizando o conhecimento do usuário em relação ao domínio, como por exemplo, o tópico "i. objeto extraído incorretamente"da dimensão *accuracy*, que refere-se na identificação de valores de objetos incorretos (Ex. nome de pessoas, data de nascimento, quantidade de lucro de um funcionário, etc.). Pode ocorrer da máquina não identificar tais erros, pois o erro não consiste no tipo de literal ou na ortografia, e sim na informação do dado. Este tipo de avaliação não será abordada no presente trabalho, pois a proposta será do *dataset* em geral, como por exemplo, a verificação de metadados de proveniência e licença, as quais dão uma maior confiabilidade ao *dataset*.

Outra limitação é a não geração de relatório. É interessante que o usuário saiba quais os pontos negativos e positivos foram identificados na avaliação do *dataset*.

3.4 Luzzu – Um *framework* para avaliação da qualidade de dados conectados

É um *framework* para avaliação de qualidade que provê:

(1) Avaliação da qualidade de dados conectados, usando uma biblioteca com métricas de qualidade genéricas e métricas de qualidade de domínio específico, providas ou pelo usuário em uma maneira escalável,

(2) Metadados de qualidade na avaliação de *datasets*

(3) Relatório de qualidade detalhando a avaliação do *dataset*

Os mesmos objetivaram criar uma infraestutura que:

- Pode ser facilmente extendida por usuários pela criação de *plugins* de métricas customizadas e métricas de domínio específico, através do emprego de uma linguagem de especificação de métricas declarativas e o emprego de pluguins convencionais necessários
- Emprego de uma ontologia abrangente para representação e mudança de todas as informações relacionadas a qualidade no fluxo de trabalho de avaliação

O fluxo de trabalho do *framework* segue três passos:

1. Começa com o processo de inicialização de métricas declarativas, que são definidas pelo Luzzu Quality Metric Language (LQML), as quais são compiladas e inicializadas junto com métricas implementadas em java.
2. Neste passo é começado o processo pela transmissão sequencial de declarações do *dataset* candidato dentro das métricas de qualidade inicializadas.
3. Uma vez que o processo no passo 2 é completado, é gerado anotações de qualidade em metadados e é compilado um relatório de qualidade abrangente.

Essa ferramenta pode ser facilmente integrada em coleções, tais como *Linked Data Stack* [1] para elevá-los com o fluxo de trabalho de avaliação de qualidade.

O *framework* é composto por três camadas:

- Camada de Comunicação (*Communication Layer*).
- Camada de Avaliação (*Assessment Layer*): que é composta por três unidades, elas manuseiam as operações relacionadas a avaliação de qualidade do *dataset*:
    i. Unidade de processamento (*Processing Unit*)
    ii. Unidade de compilação LQML
    iii. Unidade de avaliação de qualidade
- Camada de Conhecimento (*Knowledge Layer*): é composto de três unidades:
    i. Camada de esquema semântico (*Semantic Schema Layer*): consiste em dois níveis:
        a. Repesentação (representation): divido em sub-níveis
            i. Genéricos: é independente de domínio e pode ser reusado em *frameworks* similares para avaliação de qualidade. Os dois vocabulários descritos neste nível são :

1. *Dataset Quality Ontology* (daQ) (Debattista, 2014) e *Quality Problem Report Ontology*. Esses vocabulários contribuem para dois objetivos mais especificamente provendo metadados consultáveis e montagem detalhada de relatórios

ii. Específicos: um número de métricas de qualidade de *linked data* são implementadas, algumas estão no survey (Morsey, 2012). Essas métricas são usadas na avaliação da qualidade de 270[a] *datasets* estatísticos. Data Quality Metric (DQM) Ontology representa essas métricas em um estilo semântico baseado no vocabulário daQ.

b. Nível Operacional: este nível (mais alto da pilha do esquema) possui um número de ontologias que auxiliam o *framework* na realização de certas tarefas, tais como o Luzzu Metric Implementation ontology (LMI) que é um vocabulário que permite que métricas especificadas com termos da daQ se conectem com a implementação em java. Neste nível possui dois subníveis. O menor possui ontologias genéricas que forma a base da avaliação de qualidade do *framework* e o maior possui ontologias específicas necessárias para várias tarefas de avaliação de qualidade

ii. Unidade de anotação (Annotation Unit): esta gera metadados de qualidade e relatórios de qualidade para cada métrica avaliada. É baseada em duas ontologias principais – daQ e QPRO. No que se refere à problemas de qualidade, a unidade compila um conjunto de triplas, que consistem em triplas problemáticas encontradas durante a execução da métrica de avaliação do *dataset*. O usuário poderá posteriormente usar o relatório de problemas de qualidade, gerado na ferramenta de limpeza do dado, que suporta a *Quality Report Ontology* para organização do *dataset*.

iii. Unidade de operações (Operations Unit): esta não interage diretamente com a avaliação de qualidade. A mesma provê algoritmos para uso no dia-a-dia dos metadados de qualidade, incluindo um algoritmo para classificação orientada a qualidade dos *datasets*. O algoritmo (user-driven ranking algorithm) permite o usuário definir pesos para suas categorias preferidas, dimensões ou métricas, que são consideradas para sua tarefa em mãos.

3.4.1 Limitações deste trabalho

Uma das limitações deste trabalho é em relação ao relatório gerado, pois o mesmo não disponibiliza um relatório detalhado, informando as notas nos pontos existentes dentro de cada métrica, ou seja, apenas a nota da métrica em geral, dessa maneira fica um pouco difícil para o usuário saber o que mudar no *dataset* para melhorar a pontuação de cada métrica, e assim, o mesmo ter uma qualidade melhor.

E por último, outra limitação consiste na avaliação não ser aplicada a página *Web* e API do *dataset* (caso existam), pois como o *dataset* segundo o W3C, deve ser acessado e entendido por humanos e consumido por aplicações que não consumam necessáriamente o RDF e sim utilizam a API. A qualidade existente no RDF do mesmo deve ser espelhada tanto na página *Web* quanto na API, como por exemplo, deve estar disponível em ambos os recursos informações de proveniência, o *dataset* em outras linguagens, a licença utilizada pelo mesmo, etc.

3.5 QualityStamp – Avaliando a Qualidade de *Linked Datasets* para Aplicações de Domínio Específico (Travassos, 2014)

Travassos et al. (2014) buscou desenvolver uma ferramenta, denominada QualityStamp, para avaliar a qualidade de *linked datasets*. Em seu trabalho o termo "qualidade" é associado à descrição geral de "adequação ao uso" proveniente da definição de Qualidade da Informação (Wang e Strong, 1996). A abordagem de avaliação, proposta por Travassos tem como finalidade auxiliar a escolha dos *datasets* mais adequados para uma determinada aplicação dentre um conjunto de *datasets* candidatos predefinidos, ou seja, identificar os *datasets* que mais contribuem com dados relevantes.

O mesmo buscou avaliar três características dos *linked datasets* por meio de seus respectivos *endpoints*:

- Desempenho: diz respeito à capacidade do *dataset* em responder às solicitações da aplicação. Nesta categoria estão associados os critérios de qualidade Disponibilidade, Tempo de Resposta e Variação de Carga.
- Completude: está relacionada a quanto um *dataset* contribui para responder a um conjunto de consultas da aplicação, por exemplo, o quanto da informação presente nos *datasets* candidatos é útil do ponto de vista dos requisitos da aplicação. Nesta categoria estão contidos os critérios Completude de Esquema, Completude de Literal e Completude de Instâncias.
- *Interlinking*: diz respeito ao grau de interligação de um *dataset* com outros *datasets* relacionados.

Ao final da avaliação será apresentado um percentual de cada critério de qualidade, bem como uma classificação dos *datasets* por ordem de qualidade.

3.5.1 Limitações deste trabalho

Uma das limitações do trabalho em questão consiste no relatório que é gerado pela aplicação, que não informa o percentual obtido nos critérios de qualidade em cada *dataset*, consistindo

na pouca transparência do mesmo. Outra limitação é que os *datasets* considerados na avaliação devem estar em *EndPoints*, ou seja, o mesmo não considera avaliação de um *dataset* contido em um arquivo de *download* disponibilizado em uma página *Web*. E por último, a avaliação depende fortemente de consultas SPARQL, ou seja, não havendo consultas, não há avaliação.

3.6 Sumarização dos trabalhos relacionados

Nesta subseção será apresentado, de forma resumida, as contribuições e limitações dos trabalhos relacionados e também da abordagem proposta. Os resultados estão dispostos na Tabela 12

Tabela 12 – Resumo dos Trabalhos Relacionados

	Abordagem				
Trabalho	Manual	Semiautomática	Automática	Relatório	Qtd de Dimensões
Abordagem dirigida à métricas para avaliar a qualidade de dados abertos conectados (Behkamal et. al, 2013) (Behkamal et. al, 2013)	Sim	Não	Não	Não	5
Avaliação de qualidade dirigida à usuário do DBpedia (Zaveri, 2013)	Sim	Sim	Sim	Não	1
Luzzu – Um *framework* para avaliação da qualidade de dados conectados	Não	Não	Sim	Sim	10
QualityStamp – Avaliando a Qualidade de *Linked Datasets* para Aplicações de Domínio Específico (Travassos, 2014)	Não	Não	Sim	Sim	6
Abordagem Proposta – *Linked Dataset Evaluation*	Sim	Sim	Não	Sim	8

Fonte: Elaboração própria

4 Uma abordagem semiautomática dirigida à métricas para avaliação da qualidade de *datasets* conectados

A proposta apresentada nesta dissertação é uma abordagem para avaliação da qualidade de *datasets* conectados disponíveis na *Web*, com base em dimensões de qualidade no contexto de melhores práticas para publicação de dados conectados. Esta abordagem, como descrito nas seções anteriores, oferece uma solução que beneficia publicadores e consumidores de *datasets* conectados, pois executa, com base na abordagem GQM a avaliação semiautomática, reduzindo a carga de trabalho do usuário.

Nas subseções deste capítulo, será explicado mais detalhamente o desenvolvimento da proposta, iniciando com concepção da proposta na subseção 4.1. Em seguida, na seção 4.2, será mostrada a implementação da proposta desenvolvida.

Como visto nos capítulos anteriores, *datasets* disponibilizados na *Web* sofrem de vários problemas de qualidade de dados, como problemas representacionais, inconsistências e questões de interoperabilidade (Hogan et al, 2012). Esses problemas dificultam a interpretação dos dados em casos de uso particulares e afetam a qualidade dos resultados, onde esta é propagada nos *datasets* agregados.

Um particular desafio da área é analisar a qualidade de *datasets* conectados e deixá-la explícita (Zaveri et al., 2013). Com isso, essa seção tem como finalidade apresentar uma abordagem para avaliação de *datasets* conectados. A abordagem foi desenvolvida com base no GQM, a qual é composta por dimensões de qualidade selecionadas de acordo com a similaridade de conceito com as melhores práticas para publicação de dados na *Web* e dados conectados disponibilizadas pelo W3C.

4.1 Seleção de dimensões e métricas

Nesta seção será apresentado o processo de seleção das dimensões e métricas que fazem parte do sistema proposto neste trabalho.

Inicialmente houve o levantamento de dimensões de qualidade utilizadas pela literatura para avaliar a qualidade de *datasets* conectados. Desta forma, foram identificadas 26 dimensões de qualidade de dados, as quais foram descritas por Zaveri (2012) em um survey, como mostrado na Figura 4.

Posteriormente, melhores práticas para a publicação de dados na *Web* foram identificadas, mediante o documento "*Data on the Web Best Practices*", bem como melhores práticas para a publicação de dados conectados, utilizando o documento "*Best Practices for Publishing Linked Data*", tais melhores práticas serão detalhadas na seção 2.2.

Ao analisar as dimensões, suas métricas e as melhores práticas, foi percebido que há uma similaridade de conceitos entre as mesmas, por exemplo, a dimensão licenciamento (*licen-*

sing) busca avaliar se uma licença legível por humanos está disponível no *dataset*, bem como uma licença legível por máquina (Zaveri, 2012), e neste sentido, a melhor prática BP 4 do documento "*Data on the Web Best Practices*" (Hyland et. Al, 2014) orienta que seja provido um *link* ou cópia da licença que controla o uso dos dados. Com o objetivo de identificar a similaridade de conceitos entre as dimensões, métricas e melhores práticas, as mesmas foram estudadas. Desta forma, foram selecionadas as dimensões métricas e melhores práticas que tinham convergência de conceitos, este processo de seleção está descrito nas subseções posteriores. A Figura 4 mostra, de forma geral, as dimensões que foram selecionadas (circuladas de vermelho).

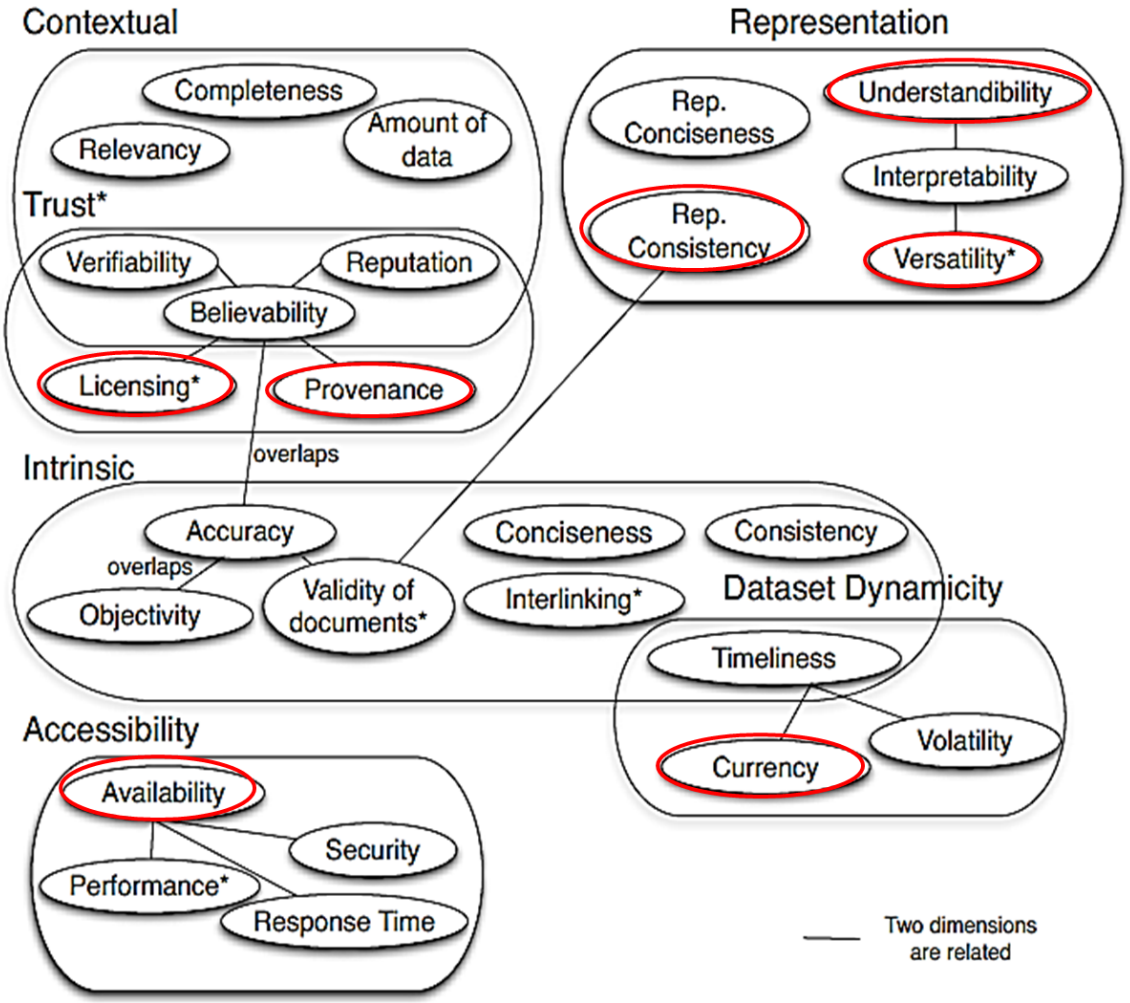

Figura 4 – Dimensões identificadas por Zaveri (2012)
Fonte: Zaveri (2012)

Seleção da dimensão disponibilidade

A dimensão Disponibilidade (*Availability*), que busca avaliar o grau em que a informação está fácil de ser obtida, possui sete métricas (Tabela 13), dentre elas a métrica "Acessibili-

dade do *endpoint* SPARQL", que consiste em, checar se o *endpoint* SPARQL responde as consultas.

Ao analisar a Melhor Prática 8: Prover máquina de acesso para os dados (Hyland et. al 2014), identificou-se que, disponibilizar um *endpoint* SPARQL é uma orientação desta melhor prática, sendo considerado um tipo de máquina de acesso, neste sentido outros tipos de máquinas de acesso para dados são orientados a serem disponibilizados, como API *RESTful* e arquivo de *download*. Devido ao *endpoint* SPARQL estar presente tanto na referida dimensão, quanto na melhor prática, as mesmas foram selecionadas e as máquina de acesso API *RESTful* e arquivo de *download* foram consideradas na abordagem de avaliação deste trabalho e inseridas na dimensão, como métricas, complementando-a.

Tabela 13 – Dimensão disponibilidade

Dimensão	***Métrica***
Disponibilidade	Acessibilidade do servidor Acessibilidade do *endpoint* SPARQL Acessibilidade de RDF *dumps* questões de De-referenciamento Disponibilidade de dados não estruturados Tipos de conteúdo incorretamente relatados Não de-referenciamento de back-links

Fonte: Zaveri (2012)

Outra forma de promover a disponibilidade do *dataset*, é torná-lo fácil de ser acessado. Para isso, o W3C fornece a orientação prover *bulk download* [BP 20; (Lóscio et. al 2016)], que consiste em disponibilizar um arquivo contendo *dataset* completo, ou seja, se o *dataset* é disponibilizado desde o ano de 1990 até 2016, o *bulk download* corresponde aos dados de todos esses anos em um só arquivo. No entanto, não existe métrica para avaliar a existência de *bulk download*, com isso, esta orientação foi adicionada a solução proposta.

Seleção da dimensão Proveniência

A dimensão Proveniência, definida como, o grau em que as informações sobre a origem dos dados estão disponíveis, é composta por 11 (onze) métricas, dentre elas, "Indicação de metadados sobre o *dataset*", que consiste em verificar a existência do título, conteúdo e URI do *dataset* (Zaveri, 2013). Considerando esta métrica, algumas melhores práticas, como "Melhor prática 1: Prover metadados" (Lóscio et. al 2016), "Melhor prática 2: Prover metadados descritivos" (Lóscio et. al 2016) e "Melhor prática 6: Provê informações de Proveniência do dado" (Hyland et. Al, 2014) orientam a disponibilização de metadados sobre o *dataset*, coincidindo com a métrica em questão.

Com isso, a métrica "Indicação de metadados sobre o *dataset*" foi selecionada para fazer parte da abordagem desenvolvida, e as melhores práticas "Melhor prática 2: Prover metadados descritivos" (Lóscio et. al 2016) e "Melhor prática 6: Provê informações de Proveniência

do dado" (Hyland et. Al, 2014) foram convertidas em métricas. A orientação "Prover metadados" não foi convertida em métrica explicitamente porque está contemplada na orientação "Prover metadados descritivos".

Tabela 14 – Dimensão Proveniência

Dimensão	***Métrica***
Proveniência	indicação de metadados sobre o *dataset*
	computação personalizada de recomendações confiáveis
	computação da confiabilidade de declarações RDF
	detectar a confiabilidade e a credibilidade de uma pessoa (publicador)
	computar a confiabilidade de uma entidade
	precisão da computação de confiabilidade entre duas entidades
	aquisição de conteúdos confiáveis a partir de usuários
	detecção de confiabilidade e credibilidade de uma fonte de dados
	Atribuição valores de confiança para dados/-fontes/regras
	Determinação do valor da confiança para dados

Fonte: Zaveri (2012)

Seleção da dimensão Licenciamento

A dimensão Licenciamento, que objetiva possibilitar que consumidores usem dados sob termos legais, busca avaliar se o *dataset* explicita uma licença sobre a qual o conteúdo pode ser usado. Com isso, esta licença possui cinco métricas, dentre elas, "Indicação de licença legível por humanos" e "Indicação de licença legível por máquinas", que ao analisar a melhor prática "Especificar uma licença apropriada" [BP 4; (Hyland et. Al, 2014)], percebemos que estas métricas são orientações contidas nesta melhor prática. Devido a esta correlação, esta dimensão e estas métricas foram selecionadas.

Tabela 15 – Dimensão Licenciamento

Dimensão	***Métrica***
Licenciamento	Indicação de uma licença legível por máquina
	Indicação de uma licença legível por humanos
	Permissões de uso para conjuntos de dados
	Indicação de atribuição
	Indicação de Copyleft ou ShareAlike

Fonte: Zaveri (2012)

Seleção da dimensão Compreensibilidade

A dimensão Compreensibilidade, que visa avaliar o grau em que os dados são facilmente compreendidos pelos consumidores de informações, possui as métricas, "indicação de vocabulários usados no *dataset*" e "Prover quadros de mensagens" (Tabela 16). Desta forma, considerando a métrica "indicação de vocabulários usados no *dataset*" foi identificado que

ela não está relacionada a nenhuma melhor prática, porque o objetivo desta métrica é apenas identificar uma lista de vocabulários usados, no entanto, há duas melhores práticas que envolvem a utilização de vocabulários padrão, que são, "Melhor prática 17: reuso de vocabulários"(Lóscio et. al 2016) e "Melhor prática 06: Vocabulários padrão"(Hyland et. Al, 2014). Devido à importância do uso de vocabulários, bem como, sua listagem explícita, esta métrica foi selecionada.

A métrica “Prover quadros de mensagens”, busca verificar a eficácia e a eficiência da utilização da lista de discussão e/ou os quadros de mensagens. Neste sentido, quadros de mensagens são os locais que permitem o usuário fornecer *feedbacks* sobre o *dataset*. Capturar *feedbacks* dos usuários, ou seja, os consumidores de dados, é uma da orientação da melhor prática BP 30: “Capturar *feedback* de consumidores de dados” (Lóscio et. al 2016). Com isso, esta melhor prática foi selecionada e considerada uma métrica.

Uma vez que os *feedbacks* são capturados, é necessário deixá-los explícitos para a comunidade. Desta forma, melhor prática que orienta esta ação é a BP 31: “Prover informações sobre a disponibilidade dos *feedback* dos usuários” [BP 31; (Lóscio et. al 2016)]. Esta melhor prática, foi também selecionada e utilizada como métrica. As duas métricas criadas complementam a dimensão compreensibilidade. Vale ressaltar, que a métrica “Prover quadros de mensagens” está contemplada na nova métrica “Capturar *feedback* de consumidores de dados”.

Tabela 16 – Dimensão Compreensibilidade

Dimensão	***Métrica***
Compreensibilidade	Nome legível por humanos de classes, propriedades e entidades, providas através de rdfs: label
	Indicação de metadados sobre um conjunto de dados
	Representações de-referenciadas: prover metadados legíveis por humanos
	Indicação de um ou mais exemplos de URIs
	Indicação de uma expressão regular que corresponde aos URIs de um conjunto de dados
	Indicação de um exemplo de consulta SPARQL
	Indicação dos vocabulários utilizados no conjunto de dados
	Fornecimento de quadros de mensagens e listas de discussão

Fonte: Zaveri (2012)

Seleção da dimensão Atualidade

A dimensão Atualidade refere-se à idade do dado, que é a diferença entre a data atual e a data de última modificação do dado. Para que esta dimensão seja executada, é necessário disponibilizar a data de modificação do *dataset*. Caso a data modificação não esteja disponível, foi considerada também a data de criação do *dataset*.

Tabela 17 – Dimensão Atualidade

Dimensão	*Métrica*
Currency	Atualidade das declarações
	Atualidade da fonte de dados
	Idade dos dados

Fonte: Zaveri (2012)

Seleção da dimensão Versatilidade

A dimensão Versatilidade, definida como "representações alternativas do dado e o manuseio dos mesmos", possui a métrica "Verificar se o *dataset* é fornecido em diferentes formatos de serialização", que coincide com a melhor prática "Prover dados e recursos em diferentes formatos de serialização" [BP 17; (Lóscio et. al 2016)]. Com isso, esta métrica foi selecionada.

Outro fator interessante, que facilita a compreensão e utilização do *dataset* por pessoas de diversos países, é disponibilizá-lo em outras linguagens, desta forma, a métrica "Prover o *dataset* em várias linguagens" referente a esta necessidade, foi selecionada.

Tabela 18 – Dimensão Versatilidade

Dimensão	*Métrica*
Versatilidade	Fornecimento de dados em diferentes formatos de serialização
	Fornecimento de dados em vários idiomas
	Aplicação da negociação de conteúdo
	Diferentes maneiras de acessar os dados

Fonte: Zaveri (2012)

Seleção da dimensão Consistência representacional

A dimensão Consistência representacional, refere-se ao uso de representações de dados existentes, como termos e vocabulários (Zaveri, 2012). Uma das métricas desta dimensão é "reusar vocabulários existentes", que coincide com prática "reusar vocabulário" [BP 17; (Lóscio et. al 2016). Com isso, esta dimensão e esta métrica foram consideradas.

Tabela 19 – Dimensão Consistência Representacional

Dimension	*Metric*
Consistência Representacional	Reutilização termos existentes
	Reuso de vocabulários existentes

Fonte: Zaveri (2012)

4.1.1 Dimensões e métricas para verificação e validação da qualidade de *datasets* conectados

Como visto anteriormente, Zaveri (2012) em sua revisão sistemática fez o levantamento de um conjunto de 26 dimensões e suas definições formalizadas, bem como das métricas para cada uma das dimensões. Na abordagem proposta neste trabalho, foram selecionadas sete dimensões de qualidade, as quais possuem similaridade de conceitos com as "Melhores Práticas Para Publicação de Dados Conectados" e as "Melhores Práticas de Dados na *Web* " providas pelo W3C, como descrito na seção anterior.

As dimensões e as métricas foram organizadas de acordo com a abordagem GQM. Para cada métrica do GQM foram associados critérios que correspondem aos recursos que serão avaliados na métrica, como, API *Json* (*dataset* no formato *Json*), Arquivo RDF (*dataset* no formato RDF) e HTML (*dataset* no formato HTML). Tais recursos foram selecionados para fazer parte da avaliação, porque são orientações da melhor prática "Prover Acesso Automatizado dos Dados" [BP 8; (Hyland et. al 2014)]. Desta forma, foram considerados como fundamentais no *dataset*.

As métricas do GQM contém o objeto que será avaliado em cada recurso, como uma licença, informações de proveniência, lista de vocabulários usados, etc. A seleção das métricas está descrita na seção 4.1. Desta forma, o GQM está configurado da seguinte maneira, tendo como objetivo (Goal):

Analisar *datasets conectados,* ***com a intenção de*** *avaliá-los com respeito a sua qualidade, através de critérios de qualidade da informação, com foco no* ***ponto de vista do*** *publicador ou consumidor de dados, no contexto de datasets disponíveis na Web.*

As questões, bem como, as métricas serão descritas a seguir:

A primeira questão tem como finalidade avaliar o *dataset* em relação à sua Disponibilidade, ou seja, medir o grau em que a informação está presente e pronta para usar (Zaveri, 2012). Desta forma, a primeira questão e suas métricas configuram-se da seguinte maneira:

Q1: Qual a qualidade do *dataset* considerando a dimensão Disponibilidade?

Tabela 20 – Métricas da dimensão Disponibilidade

Nome da Métrica	***Tópicos de verificação da métrica***	***Notas***	***Cálc. M***	***Nota questão***
M1. Verificar a existência de Acesso Automatizado dos Dados [BP 8; (Hyland et. al 2014)]	A1. API	1 ou 0	$\frac{\sum_{i=1}^{n} A_i}{n}$	$\frac{\sum_{i=1}^{n} M_i}{n}$
	A2. *endpoint* SPARQL	1 ou 0		
	A3. *Download* de arquivo contendo o *dataset*: RDF (formato selecionado para este trabalho)	1 ou 0		
M2. Verificar a acessibilidade do *endPoint* SPARQL (ZAVERI, 2014)	B. Verificar se o servidor responde a consultas SPARQL (ZAVERI, 2014)	QCC e QCF	QCF/QCC	
M3. Verificar se é provido *bulk download* [BP 20; (Lóscio et. al 2016)]	C. Checar se pode ser feito *bulk download*	1 ou 0	0 ou 1	

Fonte: Elaboração própria

A primeira métrica da dimensão Disponibilidade "M1: Verificar a existência de Acesso Automatizado dos Dados [BP 8; (Hyland et. al 2014)]" possui 3 critérios, os quais são *A1. API*, *A2. endpoint SPARQL*, *A3. Download de arquivo contendo o dataset: RDF*. Desta forma, cada critério será avaliado quanto à sua existência, podendo receber a pontuação de 0 (caso o critério não seja identificado no *dataset*) ou 1 (caso exista).

Com isso, a nota da métrica (*Cálc. M*) será obtida através da média aritmética $\frac{\sum_{i=1}^{n} A_i}{n}$ dos três critérios. Onde o *n* corresponde ao total de critérios, que são 3, o *i* é o número do critério, que vai de 1 a *n* e A_i corresponde ao critério.

Posteriormente, será avaliada a métrica "M2. Verificar a acessibilidade do *endPoint* SPARQL (ZAVERI, 2014)", no sentido de identificar se o *endPoint* SPARQL responde à três consultas SPARQL que serão submetidas. Desta forma, a métrica M2 possui um critério "B. Verificar se o servidor responde a consultas SPARQL (ZAVERI, 2014)".

A nota do critério B será obtida mediante as consultas que serão submetidas (QCF – Quantidade de Consultas Feitas) e as consultas que resultaram em uma resposta sem erros (QCC – Quantidade de Consultas Conseguidas). Com isso, a nota da métrica M2 consiste no percentual da divisão entre as consultas conseguidas (QCC) e quantidade de consultas feitas (QCF).

Calculado a métrica M2, a próxima métrica a ser avaliada será "M3. Verificar se é provido *bulk download* [BP 20; (Lóscio et. al 2016)]", que busca identificar se o *dataset* completo está disponível para ser baixado. A métrica M3 possui um critério "C. Checar se pode ser feito *bulk download*" que pode receber duas notas, que são, 0 (caso o *bulk download* não seja identificado no *dataset*) e 1 (caso o *bulk download* seja identificado no *dataset*). A nota da métrica M3 será o percentual da nota do critério C.

Desta forma, após calculado a nota das métricas M1, M2 e M3, será calculado a nota da questão Q1, que consiste na média aritmética $\frac{\sum_{i=1}^{n} M_i}{n}$ das notas das referidas métricas

(Tabela 20). Onde o *n* corresponde ao total de métricas, que são 3, o *i* é o número da métrica, que vai de 1 a *n* e M_i corresponde a métrica.

A segunda questão tem como objetivo avaliar o *dataset* em relação as suas informações de proveniência, ou seja, tem como finalidade identificar informações como título do *dataset*, descrição do responsável pela publicação, etc. essas informações focam em como representar, gerenciar e usar informações sobre a origem da fonte. Proveniência ajuda a descrever entidades aumentando a confiabilidade, concernindo assim a avaliação da autenticidade e permitindo a reprodução (Zaveri, 2012). A segunda questão e duas métricas apresentam-se da seguinte maneira:

Q2: Qual a qualidade do *dataset* considerando Proveniência?

Desta forma, a questão Q2 possui duas métricas, como apresentado nas tabelas 21 e 22. A métrica *M1. Verificar a existência de metadados descritivos sobre o dataset (ZAVERI, 2014) [BP 1 e BP2 (Lóscio et. al 2016)]* possui três critérios que são *"X1. Título do dataset (dct:title) (BP2)"*, *X2. Descrição do dataset (dct:description)*, e *X3. Data de emissão (dct: issued)*. Os três critérios serão avaliados em três recursos, as quais são, *"A1. Verificação no dataset obtido da API"*, que receberá a pontuação 1 caso o critério avaliado exista neste recurso e 0 caso contrário (o mesmo vale para os critérios A2 e A3), *"A2. HTML busca identificar"* e *A3. recursos RDF*.

Após a avaliação e estabelecido as notas de cada recurso associado a cada critério, a nota dos critérios serão obtidas. Desta forma, a nota do critério X1 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} A_i}{n}$; A nota do critério X2 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} B_i}{n}$; e a nota do critério X3 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} C_i}{n}$. Onde o *n* corresponde ao total de recursos, que são 3 em cada critério, o *i* é o número do recurso, que vai de 1 a *n* e A_i, B_i e C_i correspondem aos recursos.

Com isso, após calculada a nota de cada critério (X1, X2 e X3), será calculado a nota da métrica M1 (*Cálc. M*) que consiste em uma média aritmética das notas de cada critério $\frac{\sum_{i=1}^{n} X_i}{n}$. Onde o *n* corresponde ao total de critérios, que são 3, o *i* é o número do critério, que vai de 1 a *n* e X_i, corresponde aos critérios.

Tabela 21 – Métrica 01 da dimensão Proveniência

M1. Verificar a existência de metadados descritivos sobre o dataset (ZAVERI, 2014) [BP 1 e BP2 (Lóscio et. al 2016)]			
Tópicos de verificação da métrica		***Notas***	***Cálc. M***
X1. Título do *dataset* (dct:title) (BP2)	A1. Verificação no *dataset* obtido da API	1 ou 0	$\frac{\sum_{i=1}^{n} X_i}{n}$
	A2. HTML	1 ou 0	
	A3. recursos RDF	1 ou 0	
X2. Descrição do *dataset* (dct:description)	B1. Verificação no *dataset* obtido da API	1 ou 0	
	B2. HTML	1 ou 0	
	B3. Recursos RDF	1 ou 0	
X3. Data de emissão (dct: issued)	C1. Verificação no *dataset* obtido da API	1 ou 0	
	C2. HTML	1 ou 0	
	C3. Recursos RDF	1 ou 0	

Fonte: Elaboração própria

Tabela 22 – Métrica 02 da dimensão Proveniência

M2. Verificar a existência de informações de Proveniência do dataset [BP 6; (Lóscio et. al, 2016)]			
Tópicos de verificação da métrica		***Notas***	***Cálc. M***
X1. Entidade responsável por tornar o *dataset* disponível (*dct:publisher*)	D1. Verificação no *dataset* obtido da API	1 ou 0	$\frac{\sum_{i=1}^{n} X_i}{n}$
	D2. HTML	1 ou 0	
	D3. recursos RDF	1 ou 0	
X2. *E-mail* do publicador (*foaf:mbox*)	E1. Verificação no *dataset* obtido da API	1 ou 0	
	E2. HTML	1 ou 0	
	E3. Recursos RDF	1 ou 0	
X3. Nome da entidade ou pessoa responsável por criar o *dataset*	F1. Verificação no *dataset* obtido da API	1 ou 0	
	F2. HTML	1 ou 0	
	F3. Recursos RDF	1 ou 0	
X4. Nome da entidade ou pessoa responsável por criar o *dataset*	G1. Verificação no *dataset* obtido da API	1 ou 0	
	G2. HTML	1 ou 0	
	G3. Recursos RDF	1 ou 0	
X5. *E-mail* da entidade ou pessoa responsável por criar do *dataset*	H1. Verificação no *dataset* obtido da API	1 ou 0	
	H2. HTML	1 ou 0	
	H3. Recursos RDF	1 ou 0	

Fonte: Elaboração própria

A métrica *M2. Verificar a existência de informações de Proveniência do dataset [BP 6; (Lóscio et. al, 2016)]* possui três critérios, os quais são *"X1. Entidade responsável por tornar o dataset disponível (dct:publisher)"*, *"X2. E-mail do publicador (foaf:mbox)"*, *"X3. Nome da entidade ou pessoa responsável por criar o dataset"*, *"X4. Nome da entidade ou pessoa responsável por criar o dataset"* e *"X5. E-mail da entidade ou pessoa responsável por criar do dataset"*.

Desta forma, os critérios X1, X2, X3, X4 e X5 serão avaliados também em três recursos, que são *"Verificação no dataset obtido da API"*, *"HTML"*, e *"recursos RDF"*, assim como os critérios da métrica M1. Com isso, cada recurso receberá uma pontuação de 1, quando o critério avaliado existir no recurso e 0 caso contrário.

Com isso, após executada a avaliação dos critérios e também estabelecido as notas de cada recurso associado a cada critério, a nota dos critérios serão obtidas. Desta forma, a nota do critério X1 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} D_i}{n}$; a nota do critério X2 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} E_i}{n}$; a nota do critério X3 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} F_i}{n}$; a nota do critério X4 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} G_i}{n}$; e a nota do critério X5 será obtida através da média aritmética dos recursos $\frac{\sum_{i=1}^{n} H_i}{n}$. Em cada média aritmética o *n* corresponde ao total de recursos, que são 3 em cada critério, o *i* é o número do recurso, que vai de 1 a *n* e D_i, E_i, F_i, G_i e $H_i$correspondem aos recursos.

Com isso, após calculada a nota de cada critério (X1, X2, X3, X4 e X5), será calculado a nota da métrica M2 que consiste em uma média aritmética das notas de cada critério $\frac{\sum_{i=1}^{n} X_i}{n}$. Onde o *n* corresponde ao total de critérios, que são 5, o *i* é o número do critério, que vai de 1 a *n* e X_i, corresponde a cada critério.

Concluído a nota de cada métrica, posteriormente será obtida a nota da questão Q2, mediante a seguinte equação:

$$\frac{\sum_{i=1}^{n} M_i}{n}$$

Na equação apresentada o *n* corresponde ao total de métricas, que são 2 (M1 e M2), o *i* é o número de métricas, que vai de 1 a *n* e M_i correspondem a cada métrica.

A terceira questão tem como finalidade avaliar o *dataset* em relação à Licença atribuída, ou seja, se a mesma está de fácil identificação tanto por humanos quanto por um agente de *software*. A terceira questão, bem como suas métricas, estão descritas na tabela abaixo.

Q3: Qual a qualidade do *dataset* considerando a dimensão Licenciamento?

Nome da Métrica	***Tópicos de verificação da métrica***	***Notas***	***Cálc. M.***	***Cálc. Q***
M1. Verificar se há a indicação de uma licença legível por humanos (Zaveri, 2014) [BP 4; (Hyland et. Al, 2014)]	A. Nome da licença em uma página	0 ou 1	0 ou 1	$\frac{\sum_{i=1}^{n} M_i}{n}$
M2. Verificar se há a indicação de uma licença legível por máquina (Zaveri, 2014) [BP 5; (Lócio et. al, 2016)]	B1. Nome da licença em uma API B2. Nome da licença nos recursos RDF	0 ou 1 0 ou 1	$\frac{\sum_{i=1}^{n} B_i}{n}$	

Fonte: Elaboração própria

Com isso, a métrica *M1. Verificar se há a indicação de uma licença legível por humanos (Zaveri, 2014) [BP 4; (Hyland et. Al, 2014)]* possui um critério, *a. Nome da licença em uma página*, que consiste em identificar uma licença legível por humanos, na página *Web* do *dataset*. Desta forma, o critério *a* possui duas notas, que são 1, caso a licença exista na página *Web* do *dataset* e 0 caso contrário. Após avaliar o referido critério, a nota da métrica M1 será a mesma nota do critério *A*. Para obter o percentual da métrica, a nota da mesma deve ser multiplicada por 100.

A métrica *M2. Verificar se há a indicação de uma licença legível por máquina (Zaveri, 2014) [BP 5; (Lócio et. al, 2016)]* possui dois critérios, que são *B1. Nome da licença em uma API*, que poderá ter nota 1, caso a licença legível por máquina seja encontrada na API e 0 caso contrário; e *B2. Nome da licença nos recursos RDF*, que pode ter a nota 1 caso a licença legível por máquina seja encontrada no arquivo RDF, e 0 caso não seja encontrada.

Com isso, após calculada a nota de cada critério, será calculado a nota da métrica M2 (*Cálc. M*) que consiste em uma média aritmética das notas de cada critério $\frac{\sum_{i=1}^{n} B_i}{n}$. Onde o *n* corresponde ao total de critérios, que são 2, o *i* é o número do critério, que vai de 1 a *n* e B_i, corresponde aos critérios.

Após executado o calculo de cada métrica, posteriormente será obtida a nota da questão Q3 (*Cálc. Q*), através a equação:

$$\frac{\sum_{i=1}^{n} M_i}{n}$$

Na equação para obtenção da nota Q3, o *n* corresponde ao total de métricas, que são 2 (M1 e M2), o *i* é o número de métricas, que vai de 1 a *n* e M_i correspondem a cada métrica.

A quarta questão tem como finalidade avaliar o *dataset* em relação à Compreensibilidade, ou seja, medir o grau em que os dados são facilmente compreendidos pelos consumidores de informações. A quarta questão, e suas métricas podem ser vistas na tabela a seguir.

Q4: Qual a qualidade do *dataset* considerando a dimensão Compreensibilidade?

Nome da Métrica	***Tópicos de verificação da métrica***	***Notas***	***Cálc. M.***	***Cálc. Q***
M1. Verificar se há mecanismo para capturar *feedback* de consumidores de dados [BP 30; (Lóscio et. al 2016)]	X1. Formulário de registro	0 ou 1	$\sum_{i=1}^{n} \frac{M_i}{n}$	$\sum_{i=1}^{n} \frac{M_i}{n}$
	X2. Formulário de Contato (apenas o e-mail)	0 ou 1		
	X3. Botões para informar a qualidade do dado (estrelas)	0 ou 1		
	X4. Caixas de comentários	0 ou 1		
M2. Verificar se existe informações sobre a disponibilidade dos *feedback* dos usuários [BP 31; (Lóscio et. al 2016)]	X5. Verificar os *feedbacks* estão disponíveis para os usuários	0 ou 1	0 ou 1	
M3. Verificar se é provido no *dataset* uma lista de vocabulários usados (ZAVERI, 2014)	X6. lista de vocabulários usados providos	0 ou 1	0 ou 1	

Fonte: Elaboração própria

A questão Q4 possui três métricas, que são, *M1. Verificar se há mecanismo para capturar feedback de consumidores de dados [BP 30; (Lóscio et. al 2016)]*, *M2. Verificar se existe informações sobre a disponibilidade dos feedbacks dos usuários [BP 31; (Lóscio et. al 2016)]* e *M3. Verificar se é provido no dataset uma lista de vocabulários usados (ZAVERI, 2014)*. A métrica M1, possui quatro critérios, que são, *X1. Formulário de registro*, *X2. Formulário de Contato (apenas o email)*, *X3. Botões para informar a qualidade do dado (estrelas)* e *X4. Caixas de comentários*, os quais podem, cada um, ter a nota 1, caso o critério avaliado seja encontrado na página *Web* do *dataset*, e 0 caso não seja encontrado.

Com isso, após calculada a nota de cada critério (X1, X2, X3 e X4), será calculado a nota da métrica M1 (*Cálc. M*) que consiste em uma média aritmética das notas de cada critério $\frac{\sum_{i=1}^{n} X_i}{n}$. Onde o *n* corresponde ao total de critérios, que são 4, o *i* é o número do critério, que vai de 1 a *n* e X_i, corresponde aos critérios.

A métrica M2 possui apenas um critério, que é *X5. Verificar os feedbacks estão disponíveis para os usuários*, que pode ter, também a nota 1, se os *feedbacks* estiverem disponíveis publicamente, e 0 se não estiver. Com isso, a nota da métrica será a mesma nota do respectivo critério.

A métrica M3 possui o critério *X6. lista de vocabulários usados providos* que será avaliado no recurso RDF. Desta forma, se uma lista de vocabulário for identificada no arquivo

RDF, o critério X6 receberá a nota 1, caso contrário, receberá 0. A nota da métrica M3 será igual a nota do critério X6.

Após calculado a nota das métricas M1, M2 e M3, será calculado a nota da questão Q4 (*Cálc. Q*), utilizando a média aritmética $\frac{\sum_{i=1}^{n} M_i}{n}$ das notas das referidas métricas. Onde o *n* corresponde ao total de métricas, que são 3, o *i* é o número da métrica, que vai de 1 a *n* e M_i corresponde a métrica.

A quinta questão tem como finalidade medir o *dataset* em relação à Atualidade, ou seja, refere-se à idade do dado, que é a diferença entre a data atual e a data de última modificação do dado, ou seja, refere-se à velocidade em que a informação é atualizada.

Q5: Qual a qualidade do *dataset* considerando a dimensão Atualidade?

Tabela 23 – Métrica 02 da dimensão Atualidade

M1. Verificação de informações sobre a atualidade de declarações (ZAVERI, 2014)			
Tópicos de verificação da métrica		***Notas***	***Cálc. M.***
A. Tempo de criação	F1. Verificação na API	0 ou 1	$\sum_{i=1}^{n} \frac{F_i}{n}$
	F2. Verificação na página *Web*	0 ou 1	
	F3. Verificação nos recursos RDF	0 ou 1	

Fonte: Elaboração própria

A métrica M1 possui o critério *"A. Tempo de criação"*, que consiste na data de criação do *dataset*. Este critério será avaliado em três recursos, que são, *F1. Verificação na API*, *F2. Verificação na página Web* e *F3. Verificação nos recursos RDF*. Desta forma, cada recurso verificado, com o intuito de identificar a data de criação do *dataset*, receberá a nota 1, se a referida data seja encontrada e 0 caso contrário. Após concluída a avaliação de cada recurso, a nota da métrica M1 (*Cálc. M*) será obtida utilizando a equação $\frac{\sum_{i=1}^{n} F_i}{n}$. Onde o *n* corresponde ao total de recursos, que são 3, o *i* é o número do recurso, que vai de 1 a *n* e F_i, corresponde aos recursos.

Desta forma, para obter a nota desta questão, deve-se utilizar a seguinte fórmula:

$$\sum_{i=1}^{n} \frac{M_i}{n}$$

A sexta questão, tem como finalidade avaliar o *dataset* quanto a sua Versatilidade, ou seja, busca identificar representações alternativas do *dataset*. A sexta questão e suas métricas podem ser vistas na tabela abaixo.

Q6: Qual a qualidade do *dataset* considerando a dimensão Versatilidade?

Tabela 24 – Métricas da dimensão Versatilidade

Nome da Métrica	***Notas M.***	***Cálc. Q.***
M1. Verificar se o *dataset* é fornecido em diferentes formatos de serialização (ZAVERI, 2014) [BP 22 (Lóscio et. al 2016)]	1 ou 0	$\frac{\sum_{i=1}^{n} M_i}{n}$
M2. Verificar se o *dataset* é fornecido em várias linguagens (ZAVERI, 2014)	1 ou 0	

Fonte: Elaboração própria

A questão Q6 possui duas métricas, que são, *M1. Verificar se o dataset é fornecido em diferentes formatos de serialização (ZAVERI, 2014) [BP 22 (Lóscio et. al 2016)]* e *M2. Verificar se o dataset é fornecido em várias linguagens (ZAVERI, 2014)*.

Desta forma, M1 receberá a nota 1 (Notas M.), se outros formatos de serialização, como por exemplo, *Json* e/ou RDF forem encontrados, e receberá 0 caso nenhum outro formato do *dataset* não seja encontrado. A métrica M2 também receberá a nota 1, caso o *dataset* seja disponível em outras linguagens e 0 caso seja disponível em apenas uma linguagem.

Desta forma, após estabelecido a nota das métricas M1 e M2, será calculado a nota da questão Q6 (*Cálc. Q*), que consiste na média aritmética $\frac{\sum_{i=1}^{n} M_i}{n}$ das notas das referidas métricas. Onde o *n* corresponde ao total de métricas, que são 2, o *i* é o número da métrica, que vai de 1 a *n* e M_i corresponde a métrica.

A sétima questão visa avaliar o *dataset* com relação á Consistência Representacional, ou seja, identificar o grau em que a estrutura da informação está em conformidade com a informação anteriormente retornada.

Q7: Qual a qualidade do *dataset* considerando a dimensão Consistência representacional?

Tabela 25 – Métricas da dimensão Consistência Representacional

Métrica	***Pontuação***
M1. Verificar se há reuso de vocabulários existentes [BP 17; (Lóscio et. al 2016)](ZAVERI, 2014)	1 ou 0

Fonte: Elaboração própria

A questão Q7 possui a métrica *M1. Verificar se há reúso de vocabulários existentes [BP 17; (Lóscio et. al 2016)](ZAVERI, 2014)*, que receberá a pontuação 1 se pelo menos um vocabulário existente for identificado como reusado, e 0 caso nenhum vocabulário existente

seja reusado. Devido a questão Q7 possuir apenas uma métrica, a nota da questão será igual a nota da métrica.

QG: Qual a qualidade do *dataset* considerando o escore Global?

Após concluída a avaliação, atribuição de notas e cálculos das questões descritas anteriormente, a qualidade (score global) do *dataset* pode ser medido com a seguinte equação:

$$\frac{\sum_{i=1}^{n} Q_i}{n}$$

Com isso, Onde o *n* corresponde ao total de questões, que são 7, o *i* é o número da questão, que vai de 1 a *n* e Q_i corresponde a questão.

Foi selecionada a média aritmética como fórmula principal para medição das questões, devido à mesma ser uma medida de tendência central e assim refletir quantitativamente uma nota considerada adequada à necessidade do presente trabalho. Outro aspecto que contribuiu para utilizá-la é que a mesma pode ser facilmente compreendida pelas pessoas, devido ser muito utilizada no cotidiano, facilitando a expansão do trabalho por parte de terceiros ou replicação do mesmo.

4.2 Implementação da abordagem semiautomática para avaliação da qualidade de *datasets* conectados

Na seção 4.1 foi apresentado a concepção da solução proposta neste trabalho para avaliação da qualidade de *datasets* conectados. A proposta foi implementada com a finalidade de contribuir com a redução da carga de trabalho do avaliador, bem como, com a transparência do resultado final da avaliação, através do relatório detalhado gerado pela ferramenta desenvolvida.

Com isso, nas seções a seguir será apresentado a implementação da abordagem semiautomática, sendo que, na seção na seção 4.2.1 será descrito os Cenários de funcionalidades da solução, ou seja, os casos de uso, na seção 4.2.2 será mostrado a expansão da ontologia *Data Quality Vocabulary* (DQV), na seção 4.2.3 será mostrado a Arquitetura do *LinkedDatasetEvaluation* (ferramenta semiautomática) e finalmente, na seção 4.2.4 será mostrado o sistema na visão do usuário.

A abordagem manual, que também é uma das contribuições deste trabalho, será mostrada somente no Apêndice A, pois nela é executado a mesma avaliação da abordagem descrita na subseção 4.2, no entanto, as atividades que foram executadas de forma automática serão executadas manualmente pelo usuário.

4.2.1 Cenários de funcionalidades da solução

Esta subseção tem como finalidade descrever os cenários de execução da avaliação de cada questão apresentada na descrição do GQM mostrada anteriormente, bem como de suas respectivas métricas, no sentido de explicitar quais métricas serão avaliadas automaticamente e quais métricas serão avaliadas pelo usuário.

Para um entendimento fácil dos cenários de funcionalidades da ferramenta, foram construídos diagramas de casos de uso utilizando a ferramenta *Astah*[1] utilizando uma licença gratuita para estudantes. Pois o Diagrama de Casos de Uso tem o objetivo de auxiliar a comunicação entre o analista e o cliente. Facilitando assim, a comunicação e o entendimento do sistema por parte dos envolvidos.

Com isso, nos Diagramas que serão descritos posteriormente existem dois atores, o **Usuário**, o qual representa o humano que irá fazer a parte da avaliação que é executada de forma manual e o **Sistema** que representará o computador que fará a parte da avaliação de forma automática. Nos Diagramas há um relacionamento que estabelece a relação entre as dimensões e suas respectivas métricas, chamado de *Include*. O relacionamento *Include* representa que um caso de uso (podemos considerar as métricas) é parte de outro (neste caso são as dimensões).

Desta forma, na Figura 5 é mostrado o Diagrama de Casos de Uso da dimensão Disponibilidade. Neste diagrama existem dois Atores, que são **Usuário** e **Sistema** que se relacionam com o Caso de Uso Avaliar *dataset*. Neste diagrama existem dois casos de uso que representam a dimensão Disponibilidade, que são Avaliar Disponibilidade - parte 1, que possui como parte os caso de uso **M1. Verificar a existência de Acesso Automatizado dos Dados [BP 8; (Hyland et. al 2014)]** e **M3. Verificar se é provido *bulk download* [BP 20; (Lóscio et. al 2016)]**, e Avaliar Disponibilidade - parte 2 que possui como parte **M2. Verificar a acessibilidade do *endPoint* SPARQL (ZAVERI, 2014)**.

Desta forma, o referido Diagrama representa que o usuário irá executar manualmente as métricas M1 e M3, por conseguinte, o Sistema (representando o computador) irá avaliar automaticamente a métrica M2 (caso necessário).

Da mesma forma, na Figura 6 é apresentado o Caso de Uso da dimensão Proveniência. Com isso, podemos ver que, semelhante ao Diagrama anterior, os Atores **Usuário** e **Sistema** interagem com o Caso de Uso **Avaliar *dataset*** (este comportamento se repete para todos os diagramas posteriores) e este interage com três Casos de Uso: **Página *web***, que representa a avaliação na página *web* do *dataset*, **Arquivo RDF**, que representa a avaliação do *dataset* no formato RDF (caso exista) e **API Json** que representa a avaliação do *dataset* no formato *Json* (também caso exista). Com isso, o Ator **Sistema** irá executar a avaliação nos casos de uso **Arquivo RDF** e **API *Json***, e o Ator **Usuário** irá executar a avaliação na **Página *web***.

Outra funcionalidade representada pelo referido diagrama é que os três casos de uso

[1] http://astah.net/

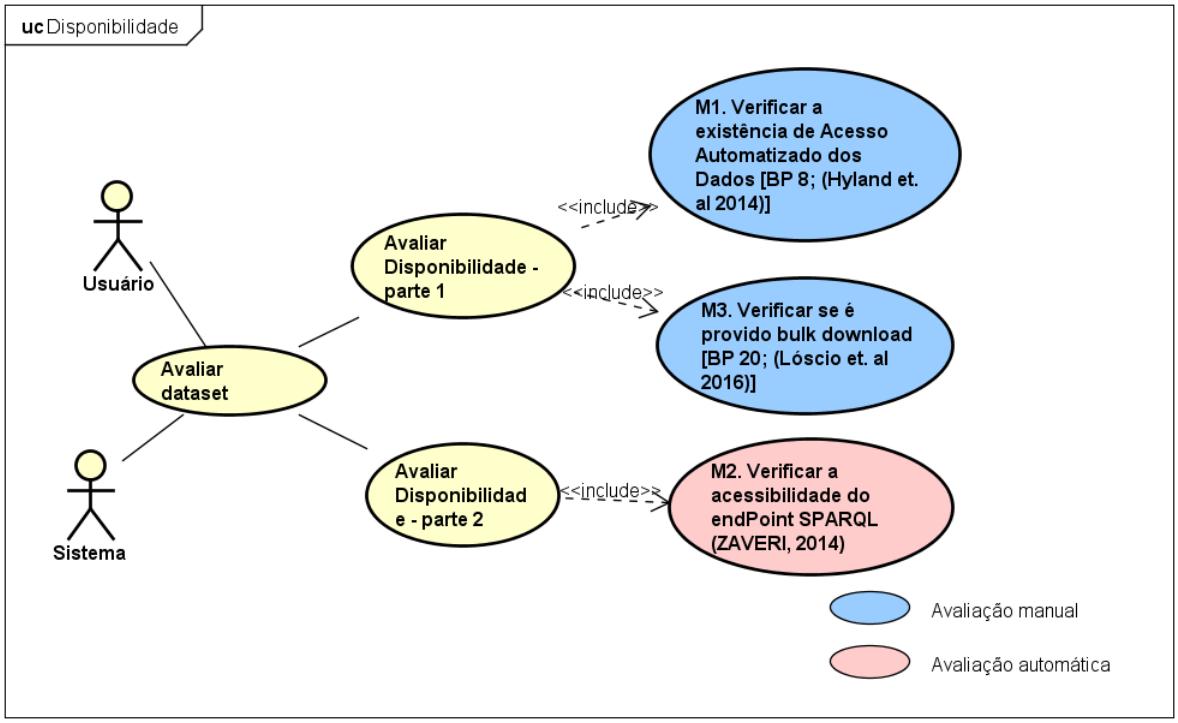

Figura 5 – Diagrama de casos de uso da dimensão Disponibilidade
Fonte: Elaboração própria

citados incluem o caso de uso **Avaliar Proveniência**, que representa a referida Dimensão, e este inclui dois casos de uso, que são, **M1. Verificar a existência de informações de Proveniência do *dataset* [BP 6; (Lóscio et. al, 2016)]** e **M2. Verificar a existência de metadados descritivos sobre o *dataset* (ZAVERI, 2014) [BP 1 e BP2 (Lóscio et. al 2016)]** que representam as referidas métricas contidas na Dimensão Proveniência.

Desta forma, este diagrama representa que o Usuário e o Sistema irão executar a mesma avaliação, ou seja, irão executar a mesma Dimensão e as mesmas métricas, no entanto, o Usuário irá executar a avaliação na Página *web* e o Sistema irá executar a avaliação do arquivo RDF e na API *Json*, ficando com a maior parte do trabalho.

O próximo Diagrama, apresentado na Figura 7, representa os requisitos da avaliação executando a dimensão Licenciamento. Desta forma, podemos perceber que neste, os Atore **Usuário** e **Sistema** interagem com o caso de uso **Avaliar *dataset***. Em seguida caso de uso **Avaliar *dataset*** interage com **Página *web***, **API *Json*** e **Arquivo RDF**.

Uma funcionalidade diferenciada deste diagrama é que o caso de uso **Página *web*** interage com o caso de uso **Avaliar Licenciamento - parte 1**, que possui como parte o caso de uso **M1. Verificar se há a indicação de uma licença legível por humanos (Zaveri, 2014) [BP 4; (Hyland et. Al, 2014)]** (será chamada de M1), e **API *Json*** e **Arquivo RDF** interagem com **Avaliar Licenciamento - parte 2**, que possui como parte **M2. Verificar se há a indicação de uma licença legível por máquina (Zaveri, 2014) [BP 5; (Lócio et. al, 2016)]** (será chamada de M2). Dessa forma, os casos de uso M1 e M2 foram criados com o intuito

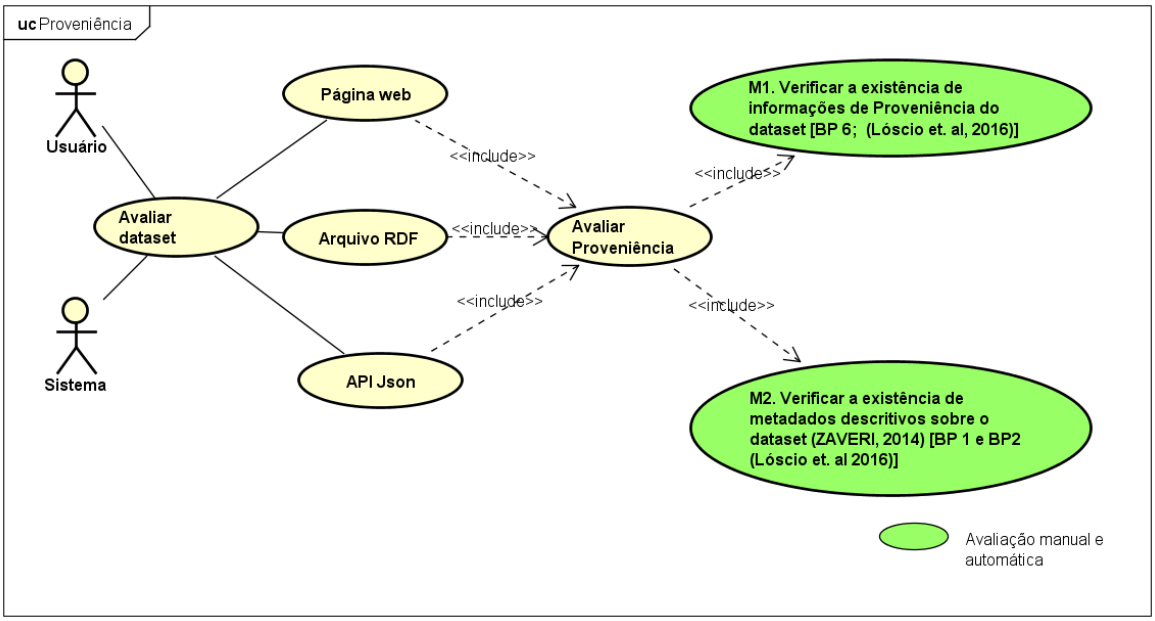

Figura 6 – Diagrama de casos de uso da dimensão Proveniência
Fonte: Elaboração própria

de representar as métricas da dimensão Licenciamento.

Portanto, este Diagrama representa que o Usuário irá executar a avaliação da página *web* do *dataset* utilizando a métrica M1 da dimensão Licenciamento, e o Sistema (computador) irá executar a avaliação da API *Json* e do Arquivo RDF de forma automática, utilizando a métrica M2 também da dimensão Licenciamento.

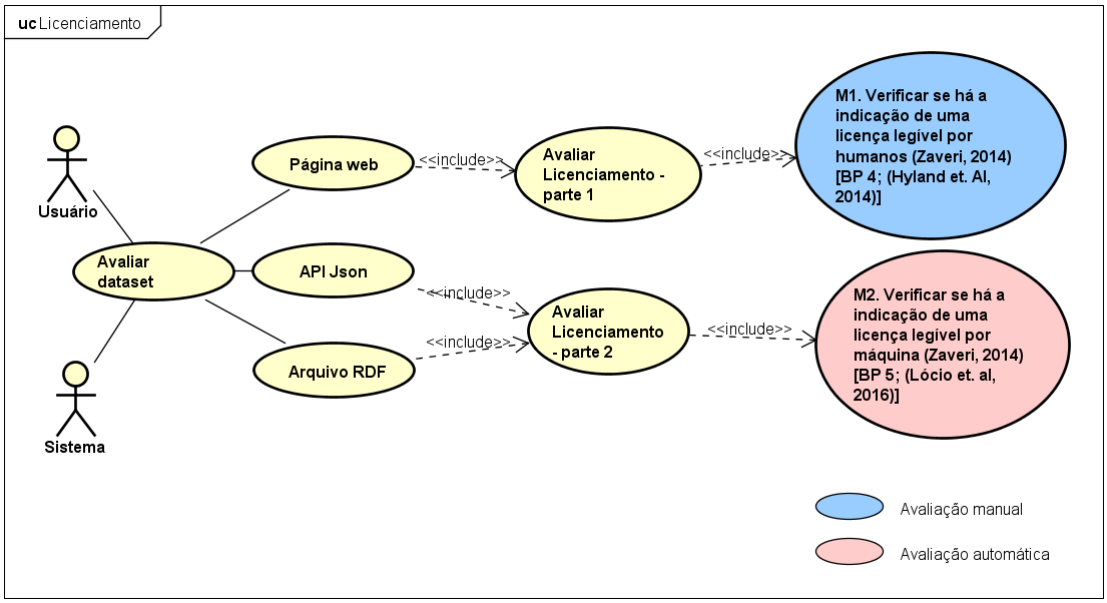

Figura 7 – Diagrama de casos de uso da dimensão Licenciamento
Fonte: Elaboração própria

Ademais, a próxima dimensão a ser representada é Compreensibilidade. Seu caso de uso,

assim como os anteriores possui **Usuário** e **Sistema** como Atores, e do mesmo modo, o caso de uso **Avaliar dataset** relacionado a eles. O caso de uso **Avaliar dataset** neste diagrama também está relacionado com os casos de uso **Página web, API *Json*** e **Arquivo RDF**.

Como podemos ver na Figura 8, há dois casos de uso que representam a dimensão Compreensibilidade, que são, **Avaliar Compreensibilidade - parte 1** que está relacionado com o caso de uso **Página *web***, e **Avaliar Compreensibilidade - parte 2** que está relacionado com os casos de uso **API Json** e **Arquivo RDF**. Além disso, o caso de uso **Avaliar Compreensibilidade - parte 1** possui como parte **M1. Verificar se há mecanismo para capturar *feedback* de consumidores de dados [BP 30; (Lóscio et. al 2016)]** (chamaremos apenas de M1) , **M2. Verificar se existe informações sobre a disponibilidade dos *feedback* dos usuários [BP 31; (Lóscio et. al 2016)]** (chamaremos de M2), através do relacionamento *include*. Ademais, o caso de uso **Avaliar Compreensibilidade - parte 2** possui como parte o caso de uso **M3. Verificar se é provido no *dataset* uma lista de vocabulários usados (ZAVERI, 2014)** (chamaremos de M3), também através do relacionamento *include*. Vale ressaltar que M1, M2 e M3 representam as métricas da dimensão Compreensibilidade.

Em síntese, o diagrama da Figura 8, representa a avaliação do *dataset* considerando a dimensão Compreensibilidade, onde o usuário irá avaliar a página *web* utilizando as métricas M1, M2 e M3, que são executadas de forma manual, e o computador irá avaliar a API *Json* e o Arquivo RDF utilizando a métrica M3 automaticamente. Como se pode perceber, a métrica M3 será utilizada de duas formas (manual e automática).

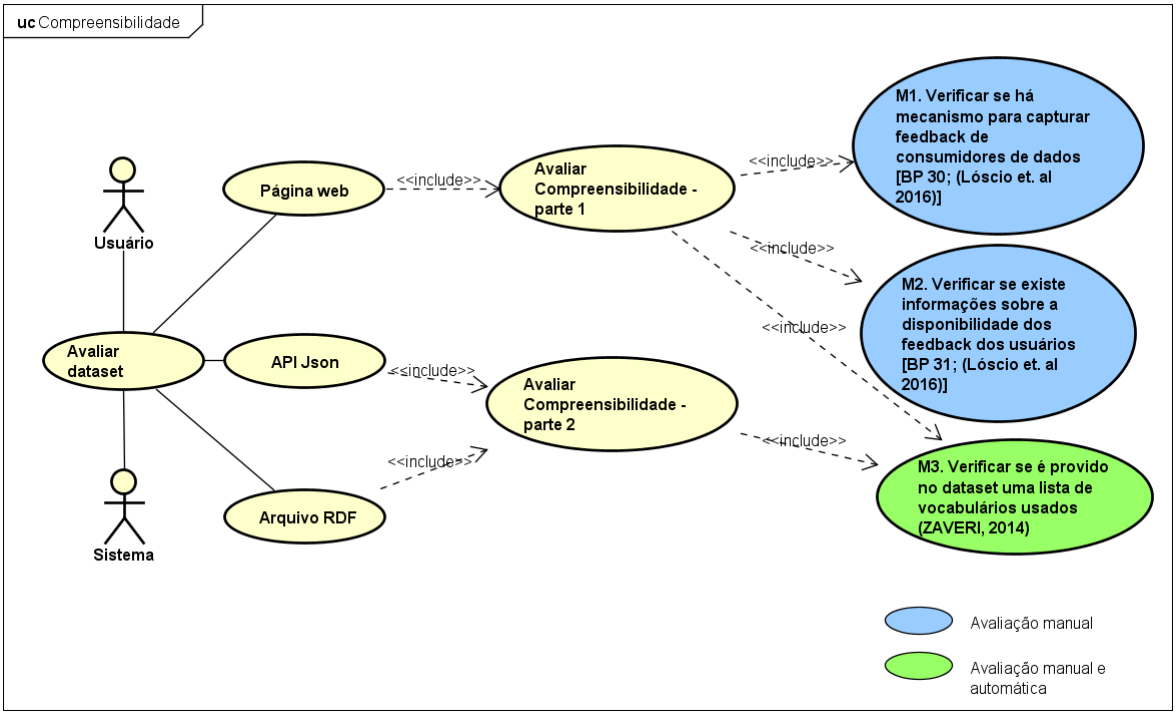

Figura 8 – Diagrama de casos de uso da dimensão Compreensibilidade
Fonte: Elaboração própria

O próximo diagrama, ilustrado na Figura 9 representa a avaliação do *dataset* conside-

rando a dimensão Atualidade. Com isso, da mesma forma como os anteriores, os Atores **Usuário** e **Sistema** se relacionam com o caso de uso **Avaliar dataset** que se relaciona com os casos de uso **Página *web*, API Json** e **Arquivo RDF**.

Por outro lado, a dimensão Atualidade está representada pelo caso de uso **Avaliar Atualidade**, que possui como parte **M1. Verificar se existe metadados de parâmetros de localidade [BP 3; (Lóscio et. al 2016)]** (chamada também de M1) e **M2. Verificação de informações sobre a atualidade de declarações (ZAVERI, 2014)** (chamada também de M2). Lembrando que os casos de uso M1 e M2 representam as métricas da dimensão Licenciamento.

Desta forma, o referido diagrama quer dizer que o Usuário (humano) irá avaliar a página *web* do *dataset* utilizando as métricas M1 e M2, e o Sistema (Computador) irá avaliar a API *Json* e o Arquivo RDF também utilizando as métricas M1 e M2.

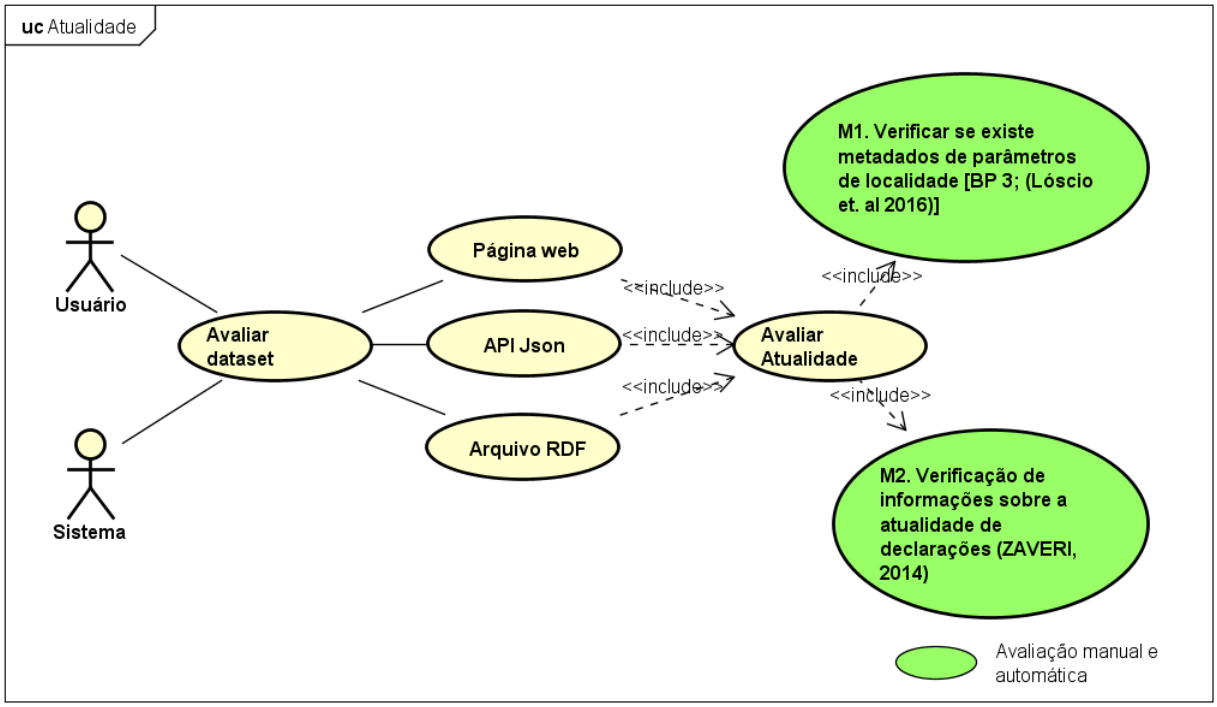

Figura 9 – Diagrama de casos de uso da dimensão Atualidade
Fonte: Elaboração própria

Com isso, a próxima dimensão representada é Versatilidade, ilustrada na Figura 10. Desta forma, podemos ver que neste Diagrama há apenas um Ator, que é o **Usuário**, o qual está relacionado com o caso de uso **Avaliar *dataset***. Por conseguinte, o caso de uso **Avaliar *dataset*** está relacionado com o caso de uso **Avaliar Versatilidade**, o qual possui como parte os casos de uso **M1. Verificar se o *dataset* é fornecido em diferentes formatos de serialização (ZAVERI, 2014) [BP 22 (Lóscio et. al 2016)]** e **M2. Verificar se o *dataset* é fornecido em várias linguagens (ZAVERI, 2014)** através do relacionamento *include*.

Em suma, este diagrama quer dizer que a dimensão Versatilidade será utilizada para avaliar o *dataset* apenas de forma manual, pelo usuário.

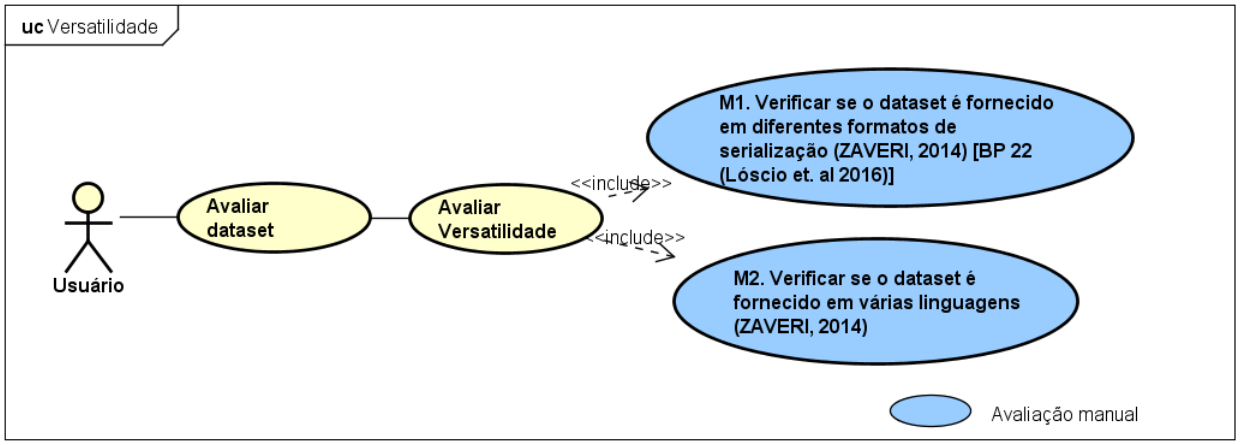

Figura 10 – Diagrama de casos de uso da dimensão Versatilidade
Fonte: Elaboração própria

Ademais, a ultima dimensão a ser representada é Consistência Representacional, como mostrado no Diagrama contido na Figura 11. Como podemos ver, neste diagrama, também há apenas um Ator, que é **Sistema**, o qual interage com o caso de uso **Avaliar *dataset***. Posteriormente, o caso de uso **Avaliar *dataset*** interage com o caso de uso **Arquivo RDF**.

Desta forma, representando a dimensão Consistência representacional, foi criado o caso de uso **Avaliar Consistência representacional** que possui como parte **M1. Verificar se há reuso de vocabulários existentes [BP 17; (Lóscio et. al 2016)](ZAVERI, 2014)** (chamada de M1). Com isso, M1 está representando a métrica da referida dimensão.

Portanto, podemos perceber que a dimensão Consistência Representacional será executada apenas automaticamente pelo Sistema no arquivo RDF, ou seja, pelo computador.

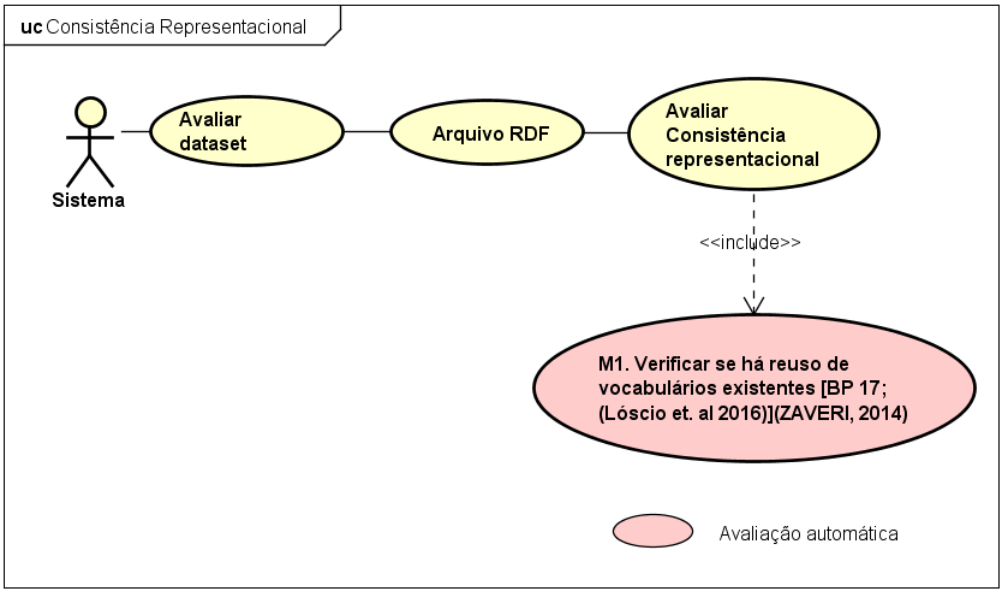

Figura 11 – Diagrama de casos de uso da dimensão Consistência Representacional
Fonte: Elaboração própria

4.2.2 Reuso e expansão da ontologia *Data Quality Vocabulary* (DQV)

O *Data Quality Vocabulary* é um vocabulário criado pelo W3C que descreve conceitos relacionados à qualidade de dados (Debattista et. al, 2015). No entanto, o referido vocabulário não provê uma definição formal e completa sobre qualidade, mas provê informações de fundamental importância para *datasets* e pode ser adaptado para propósitos específicos.

Com a finalidade de agregar as dimensões, métricas e critérios das métricas à conceitualização de qualidade fornecida pelo W3C, o vocabulário DQV foi reutilizado e expandido para a nossa necessidade. Devido às mudanças feitas no vocabulário original, o vocabulário novo vocabulário resultante passou a ser chamado de Vocabulário de *Datasets* Conectados (VDC). O vocabulário VDC pode ser baixado no *link*: `https://dl.dropboxusercontent.com/u/57888679/vdc7.owl`.

Portanto, foi adicionado ao vocabulário as classes Criteria para conceitualizar os critérios de cada métrica, CriteriaEvaluation que define o valor de cada critério e *DataSet* que conceitualiza o *dataset* que será utilizado na avaliação de qualidade. Na Figura 12 é mostrado a ontologia no programa Protégé[2]

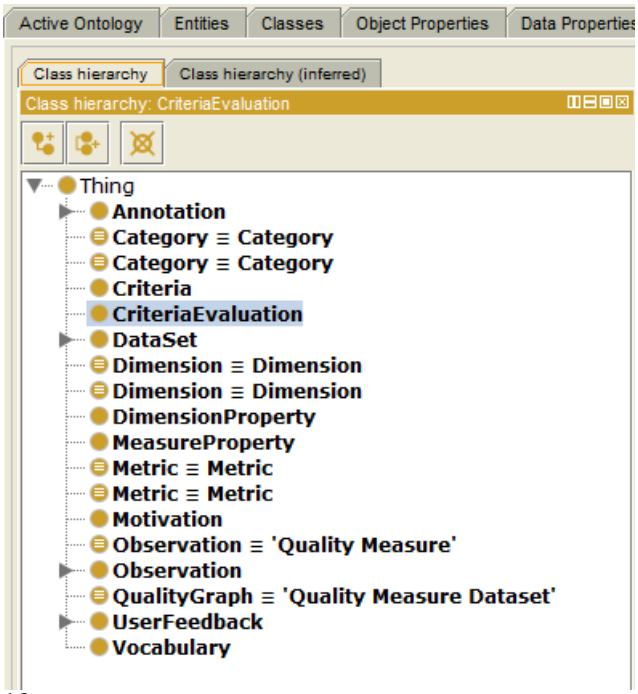

Figura 12 – Ontologia Vocabulário de *Datasets* Conectados (VDC)
Fonte: Elaboração própria

4.2.2.1 Especificações do vocabulário

Nesta seção será abordado sobre as especificações técnicas do vocabulário VDC, como as definições das classes e propriedades. Devido ao vocabulário DQV ser reutilizado neste

[2]http://protege.stanford.edu/

trabalho, parte desta documentação pode ser vista em Debattista et. al (2015).

Com isso, começaremos pela Classe Categoria (Class Category) que representa um grupo de dimensões de qualidade. As Categorias das dimensões de qualidade, segundo Zaveri (2012) são, Contextual, Confiabilidade, Intrínseca, Acessibilidade, Representacional e Dinamicidade do *dataset* (descritas na seção 13).

Desta forma, na Figura 13 podemos ver que existem duas Class Category tem a propriedade EquivalentTo, ou seja, as duas classes são equivalentes ou iguais. Essa implementação foi herdada da ontologia DQV.

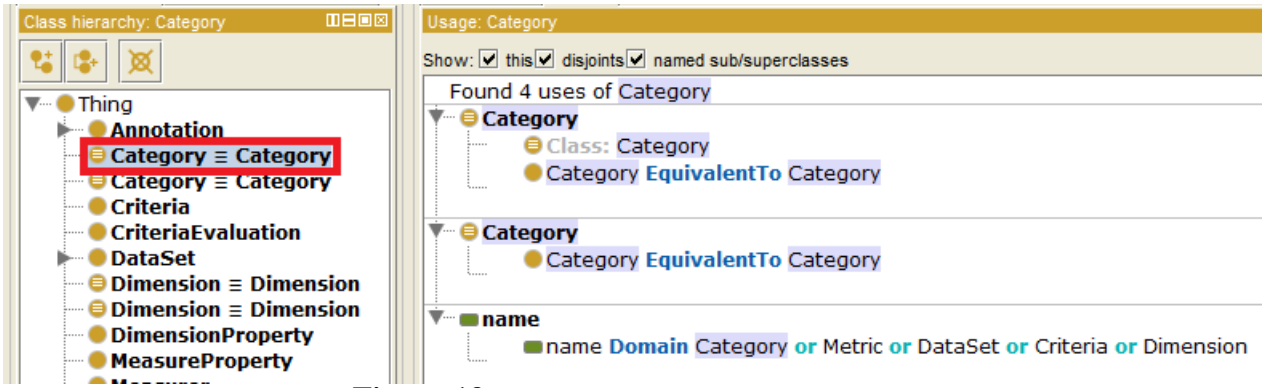

Figura 13 – Uso da Classe Categoria
Fonte: Elaboração própria

Prosseguindo, a classe **Categoria** (Figura 13) tem o **DataProperty** (propedade de dados) **name** que possui *Range* (tipo) *string* (texto) e possui como *Domains Category, Metric, DataSet*, ou *Dimension*, ou seja, Categoria, Métrica, *Dataset* ou Dimensão possuirão um nome no tipo texto. Essas implementações podem ser vistas na Figura 14.

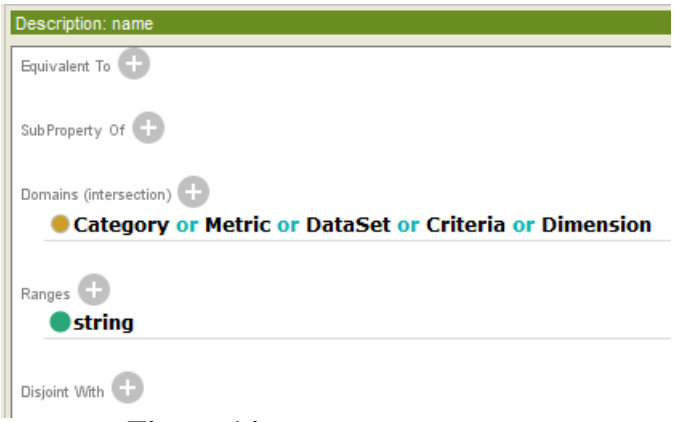

Figura 14 – DataProperty name
Fonte: Elaboração própria

Desta forma, visto que uma Categoria tem dimensões e vice-versa, foi implementado as propriedades **hasDimension** e **hasCategory**, como mostrado na Figura 15. Com isso, Dimension hasCategory Category representa que uma Dimensão tem Categoria e a inversa Category hasDimension Dimension representa que uma Categoria tem Dimensão.

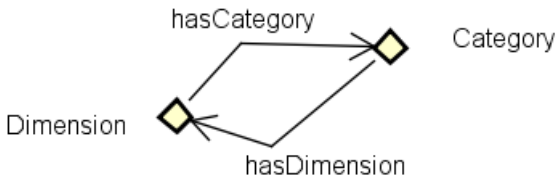

Figura 15 – Propriedades hasCategory e hasDimension
Fonte: Elaboração própria

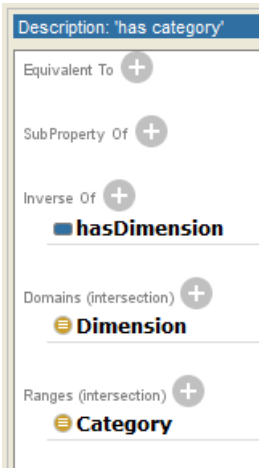

Figura 16 – Propriedades hasCategory no Protégé
Fonte: Elaboração própria

A Figura 15 mostra a implementação da Figura 16 no protégé.

Ademais, para representar que uma dimensão tem métricas e vice-versa, como mostrado na Figura 17, condizendo com o domínio da qualidade de dados, foi implementado as propriedades **hasMetric** e **hasDimension**.

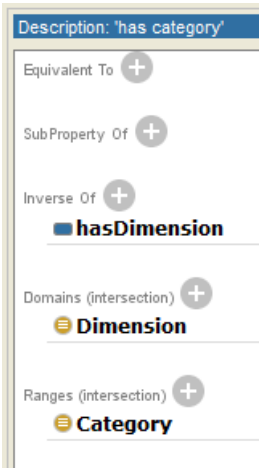

Figura 17 – Propriedades hasMetric e hasDimension
Fonte: Elaboração própria

Desta forma, percebemos que uma Dimensão tem Métricas e uma Métrica tem Dimensão, ou seja, *hasDimension* e *hasMetric* são inversas, como mostrado na Figura 17 da implementação no *Protégé*.

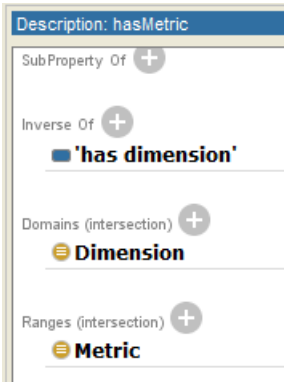

Figura 18 – Propriedades hasMetric e hasDimension no Protégé
Fonte: Elaboração própria

Analisando o domínio das dimensões de qualidade de dados trabalhados nesta dissertação, percebemos que as métricas possuem pequenos tópicos de verificação, como por exemplo, na Tabela 45 a Dimensão Licença possui o tópico que será verificado "**a. Nome da licença em uma página**", que é o tópico que compõe a métrica, da mesma forma, a Métrica 2 desta dimensão possui dois tópicos de verificação que são, "**b. Nome da licença em uma API" e "c. Nome da licença nos recursos RDF**". Visando conceitualizar esta realidade na ontologia, os tópicos de verificação das métricas foram denominados no Vocabulário de Criteria, com isso, foram criadas as propriedades belongsToMetric e hasCriteria, como mostrado na Figura 19

Desta forma, Criteria belongsToMetric Metric representa que um Critério pertence a uma Métrica e Metric hasCriteria Criteria informa que uma métrica tem Critérios. Sendo que as propriedades da figura 19 são inversas, como podemos ver na implementação mostrada na Figura 20

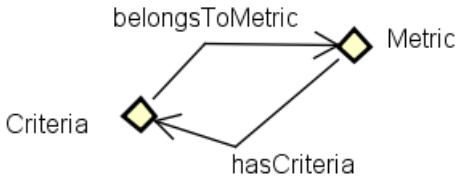

Figura 19 – Propriedades belongsToMetric e hasCriteria
Fonte: Elaboração própria

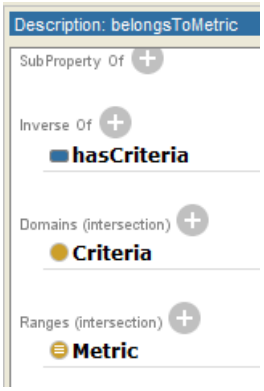

Figura 20 – Propriedades belongsToMetric e hasCriteria no Protégé
Fonte: Elaboração própria

Por conseguinte, um Critério (tópico de verificação da métrica) deve possui um valor e através do cálculo mostrado na subseção ? o valor da métrica é obtido. Com isso, foi criado a classe **CriteriaEvaluation** que representa o critério de avaliação. Com isso, a conceitualização no vocabulário ficou da seguinte maneira, **CriteriaEvaluation belongsToCriteriaEvaluation Criteria**, que representa que um Critério de avaliação pertence a um critério, e este critério de avaliação conterá um valor. Da mesma forma, **Criteria hasCriteriaEvaluation CriteriaEvaluation**, ou seja, um Critério tem um Critério de avaliação.

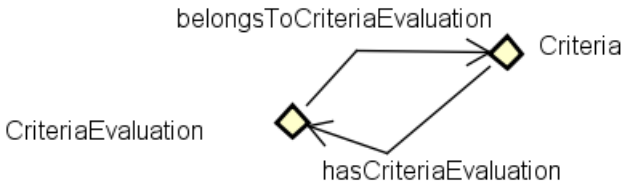

Figura 21 – Propriedades belongsToCriteriaEvaluation e hasCriteriaEvaluation
Fonte: Elaboração própria

Com isso, as propriedades **belongsToCriteriaEvaluation** e **hasCriteriaEvaluation** são propriedades inversas, como mostrado na implementação contida na Figura 22.

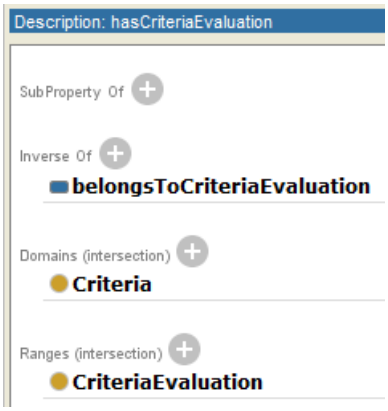

Figura 22 – Propriedades belongsToCriteriaEvaluation e hasCriteriaEvaluation no Protégé.
Fonte: Elaboração própria

4.2.2.2 Exemplos de uso da ontologia Vocabulário de Dados Conectados

Esta seção, será mostrado as partes do Vocabulário de Dados Conectados utilizadas para armazenar os dados obtidos da avaliação do usuário.

O primeiro bloco de códigos é referente a criação das classes.

```
###  http://purl.org/linked-data/cube#DataSet
<http://purl.org/linked-data/cube#DataSet> rdf:type owl:Class .

###  http://www.nees.com.br/ontologies/vdc#Criteria
:Criteria rdf:type owl:Class .

###  http://www.nees.com.br/ontologies/vdc#CriteriaEvaluation
:CriteriaEvaluation rdf:type owl:Class .

###  http://www.nees.com.br/ontologies/vdc#Measurer
:Measurer rdf:type owl:Class .

###  http://www.w3.org/ns/dqv#Category
dqv:Category rdf:type owl:Class ;
             rdfs:label "Category"@en .

###  http://www.w3.org/ns/dqv#Dimension
dqv:Dimension rdf:type owl:Class ;
              rdfs:label "Dimension"@en .

###  http://www.w3.org/ns/dqv#Metric
```

```
dqv:Metric rdf:type owl:Class ;
           rdfs:label "Metric"@en .
```

O próximo bloco de códigos representa a inserção de tipo nos *Data Properties* nameMeasurer e *DataSet*. Esses *Data Properties* serão utilizados para armazenar o nome do avaliador "*nameMeasurer*"e a url do *Dataset*, para fins de identificação.

```
###  http://www.nees.com.br/ontologies/vdc#nameMeasurer
:nameMeasurer rdf:type owl:DatatypeProperty;
              rdfs:domain :Measurer;
              rdfs:range xsd:string.

###  http://www.nees.com.br/ontologies/vdc#url
:url rdf:type owl:DatatypeProperty;
     rdfs:domain <http://purl.org/linked-data/cube#DataSet>;
     rdfs:range xsd:string.
```

O código posterior, será utilizado para armazenar no virtuoso os dados obtidos da avaliação do usuário referentes aos critérios, métricas, e dimensões.

```
###  http://www.nees.com.br/ontologies/vdc#valueCriteria
:valueCriteria rdf:type owl:DatatypeProperty ,
                        owl:FunctionalProperty ;
               rdfs:domain :CriteriaEvaluation ;
               rdfs:range xsd:double .

###  http://www.nees.com.br/ontologies/vdc#valueMetric
:valueMetric rdf:type owl:DatatypeProperty ,
                      owl:FunctionalProperty ;
             rdfs:range xsd:double ;
             rdfs:domain dqv:Metric .

###  http://www.nees.com.br/ontologies/vdc#valueDimension
:valueDimension rdf:type owl:DatatypeProperty ,
                         owl:FunctionalProperty ;
                rdfs:domain <http://purl.org/eis/vocab/daq#Dimension>;
                rdfs:range xsd:double .
```

```
### http://www.nees.com.br/ontologies/vdc#valueQualityMeasurementDataset
:valueQualityMeasurementDataset rdf:type owl:DatatypeProperty ,
                                        owl:FunctionalProperty ;
                                rdfs:range xsd:double .
```

A ontologia VDC foi inserida no virtuoso[3], com a finalidade de armazenar os resultados da avaliação de cada usuário. Com isso, os resultados armazenados são consultados e exibidos na tela do relatório detalhado, apresentada na Figura 32 da subseção 4.2.4

4.2.3 Arquitetura do *LinkedDatasetEvaluation*

Nesta sessão será abordado sobre a arquitetura de componentes modelada para implementação da abordagem semiautomática. Arquitetura baseada em componentes é uma abordagem da engenharia de *software* para estruturação e desenvolvimento de sistemas, que foca na decomposição da estrutura da funcionalidade de componentes lógicos, expondo suas interfaces de comunicação (Baccaro, 2017).

Desta forma, um componente pode ser visto como um elemento executável do sistema, que se conecta com outros componentes para integrar a composição do sistema. O relacionamento entre os componentes é feito por meio de interface, onde a interface fornecida é responsável por dar acesso ao componente, e a interface requerida é responsável por permitir o acesso.

Com isso, a arquitetura de componentes permite uma visualização geral do sistema, facilitando o entendimento do mesmo e possibilitando a atualização e expansão, quando necessário. Na Figura 23 é apresentada a arquitetura desenvolvida para a abordagem de avaliação semiautomática. Desta forma, a arquitetura foi organizada em cinco camadas, as quais são, Apresentação, Fachada, Sistema, Negócio e Persistência.

A camada de **Apresentação** é responsável pela apresentação das telas de navegação para o usuário, ou seja, são as telas do sistema onde o usuário interage. Desta forma, nesta camada é provida a interface **IDados** que é responsável por obter os dados informados pelo usuário em uma determinada página do sistema e repassá-los para a Camada **Fachada**, que é a camada inferior.

Ademais, a camada **Fachada** é responsável por gerenciar a obtenção dos dados enviados pela Interface aos componentes da camada **Sistema** os quais são **LerHTML, LerAPI** e **LerRDF**, através das interfaces providas **IHTML, IAPI** e **IRDF**, respectivamente. Com isso, a camada sistema contém os componentes relacionados as dimensões implementadas na ferramenta, ou seja, as dimensões utilizadas pelo usuário via página HTML e as dimensões processáveis por máquina que são relacionadas a API e ao RDF.

Além disso, a seguinte camada **Negócio** é responsável por gerenciar os dados dos cálculos das dimensões obtidos através das Interfaces **IDadosAPI, IDadosRDF** e **IDadosHTML**

[3]http://docs.openlinksw.com/virtuoso/

e enviá-los a camada posterior chamada **Persistência** através da Interface **IDadosDimensoes**. A camada **Negócio** também é responsável por obter os dados das dimensões armazenados no componente persistência através da interface **IDadosDimensoes** e envia-los para o componente **GeradorRelatório** da camada sistema, através da interface **IDadosDimensoes**.

A camada **Sistema** obtém os dados dos resultados das dimensões e os envia para o componente **Fachada** da camada superior, e este envia os dados das dimensões para o componente **Interface** que exibe os resultados para o usuário através do relatório gerado.

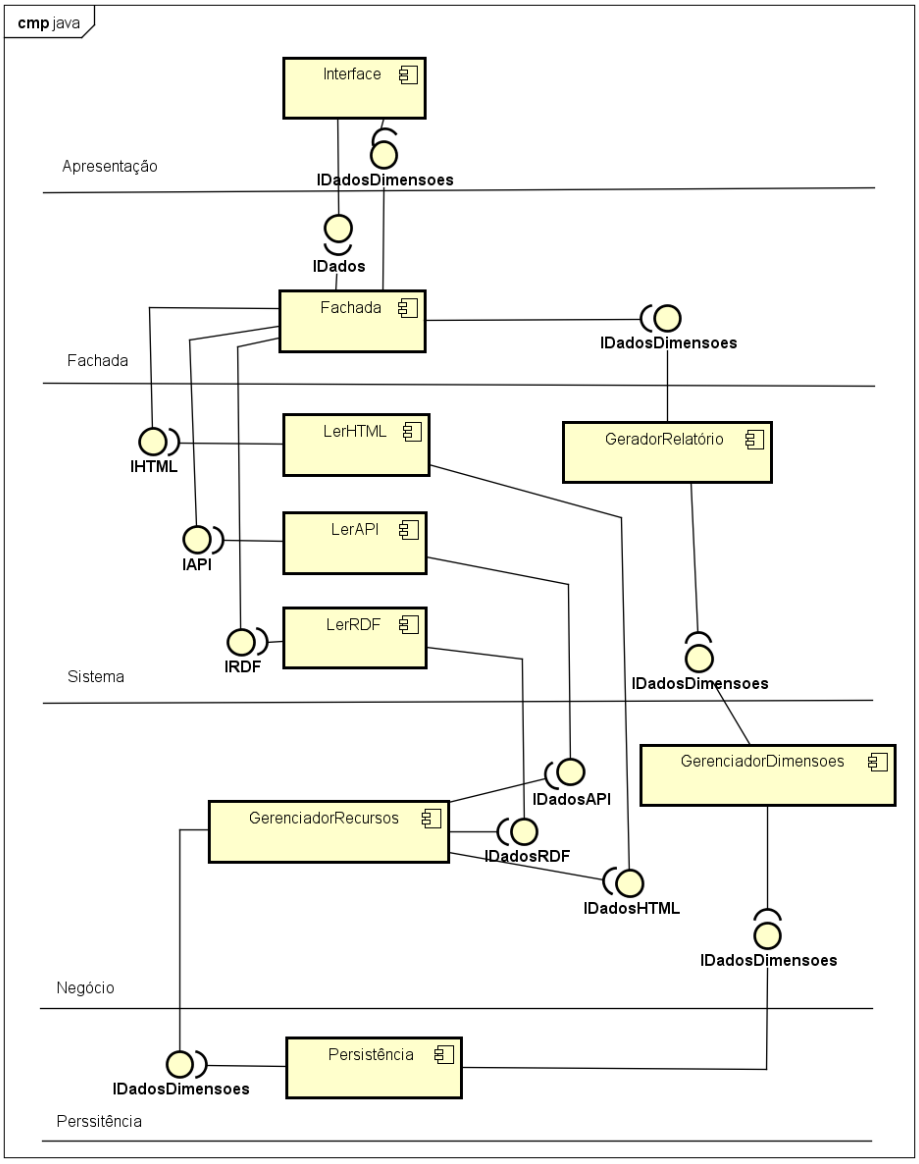

Figura 23 – Arquitetura de componentes da ferramenta semiautomática
Fonte: Elaboração própria

4.2.4 Apresentação do sistema

Nessa sessão, será apresentado o funcionamento do sistema como é apresentado para o usuário. Com isso, será mostrado as telas do sistema na sequência de apresentação para o usuário conforme as respostas fornecidas. Desta forma, a primeira tela corresponde na apresentação do objetivo da avaliação e obtenção de algumas informações, como primeiro e último nome, e o *link* do *dataset* a ser avaliado, como mostrado na Figura 24.

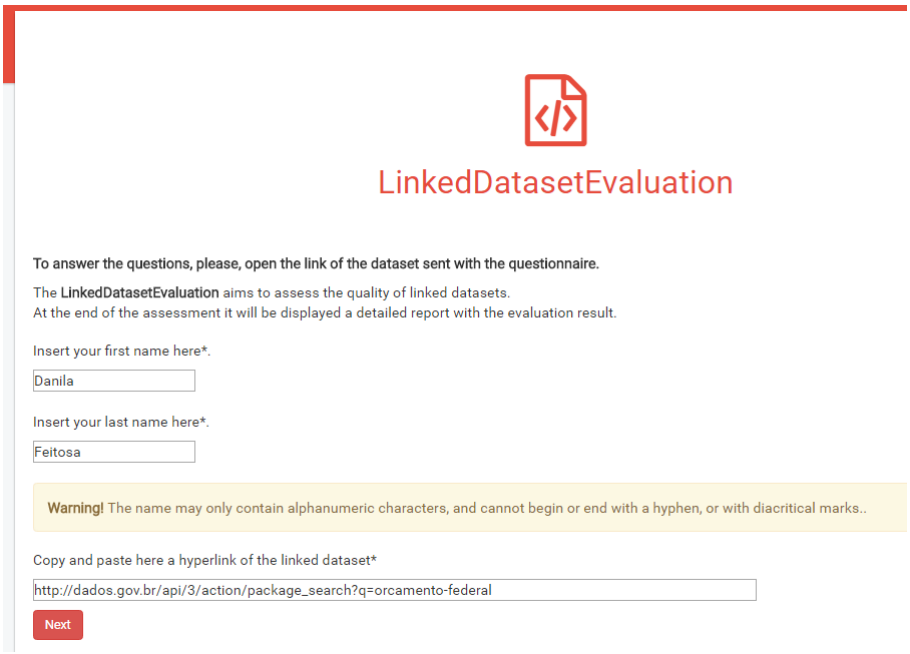

Figura 24 – Tela Inicial
Fonte: Elaboração própria

Após o usuário fornecer as informações requeridas, o sistema redirecionará para a tela de avaliação da dimensão *Disponibilidade*, como mostrado Figura 25. Nesta tela aparecerão as perguntas relacionadas aos critérios que compõe a métrica **M1. Verificar a existência de Acesso Automatizado dos Dados [BP 8; (Hyland et. al 2014)]** da referida dimensão. Os critérios podem ser vistos na Tabela 20.

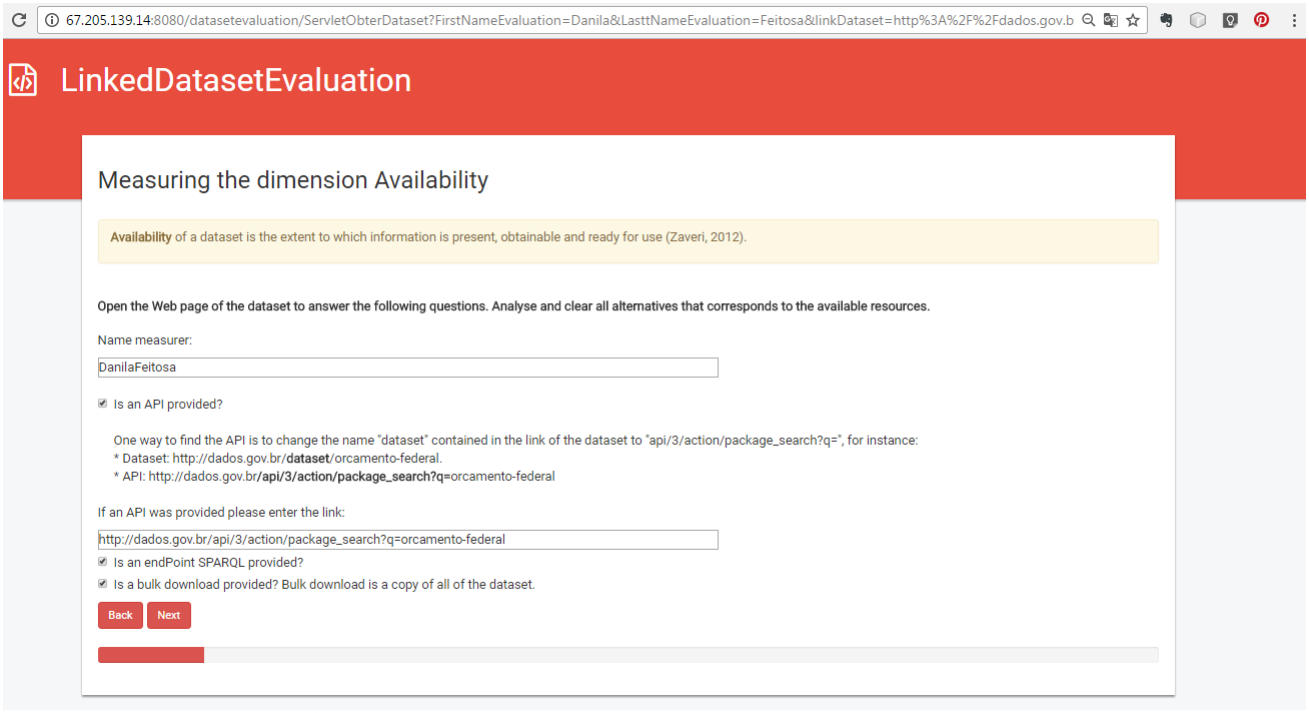

Figura 25 – Tela da Dimensão Disponibilidade
Fonte: Elaboração própria

Desta forma, o usuário irá analisar o *dataset* e buscar identificar as informações requeridas na tela da referida dimensão, como por exemplo, verificar a existência de uma API, caso o usuário não identifique a API, o mesmo poderá utilizar a dica apresentada na tela. Vale ressaltar que essa dica não é genérica, a mesma funcionará apenas se a API foi desenvolvida utilizando o CKAN[4]. Dessa forma, se o usuário identificar a API, ele deverá inserir o *link* da API no devido local na página apresentada, como mostrado na Figura 25, e posteriormente a API será avaliada automaticamente. A avaliação automática da API será feita com base na tabela 20: Métricas da dimensão Disponibilidade.

Ademais, o usuário tendo selecionado a opção referente à disponibilidade de um *endPoint* SPARQL e tendo concluído as respostas da página apresentada, o mesmo ao clicar em next será redirecionado para a página de verificação da disponibilidade do *endPoint* SPARQL encontrado. Nesta página, como mostrado na Figura 26, o usuário deverá inserir o *link* do *endPoint* e três consultas SPARQL. Em seguida, ao inserir as informações requeridas e clicar em *next*, as respostas destas consultas serão enviadas para a ferramenta, e o usuário prosseguirá com a avaliação da dimensão Disponibilidade.

[4]http://docs.ckan.org/en/latest/user-guide.html

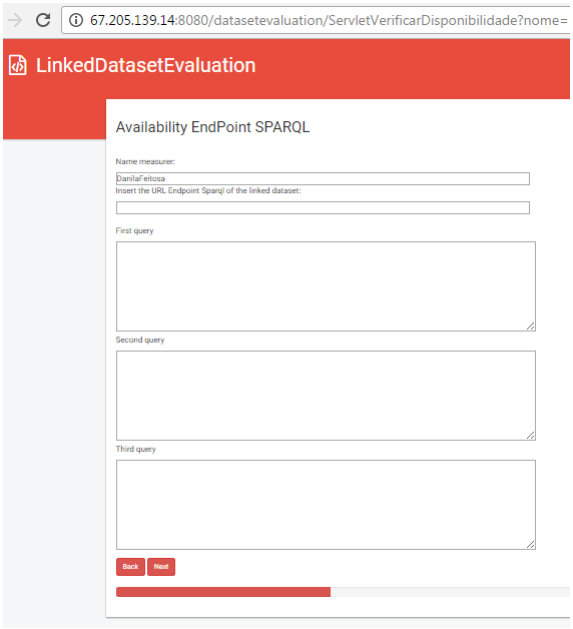

Figura 26 – Tela de avaliação da disponibilidade de um *endPoint* SPARQL
Fonte: Elaboração própria

A próxima página é uma continuidade da avaliação da Dimensão Disponibilidade, que é referente à disponibilidade do *dataset* no formato RDF. Como apresentado na tabela 20: Métricas da dimensão Disponibilidade. Desta forma, o usuário irá verificar no *dataset* a existência do *dataset* no formato RDF, caso o usuário encontre o arquivo, o mesmo deverá selecionar a opção referente a existência do *dataset* no formato RDF na página de avaliação, posteriormente o usuário será redirecionado para a tela de *Upload* do RDF. Caso o arquivo RDF não seja encontrado, o usuário deverá desselecionar a opção referende à disponibilidade do arquivo RDF e o mesmo será redirecionado para a tela de avaliação da Dimensão Proveniência.

A tela de avaliação da Dimensão Proveniência (Figura 27) apresenta as perguntas relacionadas aos critérios das métricas M1. Verificar a existência de metadados descritivos sobre o *dataset* (ZAVERI, 2014) [BP 1 e BP2 (Lóscio et. al 2016)] e M2. Verificar a existência de informações de Proveniência do *dataset* [BP 6; (Lóscio et. al, 2016)], referentes a avaliação do *dataset* na Página *Web*. Os critérios referentes a API e ao RDF serão verificados automaticamente, mediante o *link* da API e o arquivo RDF informados anteriormente.

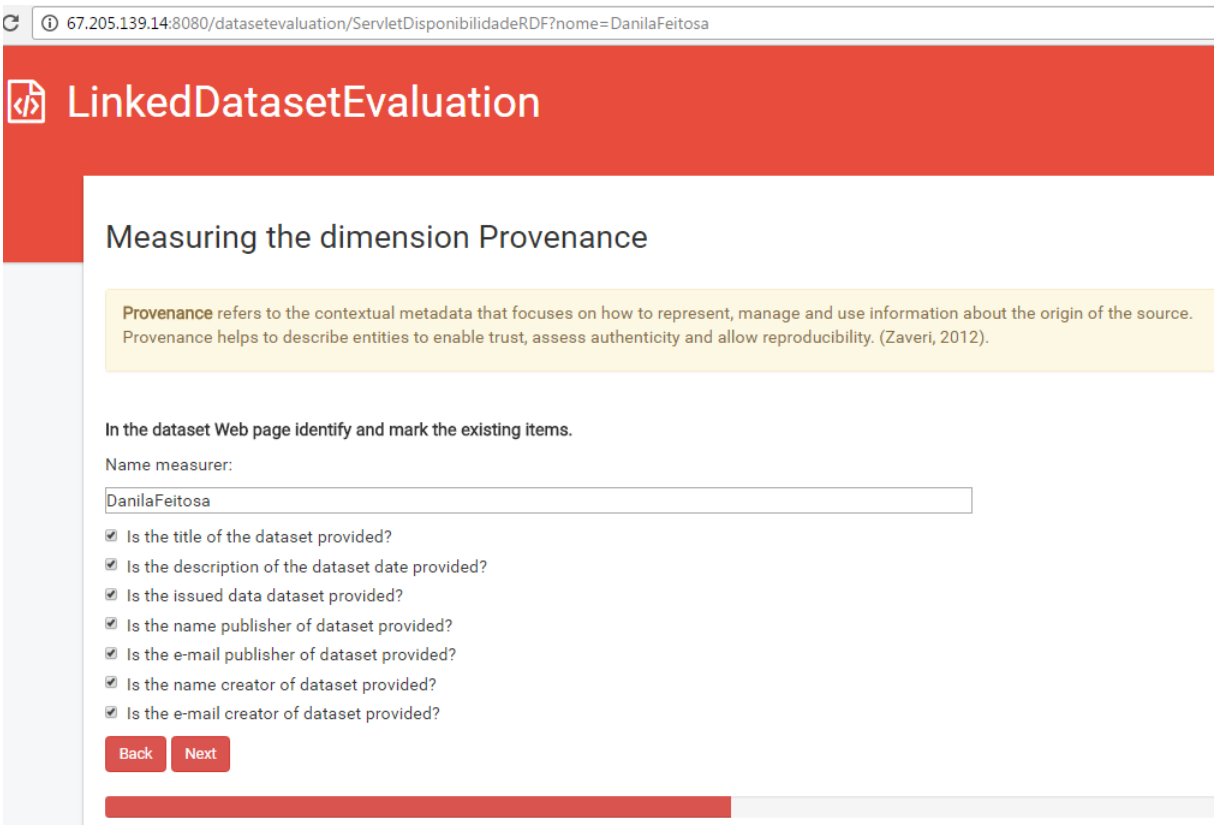

Figura 27 – Tela de avaliação da dimensão Proveniência
Fonte: Elaboração própria

Desta forma, quando o usuário terminar de responder as questões apresentadas e clicar em *next*, o mesmo será redirecionado para a tela de avaliação da dimensão Licença (Figura 28). Nesta tela, o usuário responderá questões relacionadas ao critério da métrica *M1. Verificar se há a indicação de uma licença legível por humanos (Zaveri, 2014) [BP 4; (Hyland et. Al, 2014)]*. Com isso, a métrica *M2. Verificar se há a indicação de uma licença legível por máquina (Zaveri, 2014) [BP 5; (Lócio et. al, 2016)]* será avaliada automaticamente pela ferramenta, que irá verificar na API e no RDF a licença legível por máquina.

Com isso, após concluir a avaliação da dimensão Licença o usuário será redirecionado para a tela de avaliação da dimensão Compreensibilidade, como mostrado na Figura 29. As questões que serão apresentadas para o usuário são referentes aos critérios de avaliação das métricas M1. Verificar se há mecanismo para capturar *feedback* de consumidores de dados [BP 30; (Lóscio et. al 2016)], M2. Verificar se existe informações sobre a disponibilidade dos *feedback* dos usuários [BP 31; (Lóscio et. al 2016)] e M3. Verificar se é provido no *dataset* uma lista de vocabulários usados (ZAVERI, 2014). Ambas as métricas são avaliadas apenas manualmente, pois nesta dimensão é fundamental a percepção do usuário em relação aos recursos que facilitam a compreensão do *dataset*.

Ademais, a próxima tela apresentada corresponde à avaliação da dimensão Atualidade. Assim como as telas e perguntas anteriores, a apresentação das questões da dimensão Atualidade são referentes aos critérios das métricas M1. Verificar se existe metadados de parâ-

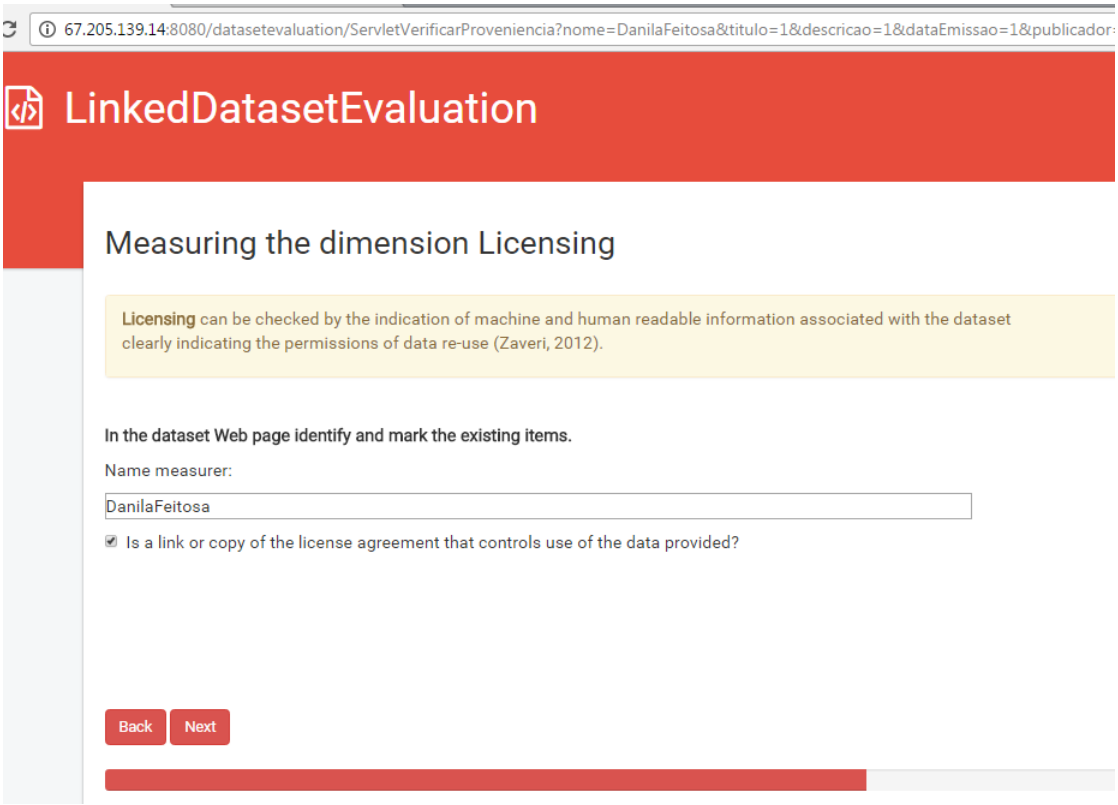

Figura 28 – Tela de avaliação da dimensão Licença
Fonte: Elaboração própria

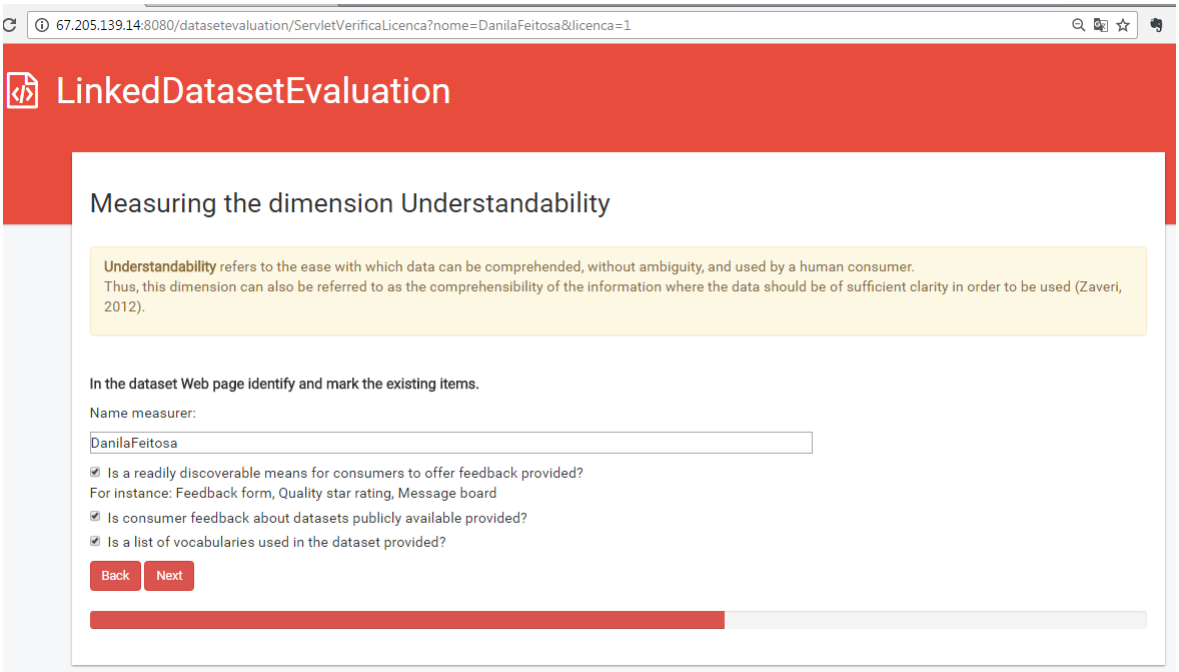

Figura 29 – Tela de avaliação da dimensão Compreensibilidade
Fonte: Elaboração própria

metros de localidade [BP 3; (Lóscio et. al 2016)] e M2. Verificação de informações sobre a idade do dado (ZAVERI, 2014). Neste caso, o usuário irá responder as perguntas apenas referentes à avaliação da página *Web*, e a avaliação da API e do RDF serão feitas automaticamente.

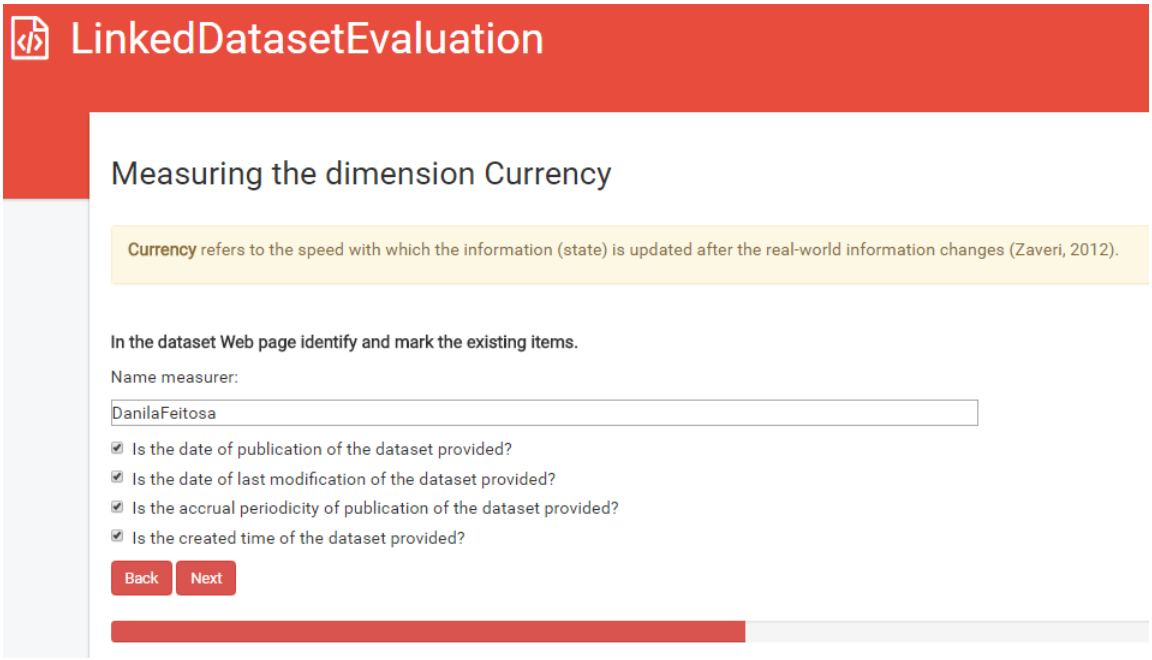

Figura 30 – Tela de avaliação da dimensão Atualidade
Fonte: Elaboração própria

Finalmente, após concluir as questões apresentadas, o usuário será redirecionado para a tela de avaliação da dimensão Versatilidade (Figura 31). Nesta tela, será apresentado para o usuário apenas uma questão que é relacionada à métrica M2. Verificar se o *dataset* é fornecido em várias linguagens (ZAVERI, 2014) e a métrica M1. Verificar se o *dataset* é fornecido em diferentes formatos de serialização (ZAVERI, 2014) [BP 22 (Lóscio et. al 2016)] foi calculada no decorrer da execução da ferramenta.

Posteriormente, finalizada a resposta desta questão, o sistema exibirá um relatório contendo detalhadamente o resultado da avaliação de cada dimensão. Parte do relatório pode ser visto na Figura 32. Ao clicar na aba de cada dimensão será mostrando o resultado de cada critério, de cada métrica e o valor de qualidade geral do *dataset*. Vale ressaltar, que os resultados apresentados não são verdadeiros, pois foram utilizados apenas para fins didáticos de escrita deste trabalho.

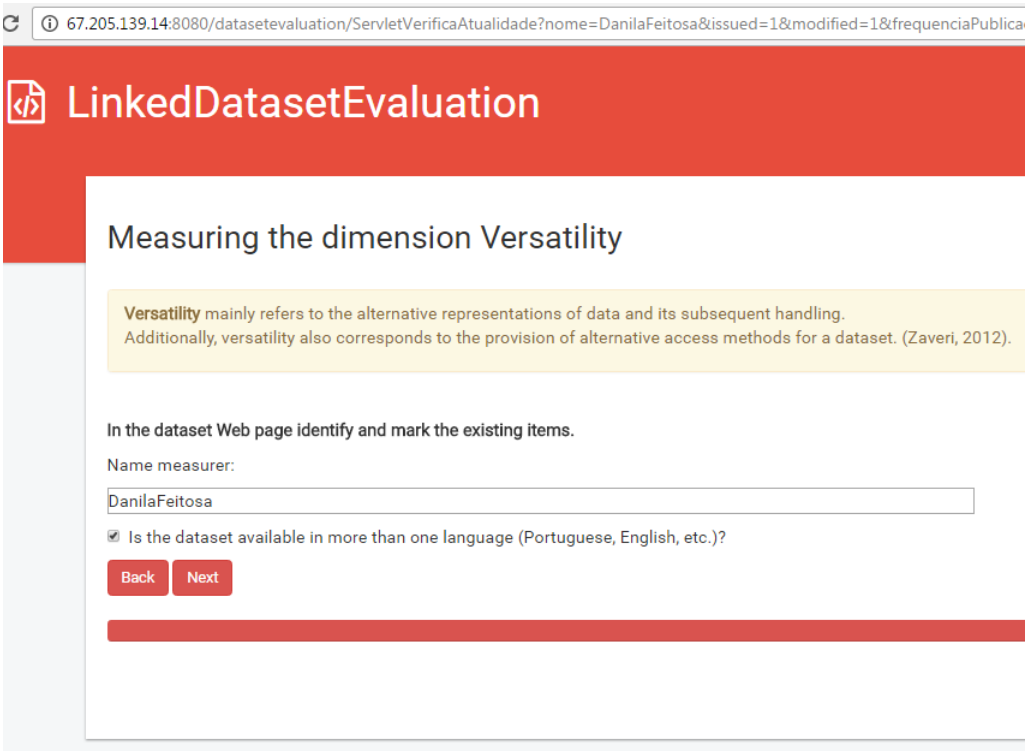

Figura 31 – Tela de avaliação da dimensão Versatilidade
Fonte: Elaboração própria

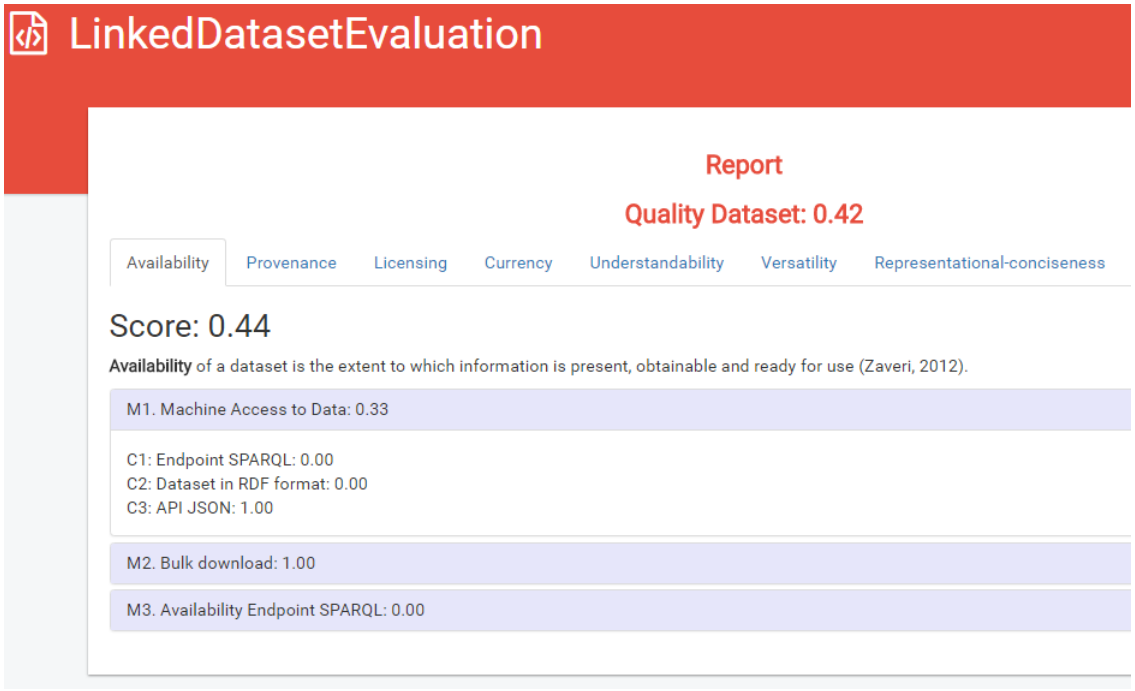

Figura 32 – Relatório final da avaliação da qualidade de *datasets* conectados
Fonte: Elaboração própria

5 Validação - Experimento da abordagem para avaliação da qualidade de *datasets* conectados

5.1 Definição do problema

5.1.1 Contextualização

A *Web* de dados visa consolidar conhecimento através da conexão de dados existentes e dispersos. No entanto, a utilidade desse conhecimento depende fortemente dos dados publicados. A atual *Web* de dados consiste em mais de 50 bilhões de fatos representados por triplas RDF que cobrem diversos domínios, como saúde, governo, música, etc.

Essa grande quantidade de dados oferecem diversos desafios em relação à qualidade, como por exemplo, *datasets* que contém informações a partir de fontes estruturadas, como o DBpedia[5] , New Castle[6] , RAE2001[7] , Kisti[8] e Yasgui[9] frequentemente oferecem informações incompletas e inconsistentes (Travassos, 2014).

A qualidade de dados é um importante requisito para o crescente sucesso de LD, no entanto, somente recentemente a mesma tem recebido atenção da comunidade *Web* Semântica, por este motivo há um número limitado de iniciativas de pesquisas existentes que focam na qualidade de *datasets* voltados para a *Web* Semântica (Behkamal, 2014).

5.1.2 Relevância do problema

O consumo de *datasets* conectados, de baixa qualidade, pode ocasionar na interpretação incorreta de dados e consequentemente na tomada de decisão incorreta.

Com isso, nos deparamos com os seguintes questionamentos:

- O que o *dataset* não possui que o torna de má qualidade?
- O que o *dataset* possui que o torna de má qualidade?

Segundo O. Harting et al. (2009), qualidade da informação (QI) consiste em um agregado de valores de múltiplos critérios, como por exemplo *accuracy*, *completeness, believability* e *timeliness*. Loshin (2011) corrobora que o problema de qualidade de dados pode ser resolvido com base no conceito de "adequado ao uso para uma proposta específica que depende de características e necessidades de negócios do usuário". A partir de informações relacionadas a este conceito, Loshin (2011) descreve dez dimensões utilizadas para medir a qualidade

[5] dbpedia.org/sparql
[6] http://newcastle.rkbexplorer.com/sparql/
[7] http://rae2001.rkbexplorer.com/sparql/
[8] http://kisti.rkbexplorer.com/sparql/
[9] http://yasgui.org/

de dados, as quais são *accuracy, lineage, semantic, structure, completeness, consistency, currency, timeliness, reasonableness* e *identifiability*.

Outros fatores que agregam valor a um *dataset* conectado é o mesmo ser publicado segundo as melhores práticas recomendadas pelo W3C para a publicação de *datasets* conectados, pois as mesmas são consideradas padrões *Web* internacionais de publicação de dados conectados (Wood, 2014).

O W3C disponibiliza diversas orientações para publicação, tanto para dados na *Web* de uma forma geral mediante o documentos "*Data on the Web Best Practices*", que provê melhores práticas relacionadas à publicação e uso dos dados na *Web*, quanto para dados conectados, através do documento "*Best Practices for Publishing Linked Data*".

5.1.3 Problema técnico

Problema: Visando resolver o problema citado anteriormente, no sentido de avaliar a qualidade de um *dataset* conectado com base em Dimensões de qualidade (completamente ou parcialmente), deseja-se também evidenciar para o usuário a qualidade do *dataset* conectado.

Perguntas de pesquisa:

- Quais dimensões devem ser consideradas para executar a verificação da qualidade de *datasets* conectados?
- Quais boas práticas devem ser utilizadas para executar a verificação de *datasets* conectados?

5.2 Objetivos da investigação

A validação é do tipo Experimental, onde a abordagem sugerida será submetida à avaliação através de um experimento. O experimento tem como objetivo avaliar a solução desenvolvida visando identificar se a mesma contribui com a avaliação da qualidade de *datasets* conectados, no sentido facilitar a avaliação (fazendo com que a nota de avaliação de usuários de *datasets* conectados seja o mais próxima possível da nota de avaliação de especialistas em *datasets* conectados) e diminuir a carga de trabalho do usuário.

Ao final do experimento, as notas alcançadas serão analisadas estatisticamente, com o intuito de identificar se a abordagem cumpre sua finalidade.

Objetivo geral

Analisar a abordagem semiautomática **com a intenção de** avaliá-la **a respeito** de sua validade, através de análises estatísticas de um score global e scores parciais **do ponto de vista** dos envolvidos, avaliando *datasets* conectados no **contexto de** *datasets* com diferentes níveis de qualidade.

Objetivos específicos

O experimento fará uma análise das notas da avaliação semiautomática (que são as notas obtidas de pessoas que utilizaram a ferramenta para fazer a avaliação do *dataset*) e manual (são as notas das pessoas que utilizaram a abordagem manual para fazer a avaliação do *dataset*) com as notas base (são as notas obtida da avaliação de pessoas especialistas em *datasets* conectados que utilizaram a ferramenta).

As notas obtidas são os escores globais, que são a nota de qualidade geral do *dataset* (podem ser vistas na subseção 4.1.1) e os escores parciais, que são as notas de cada dimensão de qualidade obtida da avaliação do *dataset* (podem ser vistas na subseção 4.1.1). Desta forma, no experimento será feito as seguintes comparações:

- Comparação do score global:
  - Comparar o score global da avaliação semiautomática com o escore global da avaliação manual.
  - Comparar o score global da avaliação semiautomática com o escore global base.
  - Comparar o score global da avaliação manual com o escore global base.
- Comparação dos scores parciais:
  - Comparar os scores parciais da avaliação semiautomática com os scores parciais obtidos da avaliação manual.
  - Comparar os scores parciais da avaliação semiautomática com os scores parciais base.
  - Comparar os scores parciais da avaliação manual com os scores parciais base.
- Comparação do tempo:
  - Comparar o tempo da avaliação semiautomática com o tempo da avaliação manual.
  - Comparar o tempo da avaliação semiautomática dos avaliadores com o tempo obtido da nota base.
  - Comparar o tempo da avaliação manual dos avaliadores com o tempo obtido da nota base.

5.3 Planejamento do experimento

5.3.1 Questões de pesquisa e hipóteses

5.3.1.1 Hipóteses

A principal questão de pesquisa do design de experimento, objetiva determinar e comparar a eficiência da validação e verificação da qualidade um *dataset* conectado, utilizando a so-

lução proposta. A eficiência da solução proposta está em fazer com que o usuário faça uma avaliação semelhante a avaliação de um especialista em *datasets* conectados e em temos semelhantes. Desta forma, a principal questão de pesquisa é:

P1 – Como prover uma solução computacional que facilite a avaliação da qualidade de *datasets* conectados considerando práticas para publicação de *datasets* conectados e dimensões de qualidade?

Levando-nos às seguintes hipóteses:

H1-0: O score global (qualidade geral do dataset) com o uso da ferramenta é igual ao score global da avaliação manual.

H1-1: O score global com o uso da ferramenta é diferente do score global da avaliação manual.

H2-0: O score global com o uso da ferramenta é igual ao score global base.

H2-1: O score global com o uso da ferramenta é diferente do score global base.

H3-0: O score global da avaliação manual é igual ao score global base.

H3-1: O score global da avaliação manual é diferente do score global base.

H4-0: Os scores parciais considerando a dimensão Disponibilidade com o uso da ferramenta são iguais aos scores parciais da mesma dimensão da avaliação manual.

H4-1: Os scores parciais considerando a dimensão Disponibilidade com o uso da ferramenta são diferentes aos scores parciais da mesma dimensão da avaliação manual.

H5-0: Os scores parciais considerando a dimensão Disponibilidade com o uso da ferramenta são iguais dos scores parciais base desta dimensão.

H5-1: Os scores parciais considerando a dimensão Disponibilidade com o uso da ferramenta são diferentes dos scores parciais base desta dimensão.

H6-0: Os scores parciais considerando a dimensão Disponibilidade da avaliação manual são iguais aos scores parciais base desta dimensão.

H6-1: Os scores parciais considerando a dimensão Disponibilidade da avaliação manual são diferentes aos scores parciais base desta dimensão.

H7-0: Os scores parciais considerando a dimensão Proveniência com o uso da ferramenta são iguais aos scores parciais da mesma dimensão da avaliação manual.

H7-1: Os scores parciais considerando a dimensão Proveniência com o uso da ferramenta são diferentes aos scores parciais da mesma dimensão da avaliação manual.

H8-0: Os scores parciais considerando a dimensão Proveniência com o uso da ferramenta são iguais dos scores parciais base desta dimensão.

H8-1: Os scores parciais considerando a dimensão Proveniência com o uso da ferramenta são diferentes dos scores parciais base desta dimensão.

H9-0: Os scores parciais considerando a dimensão Proveniência da avaliação manual são iguais aos scores parciais base desta dimensão.

H9-0: Os scores parciais considerando a dimensão Proveniência da avaliação manual são diferentes aos scores parciais base desta dimensão.

H10-0: Os scores parciais considerando a dimensão Licenciamento com o uso da ferramenta são iguais aos scores parciais da mesma dimensão da avaliação manual.

H10-1: Os scores parciais considerando a dimensão Licenciamento com o uso da ferramenta são diferentes aos scores parciais da mesma dimensão da avaliação manual.

H11-0: Os scores parciais considerando a dimensão Licenciamento com o uso da ferramenta são iguais dos scores parciais base desta dimensão.

H11-1: Os scores parciais considerando a dimensão Licenciamento com o uso da ferramenta são diferentes dos scores parciais base desta dimensão.

H12-0: Os scores parciais considerando a dimensão Licenciamento da avaliação manual são iguais aos scores parciais base desta dimensão.

H12-0: Os scores parciais considerando a dimensão Licenciamento da avaliação manual são diferentes aos scores parciais base desta dimensão.

H13-0: Os scores parciais considerando a dimensão Compreensibilidade com o uso da ferramenta são iguais aos scores parciais da mesma dimensão da avaliação manual.

H13-1: Os scores parciais considerando a dimensão Compreensibilidade com o uso da ferramenta são diferentes aos scores parciais da mesma dimensão da avaliação manual.

H14-0: Os scores parciais considerando a dimensão Compreensibilidade com o uso da ferramenta são iguais dos scores parciais base desta dimensão.

H14-1: Os scores parciais considerando a dimensão Compreensibilidade com o uso da ferramenta são diferentes dos scores parciais base desta dimensão.

H15-0: Os scores parciais considerando a dimensão Compreensibilidade da avaliação manual são iguais aos scores parciais base desta dimensão.

H15-0: Os scores parciais considerando a dimensão Compreensibilidade da avaliação manual são diferentes aos scores parciais base desta dimensão.

H16-0: Os scores parciais considerando a dimensão Atualidade com o uso da ferramenta são iguais aos scores parciais da mesma dimensão da avaliação manual.

H16-1: Os scores parciais considerando a dimensão Atualidade com o uso da ferramenta são diferentes aos scores parciais da mesma dimensão da avaliação manual.

H17-0: Os scores parciais considerando a dimensão Atualidade com o uso da ferramenta são iguais dos scores parciais base desta dimensão.

H17-1: Os scores parciais considerando a dimensão Atualidade com o uso da ferramenta são diferentes dos scores parciais base desta dimensão.

H18-0: Os scores parciais considerando a dimensão Atualidade da avaliação manual são iguais aos scores parciais base desta dimensão.

H18-0: Os scores parciais considerando a dimensão Atualidade da avaliação manual são diferentes aos scores parciais base desta dimensão.

H19-0: Os scores parciais considerando a dimensão Versatilidade com o uso da ferramenta são iguais aos scores parciais da mesma dimensão da avaliação manual.

H19-1: Os scores parciais considerando a dimensão Versatilidade com o uso da ferramenta são diferentes aos scores parciais da mesma dimensão da avaliação manual.

H20-0: Os scores parciais considerando a dimensão Versatilidade com o uso da ferramenta são iguais dos scores parciais base desta dimensão.

H20-1: Os scores parciais considerando a dimensão Versatilidade com o uso da ferramenta são diferentes dos scores parciais base desta dimensão.

H21-0: Os scores parciais considerando a dimensão Versatilidade da avaliação manual são iguais aos scores parciais base desta dimensão.

H21-0: Os scores parciais considerando a dimensão Versatilidade da avaliação manual são diferentes aos scores parciais base desta dimensão.

H22-0: Os scores parciais considerando a dimensão Consistência Representacional com o uso da ferramenta são iguais aos scores parciais da mesma dimensão da avaliação manual.

H22-1: Os scores parciais considerando a dimensão Consistência Representacional com o uso da ferramenta são diferentes aos scores parciais da mesma dimensão da avaliação manual.

H23-0: Os scores parciais considerando a dimensão Consistência Representacional com o uso da ferramenta são iguais dos scores parciais base desta dimensão.

H23-1: Os scores parciais considerando a dimensão Consistência Representacional com o uso da ferramenta são diferentes dos scores parciais base desta dimensão.

H24-0: Os scores parciais considerando a dimensão Consistência Representacional da avaliação manual são iguais aos scores parciais base desta dimensão.

H24-1: Os scores parciais considerando a dimensão Consistência Representacional da avaliação manual são diferentes aos scores parciais base desta dimensão.

H25-0: O tempo de avaliação da qualidade de *datasets* conectados com o uso da ferramenta é igual ao tempo de avaliação com o uso da abordagem manual.

H25-1: O tempo de avaliação da qualidade de *datasets* conectados com o uso da ferramenta é diferente do tempo de avaliação com o uso da abordagem manual.

H26-0: O tempo de avaliação da qualidade de *datasets* conectados com o uso da ferramenta é igual ao tempo base de avaliação.

H26-1: O tempo de avaliação da qualidade de *datasets* conectados com o uso da ferramenta é diferente do tempo base de avaliação.

H27-0: O tempo de avaliação da qualidade de *datasets* conectados da avaliação manual é igual ao tempo base de avaliação.

H27-1: O tempo de avaliação da qualidade de *datasets* conectados da avaliação manual é diferente do tempo base de avaliação.

Caso a hipótese nula seja refutada, a média do tempo será analisada para entendimento do motivo pelo qual houve discrepância. Dessa forma será possível estabelecer qual tipo de verificação e validação que possui melhor tempo de execução.

5.3.1.2 Fatores e variáveis de resposta

Após a elaboração das hipóteses de pesquisa, o próximo passo do planejamento do experimento consiste em definir as variáveis que serão manipuladas. Desta forma, o presente trabalho possui dois tipos de variáveis: variáveis independentes chamadas de fatores e as variáveis dependentes, chamadas de métricas de avaliação.

As variáveis independentes do presente experimento são apresentadas na tabela 26.

Tabela 26 – Variáveis independentes e níveis de fatores

Fator	*Nível*	*Descrição*
Abordagem Semiautomática	AS	Consiste em uma ferramenta computacional que contém computadas algumas dimensões de qualidade, as quais serão avaliadas automaticamente, e também obterá as respostas necessárias do usuário em avaliações manuais.
Abordagem Manual	AM	Consiste em um formulário que contém questões baseadas nas dimensões propostas por Zaveri (2012), as mesmas utilizadas na abordagem semiautomática.

Fonte: Elaboração própria

As variáveis dependentes, ou seja, que dependem da aplicação das variáveis independentes, são:

- Score global base: obtido da média geral (envolvendo todas as dimensões de qualidade) da avaliação de especialistas.
- Scores parciais base: obtido da média geral de cada dimensão de qualidade da avaliação dos avaliadores envolvidos no experimento.
- Score global: obtido da média (envolvendo todas as dimensões de qualidade) da avaliação dos avaliadores envolvidos no experimento. Este escore será gerado tanto para a avaliação manual, quanto da avaliação semiautomática.
- Scores parciais: obtido da média de cada dimensão de qualidade da avaliação dos avaliadores envolvidos no experimento. Este escore será gerado tanto para a avaliação manual, quanto da avaliação semiautomática.
- Tempo: obtido do tempo de duração da avaliação. Esta variável será obtida tanto para a avaliação manual, quanto da avaliação semiautomática.
- Tempo de avaliação: é o tempo de avaliação utilizando a abordagem manual e semiautomática.

5.3.1.3 Definição formal das hipóteses

Formalmente, as hipóteses descritas anteriormente podem ser definidas conforme a Tabela 27:

Tabela 27 – Definição formal das hipóteses

Hip.	Hipótese nula	Hipótese alternativa
H1	Global(AS) = Global(AM)	Global(AS) $\neq$ Global(AM)
H2	Global (AS) Global(base)	Global (AS) $\neq$ Global(base)
H3	Global (AM) = Global(base)	Global (AM) $\neq$ Global(base)
H4	Dispon.(AS) = Dispon.(AM)	Dispon.(AS) $\neq$ Dispon.(AM)
H5	Dispon.(AS) = Dispon.(base)	Dispon.(AS) $\neq$ Dispon.(base)
H6	Dispon.(AM) = Dispon.(base)	Dispon.(AM) $\neq$ Dispon.(base)
H7	Prov.(AS) = Prov.(AM)	Prov.(AS) $\neq$Prov.(AM)
H8	Prov.(AS) = Prov.(base)	Prov.(AS) $\neq$Prov.(base)
H9	Prov.(AM) = Prov.(base)	Prov.(AM) $\neq$ Prov.(base)
H10	Licenc.(AS) = Licenc.(AM)	Licenc.(AS) $\neq$ Licenc.(AM)
H11	Licenc.(AS) = Licenc.(base)	Licenc.(AS) $\neq$ Licenc.(base)
H12	Licenc.(AM) = Licenc.(base)	Licenc.(AM) $\neq$ Licenc.(base)
H13	Comp.(AS) = Comp.(AM)	Comp.(AS) $\neq$Comp.(AM)
H14	Comp.(AM) = Comp.(base)	Comp.(AM) $\neq$ Comp.(base)
H15	Comp.(AS) = Comp.(base)	Comp.(AS) $\neq$ Comp.(base)
H16	Atual.(AS) = Atual.(AM)	Atual.(AS) $\neq$ Atual.(AM)
H17	Atual.(AS) = Atual.(base)	Atual.(AS) $\neq$ Atual.(base)
H18	Atual.(AM) = Atual.(base)	Atual.(AM) $\neq$ Atual.(base)
H19	Versat.(AM) = Versat.(AS)	Versat.(AM) $\neq$ Versat.(AS)
H20	Versat.(AS) = Versat.(base)	Versat.(AS) $\neq$ Versat.(base)
H21	Versat.(AM) = Versat.(base)	Versat.(AM) $\neq$ Versat.(base)
H22	Cons. Repres.(AM) = Cons. Repres. (AS)	Cons. Repres.(AM) $\neq$ Cons. Repres. (AS)
H23	Cons. Repres.(AS) = Cons. Repres. (base)	Cons. Repres.(AS) $\neq$ Cons. Repres. (base)
H24	Cons. Repres.(AM) = Cons. Repres. (base)	Cons. Repres.(AM) $\neq$ Cons. Repres. (base)
H25	Tempo(AM) = Tempo(AS)	Tempo(AM) $\neq$ Tempo(AS)
H26	Tempo(AS) = Tempo(base)	Tempo(AS) $\neq$ Tempo(base)
H27	Tempo(AM) = Tempo(base)	Tempo(AM) = Tempo(base)

Fonte: Elaboração própria

Onde Global refere-se ao valor do Score Global, Tempo refere-se ao Tempo de avaliação aplicadas na abordagem semiautomática (AS) e abordagem manual (AM), e Dispon., Prov., Licenc., Comp., Atual., Versat., Cons. Repres, são funções que retornam respectivamente os valores das dimensões: Disponibilidade, Proveniência, Licenciamento, Compreensibilidade, Atualidade, Versatilidade e Consistência Representacional.

5.4 Unidades experimentais

O experimento é do tipo comparativo, pertencente ao grupo fatorial, onde o grupo de controle consiste na avaliação utilizando a abordagem manual, e o grupo experimental será usando a abordagem semiautomática, e o tempo nos dois momentos da avaliação serão utilizados para medir a demora em executar a avaliação com as duas abordagens.

Serão utilizados dois *datasets*, os quais foram extraídos de um documento disponibilizado pelo W3C:

- *dataset* 1 - *http://catalog.data.gov/dataset/crimes-2001-to-present-398a4*: este *dataset* foi obtido do *web site* data.gov do governo dos Estados Unidos que disponibiliza 1000 *datasets* como dados conectados;
- *dataset* 2 - *https://data.gov.uk/dataset/land-registry-monthly-price-paid-data*: este *dataset* foi obtido do *web site* data.gov.uk que é oficial do governo do Reino Unido e disponibiliza 3600 *datasets* como dados conectados;

A amostra consistiu de vinte e duas pessoas, as quais são os avaliadores que fizeram parte do experimento, onde:

- 4 pessoas avaliaram o *dataset* 1 - utilizando a abordagem manual;
- 4 pessoas avaliaram o *dataset* 1 utilizando a abordagem semiautomática;
- 5 pessoas avaliaram o *dataset* 2 utilizando a abordagem manual;
- 5 pessoas avaliaram o *dataset* 2 utilizando a abordagem semiautomática;
- 2 especialistas do NEES (Núcleo de Excelência para Tecnologias Sociais), que trabalham com dados conectados, avaliaram o *dataset* 1 utilizando a abordagem semiautomática;
- 2 especialistas do NEES, que também que trabalham com dados conectados, avaliaram o *dataset* 2 utilizando a abordagem semiautomática;

5.5 Plano de execução

Esta seção descreve como será a execução do experimento que foi planejado e descrito ao longo das seções anteriores. A execução do experimento envolve os seguintes passos:

1. Seleção dos *datasets*: foram selecionados dois *datasets* contidos na página TaskForces/CommunityProjects/LinkingOpenData/DataSets do W3C. Esses *datasets* foram escolhidos devido a serem os que mais se aproximam do conceito de *datasets* conectados estabelecido pelo W3C.
2. Seleção dos participantes: foram selecionadas listas de especialistas em dados conectados para a realização do experimento. As listas de especialistas selecionadas foram:
    - *Linked Geo Data*: *https://groups.google.com/forum/#!forum/linked-geo-data*

- *Linked Data API Discuss*: *https://groups.google.com/forum/#!forum/linked-data-api-discuss*
- *Europeana Linked Data Pilot*: *https://groups.google.com/forum/#!forum/europeana-lod*
- *cerif-linked-data*: *https://groups.google.com/forum/#!newtopic/cerif-linked-data*
- *UK Government Linked Data Working Group*: *https://groups.google.com/forum/#!forum/ukgovld*
- *Linked Data And Services*: *https://groups.google.com/forum/#!forum/linkeddataandservices*
- *RapidMiner*: Linked Open Data Extension *https://groups.google.com/forum/#!forum/rmlod*
- *DyLDO*: *https://groups.google.com/forum/#!forum/dyldo*
- *LinkedData.tw*: *https://groups.google.com/forum/#!forum/linkeddatatw*
- *annalist-discuss*: *https://groups.google.com/forum/#!forum/annalist-discuss*
- *LinkedData*: *https://groups.google.com/forum/#!forum/linkeddata*
- *Callimachus Discussion*: *https://groups.google.com/forum/#!forum/callimachus-discuss*
- *LOD Gazetteer Consortium*: *https://groups.google.com/forum/#!forum/lod-gc*
- *Linked Open Data in Libraries, Archives, & Museums*: *https://groups.google.com/forum/#!forum/lod-lam*
- *linkeddatastack*: *https://groups.google.com/forum/#!forum/linkeddatastack*
- *NeoGeo Semantic Web Vocabs*: *https://groups.google.com/forum/#!forum/neogeo-semantic-web-vocabs*
- *LDSpider*: *https://groups.google.com/forum/#!forum/ldspider*

3. Preparação de um documento para avaliação manual: foi elaborado um documento para avaliação manual, o qual continha um questionário abordando as dimensões e métricas utilizadas no presente trabalho.

4. Execução do experimento: envio dos *links* das abordagens manual (questionário *online*) e semiautomática (ferramenta), juntamente com os *datasets* para os e-mails dos especialistas (ver mensagem enviada no Apêndice C).

5. Coleta de dados: as notas da ferramenta semiautomática foram armazenadas no *Virtuoso Universal Server* utilizando a ontologia *Data Quality Vocabulary*. As notas do questionário foram armazenadas pela ferramenta *Qualtrics* (utilizada para criar a abordagem manual).

6. Análise dos dados: Foi feita uma análise estatística dos dados, com a criação de um histograma das notas (também chamadas de scores) e análise de Boxplots desses scores para responder as hipóteses.

Após a execução dos passos anteriores, os resultados poderão ser obtidos e comparados entre si.

5.6 Instrumentação

- Na abordagem manual, foi utilizado um questionário feito na ferramenta Qualtrics (ver detalhamento do questionário no Apêndice A),
- Na abordagem semiautomática, foi utilizado uma ferramenta desenvolvida no presente trabalho,
- As análises dos resultados foram feitas utilizando a ferramenta R.

5.7 Ameaças à validade

5.7.1 Ameaças à validade interna

- Instrumentação: está relacionada com a diferença de resultados, que pode ser consequência de uma medição executada de maneira incorreta ou pelo fato de ter usado instrumento inadequado ao experimento (LIMA, 2014);
- Maturação: refere-se desmotivação das pessoas envolvidas no decorrer do experimento ou ao aumento da capacidade, com o passar do tempo (LIMA, 2014);
- Mortalidade seletiva: refere-se à evasão de participantes com características específicas e relevantes durante o experimento (LIMA, 2014);
- Contaminação: por exemplo, participantes que fazem parte de um grupo ensinam, ou aprendem, com participantes do outro grupo, enviesam o resultado desejado da amostra (LIMA, 2014);

5.7.2 Ameaças à validade de constructo

- É possível que a quantidade mínima de avaliadores e critérios de validação não seja suficiente para observar diferenças significativas de eficácia e eficiência na avaliação.
- Número baixo de *datasets*, embora tenham sido feitas muitas observações, apenas dois *datasets* pode ser considerado um número baixo. Talvez um número maior de *datasets*, consequentemente com mais notas coletadas, os resultados poderiam ser mais precisos.

- Cansaço dos avaliadores durante o experimento: O experimento envolveu a verificação de três *datasets* com e sem a abordagem semiautomática, ou seja, foram feitos seis avaliações, tornando o experimento um pouco exaustivo, o que pode ter influenciado em algumas das notas dos participantes.

5.8 Análise estatística

Nesta seção é apresentada a análise estatística dos dados obtidos da amostra definida na seção 5.6. Desta forma, a análise que será apresentada foi feita com base do *design* de experimento descrito anteriormente. A análise estatística visa evidenciar através de gráficos e números os resultados da análise que mostram qual das abordagens desenvolvidas (manual e semiautomática) contribui com uma melhor avaliação da qualidade de *datasets* conectados.

Com isso, será analisado o comportamento estatístico dos *Scores* parciais (dimensões): *Availability, Provenance, Licensing, Currency, Understandibility, Versatility* e *Representational consistency*, e também do *Score* Global (nota geral do *dataset*) construído com base nos *scores* das dimensões. Desta forma, inicialmente vamos investigar graficamente a distribuição de probabilidades destas variáveis, inclusive no sentido de identificar possível normalidade. Neste sentido, apresentamos nas Figuras 33 e 34 histogramas e *QQplots* (Barbetta, Reis e Bornia, 2008), respectivamente.

Nos histogramas as curvas em linha são uma estimativa de distribuição normal para cada *Score* parcial. Como é possível observar, essas linhas não se ajustam bem aos histogramas para a maioria das variáveis, inclusive com bases nestas curvas em alguns casos ocorreria até a possibilidade das notas assumirem valores negativos, o que não é possível, dado que os valores das variáveis se encontram no intervalo $(0, 1)$.

Para o *Score* parcial *Currency* e o *Score* Global a aderência da curva ao histograma é relativamente melhor que as outras variáveis, mas ainda a curva em linha estimaria valores negativos, o que não seria adequado. Logo, com base nos histogramas, a suposição de normalidade das variáveis envolvidas no problema deveria ser descartada.

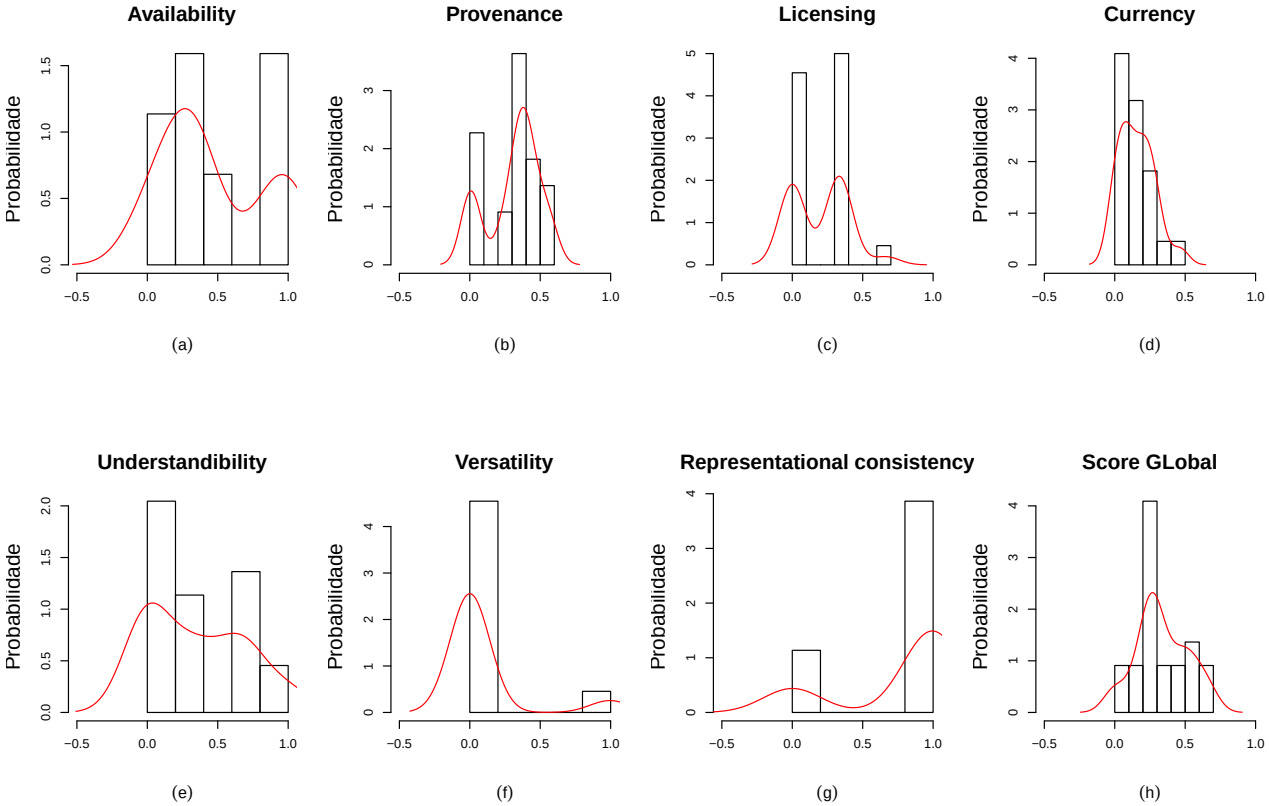

Figura 33 – Histogramas das variáveis: *Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency* e Score *Global*
Fonte: Elaboração própria

Para corroborar com a assertiva acima, foram construídos *QQplots*. Com isso, esses gráficos compararam os valores das variáveis com que seriam seus respectivos valores, caso de fato a variável fosse normal. Assim, esses gráficos são ferramentas poderosas na investigação de normalidade.

É considerado que a distribuição da variável está bem representada pela distribuição normal quanto mais os pontos aderem à reta. Com base na Figura 34, nota-se que para a maioria das variáveis relativas à dimensão não apresenta boa aderência à reta.

Desta forma, o *Score* parcial *Currency* e o *Score* Global são as que apresentam uma certa aderência, mas não ao ponto de considerar que a distribuição normal é a que descreve melhor o comportamento probabilístico destas duas variáveis.

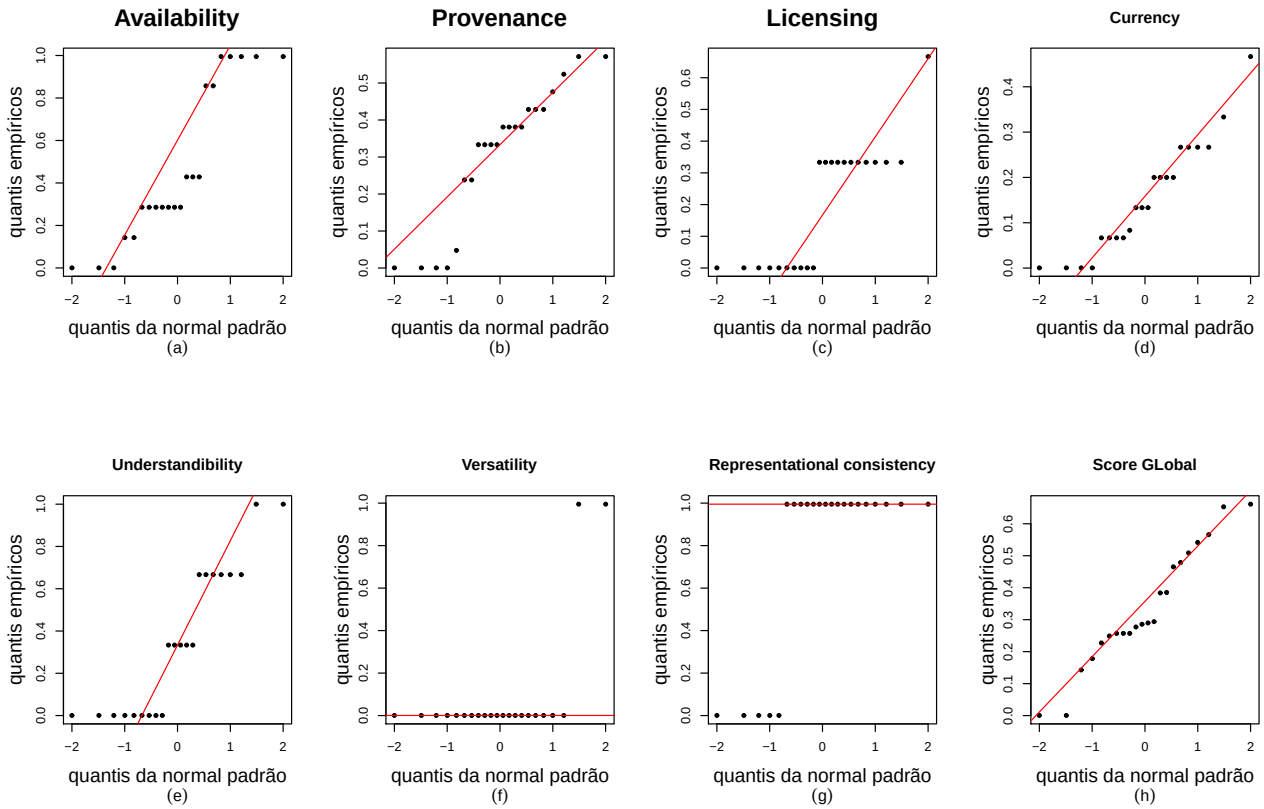

Figura 34 – QQplots das variáveis: *Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency* e *Score* Global considerando todas as observações.

Fonte: Elaboração própria

Em seguida, foi analisado separadamente o *Score* parcial *Time*. Com base na Figura 35 pode-se notar que a suposição de normalidade não é adequada para este *Score* parcial , uma vez que, tanto no histograma quanto no *QQplot* não existe aderência à curva (histograma) e à reta (*QQplot*).

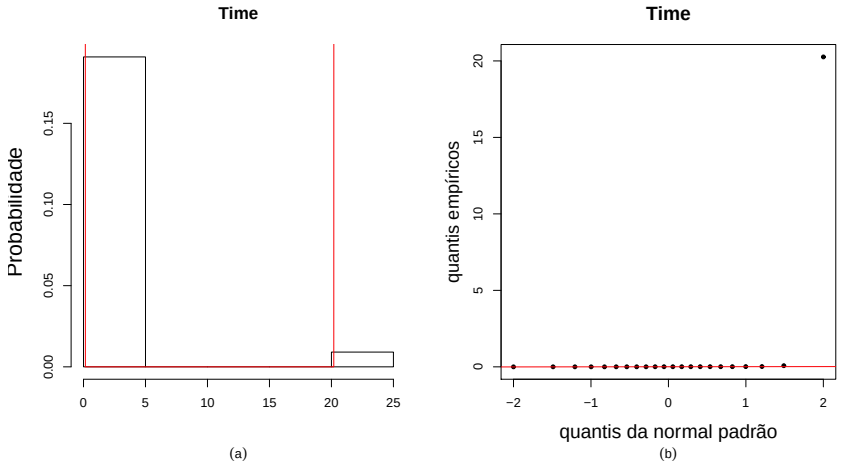

Figura 35 – Histograma e QQplots da variável: Time, considerando todas as observações.
Fonte: Elaboração própria

Os gráficos anteriores são ferramentas bastante robustas para avaliação de normalidade de variáveis. Pode-se também utilizar testes de hipóteses tradicionais para comparar com os resultados obtidos graficamente.

Com base no comando *shapiro.test* do programa R, realizamos o teste *Shapiro-Wilk* (Barbetta, Reis e Bornia,. 2008); (Thas, O. 2010) para normalidade, neste caso, considerando todos os *Scores*:

- **Time:** W = 0.22335, p–value = 7.646e-10;
- **Availability:** W = 0.83901, p–value = 0.002183;
- **Provenance:** W = 0.877, p–value = 0.01059;
- **Licensing:** W = 0.73811, p-value = 6.227e-05;
- **Currency:** W = 0.92734, p-value = 0.108;
- **Understandibility:** W = 0.8323, p–value = 0.001678;
- **Versatility:** W = 0.33222, p–value = 4.875e-09;
- **Representational consistency:** W = 0.52227, p–value = 2.142e-07;
- **Score Global:** W = 0.95411, p–value = 0.38.

A notação, por exemplo, 0.2e-03 significa $0.2 \times 10^{-03} = 0.0002$. Como os p–valores são inferiores a 5% rejeita-se a suposição de normalidade (H_0) para todos os *Scores*.

Visto que os *Score* envolvidos não apresentam boa aproximação para a distribuição normal, os testes usuais de comparação de médias, testes t–*Student* não podem ser utilizados.

Uma opção são os testes não paramétricos. Neste caso, o teste utilizado é o *kruskal–Wallis* (Thas, O. 2010). Utilizando o comando *kruskal.test* do programa R, Tabela 28. Com base nesta tabela, quando comparado os resultados dos grupos semiautomático e nota base, não existe diferença significativa em relação a qualquer critério avaliado, uma vez que os p–Valores são superiores a 5% indicando a NÃO rejeição de H_0: Com isso, Não existem diferenças entre as duas populações.

Quanto aos grupos Nota base e Manual, existem diferenças significativas, ou ao nível de 5%, ou ao nível de 10%, quanto aos *Scores* parciais *Availability, Representational* e o *Score* Global. Isto quer dizer, por exemplo, que o *Score* global para o grupo base difere significativamente do *Score* global do grupo manual, neste caso, a um nível de 10%, pois o p–Valor$= 0.09 < 0.10$.

Para alguns *Scores* parciais, como *Representational* (Nota Base e Semiautomático) não foi possível realizar o teste de *Krukal–Wallis*, pois essa variável não é aleatória quando subdividida em dois grupos, visto que apresentam todos os valores iguais a um. Finalmente, comparando os grupos Manual e semiautomático existem diferenças significativas, ou ao nível de 5%, ou ao nível de 10%, mais uma vez quanto as variáveis: *Availability, Representational* e *Score* Global.

Tabela 28 – Teste *Kruskal–Wallis*. Comparação das variáveis: *Time, Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency* e *Score* Global. H_0: Não existe diferença entre o comportamento da variável entre os dois grupos. Pares de grupos: Nota Base e Semiautomático, Nota Base e Manual e Manual e Semiautomático.

Nota Base e Semiautomático								
Time	Availability	Provenance	Licensing	Currency	Understan-dability	Versatility	Representa-tional	Score Global
0.5371	0.3639	1.000	0.8586	0.3458	0.5479	0.3248	-	1.000
Nota Base e Manual								
Time	Availability	Provenance	Licensing	Currency	Understan-dability	Versatility	Representa-tional	Score Global
0.7576	0.07035	0.4334	0.7313	0.3402	0.6231	-	0.06789	0.08919
Manual e Semautomático								
Time	Availability	Provenance	Licensing	Currency	Understan-dability	Versatility	Representa-tional	Score Global
0.4015	0.07037	0.1075	0.8038	0.6541	0.9631	0.1449	0.01056	0.05699

Fonte: Elaboração própria

Também é possível utilizar ferramentas gráficas para comparar o comportamento de variáveis entre grupos. Para tal, é interessante o uso de *boxplots*, como mostrado nas Figuras 36– 38. Com base na Figura 36, podemos notar a diferença entre os grupos Nota base *vs* Manual e Manual *vs* Semiautomático quanto à variável *Availiability*, de fato, o grupo Manual apresenta uma média $(\bar{Y})$ inferior aos demais grupos.

Quanto *Representational consistency*, fica evidente que não existe diferença entre os grupos Nota base *vs* Semiautomático, e que o grupo Manual difere drasticamente dos outros dois apresentando uma média substancialmente inferior. Mais uma vez, o grupo manual apresenta

uma média inferior às médias dos grupos semiautomático e base quanto ao *score* global, enquanto que, semiautomático e base não diferem entre si.

Nas Figuras 37– 38 são apresentados os são apresentados os *boxplots* das variáveis *Provenance, Licensing, Currency, Understandibility* e *Versatility* em que as médias não apresentaram diferença estatisticamente significante entre os três grupos.

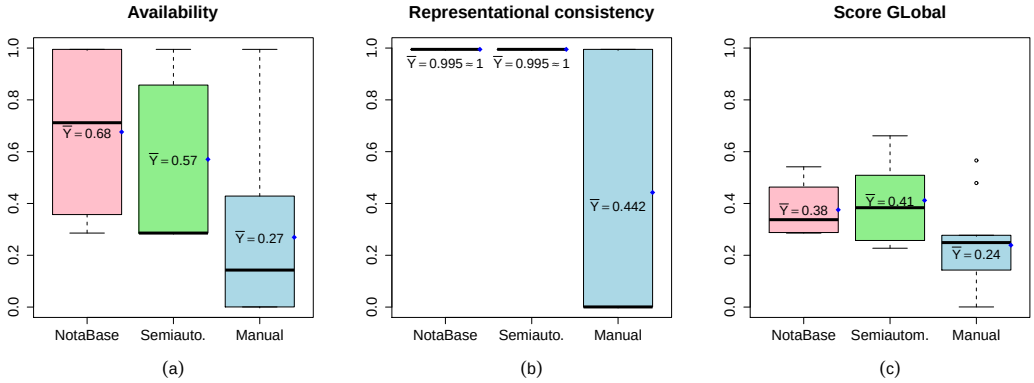

Figura 36 – Boxplot de *Avaliability, representational consistency* e *score* global considerando os grupos: Nota base, Semiautomático e Manual

Fonte: Elaboração própria

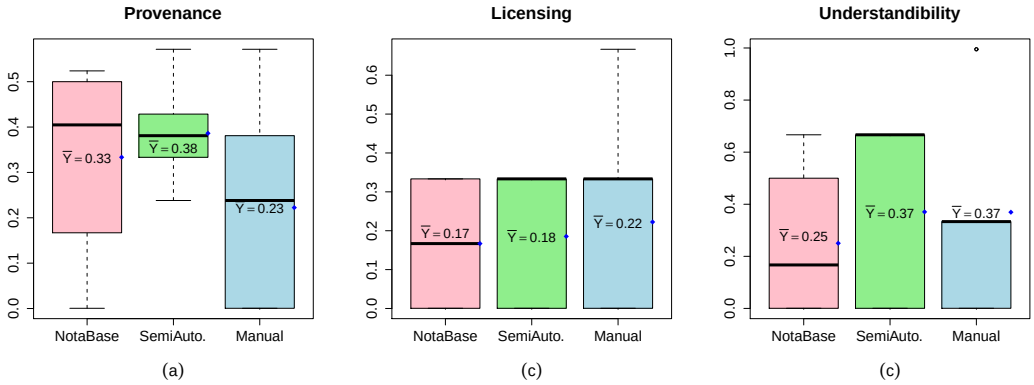

Figura 37 – Boxplot de *Provenance, Licensing* e *Understandibility* considerando os grupos: Nota base, Semiautomático e Manual.

Fonte: Elaboração própria

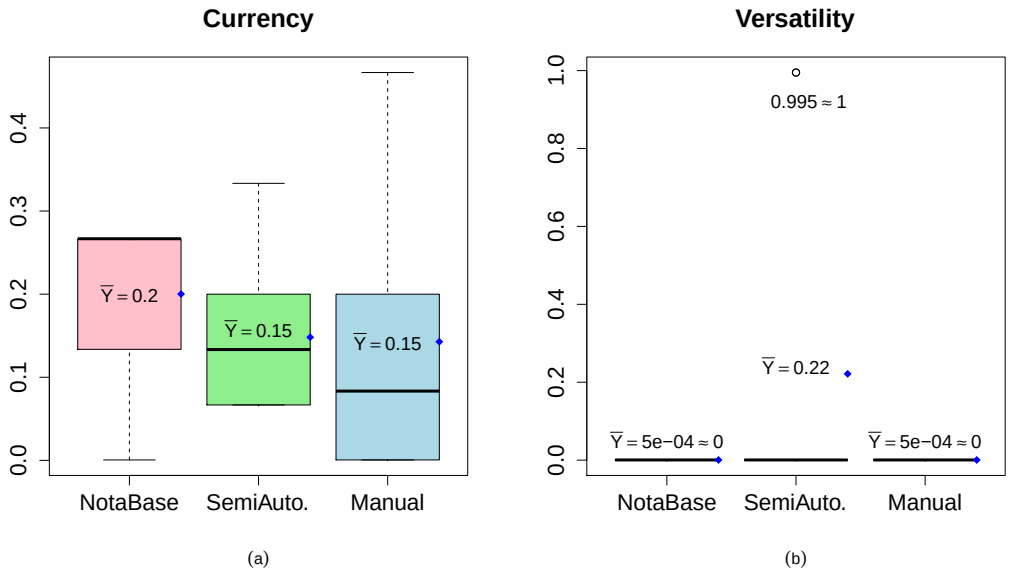

Figura 38 – Boxplot de *Currency* e *Versatility*, considerando os grupos: Nota base, Semiautomático e Manual.
Fonte: Elaboração própria

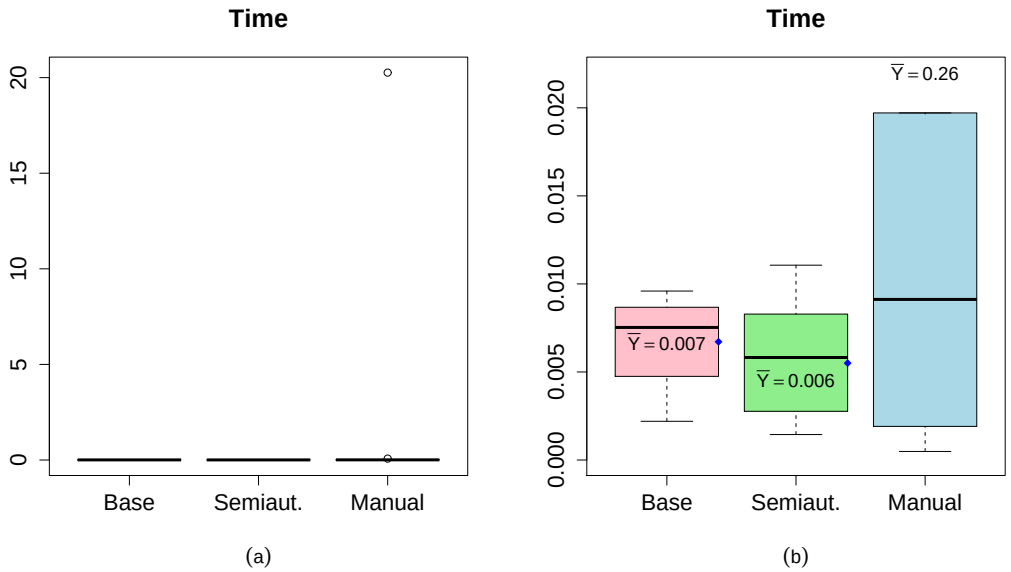

Figura 39 – Boxplot de *Time* considerando os grupos: Nota base, Semiautomático e Manual.
Fonte: Elaboração própria

No que diz respeito à variável *Time*, existe a ocorrência de um *outlier* para o grupo Manual que não permite a visualização adequada do comportamento comparativo entre os três grupos (Figura 38a). Na Figura 37b os *boxplots* foram construídos sob uma nova escala, sem a retirada do *outlier*, mas uma escala que permite avaliar os três grupos em um intervalo inferior ao valor do *outlier*. Ou seja, o *outlier* é considerado para a construção do *boxplot*, para o cálculo da média, no entanto, é feito um corte no gráfico para poder verificar o que ocorre para valores comparáveis da variável *Time* entre os três grupos. Com base nesta figura, nota-se uma proximidade entre as medianas de *Time* considerando os grupos. Por outro lado, o outlier presente no grupo manual eleva consideravelmente o valor da média de time deste grupo. No entanto, os testes de hipótese conseguem ser robustos a este ponto extremo e

concluem que de fato, as médias de time não diferem estatisticamente para os três grupos avaliados. Mais uma vez deve ser ressaltado que o outlier não foi retirado da amostra.

Tabela 29 – Medidas resumo das variáveis Time Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency e Score GLobal.

	Variáveis				
Estatística Resumo	Time	Availability	Provenance	Licensing	Currency
Mínimo	0.000486	0.0005	0.0005	0.0005	0.00050
Primeiro quartil	0.002341	0.2857	0.2381	0.0005	0.06667
Madiana	0.007216	0.2857	0.3571	0.3333	0.13333
Média	0.930211	0.4665	0.3096	0.1972	0.15539
Terceiro Quartil	0.009476	0.8571	0.4286	0.3333	0.25000
Máximo	20.261898	0.9950	0.5714	0.6667	0.46667
Estatística Resumo	Understandibility	Versatility	Representational consistency	Score GLobal	
Mínimo	0.0005	0.00050	0.0005	0.0005	
Primeiro quartil	0.0005	0.00050	0.9950	0.2510	
Madiana	0.3333	0.00050	0.9950	0.2878	
Média	0.3482	0.09091	0.7690	0.3345	
Terceiro Quartil	0.6667	0.00050	0.9950	0.4755	
Máximo	0.9950	0.99500	0.9950	0.6612	

Fonte: Elaboração própria

Na Tabela 29 estão as estatísticas descritivas de todas as variáveis considerando os três grupos como um todo. Com base nesta tabela é possível verificar o que ocorre com variáveis como *Licensing, Versatility* e *Represational*, as quais apresentam valores extremos iguais a 0.0005 ou iguais a 0.995. Também é possível verificar que só existe um valor extremo para *Time* igual a 20.26. No caso da variável *Representational Consistency* para o grupo manual o conjunto de observações é igual a $(0.0005, 0.0005, 0.0005, 0.0005, 0.0005, 0.9950, 0.9950, 0.9950, 0.9950)$ em que $0.0005 \approx 0$ e $0.9950 \approx 1$. Desta forma é possível notar que o mínimo e a mediana são iguais a 0.0005, a média aritmética fica igual a 0.442 exatamente devido aos valores extremos altos 0.995 que puxam a média para um valor muito superior à mediana.

5.9 Discussão

Nesta seção é apresentada uma discussão sobre os resultados apresentados na seção 5.8, sobre a análise estatística dos *Scores* parciais e os *Scores* globais obtidos da avaliação utilizando as abordagens manual e semiautomática.

Desta forma, com a finalidade de verificar a normalidade das variáveis foram gerados histogramas dos *Scores* parciais (dimensões): *Availability, Provenance, Licensing, Currency, Understandibility, Versatility* e *Representational consistency*, e também do *Score* Global (nota geral do *dataset*), como apresentado nas Figuras 33 e 35. Desta forma, ao observar os histogramas, percebemos que as linhas não se ajustam bem aos histogramas para todos os *Scores*, constatando a não normalidade das variáveis.

Com isso, para estudar melhor o comportamento, foram gerados também *QQplots* dos *Scores*, como apresentado nas Figuras 34 e 35, e podemos perceber que também não há boa adequação da linha à reta, comprovando mais uma vez a não normalidade das variáveis.

Outra forma utilizada neste trabalho para verificar a normalidade dos *Scores*, foi o teste de hipóteses *Shapiro-Wilk* (Barbetta, Reis e Bornia,. 2008); (Thas, O. 2010) e constatou-se mais uma vez a não normalidade das variáveis, pois os p–value obtidos dos *Scores* foram inferiores a 5%, rejeita-se mais uma vez a suposição de normalidade para todos os *Scores*.

Com isso, após a identificação de não normalidade dos *Scores* de uma forma geral, foi utilizado o teste *kruskal–Wallis* (Thas, O. 2010), com a finalidade de identificar o comportamento dos *Scores* entre os grupos semiautomático, nota base e manual.

Como podemos ver na Tabela 28, os p–Valores são superiores a 5% indicando a NÃO rejeição de H_0, ou seja, os *Scores* parciais *Availability, Provenance, Licensing, Currency, Understandibility, Versatility* e *Representational consistency*, e também do *Score* Global dos grupos Nota base e Semiautomático são idênticos. Isso significa que as notas dos especialistas (grupo nota base) não diferem das notas dos avaliadores que utilizaram a abordagem semiautomática.

Quanto aos grupos Nota base e Manual, os *Scores* parciais *Availability, Representational* e o *Score* Global possuem os p–Valores inferiores a 10%, ocasionando na rejeição de H_0, ou seja, os valores desses *Scores* para os grupos Nota base e Manual são bastantes diferentes. Em outras palavras, as notas dos especialistas (grupo nota base) são diferentes das notas dos avaliadores que utilizaram a abordagem manual.

Desta forma, com a finalidade de comparar graficamente o comportamento das variáveis também entre os grupos foi utilizado o gráfico *boxplot*. Com isso, na Figuras 36, 37 e 38, as medianas dos *Scores* parciais *Availability, Representational, Provenance, Licensing, Time* e o *Score* Global, considerando o grupo Nota base e Manual se aproximam consideravelmente. No entanto, os *Scores* parciais *Understandibility e Currency* possuem as medianas iguais considerando os grupos semiautomático e manual.

Portanto, mediante os resultados apresentados, foi evidenciado que a abordagem semiautomática contribui para uma boa avaliação, dado que as notas dos avaliadores que utilizaram tal abordagem se aproximaram consideravelmente das notas dos especialistas.

6 Conclusão

Conforme exposto ao longo deste trabalho, a transparência da qualidade de dados para os usuários é uma necessidade elencada pela literatura e que tem sido o objetivo de muitos trabalhos (Behkamal et. al (2014), Debattista et. al (2016), Loshin (2010) e Zaveri et. al (2013)). Além disso, evidenciar a qualidade de um *dataset* conectado, bem como, os aspectos positivos e negativos contribuem para um aperfeiçoamento do mesmo, para que haja uma boa reutilização por parte dos consumidores interessados. Facilitar o entendimento de *datasets* conectados para seu fácil consumo, seja por humanos, ou por agentes computacionais, é o principal objetivo de publicação dos mesmos, senão para que haveria a necessidade de publicá-los?

Com isso, para um melhor entendimento deste trabalho, no capítulo 2 foi apresentado os principais conceitos relacionados a este trabalho, possibilitando ao leitor ter uma melhor compreensão das abordagens desenvolvidas, com base na metodologia GQM utilizando dimensões de qualidade juntamente com melhores práticas para publicação de *datasets* conectados.

O capítulo 3 apresenta os trabalhos relacionados que também buscaram desenvolver soluções para o problema de pouca qualidade de *datasets* conectados. No entanto, cada um dos trabalhos adentram em um aspecto de qualidade diferente e também implementam soluções diferentes. Todos os trabalhos apresentados são importantes, e a utilização deles depende da necessidade do usuário.

No capítulo 4 é apresentada a solução desenvolvida com base do GQM mediante o uso de dimensões de qualidade de *datasets* conectados. Na seção 4.2.4 foi apresentado a abordagem semiautomática, que consiste na implementação computacional do GQM desenvolvido neste trabalho; No Apêndice A foi apresentado a abordagem manual, que consiste em um questionário construido também com base no GQM, que possui uma avaliação idêntica ao da abordagem semiautomática, no entanto a avaliação é totalmente manual. Ambas as abordagens podem ser utilizadas para avaliar a qualidade de *datasets* conectados, no entanto, a semiautomática diminui consideravelmente a carga de trabalho e o tempo de avaliação do usuário.

A criação dessas abordagens deu-se em necessidade de evidenciar que, o auxílio de uma ferramenta semiautomática é vantajoso em relação à utilização da abordagem manual (utilizada pela maior parte dos trabalhos relacionados), visto que o principal diferencial da abordagem semiautomática é a execução automática de algumas etapas da avaliação e o relatório detalhado que é gerado e disponibilizado ao final da avaliação, que exibe de forma transparente os resultados da avaliação para o usuário, de cada dimensão de qualidade, bem como, de cada métrica e critério.

A solução proposta foi criada a partir de dimensões utilizadas pela literatura elencadas por Zaveri (2012). Atrelado ás dimensões, foram utilizadas também melhores práticas para

publicação de *datasets* conectados consideradas padrões *Web* internacionais de publicação de dados no âmbito da *Web* Semântica.

Ainda no capítulo 4 é mostrada a implementação da solução, abordando o desenvolvimento da ontologia VDC, através da reutilização da ontologia DQV desenvolvida pelo W3C. Também é apresentado, os casos de uso das funcionalidades do sistema possibilitando o entendimento das dimensões e métricas executadas de forma manual e automática. Com o objetivo de contribuir com o desenvolvimento de outros sistemas que implementam também dimensões de qualidade ou a expansão futura do presente trabalho, foi apresentado a arquitetura de componentes desenvolvida para implementação do código do sistema.

Desta forma, considerando o lado do avaliador, que poderá usufruir da abordagem semiautomática desenvolvida, no capítulo 4 é mostrado o passo a passo de utilização da ferramenta na perspectiva do usuário. Com isso, todas as telas de avaliação das dimensões, contendo suas respectivas perguntas são delineadas para facilitar a utilização da ferramenta.

O capítulo 5 detalha a validação das abordagens desenvolvidas, com o intuído de buscar identificar qual das abordagens é mais vantajosa ao ser utilizada, ou se não há diferença em utilizá-las. Este capítulo também contribui para a comunidade científica no que se refere à compreensão da metodologia utilizada para a avaliação do experimento de um trabalho científico no contexto do trabalho apresentado.

Desta forma, no capítulo 6 é constatado que a solução criada (abordagem semiautomática) atingiu o objetivo proposto, que foi além de diminuir a carga de trabalho dos usuários de *datasets* conectados, diminuir o tempo de avaliação. Desta forma, esta evidência foi comprovada através de análises estatísticas que mostraram que é vantajoso utilizar a ferramenta desenvolvida para fazer a avaliação de qualidade de *datasets* conectados.

Um resultado interessante, consideravelmente bom, é que não foi possível utilizar o teste de kruskal–Wallis para a dimensão Representational consistency, considerando os grupos Nota Base e Semiautomático, visto que todos os valores dessas variáveis são iguais, ou seja, a nota dos avaliadores que utilizaram a ferramenta (que desconheciam a ferramenta, e são pessoas espacialmente dispostas no Mundo) foi igual à nota dos avaliadores especialistas envolvidos na nota base que conheciam a ferramenta.

6.1 Trabalhos futuros

Como trabalhos futuros esta pesquisa propõe:

- Contribuir com o entendimento da área de qualidade de *datasets* conectados, para que novos trabalhos venham a ser desenvolvidos, tanto ampliando a pesquisa apresentada, conforme novas práticas e novas dimensão possam ser utilizadas, quanto na criação de novas soluções que possam abordar outros aspectos de qualidade.

- Utilizar algumas dimensões de qualidade implementadas neste trabalho em uma aplicação governamental real e ver na prática a qualidade de *datasets* conectados e seu

continuo aprimoramento, através de monitoramento que será feito em intervalos de tempo contínuos. Também poderão ser criadas novas dimensões necessárias que ainda não foram previstas pela literatura, bem como, reutilizadas dimensões elencadas pela Zaveri (2012) que não foram abordadas neste estudo.

- Publicar artigos em periódicos relevantes relacionados a área de pesquisa pertencente ao contexto deste trabalho, com a finalidade de disseminar o conhecimento obtido para a comunidade científica e fomentando o desenvolvimento de novas pesquisas.

7 REFERÊNCIAS

A. Bröckers, C. Differding and G. Threin. The Role of Software Process Modeling in Planning Industrial Measurement Programs. In Proceedings of the METRICŚ96. ICSE, 1996.

Auer, S. (2011). Creating knowledge out of interlinked data: making the a data washing machine. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics (p. 4). ACM.

BASSMAN, Mitchell J.; MCGARRY, Frank; PAJERSKI, Rose. Software measurement guidebook. 1994.

Behkamal, B., Kahani, M., Bagheri, E., & Jeremic, Z. (2014). A metrics-driven approach for quality assessment of linked open data. Journal of theoretical and applied electronic commerce research, 9(2), 64-79.

Bernadette Farias Lóscio, C. B. and Calegari, N. Data on the Web best practices. (2016). Disponível em: <https://www.w3.org/TR/2016/WD-dwbp-20160112/>. Acessado em Março de 2016.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, 2001. Disponível em: <http://www.cs.umd.edu/ golbeck/LBSC690/SemanticWeb.html>. Acesso em: 04 maio 2014.

Bizer, C., Heath, T., and Berners-Lee, T. Linked data - the story so far. Int. J. Semantic Web Inf. Syst., 5(3):122 (2009)

BRIAND, Lionel C.; DIFFERDING, Christiane M.; ROMBACH, H. Dieter. Practical guidelines for measurement-based process improvement. Software Process Improvement and Practice, v. 2, n. 4, p. 253-280, 1996.

CALDIERA, Victor R. Basili-Gianluigi; ROMBACH, H. Dieter. Goal question metric paradigm. Encyclopedia of Software Engineering, v. 1, p. 528-532, 1994.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In New Trends of Research in Ontologies and Lexical Resources (pp. 7-25). Springer Berlin Heidelberg.

Consoli, S., Gangemi, A., Nuzzolese, A. G., Peroni, S., Presutti, V., Recupero, D. R., and Spampinato, D. (2014). Geolinked open data for the municipality of catania. In Proceedings of the 4th international conference on Web intelligence, mining and semantics (WIMS14) (p. 58). ACM.

Cordeiro, K.F., Campos, M.L.M., & Borges, M.R.S. (2011). Empowering Citizens and Government with Collaboration on Linked Open Data. In Proc. of the Extended Semantic Web Conference (ESWC). Crete, Grece.

David Wood, Marsha Zaidman, L. R. M. H. Linked Data: Structured data on the Web. (2014)

de Mendonça, R. R., da Cruz, S. M. S., De La Cerda, J. F., Cavalcanti, M. C., Cordeiro,

K. F., and Campos, M. L. M. (2013). LOP: capturing and linking open provenance on LOD cycle. In Proceedings of the Fifth Workshop on Semantic Web Information Management (p. 3). ACM.

de Souza, F. M., da Cunha, A. M., & Torres, C. (2009). Uso do GQM para avaliar documentos de utilização de *framework*.

Debattista, J., Lange, C., Auer, S. Representing *Dataset* Quality Metadata using Multi-Dimensional Views. In: 2014, pp. 92–99

Debattista, Jeremy, and Christoph Lange. "Luzzu–A Framework for Linked Data Quality Assessment."2016 IEEE Tenth International Conference on Semantic Computing (ICSC). IEEE, 2016.

Debattista, Jeremy et. al. Data on the Web Best Practices: Data Quality Vocabulary. 2015. Disponível em: <https://www.w3.org/TR/2015/WD-vocab-dqv-20151217/>. Acesso em: abril 2016.

Albertoni, Riccardo. Guéret, Christophe. Isaac, Antoine. Data Quality Vocabulary. Revisão em 15 dez. 2016. Disponível em: <https://www.w3.org/TR/2015/WD-vocab-dqv-20150625/>. Acesso em 30 dez 3016.

Marco Baccaro. Arquitetura baseada em Componentes. 2010. Disponível em <https://marcobaccaro.wordpress.com/2010/10/05/arquitetura-baseada-em-componentes/>. Acesso em: 2 jan. 2017.

Ding, L., Lebo, T., Erickson, J. S., DiFranzo, D., Williams, G. T., Li, X., ... and Flores, J. (2011). TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web, 9(3), 325-333.

Erkimbaev, A. O., Zitserman, V. Y., Kobzev, G. A., Serebrjakov, V. A., and Teymurazov, K. B. (2013). Publishing scientific data as linked open data.Scientific and Technical Information Processing, 40(4), 253-263.

Fitzgerald, A., Hooper, N., and Cook, J. S. (2013). Implementing open licensing in government open data initiatives: a review of Australian government practice. In Proceedings of the 9th International Symposium on Open Collaboration (p. 39). ACM.

Fontoura, L. M., Price, R. T., & Phil, D. (2004). Usando GQM para gerenciar riscos em projetos de software. JF de Castro, editor, 18, 39-54.

Foulonneau, M., Martin, S., & Turki, S. (2014). How open data are turned into services?. In Exploring Services Science (pp. 31-39). Springer International Publishing.

Foulonneau, M., Martin, S., and Turki, S. (2014). How open data are turned into services?. In Exploring Services Science (pp. 31-39). Springer International Publishing.

Frosterus, M., Hyvönen, E., & Laitio, J. (2011). Datafinland—a semantic portal for open and linked datasets. In The Semanic Web: Research and Applications (pp. 243-254). Springer Berlin Heidelberg.

Fuggetta, A., Lavazza, L., Morasca, S., Cinti, S., Oldano, G., & Orazi, E. (1998). Applying GQM in an industrial software factory. ACM Transactions on Software Engineering

and Methodology (TOSEM), 7(4), 411-448.

Fürber, C., and Hepp, M. (2013). Using Semantic Web Technologies for Data Quality Management. In Handbook of Data Quality (pp. 141-161). Springer Berlin Heidelberg.

GRESSE, Christiane; HOISL, Barbara; WÜST, Jürgen. A process model for GQM-based measurement. Universität Kaiserslautern, 1995.

GUENTHER, Holger; ROMBACH, H.. Dieter ; RUHE, Guenther. Kontinuierliche Qualitätsverbesserung in der Software-Entwicklung: Erfahrungen bei der Allianz Lebensversicherungs AG. Wirtschaftsinformatik, v. 38, n. 2, p. 160-171, 1996.

H. Günther, H. D. Rombach, and G. Ruhe. Kontinuierliche Qualitätsverbesserung inder Software Entwicklung - Erfahrungen bei der Allianz Lebensversicherungs-AG (in German). Wirtschaftsinformatik 38, 1994.

Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. (2014). The linked medical data access control framework. Journal of biomedical informatics, 50, 213-225.

Kaschesky, M., and Selmi, L. (2013). Fusepool R5 linked data framework: concepts, methodologies, and tools for linked data. In Proceedings of the 14th Annual

International Conference on Digital Government Research (pp. 156-165). ACM.

Lassinantti, J., Bergvall-Kåreborn, B., and Stahlbröst, A. (2014). Shaping local open data initiatives: politics and implications. Journal of theoretical and applied electronic commerce research, 9(2), 17-33.

Loshin, David. The practitioner's guide to data quality improvement. Elsevier, 2010.

M.K. Daskalantonakis. A Practical View of Software Measurement and Implementation Experiences within Motorola. IEEE Transactions on Software Engineering, Vol. 18, No. 11, 1992.

Marden, J., Li-Madeo, C., Whysel, N., and Edelstein, J. (2013). Linked open data for cultural heritage: evolution of an information technology. In Proceedings of the 31st ACM international conference on Design of communication (pp. 107-112). ACM.

Marjit, U., Sharma, K., & Biswas. U. 2012. Provenance Representation and Storage Techniques in Linked Data: A State-of-the-art Survey. Int. J. of Computer Applications 38(9):23-28.

Martins, João Vitor P. Geração de um modelo de métricas para controle de qualidade no processo de desenvolvimento de software. 2011

Mendonça, R. R., da Cruz, S. M. S., De La Cerda, J. F., Cavalcanti, M. C., Cordeiro, K. F., & Campos, M. L. M. (2013). LOP: capturing and linking open provenance on LOD cycle. In Proceedings of the Fifth Workshop on Semantic Web Information Management (p. 3). ACM.

Ne, M., Chlapek, D., Kucera, J., Maurino, A., Konecny, M., and Vanova, L. (2015). Methodology for publishing datasets as open data. pages 1–31.

Neves, Philippe Pereira. Aplicando a Abordagem GQM para Avaliar o Impacto da Adoção da Metodologia Ágil Scrum. 2012.

Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., and Yu, Y. (2011). Zhishi. me-weaving chinese linking open data. In The Semantic Web–ISWC 2011 (pp. 205-220). Springer Berlin Heidelberg.

Pabón, G., Gutiérrez, C., Fernández, J. D., and Martínez-Prieto, M. A. (2013).

Linked Open Data technologies for publication of census microdata. Journal of the American Society for Information Science and Technology, 64(9), 1802-1814.

Pathak, J., Kiefer, R. C., and Chute, C. G. (2012). Applying linked data principles to represent patient's electronic health records at Mayo clinic: a case report. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (pp. 455-464). ACM.

Plu, J., and Scharffe, F. (2012). Publishing and linking transport data on the web: extended version. In Proceedings of the First International Workshop on Open Data (pp. 62-69). ACM.

Salas, P. E. R., Martin, M., Mota, F. M. D., Auer, S., Breitman, K., and Casanova, M. (2012). Publishing statistical data on the web. InSemantic Computing (ICSC), 2012 IEEE Sixth International Conference on(pp. 285-292). IEEE.

Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C. S., Willighagen, E., Hajagos, J., ... and Stephens, S. (2011). Linked open drug data for pharmaceutical research and development. Journal of cheminformatics, 3(1), 19.

Shaon, A., Woolf, A., Crompton, S., Boczek, R., Rogers, W., and Jackson, M. (2011). An open source linked data framework for publishing environmental data under the uk location strategy. In Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web (Vol. 798).

Tao, C., Song, D., Sharma, D., & Chute, C. G. (2013). Semantator: Semantic annotator for converting biomedical text to linked data. Journal of biomedical informatics, 46(5), 882-893.

Villazon-Terrazas, B., Vila-Suero, D., Garijo, D., Vilches-Blazquez, L. M., Poveda-Villalon, M., Mora, J., ... and Gomez-Perez, A. (2012). Publishing Linked Data-There is no One-Size-Fits-All Formula.

Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., and Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. InLinking government data (pp. 27-49). Springer New York.

VON WANGENHEIM, Christiane Gresse; RUHE, Günther. Análise de Custo e Benefício de Mensuração Baseada em GQM-Um Estudo de Caso Replicado.1999.

WANGENHEIM, Christiane Gresse von. Utilização do GQM no desenvolvimento de software. Laboratório de Qualidade de Software, Instituto de Informática, Universidade do Vale do Rio dos Sinos. São Leopoldo, 2000.

Zaveri, A. et al. Quality Assessment Methodologies for Linked Open Data. In: Semantic Web Journal (2014).

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., & Hitzler, P.

(2013). Quality assessment methodologies for linked open data.Submitted to Semantic Web Journal.

Hartig, O., and Zhao, J. (2010). Publishing and consuming provenance metadata on the web of linked data. In Provenance and annotation of data and processes (pp. 78-90). Springer Berlin Heidelberg.

Galiotou, E., and Fragkou, P. (2013). Applying linked data technologies to Greek open government data: a case study. Procedia-Social and Behavioral Sciences, 73, 479-486.

A Apêndice

Este apêndice tem como finalidade mostrar as telas de avaliação da qualidade de *datasets* conectados utilizando a abordagem manual. O processo de avaliação utilizado nesta abordagem é igual ao processo de avaliação utilizado na abordagem semiautomática, no entanto, a avaliação da API e do RDF são feitos de forma automática.

Desta forma, ao iniciar a avaliação a primeira tela que será exibida é referente a dimensão Disponibilidade.

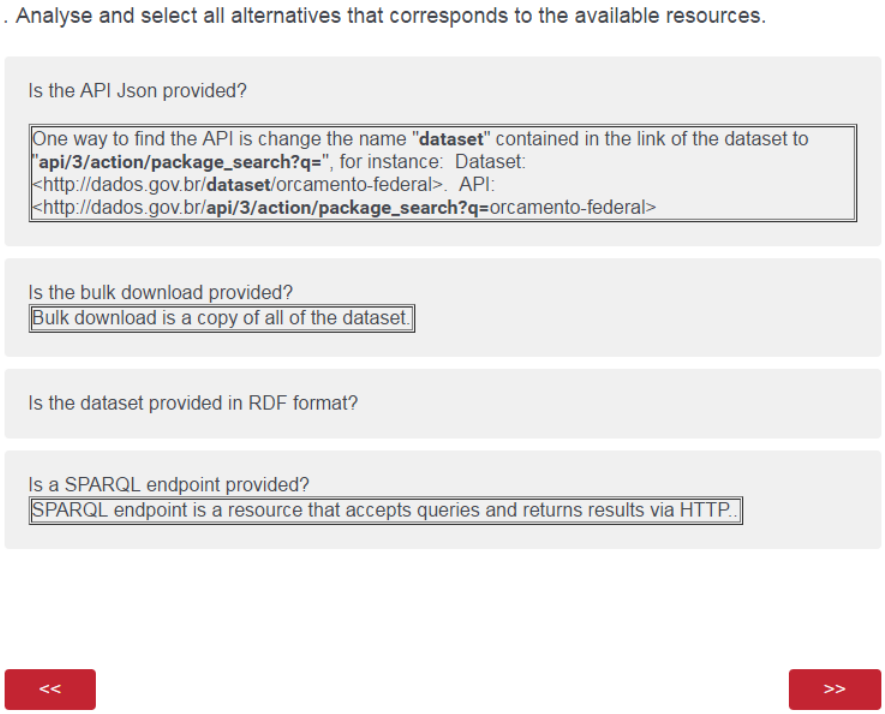

Figura 40 – Tela de avaliação da dimensão Disponibilidade
Fonte: Elaboração própria

Com isso, para selecionar as alternativas corretas, ou seja, dos recursos existentes no dataset, basta clicar em cima da alternativa desejada. Caso a alternativa relacionada ao *endpoint* SPARQL seja selecionada, a próxima tela que será exibida é relacionada a avaliação de disponibilidade do *endpoint* identificado (Figura 41)

Com isso, para responder a questão da referida tela, o usuário deverá submeter manualmente, três consultas no *endPoint* encontrado, e em seguida, selecionar na tela do questionário as consultas que retornaram resultado considerado correto.

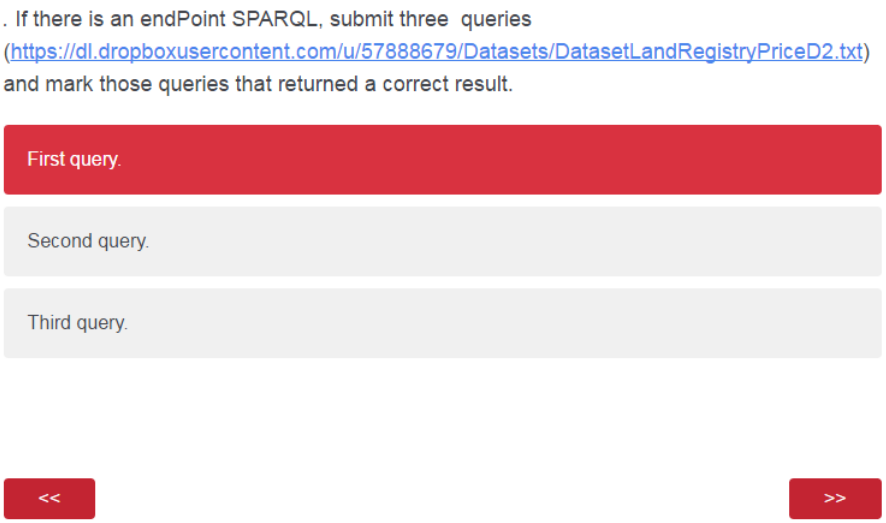

Figura 41 – Tela de avaliação da disponibilidade de um *endPoint* SPARQL
Fonte: Elaboração própria

Ao usuário concluir as respostas desta questão e clicar no botão next, aparecerá a tela de avaliação da dimensão Proveniência, como mostrado na Figura 42. Desta forma o avaliador deverá analisar a página *Web* do *dataset* buscando identificar os itens pedidos na pergunta da tela de avaliação.

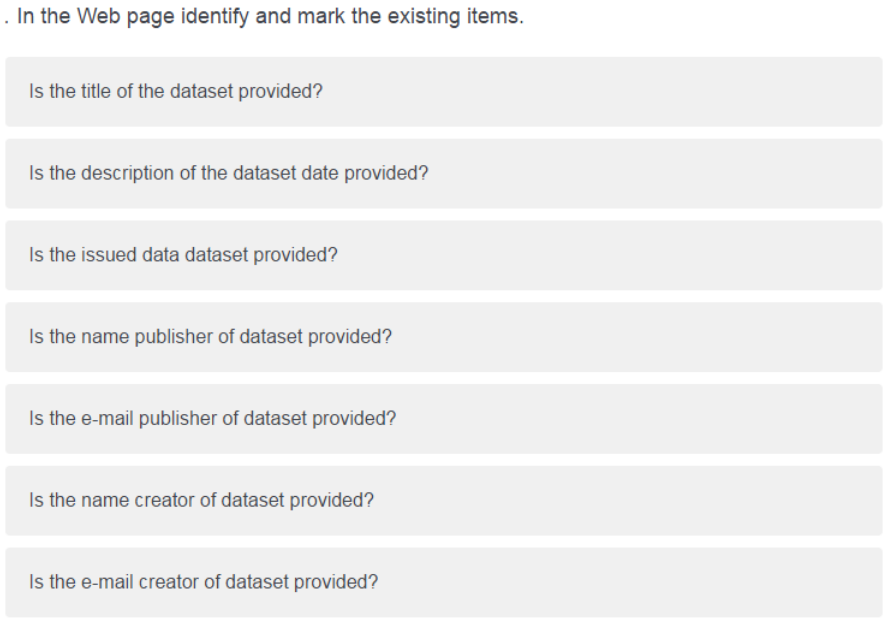

Figura 42 – Tela de avaliação da dimensão Proveniencia relacionada a página *Web*
Fonte: Elaboração própria

Ainda na mesma tela, são apresentadas as questões de avaliação da dimensão Proveniência voltadas para a API (Figura 43) e o RDF (Figura 44) (caso sejam encontrados). Os itens contidos nestas questões são os mesmos itens apresentados na Figura 42.

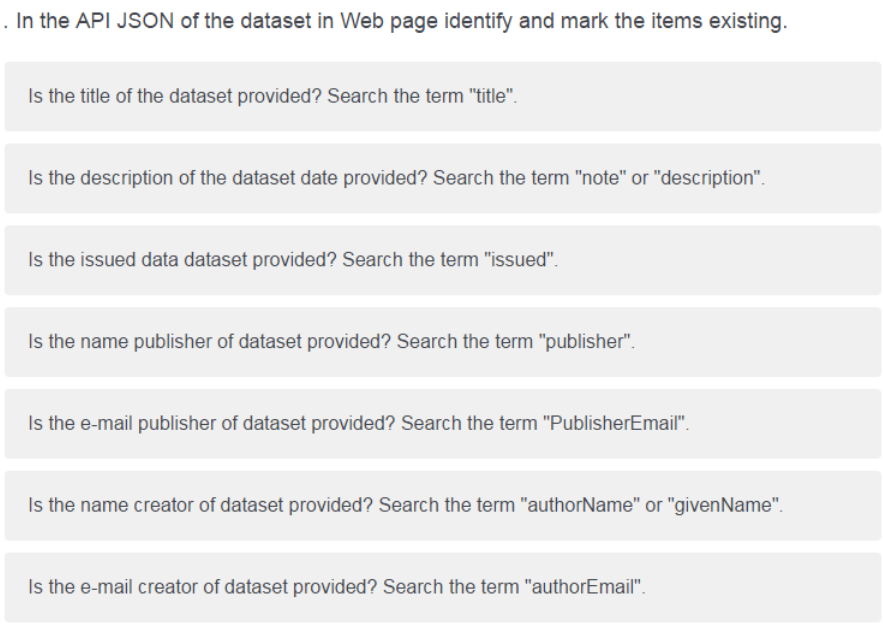

Figura 43 – Tela de avaliação da dimensão Proveniencia relacionada a API
Fonte: Elaboração própria

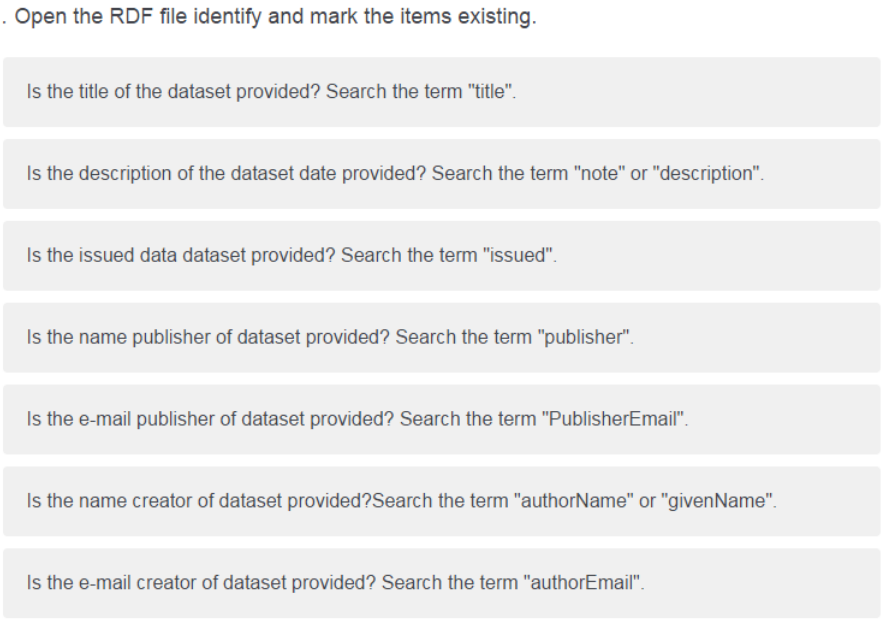

Figura 44 – Tela de avaliação da dimensão Proveniencia relacionada ao RDF
Fonte: Elaboração própria

Em seguida, ao clicar no botão next, será apresentado a tela de avaliação da dimensão Proveniência, como exibido na Figura 45.

Posteriormente, a próxima dimensão que será avaliada é Atualidade. Com isso, na referida tela, será mostrado três questões que possuem os mesmos tópicos de verificação, no entanto , as questões são voltadas para a página *Web* (Figura 46), API (Figura 47) e RDF (Figura 48), respectivamente.

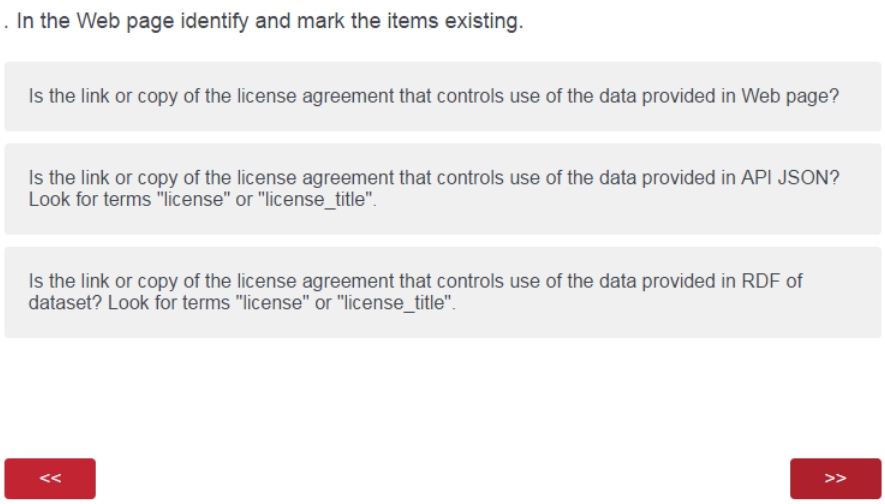

Figura 45 – Tela de avaliação da da dimensão Licença
Fonte: Elaboração própria

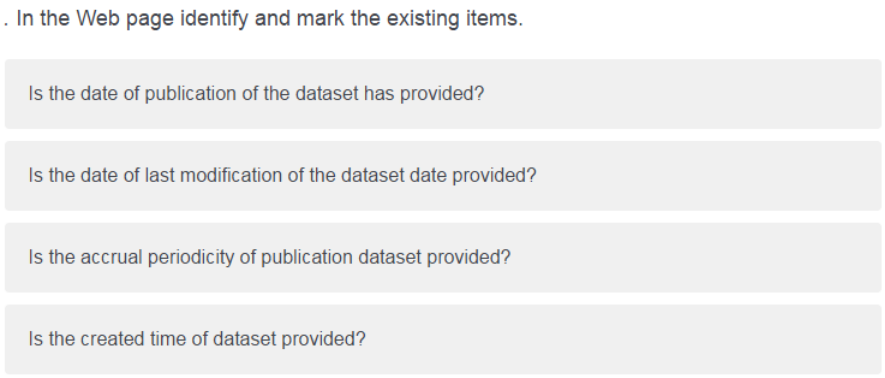

Figura 46 – Tela de avaliação da dimensão Atualidade relacionada a página *Web*
Fonte: Elaboração própria

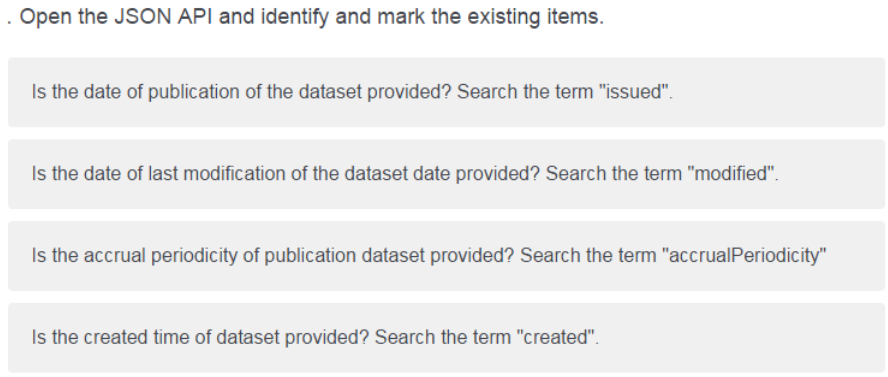

Figura 47 – Tela de avaliação da dimensão Atualidade relacionada a API
Fonte: Elaboração própria

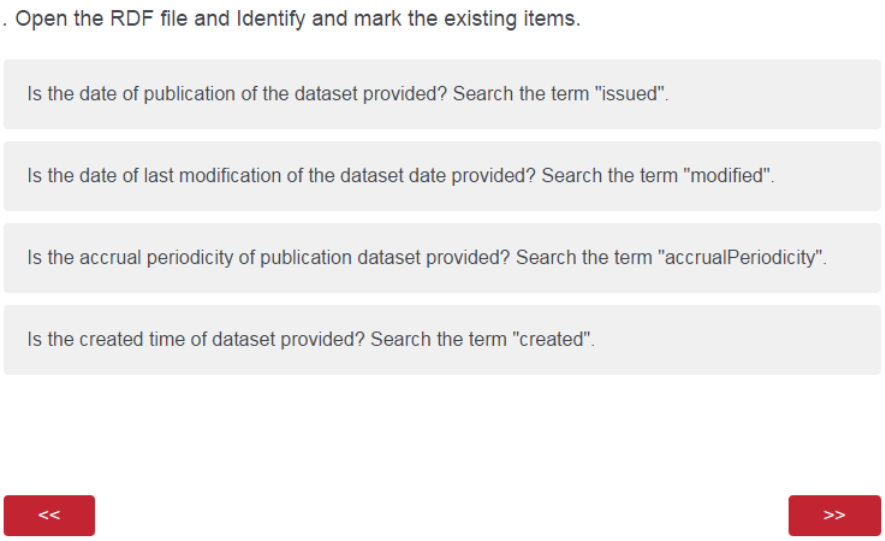

Figura 48 – Tela de avaliação da dimensão Atualidade relacionada ao RDF
Fonte: Elaboração própria

Posteriormente, será apresentado a tela de avaliação da dimensão Compreensibilidade, como mostrado na Figura 49. Finalmente, a ultima tela que será mostrada é a de avaliação da dimensão Versatilidade (Figura 50).

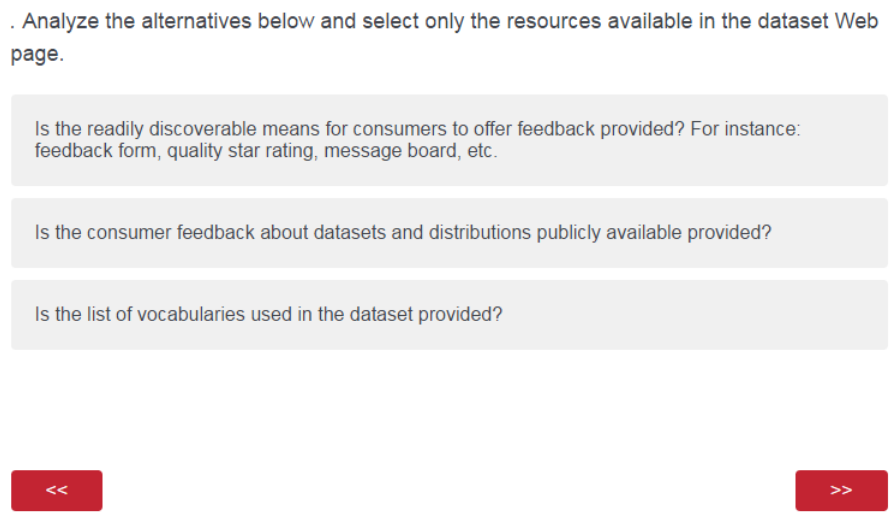

Figura 49 – Tela de avaliação da dimensão Compreensibilidadeq
Fonte: Elaboração própria

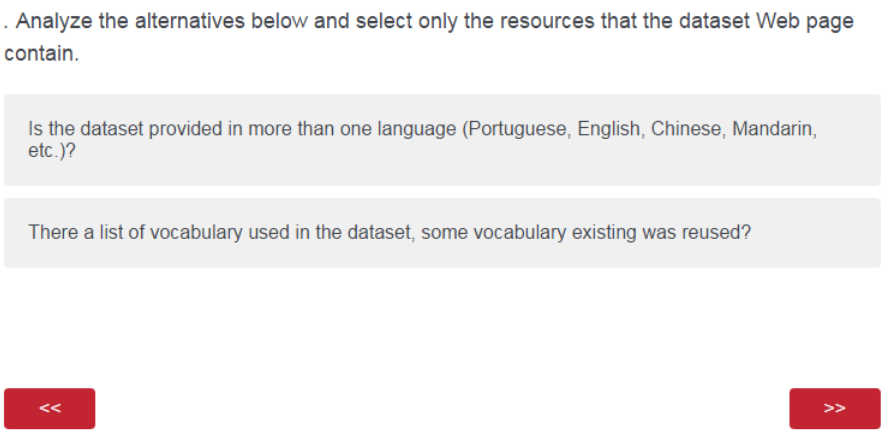

Figura 50 – Tela de avaliação da Dimensão Versatilidade
Fonte: Elaboração própria

B Apêndice

B.1 Modelos de regressão

Um modelo de regressão clássico é definido como $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i$, em que o índice i representa cada um dos componentes da amostra que tipicamente apresenta tamanho n, então $i = 1, \ldots, n$. Na prática, o modelo anterior se reduz a $\mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n$, em que $\mu_i = \mathrm{E}(y_i)$, ou seja, o valor esperado, a média populacional da resposta, por suposição $\mathrm{E}(\epsilon_i) = 0$. Na modelagem de regressão, os valores conhecidos são os valores observados da amostra, isto é, os valores da resposta $y_1, \ldots y_n$ e os valores das variáveis explicativas: $x_{i2}, x_{i3}, \ldots, x_{ik}$, $i = 1, \ldots, n$. Com base nos valores de y e nos valores de $x_2, x_3, \ldots, x_k$ são estimados $\beta_1, \beta_2, \ldots, \beta_k$ tipicamente utilizando o método de estimação por máxima verossimilhança (Lehmann e Casella, 1998) e uma vez verificados seus valores os mesmos passam a ser denotados por $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$, e finalmente tem-se o modelo de regressão o qual pode ser utilizado

$$\widehat{\mu}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \ldots + \widehat{\beta}_k x_{ik}, \quad i = 1, \ldots, n.$$

Agora é possível obter estimativas para μ_i ou para y_i, uma vez que tipicamente $\widehat{y}_i = \mu_i$. O modelo clássico se baseia na distribuição normal. No entanto, na prática essa distribuição não é adequada para diversos tipos de variáveis aleatórias. Se μ é a média da variável resposta que pode assumir tanto valores positivos quanto valores negativos e a curva de densidade de y é próxima da forma de sino, então se justifica pensar na distribuição normal. Caso contrário existem outras distribuições mais apropriadas para descrever o comportamento da variável resposta que a distribuição normal. Se $y \in (0, 1)$ tais como taxas, proporções e alguns escores uma distribuição que tem se mostrado bastante útil é a distribuição beta.

Este último exemplo é particularmente útil nesta área, pois tratamos muito aqui de escores, notas, que se encontram em um intervalo do tipo $(0, 1)$. Neste ponto, precisamos generalizar o modelo linear clássico com o objetivo de permitir o uso de diversas distribuições além da normal. Isto é feito considerando a expressão abaixo.

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}.$$

Quando $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, vimos que $g(\mu_i) = \mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}$, ou seja, g é o que chamamos de função identidade, é usada a própria μ_i. Isto acontece no modelo normal porque μ pode assumir qualquer valor positivo ou negativo consequentemente $\widehat{\mu}_i$ poder assumir qualquer valor positivo ou negativo. Assim, $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ estão livres para também assumir qualquer valor.

Isto não acontece por exemplo se a variável resposta segue uma distribuição beta. Porque,

$\mu \in (0,1)$ e deve acontecer o mesmo com $\widehat{\mu}$, ou seja, com esta restrição os $\widehat{\beta}$'s não estão livres, pois temos que garantir que $\widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \ldots + \widehat{\beta}_k x_{ik} = \widehat{\mu}_i$ só assuma valores em $(0,1)$. O processo de estimação do β's teria que para garantir que $\widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \ldots + \widehat{\beta}_k x_{ik} \in (0,1)$, o que pode ser bastante complicado.

Então, a alternativa é aplicar uma função g em μ_i de forma que $g(\mu_i)$ transforme valores que estão no $(0,1)$ em qualquer valor positivo e negativo, garantindo que os $\widehat{\beta}$'s estejam liberados. Neste caso, em que $\mu_i \in (0,1)$ é uma função de ligação que conduz essa média a todos os reais (todos valores positivos ou negativos) é a função de ligação logito, dada por

$$\mu_i \in (0,1) \leftrightarrow \log\left\{\frac{\mu_i}{(1-\mu_i)}\right\}$$

A densidade beta pode assumir formas diferentes dependendo da combinação de valores de parâmetros. Ferrari e Cribari–Neto (2004) o modelo de regressão beta em que

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

Em que $\mu_i = \mathrm{E}(y_i)$ e cada y_i, $i = 1, \ldots, n$, segue a distribuição beta com densidade:

$$f(y; \mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)} y^{\mu_i\phi_i - 1}(1-y)^{(1-\mu_i)\phi_i - 1}, \quad 0 < y < 1, \qquad (1)$$

onde $0 < \mu_i < 1$ and $\phi_i > 0$ e $\mathrm{var}(y_i) = (\mu_i(1-\mu_i))/(1+\phi_i)$.

Pesquisadores podem usar o pacote `betareg` o qual está disponível no *software* estatístico `R`.

Antes de iniciar a análise de regressão, é necessário uma análise descritiva preliminar dos dados.

B.1.1 Modelos de regressão gerados

Com o objetivo de estudar melhor as variáveis relacionadas às Dimensões de qualidade, foram criados modelos de regressão. A grande vantagem do uso dos modelos de regressão, neste caso, é que além de avaliar se existe diferença significativa em relação a uma variável resposta entre dois grupos é adicionalmente possível verificar o grau desta diferença, ou seja, quantificar essa diferença, caso ela exista.

Neste sentido foram considerados 21 modelos de regressão. Entre estes, 18 considerando a distribuição *beta*, uma vez que os valores das respostas se encontram no intervalo $(0,1)$, são as variáveis respostas: *Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency* e *Score* Global. Os modelos *beta* são definidos pela expressão: $\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2$ *IndicadoraEntreGrupos*$_i$. Ressaltando que μ é a média

populacional da variável resposta.

No que diz respeito à variável resposta *Time*, foram considerados modelos de regressão em que a distribuição da variável resposta é *gama*. Essa decisão se baseou no fato que, para o grupo Manual as notas se distribuem entre $(0, 1)$, que são valores positivos, superiores a "1", o que caracteriza como possibilidade a distribuição *gama*, em que $\mu \in (0, \infty)$. Ressaltando que já foi verificado que a variável *Time* não apresenta distribuição descrita pela distribuição normal.

No caso da distribuição *gama*, o modelo de regressão faz parte dos modelos lineares generalizados (McCullagh e Nelder, 1989) e é definido pela expressão: A covariada *IndicadoraEntreGrupos* de fato são três covariadas, a saber:

- *Indicadora*$_{Base \times Semiaut.}$, a qual assume valor igual a um ("1"), se os valores da resposta dizem respeito ao grupo base e zero ("0"), se dizem respeito ao grupo semiautomático;

- *Indicadora*$_{Base \times Manual}$, a qual assume valor igual a um ("1"), se os valores da resposta dizem respeito ao grupo base e zero ("0"), se dizem respeito ao grupo manual; e a

- *Indicadora*$_{Manual \times Semiaut.}$, a qual assume valor igual a um ("1"), se os valores da resposta diz respeito ao grupo semiautomático e zero ("0"), se diz respeito ao manual, os valores iguais a "0" foram atribuídos propositalmente ao grupo manual devido aos indícios apresentados nos *boxplots* que as notas referentes a este grupo são inferiores às notas dos outros dois grupos.

Na Tabela 30 estão apresentados os 21 modelos, as variáveis respostas, as covariadas, seus respectivos coeficientes, estimativas destes coeficientes e significância destes coeficientes (p–Valores).

Tabela 30 – Modelos de regressão Beta: *Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational consistency* e *Score* Global. Modelo de regressão gama: *Time*.

Covariada	Coeficientes	Grupos	Nota Base e Semi Automático		Nota Base e Manual		Manual e Semi Automático	
		Modelos	$\beta_1 + \beta_2 Indicadora$		$\beta_1 + \beta_2 Indicadora$		$\beta_1 + \beta_2 Indicadora$	
		Resposta	Estimativa	p–Valor	Estimativa	p–Valor	Estimativa	p–Valor
Intercepto	β_1		0.5102	0.216	-0.8417	0.0642	-0.9317	0.0325
Indic	β_2	Availability	0.5206	0.466	1.6364	**0.0498**	1.3615	**0.0309**
Intercepto	β_1		-0.3826	0.1712	-1.4338	0.00148	-1.9328	5.78e-07
Indic	β_2	Provenance	-1.1340	**0.0399**	0.3894	0.567	1.5716	**0.00105**
Intercepto	β_1		-1.7020	0.00039	-1.5241	0.00141	-1.55850	0.000517
Indic	β_2	Licensing	-0.1073	0.87000	-0.1452	0.82855	-0.04391	0.933348
Intercepto	β_1		-1.5380	1.7e-07	-1.8944	2.68e-05	-2.3219	1.69e-09
Indic	β_2	Currency	-0.4984	0.341	0.3353	0.594	0.8713	**0.0517**
Intercepto	β_1		-1.0164	0.0288	-0.4689	0.301	-0.4859	0.278
Indic	β_2	Understandibility	-0.2032	0.7783	-0.4994	0.516	-0.3359	0.582
Intercepto	β_1		-0.9608	0.0495	-7.600	<2e-16	-1.5017	0.00187
Indic	β_2	Versatility	-0.3942	0.5926	-2.955e-16	0.589	0.4175	0.46494
Intercepto	β_1		–	–	-0.3602	0.437	-0.4213	0.3574
Indic	β_2	Representational	–	–	1.2765	**0.129**	1.4653	**0.0312**
Intercepto	β_1		-0.3550	0.0643	-1.5152	9.17e-05	-1.5965	9.76e-06
Indic	β_2	Score Global	-0.1269	0.7152	1.1327	**0.0606**	1.3002	**0.0049**
Intercepto	β_1		181.77	0.0003	0.4414	0.265	2.265	0.174
Indic	β_2	Time	-32.80	0.572	148.5241	0.451	-2.260	0.330

Fonte: Elaboração própria

Com base na Tabela 30 verifica-se que, de fato as variáveis respostas que apresentam alguma diferença entre os grupos são *Availibility, Representational* e *scores* global. No caso de *Representational* os escores dos grupos base e semiautomático são identicamente iguais a “1", não variam, o que caracteriza que neste caso a variável não é uma variável aleatória, é uma constante. Assim, não é possível realizar uma analise de regressão. Ainda quanto *Representational* tem-se que o p–Valor referente ao coeficiente β_2 da covariada *Indicadora*$_{Base \times Manual}$ é superior a 10%, o que conduziria a uma conclusão de não significância de β_2, ou melhor, a hipótese $H_0 : \beta_2 = 0$ não seria rejeitada e, portanto, a covariada seria excluída do modelo. No entanto, esses dados apresentam uma inconveniência, a quantidade de observações do grupo base é muito pequeno quando comparado à quantidade de observações de manual e semiautomático, o que pode comprometer testes de hipóteses. Assim, pode ser aberto um precedente e considerar que β_2 é significativamente diferente de zero a um nível próximo de 10%, dado que o p–Valor referente ao teste de hipótese $H_0 : \beta_2 = 0$ é igual 0.13.

Com base na Tabela 30 é possível mensurar a diferença entre os grupos Nota base *vs* Manual e Manual e Semiautomático quanto a variável *Availiability*, com base nos coeficientes da covariada indicadora. Inicialmente, Nota base *vs* Manual, $\widehat{\beta_2} = 1.6364$, implicando

que $1 - \exp(1.6364) = 4.13$ e consequentemente, $4.13 \times 100 = 413\%$. Ou seja, a média de nota do grupo base é 413% maior que a média de nota do grupo manual (*Availiability*). Quanto a Manual e Semiautomático, $\widehat{\beta}_2 = 1.3615$, $1 - \exp(1.6364) = 2.90$. Ou seja, a média de nota do grupo semiautomático é 290% maior que a média de nota do grupo manual (*Availiability*).

Quanto *Representational consistency* Nota base *vs* Manual, $\widehat{\beta}_2 = 1.2765$, $1 - \exp(1.2765) = 2.58$ e então, a média de nota do grupo base é 258% maior que a média de nota do grupo manual. Comparando Semiautomático *vs* Manual tem-se que a média de notas de semiautomático é 333% superior a média de notas do grupo manual dado que $\widehat{\beta}_2 = 1.4653$ e $1 - \exp(1.4653) = 3.33$.

Finalmente, o escore global do grupo manual é 210% inferior ao do grupo base e 277% inferior ao do grupo semiautomática. Mais uma vez deve ser ressaltado que as diferenças entre as notas do grupo manual e base poderiam ser mais evidentes se o grupo base apresentasse mais observações.

Para finalizar a análise estatística foram investigados modelos de regressão beta para explicar a média da variável resposta *score* global. Na Tabela 31 são apresentados quatro modelos de regressão beta competidores utilizados para inicializar o processo de construção do melhor modelo. Foram inicialmente investigadas todas as variáveis envolvidas no problema para representar as covariadas ou variáveis explicativas do modelo. O modelo 4 foi o que apresentou todos os coeficientes significativos e, portanto, foi o modelo escolhido para a realização de uma análise de diagnóstico e uma seleção mais apropriada do melhor modelo. O Modelo 4 é definido pela expressão:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2 \text{Avail}_i + \beta_3 \text{Lic}_i + \beta_4 \text{Curr}_i + \beta_5 \text{Repre}_i + \beta_6 \text{IndicSemiAuto}_i, i = 1, \ldots, 22.$$

e sua versão estimada aparece na Tabela Tabela 31, em que todos os coeficientes das covariadas são considerados significativos. O próximo passo é realizar uma análise de diagnóstico.

Tabela 31 – Modelos de regressão Beta. Variável resposta: *Score* Global. Covariadas candidatas:*Time, Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational* e Indicadora do grupo: Nota Base e semiautomático

Modelo Beta 1				Modelo Beta 2			
Covariadas	Coeficientes	Estimativa dos Coeficientes	p–Valor	Covariadas	Coeficientes	Est.	p–Valor
Intercepto	β_1	-3.83	<2e-16***	Intercepto	β_1	-3.82	< 2e-16***
time	β_2	0.02	0.26969	time	β_2	0.03	0.12621
Avail	β_3	1.05	0.00508**	Avail	β_3	0.96	0.00016***
Prov	β_4	0.26	0.73979	Prov	β_4	0.21	0.78049
Lic	β_5	2.06	0.07435.	Lic	β_5	1.66	1.97e-05***
Curr	β_6	4.01	0.00385**	Curr	β_6	3.95	0.00385**
Under	β_7	-0.27	0.71931	Vers	β_7	-0.02	0.93101
Vers	β_8	-0.06	0.85527	Repre	β_8	1.92	9.83e-08***
Repre	β_9	1.88	1.38e-07***	IndicNotaBase	β_9	-0.44	0.05670.
IndicNotaBase	β_{10}	-0.49	0.07659 .				
Modelo Beta 3				Modelo Beta 4			
Covariadas	Coeficientes	Estimativa dos Coeficientes	p–Valor	Covariadas	Coeficientes	Est.	p–Valor
Intercepto	β_1	-3.81	< 2e-16***	Intercepto	β_1	-3.6128	< 2e-16 ***
time	β_2	0.03	0.1163	Avail	β_2	0.9382	0.000141 ***
Avail	β_3	0.96	2.63e-05 ***	Lic	β_3	1.8821	4.02e-06 ***
Lic	β_4	1.65	2.32e-05 ***	Curr	β_4	4.0394	1.64e-07 ***
Curr	β_5	4.18	2.55e-08 ***	Repre	β_5	1.4229	2.99e-07 ***
Repre	β_6	1.95	3.16e-10 ***	IndicSemiAuto	β_6	0.3327	0.046037 *
IndicNotaBase	β_7	-0.45	0.0148 *				
· siginificante a 10%. * siginificante a 5%. ** siginificante a 1%.							
*** siginificante a mais 1%							

Fonte: Elaboração própria

B.1.1.1 Análise de Diagnóstico

Verificar se um determinado modelo é uma representação adequada dos dados é um passo importante da análise estatística. A construção de um modelo de regressão envolve a definição da distribuição da variável de resposta, a escolha da função de ligação, e a escolha das covariáveis. Vários fatores podem levar um modelo ajustado pobre. Por exemplo, a função de ligação inadequado, omissão de covariadas importante, a escolha errada da distribuição da variável resposta, pontos influentes, **especificação incorreta da variância** entre outros fatores.

A avaliação da qualidade de um modelo de regressão pode ser alcançado por meio de análise de diagnóstico. Normalmente, esses diagnósticos são construídos em torno da análise gráfica dos resíduos e critérios de seleção como o R^2. A maior parte dos resíduos é baseada nas diferenças entre as respostas observadas (y) e a média estimada ($\widehat{\mu}$). Por exemplo $r_i = y_i - \widehat{\mu}_i$, em que $\widehat{\mu}_i$ pode ser visto como $\widehat{y}_i$. Ou seja, o resíduo é uma medida de discrepância entre os dados reais e os dados estimados com base no modelo de regressão. O resíduo mais

utilizado nos modelos de regressão beta é o resíduo ponderado proposto por Espinheira et al (2008). Mais recente Espinheira et al (2017) propuserem o resíduo combinado para a classe de modelos de regressão beta que ao ser utilizado em gráficos de resíduos mostrou-se bastante eficaz em identificar problemas no modelo de regressão.

Os gráficos de resíduos versus índices das observações são os mais básicos. Se um modelo está especificado corretamente, então estes gráficos não devem apresentar nenhuma tendência, os resíduos devem estar aleatoriamente distribuídos em torno do zero ou igualmente distribuídos acima e abaixo do zero. A presença de quaisquer características sistemáticas tipicamente implica uma falha de um ou mais pressupostos do modelo. Outro gráfico de resíduos importante é o gráfico de probabilidade normal com envelopes simulados, que pode ser usada mesmo quando as distribuições empíricas dos resíduos não são normais. Se o modelo está adequado aos dados, esperamos que a maioria dos resíduos estejam aleatoriamente distribuídos dentro das bandas do envelope. Na Figura 51 estão apresentados os gráficos de resíduos referentes ao modelo 4.

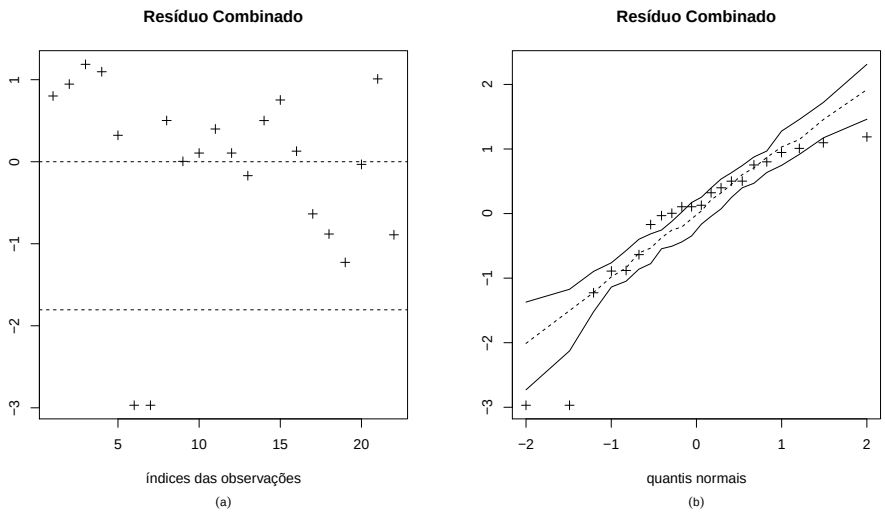

Figura 51 – Gráfico de resíduos. Modelo 4
Fonte: Elaboração própria

Os gráficos da Figura 51 mostram que os resíduos não estão aleatoriamente distribuídos em torno do zero, os resíduos deveriam apresentar a mesma forma de distribuição acima e abaixo da reta igual a zero. Isto não ocorre, claramente a parte acima do zero apresenta um padrão bem diferente da parte abaixo do zero. Isto é um indício que a variabilidade, o que também é chamada de dispersão do modelo não está adequada e, portanto, a variância suposta para os dados não está adequada. Quanto ao gráfico em (b), conhecido como gráfico normal de probabilidades com envelopes simulados, o modelo é considerado adequado se todos os resíduos se distribuem de forma aleatória dentro das bandas, o que não ocorre. Diante de tais evidências será considerado um novo modelo de regressão beta na tentativa de tornar adequada a variância estimada dos dados.

Esse novo modelo considera duas expressões matemáticas, uma para a média e outra para a dispersão (Variabilidade), o que vai alterar a variância estimada dos dados. A ideia é: uma vez que é suposto um modelo de regressão para a média populacional, também pode ser suposto um modelo de regressão para a variância populacional, ou dispersão populacional. No modelo de regressão beta a dispersão está associada ao parâmetro ϕ presente na função de densidade beta apresentada na Equação 1. No modelo da dispersão foi considerada inicialmente apenas a variável *IndicadoraManual* que representa as observações referentes ao grupo Manual. A escolha da indicadora deste grupo deve-se ao fato de algumas variáveis apresentarem diferença entre o grupo Manual e os outros dois grupos. Modelo 5:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2 \text{Avail}_i + \beta_3 \text{Lic}_i + \beta_4 \text{Curr}_i + \beta_5 \text{Repre}_i + \beta_6 \text{IndicSemiAuto}_i, i = 1, \ldots, 22.$$

e

$$\log(\phi_i) = \gamma_1 + \gamma_2 \text{IndicadoraManual}_i, i = 1, \ldots, 22.$$

Aqui, γ_1 e γ_2 são os coeficientes do modelo da dispersão e a covariada *IndicadoraManual* é considerada importante para explicar a dispersão dos dados caso a hipótese $H_0 : \gamma_2 = 0$ seja rejeitada a um nível de significancia desejado. Por exemplo, se considramos um nível de significância até 10%, H_0 é rejeitada se p–Valor< 0.10 e, neste caso, é considerado que $\gamma_2 \neq 0$ implicando na permanência de *IndicadoraManual* no modelo da dispersão. O mesmo princípio se aplica aos coeficientes do modelo da média.

Na Tabela 32 estão apresentadas as estimativas dos coeficientes tanto para o modelo da média, quanto para o modelo da dispersão referentes aos três novos modelos de regressão betas, os modelos 5, 6 e 7 e os respectivos p–Valores dos testes de hipóteses associados aos coeficientes. Uma vez que todos os coeficientes se mostraram significativos o próximo passo é a análise gráfica de resíduos. Na Figura 52 estão os gráficos de resíduos referentes ao modelo 5.

Tabela 32 – Modelos de regressão Beta. Variável resposta: *Score* Global. Covariadas candidatas: *Time, Availability, Provenance, Licensing, Currency, Understandibility, Versatility, Representational* e Indicadora do grupo: Semiautomático.

Modelo Beta 5				Modelo Beta 6				Modelo Beta 7			
Modelo da Média				Modelo da Média				Modelo da Média			
Covariada	Coef.	Est.Coef.	p–Valor	Cov.	Coef.	Est.	p–Valor	Cov.	Coef.	Est.	p–Valor
Intercepto	β_1	-3.17	0.0000	Intercepto	β_1	-2.95	0.000	Intercepto	β_1	-2.22	0.000
Avail	β_2	0.99	0.0000	Avail	β_2	0.79	0.000	Avail	β_2	0.91	0.000
Lic	β_3	1.63	0.0000	Lic	β_3	1.62	0.000	Lic	β_3	1.52	0.000
Curr	β_4	2.61	0.0000	Curr	β_4	1.71	0.000	Curr	β_4	0.87	0.008
Repre	β_5	1.18	0.0012	Repre	β_5	1.27	0.002	Repre	β_5	0.69	0.000
IndicSemiAuto	β_6	0.35	0.0000	IndicSemiA.	β_6	0.19	0.000	IndicSemiA.	β_6	0.11	0.103
				Vers	β_7	0.40	0.000	Vers	β_7	0.48	0.000
Modelo da Dispersão				Modelo da Dispersão				Modelo da Dispersão			
Covariada	Coef.	Est.Coef.	p–Valor	Cov.	Coef.	Est.	p–Valor	Cov.	Coef.	Est.	p–Valor
Intercepto	γ_1	5.74	0.0000	Intercepto	γ_1	6.78	0.000	Intercepto	γ_1	3.94	0.000
IndicManual	γ_2	-3.12	0.0000	IndicManual	γ_2	-4.45	0.000	IndicManual	γ_2	-2.75	0.000
								Curr	γ_3	6.51	0.007
								Lic	γ_3	7.08	0.000
· siginificante a 10%. * siginificante a 5%. ** siginificante a 1%.											
*** siginificante a mais 1%											

Fonte: Elaboração própria

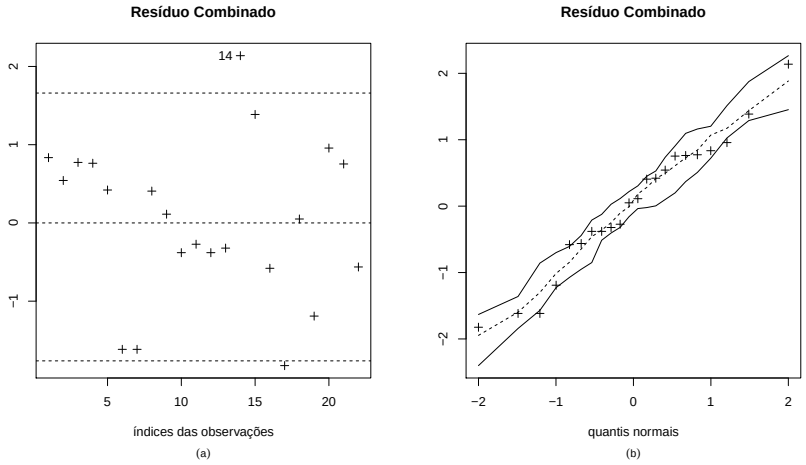

Figura 52 – Gráfico de resíduos. Modelo 5
Fonte: Elaboração própria

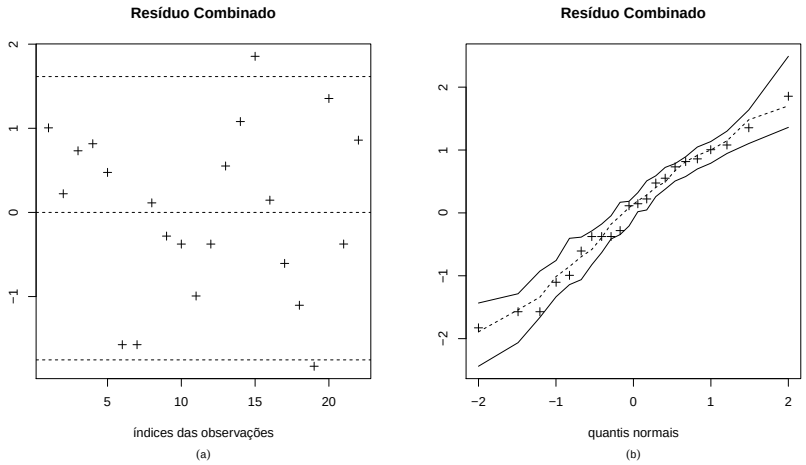

Figura 53 – Gráfico de resíduos. Modelo 6
Fonte: Elaboração própria

É possível notar uma melhoria de qualidade do ajuste ao se adotar o modelo 5, uma vez que os resíduos referentes a este modelo estão aleatoriamente distribuídos ao redor do zero, sem padrões diferenciados quando comparadas a forma em que os resíduos se distribuem acima e abaixo da reta zero (Figura 52a). No entanto, também na Figura 52a é detectado um ponto aberrante, que é um ponto em que a diferença entre a resposta verdadeira e a resposta estimada é estatisticamente maior que a das demais observações, o modelo não conseguiu estimar bem esse ponto. Então é necessário buscar um melhor modelo. Para isso é preciso entender a natureza do ponto aberrante, no caso a observação 14. Quando investigada essa observação nota-se que o valor da variável *Versatility* é igual a “1"e que em geral todos os outros valores são iguais a “0", por isso o modelo anterior apresentou dificuldades em estimar esse ponto.

Assim volta-se a investigar a entrada da covariada *Versatility* no modelo da média, sendo que agora o modelo beta também considera a modelagem da dispersão, esse é o Modelo 6. Todos os coeficientes são considerados significativamente diferentes de zero implicando que a covariada *Versatility* pode entrar no modelo. Mas, ainda falta a análise dos resíduos do Modelo 6; Figura 53. A Figura 53a evidencia que para o modelo 6 o resíduo do caso 14 diminuiu, mas ainda pode melhorar. Neste ponto pode-se investigar a inclusão de covariadas no modelo da dispersão e isto é feito no Modelo 7. Considerando a Figura 54 é possível notar como a inclusão das covariadas *Currency* e *Licencing* no modelo da dispersão conduziu a um modelo de regressão com qualidade de ajuste superior aos demais modelos investigados, dado que a maioria dos pontos estão aleatoriamente distribuídos em torno do zero, dentro dos envelopes e ao modelo 7 consegue estimar mais adequadamente o caso 14, o qual não é mais considerado aberrante.

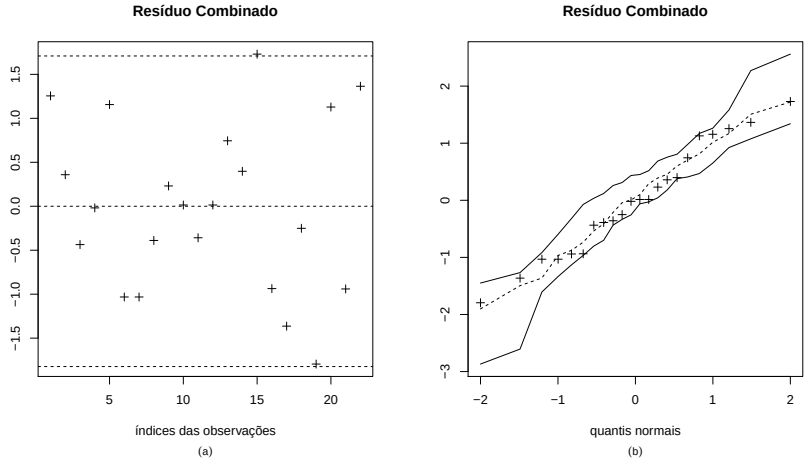

Figura 54 – Gráfico de resíduos. Modelo 7
Fonte: Elaboração própria

O modelo final apresenta as seguintes expressões matemáticas:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2 \text{Avail}_i + \beta_3 \text{Lic}_i + \beta_4 \text{Curr}_i + \beta_5 \text{Repre}_i + \beta_6 \text{IndSemiA.}_i + \beta_7 \text{Vers}_i;$$

$$\log(\phi_i) = \gamma_1 + \gamma_2 \text{IndicadoraManual}_i + \gamma_3 Curr_i + \gamma_4 Lic_i, \quad i = 1, \ldots, 22.$$

O modelo que será utilizado para fazer estimativas para a resposta é o modelo da média apenas. O modelo da dispersão é utilizado para adequar as variância dos dados, as variâncias dos estimadores dos coeficientes que influenciam diretamente nos p–Valores relacionados a significância de todos os coeficientes. Após estimado, o modelo passa a ser:

$$\log\left(\frac{\widehat{\mu}_i}{1-\widehat{\mu}_i}\right) = \widehat{\beta}_1 + \widehat{\beta}_2 \text{Avail}_i + \widehat{\beta}_3 \text{Lic}_i + \widehat{\beta}_4 \text{Curr}_i + \widehat{\beta}_5 \text{Repre}_i + \widehat{\beta}_6 \text{IndSemiA.}_i + \widehat{\beta}_7 \text{Vers}_i.$$

E com base na Tabela 31 chegamos a:

$$\log\left(\frac{\widehat{\mu}_i}{1-\widehat{\mu}_i}\right) =$$

$$-2.21 + 0.91\text{Avail}_i + 1.52\text{Lic}_i + 0.87\text{Curr}_i + 0.69\text{Repre}_i + 0.11\text{IndSemiA.}_i + 0.48\text{Vers}_i.$$

Consequentemente

$$\widehat{\mu}_i = \frac{\exp(-2.21 + 0.91\text{Avail}_i + 1.52\text{Lic}_i + 0.87\text{Curr}_i + 0.69\text{Repre}_i + 0.11\text{IndSemiA.}_i + 0.48\text{Vers}_i)}{1 + (\exp(-2.21 + 0.91\text{Avail}_i + 1.52\text{Lic}_i + 0.87\text{Curr}_i + 0.69\text{Repre}_i + 0.11\text{IndSemiA.}_i + 0.48\text{Vers}_i)}. \quad (2)$$

Os valores estimados dos coeficientes tem uma interpretação muito importante. Se o sinal da estimativa é positivo significa que ao aumentar o valor da covariada aumenta-se o valor da resposta e maior será esse aumento quanto maior for o valor da estimativa. Assim por exemplo uma das covariadas mais importantes para a qualificação global é *Licencing*. Um exemplo de valor estimado é: considerando a observação 14 os valores de cada uma das covariadas do modelo são Avail$_{14} = 0.995 \approx 1$, Lic$_{14}$ = 0.333 Curr$_{14}$ =0.2, Repre$_{14} = 0.995 \approx 1$, IndSemiA.$_{14}$ = 1, Vers$_{14} = 0.995 \approx 1$. Substituindo esses valores na expressão acima tem-se que o score global estimado $\widehat{\mu}_{14} = 0.65700$ e o valor verdadeiro é $y_{14} = 0.66122$, essa proximidade entre o valor verdadeiro e o valor estimado deve-se a qualidade do modelo escolhido. Para estimar ou prever novas observações basta utilizar a fórmula em 1.2 substituindo os valores das covariadas. A fórmula pode ser utilizada em qualquer conjunto de dados para obter o escore global.

C Apêndice

Para execução do experimento, foram enviados os *links* das abordagens manual (questionário *online*) e semiautomática (ferramenta), juntamente com os *datasets* para os e-mails de especialistas, e também para Grupos de dados conectados do Google.

Desta forma, para os grupos do Google, foi enviado a seguinte mensagem com seu respectivo formulário de consentimento:

> Dear researcher / professional,
>
> You are being invited to participate in a survey that evaluates the quality of Linked *datasets* according to dimensions: Availability, Provenance, Licensing, Currency, Understandability, Versatility and Representational-Conciseness. This study is being conducted by Danila Feitosa and Prof. Dr. Ig Ibert Bittencourt, both of them researchers of the Computing Institute of the Federal University of Alagoas (UFAL, Brazil). This research aims to validate the approach called LinkedDataSetEvaluation, composed of dimensions and metrics that increase the quality of these Linked Datasets. We define some questions related to the quality of linked data and we would like to have your feedback for the success of this research. This survey is available at consent form in attachment.

LinkedDataEvaluation

Manual Evaluation of Linked Datasets Quality

CONSENT FORM

Federal University of Alagoas (UFAL)

Computing Institute of Federal University of Alagoas

You are being invited to participate in a survey that evaluates the quality of Linked datasets according to these dimensions: Availability, Provenance, Licensing, Currency, Understandability, Versatility and Representational-Conciseness.

Title of study: Verificating and validating semi-automatic quality of linked datasets

Ethics Application Number

Researchers:

Prof. Dr. Ig Ibert Bittencourt, Computing Institute (57072-970), ig.ibert@ic.ufal.br (Supervisor)
Danila Feitosa, Computing Institute (57072-970), dfcco@ic.ufal.br (Master student)

Purpose and Procedure: this research aims to validate the approach called LinkedDataSetEvaluation, composed of dimensions and metrics that increase the quality of these Linked Datasets. We defined some questions related to the quality of linked data and we would like to have your feedback for the success of this research.

Potential Benefits: findings from the study could lead to a better understanding of the quality of linked datasets related to dimensions, metrics, criteria end best practices for publishing linked data.

Potential Risks: there are no known risks in this study.

Confidentiality: this survey is hosted by Google Inc. ("Google") that protects the users access or unauthorized alteration, revelation or unauthorized destruction of information that they hold. The Google Privacy Policy is available in https://www.google.com/intl/pt-BR/policies/privacy/

Dissemination of Results: aggregated results in this survey will appear in Masters dissertation.

Right to Withdraw: participation in this survey is voluntary, and you can decide not to participate any time, or choose not to answer any questions you don't feel comfortable to Survey responses will remain anonymous. Since the survey is anonymous, once it is submitted it cannot be removed.

This survey is available at http://67.205.139.14:8080/datasetevaluation/index.html To answer the questions open this file https://dl.dropboxusercontent.com/u/57888679/Datasets/DatasetLandRegistryPriceD2.txt

LinkedDataEvaluation

Manual Evaluation of Linked Datasets Quality

CONSENT FORM

Federal University of Alagoas (UFAL)

Computing Institute of Federal University of Alagoas

You are being invited to participate in a survey that evaluates the quality of Linked datasets according to these dimensions: Availability, Provenance, Licensing, Currency, Understandability, Versatility and Representational-Conciseness.

Title of study: Verificating and validating semi-automatic quality of linked datasets

Ethics Application Number

Researchers:

Prof. Dr. Ig Ibert Bittencourt, Computing Institute (57072-970), ig.ibert@ic.ufal.br (Supervisor)
Danila Feitosa, Computing Institute (57072-970), dfcco@ic.ufal.br (Master student)

Purpose and Procedure: this research aims to validate the approach called LinkedDataSetEvaluation, composed of dimensions and metrics that increase the quality of these Linked Datasets. We defined some questions related to the quality of linked data and we would like to have your feedback for the success of this research.

Potential Benefits: findings from the study could lead to a better understanding of the quality of linked datasets related to dimensions, metrics, criteria end best practices for publishing linked data.

Potential Risks: there are no known risks in this study.

Confidentiality: this survey is hosted by Google Inc. ("Google") that protects the users access or unauthorized alteration, revelation or unauthorized destruction of information that they hold. The Google Privacy Policy is available in https://www.google.com/intl/pt-BR/policies/privacy/

Dissemination of Results: aggregated results in this survey will appear in Masters dissertation.

Right to Withdraw: participation in this survey is voluntary, and you can decide not to participate any time, or choose not to answer any questions you don't feel comfortable to Survey responses will remain anonymous. Since the survey is anonymous, once it is submitted it cannot be removed.

This survey is available at https://qtrial2016q3az1.qualtrics.com/SE/?SID=SV_1Twt4Xi0XRtvp4x

Para os e-mail dos avaliadores foi enviado dois tipos de mensagem, onde uma era referente a avaliação utilizando a abordagem manual e a outra era referente a avaliação utilizando a abordagem semiautomática, respectivamente:

Mensagem referente a abordagem manual (questionário)

> Gostaria de convidá-lo para participar de uma pesquisa, que tem como finalidade utilizar um questionário para avaliar a qualidade de conjuntos de dados (datasets) conectados. A pesquisa está sendo conduzida por Danila Feitosa e pelo Prof. Dr. Ig Ibert Bittencourt, ambos pesquisadores do Instituto de Computação da Universidade Federal de Alagoas (IC - UFAL). O questionário é composto por dimensões e métricas utilizadas pela literatura para avaliar a qualidade de conjuntos de dados conectados.
>
> A pesquisa visa obter como resultado, uma ferramenta de avaliação de conjuntos de dados conectados, contribuindo para a melhoria da qualidade dos dados e informações a serem utilizados e reutilizados no âmbito da *Web* Semântica. Esperamos contar com sua contribuição para o aprimoramento deste trabalho.
>
> Caso aceite este convite, segue o *link* do questionário: Link: `https://qtrial2016q3az1.qualtrics.com/SE/?SID=SV_1Twt4Xi0XRtvp4x` Em caso de dúvidas, estou disponível para fazer uma vídeo conferência de 30 minutos para que possa auxiliar em tempo real a execução do questionário.
>
> E-mail: danila.fco@gmail.com
> Skype: `danila_feitosa`
> Tel: 082 9 9621-0322

Mensagem referente a abordagem semiautomática (ferramenta)

> Gostaria de convidá-lo para participar de uma pesquisa, que tem como finalidade utilizar uma ferramenta semiautomática (LinkedDataSetEvaluation) para avaliar a qualidade de conjuntos de dados (datasets) conectados. A pesquisa está sendo conduzida por Danila Feitosa e pelo Prof. Dr. Ig Ibert Bittencourt, ambos pesquisadores do Instituto de Computação da Universidade Federal de Alagoas (IC - UFAL). A ferramenta semiautomática é composta por dimensões e métricas utilizadas pela literatura para avaliar a qualidade de conjuntos de dados conectados.
>
> A pesquisa visa obter como resultado, uma ferramenta de avaliação de conjuntos de dados conectados, contribuindo para a melhoria da qualidade dos dados e informações a serem utilizados e reutilizados no âmbito da *Web*

Semântica. Esperamos contar com sua contribuição para o aprimoramento deste trabalho.

Caso aceite este convite, segue o *link* da ferramenta: Link: `http://67.205.139.14:8080/datasetevaluation/index.html` Você irá avaliar o conjunto de dados: `https://data.gov.uk/dataset/land-registry-monthly-price-paid-data` O conjunto de dados acima e o arquivo contido neste *link* (`https://dl.dropboxusercontent.com/u/57888679/Datasets/DatasetLandRegistryPriceD2.txt`) serão utilizados para responder as questões.

Em caso de dúvidas, estou disponível para fazer uma vídeo conferência de 30 minutos para que possa auxiliá-lo em tempo real na execução da ferramenta semiautomática.

E-mail: danila.fco@gmail.com
Skype: `danila_feitosa`
Tel: 082 9 9621-0322