Natacha Targino Rodrigues Simões Brasileiro

UMA ABORDAGEM PARA GERAÇÃO DE PERFIL DE CONJUNTO DE DADOS COM METADADOS ENRIQUECIDOS SEMANTICAMENTE

UNIVERSIDADE FEDERAL DE PERNAMBUCO
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

Recife
2018

Natacha Targino Rodrigues Simões Brasileiro

UMA ABORDAGEM PARA GERAÇÃO DE PERFIL DE CONJUNTO DE DADOS COM METADADOS ENRIQUECIDOS SEMANTICAMENTE

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: *Ana Carolina Brandão Salgado.*
Coorientadora: *Damires Yluska de Souza Fernandes.*

Recife
2018

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

B823a Brasileiro, Natacha Targino Rodrigues Simões
Uma abordagem para geração de perfil de conjunto de dados com metadados enriquecidos semanticamente / Natacha Targino Rodrigues Simões Brasileiro. – 2018.
118 f.: il., fig.

Orientadora: Ana Carolina Brandão Salgado.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2018.
Inclui referências e apêndices.

1. Banco de dados. 2. Metadados. 3. Enriquecimento semântico. I. Salgado, Ana Carolina Brandão (orientadora). II. Título.

025.04 CDD (23. ed.) UFPE- MEI 2018-100

Natacha Targino Rodrigues Simões Brasileiro

Uma Abordagem para Geração de Perfil de Conjunto de Dados com Metadados Enriquecidos Semanticamente

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 02/03/2018.

BANCA EXAMINADORA

_____

Profa. Dra. Bernadette Farias Lóscio
Centro de Informática / UFPE

_____

Profa. Dra. Fernanda Araujo Baião Amorim
Departamento de Informática Aplicada / UNIRIO

_____

Profa. Dra. Ana Carolina Brandão Salgado
Centro de Informática /UFPE
(**Orientadora**)

Dedico este trabalho a minha família
e minhas orientadoras.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ser tão maravilhoso e por ter me dado forças, saúde e sabedoria para adquirir conhecimentos e capacidade para superar todos os obstáculos que foram surgindo ao longo desta caminhada.

Agradeço a minha família, em especial aos meus pais, Jandui Brasileiro e Régia Targino, pelo apoio incondicional, que souberam reconhecer a importância do presente estudo na minha formação, mesmo distante estavam perto, me oferecendo carinho, orientação e incentivo, estando comigo em todos os momentos dessa caminhada. Obrigada pelas palavras de ânimo e pelas orações, que me serviram de luz e escudo para seguir nesta caminhada.

Aos meus irmãos, Caio, Shimena, Sayonara e Quênia, sinônimos de amizade e companheirismo, obrigada pelas orações e palavras de apoio nas horas em que mais preciso, pelos telefonemas de descontração, que me fazia sentir em casa e no convívio com todos. Obrigada, pelo cuidado e compreensão nesta minha caminhada.

Agradeço aos meus colegas de curso, principalmente a Marcelo e Gabrielle, pelos conselhos e ajudas ao longo da pesquisa, apontando o caminho nos momentos de dúvida e insegurança, além da companhia durante as tardes no laboratório tornando aqueles momentos descontraídos e bastantes agradáveis.

Meus sinceros agradecimentos aos professores que contribuíram de forma ímpar em minha formação, pelos conhecimentos transmitidos e pela grandeza de pessoas e profissionais que vocês são.

Agradeço de forma especial as minhas queridas orientadoras Ana Carolina Brandão Salgado e Damires Yluska de Souza Fernandes, por toda atenção, paciência e sabedoria em suas palavras e por acreditarem em meu potencial. Obrigada por todos os conselhos, sugestões, ensinamentos e paciência ao longo desses últimos anos, que, de forma brilhante, me capacitou e guiou para conclusão desta jornada.

Também não posso deixar de agradecer mais uma vez a professora amiga Damires Souza, pelos momentos de apoio, ensinamentos e orientação que me foram passados durante o período da graduação no IFPB, me capacitando para a vida pessoal e profissional, os quais me proporcionaram chegar a este momento precioso.

Por fim, agradeço a todos os demais que de uma forma ou outra contribuíram e me apoiaram para a realização deste trabalho.

Que os vossos esforços desafiem as impossibilidades, lembrai-vos de que as grandes coisas do homem foram conquistadas do que parecia impossível.

—CHARLES CHAPLIN

RESUMO

A variedade de conjuntos de dados publicados na Web possibilita um cenário ilimitado de informações. No entanto, identificar conjuntos de dados adequados para uma determinada atividade é um grande desafio. Com o intuito de facilitar essa tarefa, metadados que descrevem os conjuntos de dados podem ser disponibilizados. Adicionalmente, os metadados podem ser enriquecidos semanticamente de modo que suas informações contribuam não somente para a identificação, mas também para a compreensão e o consumo dos conjuntos de dados. É possível disponibilizar os metadados por meio de um Perfil de Conjunto de Dados (PCD), utilizando vocabulários recomendados e o representando em formato de dados estruturado. O PCD é composto de informações descritivas, estruturais e de qualidade sobre um determinado conjunto de dados. Na literatura são encontrados alguns trabalhos que abordam a geração de um perfil composto por metadados enriquecidos semanticamente, mas não foram encontrados trabalhos que considerem a geração de um perfil que inclui aspectos descritivos, estruturais e de qualidade. Neste contexto, este trabalho propõe uma abordagem para criação de um PCD, composto por metadados descritivos, estruturais e de qualidade enriquecidos semanticamente. Prover um PCD pode facilitar ainda mais a compreensão, localização, processamento e reuso dos conjuntos de dados por seus consumidores (usuários e aplicações). Para avaliar a abordagem proposta, foi desenvolvido um protótipo e foram realizados alguns experimentos, que demonstraram que a estratégia proposta produz bons resultados no que diz respeito à geração do Perfil do Conjunto de Dados e, em especial, ao enriquecimento semântico dos metadados.

Palavras-chave: Conjuntos de Dados da Web. Perfil de Conjunto de Dados. Metadados. Enriquecimento Semântico. Publicação e Consumo de Dados.

ABSTRACT

The variety of datasets published on the Web provides an unlimited scenario for information usage. However, identifying appropriate datasets for a given task is still a challenge. In order to facilitate this task, metadata describing datasets can be made available. In addition, metadata can be semantically enriched so that their information can contribute not only to identify datasets but also to better understand and consume them. It is possible to make metadata available through a Dataset Profile (DSP), by using recommended vocabularies and representing it in a structured data format. The DSP is composed of descriptive, structural and quality information about a particular dataset. In the literature, some works regarding the generation of dataset profile, composed by semantically enriched metadata, were found. However, there was no work which considered the generation of a dataset profile including descriptive, structural and quality aspects. In this light, this work proposes an approach for the creation of a DSP, consisting of descriptive, structural and quality semantic enriched metadata. Providing a DSP can further facilitates the understanding, localization, processing, and reuse of datasets by their consumers (users and applications). In order to evaluate the proposed approach, a prototype was developed and some experiments were accomplished, which demonstrated that the proposed strategy produces good results with respect to the generation of the Dataset Profile and, in particular, to the semantic enrichment of the metadata.

Keywords: Datasets on the Web. Dataset Profile. Metadata. Semantic Enrichment. Data Publication and Consumption.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ACRÔNIMOS

SUMÁRIO

1

INTRODUÇÃO

Este capítulo apresenta uma visão geral desta pesquisa e o contexto no qual este trabalho está inserido. Na Seção 1.1 é exposta a motivação deste estudo. Na Seção 1.2 é descrita a caracterização do problema abordado. Em seguida, na Seção 1.3, são elencados os objetivos deste trabalho. Uma discussão sobre as contribuições esperadas é realizada na Seção 1.4, e as hipóteses de pesquisa são apresentadas na Seção 1.5. Por fim, uma descrição da estrutura desta dissertação é apresentada na Seção 1.6.

1.1 MOTIVAÇÃO

A variedade de conjuntos de dados disponibilizados na Web possibilita um cenário ilimitado de informações, e combinações dessas informações podem trazer descobertas importantes. Os conjuntos de dados são coleções de dados publicados na Web e podem estar disponíveis por meio de distribuições, permitindo que esses conjuntos de dados sejam utilizados por vários grupos de consumidores de dados [Lóscio et al. 2017].

A produção e consumo de conjuntos de dados pode ser entendida como um ecossistema na Web, mas, nesse ecossistema, nem sempre os publicadores e consumidores se conhecem. Portanto, é necessário fornecer algumas informações sobre os conjuntos de dados e suas distribuições que possam contribuir para a sua compreensão, reutilização e, consequentemente, a comunicação entre publicadores e consumidores. Para isso, normalmente são disponibilizados metadados [Clarke et al. 2014]. Metadados fornecem informações que descrevem os dados, facilitando o processo de localização e reutilização dos conjuntos de dados por usuários ou aplicações (software), principalmente quando considerados os mais variados domínios aos quais os conjuntos de dados podem pertencer.

Os conjuntos de dados publicados na Web geralmente não disponibilizam descrições sobre seu conteúdo, pois ainda não é uma prática muito comum os publicadores fornecerem metadados que representem corretamente o conteúdo dos conjuntos de dados [Abele 2016]. Os metadados podem ajudar na compreensão e processamento dos dados por meio da disponibilização de informações descritivas sobre o conteúdo, estrutura, qualidade e outras características dos

conjuntos de dados. Como ilustração, os portais de dados abertos funcionam como um repositório para conjuntos de dados que podem ser livremente acessados e utilizados, sendo comum encontrar alguns metadados a seu respeito, como data de publicação e palavras-chave. Entretanto, nem sempre os metadados estão disponibilizados de forma estruturada e com informações suficientes que descrevem o conjunto de dados e sua estrutura, o que facilita seu processamento e entendimento [Oliveira et al., 2016].

Como uma boa prática de publicação de dados na Web, o *World Wide Web Consortium* (W3C) recomenda o uso de metadados para o fornecimento de informações que ajudem usuários e aplicações a entender os dados, bem como outros aspectos importantes que descrevem um conjunto de dados ou sua forma de disponibilização. Outra recomendação é que os dados sejam enriquecidos sempre que possível. Esse enriquecimento semântico ocorre por meio de um conjunto de processos que podem ser utilizados para aumentar, refinar ou melhorar dados brutos ou processados anteriormente, podendo resultar em novos dados com melhores descrições e significados [Lóscio et al. 2017].

O processo de enriquecimento também pode ser realizado em nível de metadados, de modo a ajudar na atribuição de seu significado e na melhoria das descrições dos conjuntos de dados. Podendo também ser realizado pela adição de informações e pela representação de metadados em formatos semânticos estruturados, como o modelo de dados RDF[1]. Essa representação facilita seu processamento, compreensão e reuso pelos usuários e aplicações (consumidores).

Para viabilizar a estruturação dos metadados de conjuntos de dados, alguns autores propuseram a criação de perfis. Abele (2016) define o perfil de um conjunto de dados como o grupo de informações descritivas e estatísticas a seu respeito. Nesta perspectiva, enfatiza-se que a criação de um perfil de conjunto de dados pode permitir que os consumidores de dados tenham acesso a metadados, possivelmente enriquecidos semanticamente, que são referenciados por termos de vocabulários padronizados e descrevem aspectos descritivos, estruturais e de qualidade do conjunto de dados. A ideia também é que o perfil do conjunto de dados esteja em formato estruturado e legível por máquinas.

1.2 PROBLEMA DE PESQUISA

Para entender melhor o problema em questão, consideremos um exemplo real a partir do Portal Brasileiro de Dados Abertos[2]. O referido portal é uma ferramenta disponibilizada pelo governo do Brasil para que a sociedade tenha acesso aos dados e informações públicas por meio

[1] https://www.w3.org/RDF/
[2] dados.gov.br

de conjuntos de dados. Essa iniciativa viabiliza a população o acesso a cerca de cinco mil conjuntos de dados em diversos formatos (e.g., XML[3], PDF, CSV), que compreendem diversas áreas, permitindo que sejam desenvolvidas novas aplicações para a análise desses dados. Além dos conjuntos de dados, também são disponibilizados para *download* dicionários de dados, que fornecem definições e representações dos elementos do conjunto de dados.

Como ilustração, entre os conjuntos de dados, é encontrado o conjunto de dados sobre Unidades Básicas de Saúde (UBS)[4] que disponibiliza dados referentes a cerca de 38 mil UBS, como sua localização, contato, situação em relação à estrutura física, acessibilidade, equipamentos e medicamentos. Este conjunto de dados, assim como outros, está disponibilizado no formato de dados estruturados CSV. Entretanto, seus metadados estão disponíveis por meio de um dicionário de dados em formato de documento de texto, que descreve as propriedades que compõem a estrutura do conjunto de dados, e de um quadro apresentado na sua página de acesso, conforme mostrado na Figura 1.

Figura 1 - Exemplo de metadados disponibilizados para um conjunto de dados no Portal Brasileiro de Dados Abertos.

Campo	Valor
Autor	Ministério da Saúde
Mantenedor	Ministério da Saúde
Última Atualização	25 de Janeiro de 2016, 16:10 (UTC-02:00)
Criado	21 de Agosto de 2013, 15:50 (UTC-03:00)
Atualidade	04/2013
Cobertura geográfica	Brasil
VCGE	Unidade básica de saúde [http://vocab.e.gov.br /2011/03/vcge#unidade-basica-saude], Atendimento ao cidadão [http://vocab.e.gov.br /2011/03/vcge#atendimento-cidadao]

Fonte: Conjunto de dados sobre Unidades Básicas de Saúde – UBS[5]

Disponibilizar os metadados por meio de uma tabela HTML[5] não é considerado uma boa forma para representá-los. Uma melhor escolha seria dispor esses metadados por meio de uma serialização RDF (e.g., RDFa[6]), que permite utilizar marcadores XHTML[7] com semântica e referenciados por vocabulários disponíveis na Web, como o Data Catalog Vocabulary[8] (DCAT). Além disso, não são disponibilizadas informações suficientes para a descrição do conjunto de dados, sendo necessária a geração de novos metadados que o descrevam, por exemplo, palavras-

[3] https://www.w3.org/XML/
[4] http://dados.gov.br/dataset/unidades-basicas-de-saude-ubs
[5] https://www.w3schools.com/html/
[6] https://rdfa.info/
[7] https://www.w3schools.com/html/html_xhtml.asp
[8] https://www.w3.org/TR/vocab-dcat/

chave e distribuição. Também podem ser disponibilizados metadados que fornecem informações adicionais, como os metadados de qualidade que representam critérios de qualidade e permitem avaliar a adequação do conjunto de dados para uma determinada tarefa.

Neste panorama, o problema alvo deste trabalho é definido como segue: *Dado o ecossistema de produção e consumo de conjuntos de dados na Web, como gerar metadados enriquecidos semanticamente que forneçam informações descritivas sobre o conteúdo, estrutura e qualidade dos conjuntos de dados, a fim de facilitar sua compreensão e seu processamento por consumidores?*

Este trabalho propõe uma estratégia para geração e disponibilização de metadados por meio da criação de um Perfil do Conjunto de Dados (PCD). O PCD é composto por metadados descritivos, estruturais e de qualidade enriquecidos semanticamente. Ao considerar a abordagem proposta, espera-se contribuir para o processo de disponibilização e geração de metadados enriquecidos semanticamente para descrever conjuntos de dados, nos quais sejam disponibilizados metadados facilmente compreensíveis por humanos e pelas máquinas.

1.3 OBJETIVOS

Este trabalho tem como principal objetivo a especificação de uma abordagem para a geração de PCD, visando facilitar a compreensão, assim como o processamento dos dados e metadados. O perfil é composto de: (i) metadados descritivos, que correspondem às características e conteúdo do conjunto de dados, como palavras-chave, título, identificação do domínio e a recomendação de vocabulários de domínio; (ii) metadados estruturais, que descrevem a estrutura interna do conjunto de dados por meio de suas propriedades; e (iii) metadados de qualidade, que dizem respeito a critérios de qualidade que contemplam a compreensibilidade e processabilidade do conjunto de dados. Os metadados do perfil são referenciados a partir de termos de vocabulários já existentes, o que permite agregar maior significado aos metadados. O perfil é representado em formato compreensível por máquina (RDF), facilitando seu processamento e manipulação. Desta forma, foram estabelecidos alguns objetivos específicos, sendo eles:

- Definição dos metadados descritivos e estruturais que compõem o perfil;
- Definição de critérios e métricas de qualidade, representados por metadados de qualidade, para compor o perfil;
- Especificação de uma abordagem que descreve conjuntos de dados na Web por meio da geração do Perfil de Conjunto de Dados (PCD);
- Implementação de um protótipo para automatizar o processo de geração do PCD;

- Realização de experimentos para avaliar a abordagem proposta.

1.4 CONTRIBUIÇÕES ESPERADAS

A principal contribuição deste trabalho é a especificação de uma abordagem para geração de PCD. De acordo com o levantamento realizado na literatura, não foram encontrados trabalhos que realizam a geração de um perfil composto de metadados enriquecidos semanticamente de forma abrangente, incluindo aspectos descritivos, estruturais e de qualidade, como proposto por esta abordagem. Entre as principais contribuições desta dissertação, destacam-se:

- Especificação de uma abordagem para geração de um Perfil de Conjunto de Dados, que considera aspectos descritivos, estruturais e de qualidade;
- Identificação do domínio do conjunto de dados;
- Recomendação de vocabulários de domínio;
- Definição de critérios de qualidade associados ao conjunto de dados;
- Criação de uma ferramenta a partir da implementação da abordagem proposta;
- Realização de experimentos com o protótipo desenvolvido.

1.5 HIPÓTESES

A dissertação possui as seguintes hipóteses de pesquisa:

H1 - Por meio das informações e metadados extraídos do conjunto de dados é possível gerar metadados descritivos, estruturais e de qualidade enriquecidos semanticamente, sem a necessidade de intervenção humana, resultando em um PCD com descrições mais completas sobre o conjunto de dados.

H2 – A identificação automática do metadado “domínio de conhecimento” do conjunto de dados apresenta um resultado similar à identificação do domínio realizado de forma manual;

H3 – A recomendação automática de vocabulários de domínio para o conjunto de dados apresenta um resultado similar à recomendação de vocabulários de domínio realizada de forma manual;

H4 - O perfil gerado melhora aspectos relacionados à descrição da qualidade do conjunto de dados, pois agrega informações para melhorar sua compreensão e/ou processamento por parte dos usuários e aplicações consumidoras de dados.

Cada hipótese será avaliada por meio de experimentos que serão apresentados ao longo deste trabalho.

1.6 ESTRUTURA DA DISSERTAÇÃO

Além do presente capítulo, este trabalho está organizado como segue:

- O Capítulo 2 introduz a fundamentação teórica referente aos principais conceitos para o desenvolvimento desta pesquisa;
- O Capítulo 3 descreve e discute os trabalhos relacionados;
- O Capítulo 4 especifica a abordagem proposta para o processo de geração do Perfil de um Conjunto de Dados (PCD);
- O Capítulo 5 destaca aspectos relacionados à implementação do protótipo e aos experimentos realizados para avaliar as hipóteses de pesquisa, apresentando os resultados obtidos;
- O Capítulo 6 apresenta algumas considerações sobre esta pesquisa, contribuições, limitações e trabalhos futuros.

2

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os principais conceitos relacionados ao desenvolvimento deste trabalho. Primeiramente, na Seção 2.1, são introduzidos conceitos acerca da Web Semântica e tecnologias relacionadas. Na Seção 2.2 são abordados tópicos associados à produção e consumo de conjuntos de dados na Web. Em seguida, na Seção 2.3, são discutidos os conceitos relacionados aos metadados. A Seção 2.4 inclui definições acerca de enriquecimento semântico. Na Seção 2.5 encontram-se os conceitos relacionados à Qualidade da Informação, detalhando os critérios e métricas relacionados aos conjuntos de dados. Na Seção 2.6 são expostos os conceitos relacionados ao perfil de conjunto de dados. Por fim, na Seção 2.7 são apresentadas considerações sobre o capítulo.

2.1 WEB SEMÂNTICA

Os benefícios compartilhados pela *World Wide Web*[9] (WWW) permitiu a disseminação da informação, fazendo com que o conhecimento não seja apenas um privilégio de poucos. Apesar dos dados estarem, em sua maior parte, disponibilizados na Web, ainda não é fácil localizá-los e extrair informações relevantes. A falta de utilização de padrões que estabeleçam esses dados são obstáculos para sua compreensão, integração e processamento, pois, para estas atividades, é necessário entender os dados. Esse entendimento deve ocorrer tanto pelos usuários quanto pelas aplicações que irão consumi-los, entretanto a maior parte das fontes disponibilizam dados em formatos que podem ser compreendidos pelos humanos, mas não pelas máquinas. A produção de dados apenas para o consumo humano dificulta seu processamento automático.

As tecnologias associadas à Web (e.g., HTML, URL) possibilitam o acesso e compartilhamento de dados, porém não é suficiente para que esses dados possam ser realmente acessíveis e compartilháveis entre seus usuários. A disponibilização de dados em formato estruturado e que permita a extração de semântica e significado torna possível

[9] https://webfoundation.org/

solucionar alguns dos problemas relacionados ao acesso e compartilhamento de dados, como a heterogeneidade e integração de dados.

A Web Semântica propõe uma estrutura para a publicação de dados em que é possível tornar o conteúdo significativo e processável pelas máquinas. Para Berners-Lee et al. (2001), a Web Semântica representa uma extensão da *World Wide Web*, que permite o compartilhamento de dados de forma ilimitada entre seus usuários. Para isso, os dados publicados devem estar em formato estruturado e utilizando padrões, o que torna possível obter o significado dos dados, mesmo que sejam provenientes de diferentes tipos de fontes de dados. Isso possibilita seu entendimento tanto pelos humanos como pelas máquinas. Com a Web Semântica é possível prover dados úteis na internet, que possam ser analisados para as mais variadas tarefas, viabilizando a obtenção do significado dos dados provenientes de diferentes tipos de fontes da Web. Assim como a WWW, a Web Semântica está diretamente relacionada a algumas tecnologias que facilitam o entendimento e acesso dos dados de forma universal (e.g., URI[10], XML, RDF, OWL[11] e SPARQL[12]), e ao uso de recursos como as ontologias e os vocabulários.

2.1.1 Ontologias e Vocabulários

A palavra ontologia vem da filosofia e representa o estudo sistemático do ser, categorias, princípios e essência. Na Ciência da Computação, uma ontologia descreve conceitos promovendo o conhecimento e compartilhamento dos dados, permitindo sua interpretação por seres humanos ou por máquinas [Gruber 1993]. As ontologias podem ser consideradas um instrumento útil para a representação e compartilhamento do conhecimento, sendo possível combinar informações provenientes de diferentes bases de dados. Também estabelecem a compreensão comum de principais conceitos e relações envolvidas, podendo servir como base para a modelagem de vocabulários em diferentes domínios [Falbo et. al. 2016].

As ontologias podem estabelecer uma definição formal das relações entre termos e a equivalência desses termos, que ocorre geralmente por meio da disponibilização de taxonomia e conjunto de regras de inferência [Berners-Lee et. al. 2001], onde:

- Taxonomia: definem classes, subclasses e propriedades de objetos, permitindo expressar um grande número de relações entre entidades;

[10] https://www.w3.org/wiki/URI
[11] https://www.w3.org/OWL/
[12] https://www.w3.org/TR/rdf-sparql-query/

- Conjunto de regras de inferência: possibilitam a manipulação dos termos de forma eficaz e significativa para o usuário humano.

OWL (*Ontology Web Language*) é uma linguagem padrão proposta pelo *World Wide Web Consortium*[13] (W3C) e utilizada para representar um conhecimento rico e complexo sobre coisas, grupos de coisas e relações entre coisas. Essa linguagem é utilizada para definição de ontologias, tornando possível que o conhecimento expresso possa ser processado por aplicações. Isso possibilita identificar nos dados seus recursos e suas relações, sem ambiguidades, que pode ser utilizado para adicionar mais significado e restrições à representação dos dados, atribuindo semântica e definindo relações.

As ontologias também podem melhorar a precisão de buscas na Web, relacionando informações de uma página com as estruturas de conhecimento associadas, podendo também responder a consultas que exigiriam o conhecimento humano. Por exemplo, ao utilizar ontologias para referenciar os termos contidos em uma base de dados, esses termos devem representar classes e propriedades que compõem um vocabulário, o que permite acesso a um significado comum, mesmo para termos de diferentes bases de dados.

Para facilitar o entendimento, significado e o uso integrado dos dados, devem ser utilizados vocabulários que possuem uma semântica bem definida, eliminando ambiguidades em termos utilizados nas diferentes bases de dados. Esses vocabulários podem utilizar ontologias para representar um conjunto de termos pertencentes a um domínio. Atualmente existem centenas de vocabulários disponíveis na Web, sendo possível encontrar os vocabulários mais apropriados para um determinado domínio (e.g., saúde, música). Para ajudar nesta tarefa, existem catálogos de vocabulários abertos, como o *Linked Open Vocabularies*[14] (LOV), que é um dos maiores catálogos de vocabulários abertos. Devem ser reutilizados vocabulários existentes e de preferência os mais populares. Entretanto, nos casos em que seja necessário criar um novo vocabulário, deve existir o cuidado em reutilizar o maior número possível de termos de ontologias já existentes, o que evita a criação de referências diferentes para os mesmos conceitos [Laufer 2015].

Os vocabulários, assim como suas classes e propriedades, são identificados por um identificador universal de recursos (URI), sendo este um identificador único que permite a eliminação de ambiguidades e possibilita o compartilhamento de conhecimentos com consumidores de dados e agentes de software. Uma URI também pode estar vinculada a uma

[13] https://www.w3.org/
[14] http://lov.okfn.org/dataset/lov/

definição exclusiva para um conceito e pode ser encontrada na Web [Berners-Lee et al. 2001]. Por exemplo, o vocabulário FOAF é identificado pela URI "http://xmlns.com/foaf/0.1/", e suas classes e propriedades recebem uma URI que é construída a partir da concatenação da URI do vocabulário com o nome da respectiva classe ou propriedade. A propriedade "*name*" do vocabulário FOAF tem a URI correspondente: "http://xmlns.com/foaf/0.1/name" [Laufer 2015].

2.1.2 Modelo de Dados RDF

A Web Semântica utiliza algumas tecnologias padronizadas pelo W3C, entre elas o *Resource Description Framework* (RDF). O RDF é um modelo padrão para intercâmbio de dados na Web, que facilita o cruzamento de dados, utiliza URI para nomear a relação entre as coisas e uma estrutura de vinculação que forma um grafo rotulado e direcionado [W3C 2014].

Inicialmente, RDF foi desenvolvido para representar metadados sobre recursos da Web, mas também para prover a disponibilização dos dados de forma estruturada, sendo possível atribuir significado a esses dados, representando e relacionando recursos, tornando-se o padrão para a modelagem de informações na Web Semântica [Berners-Lee et al. 2001]. RDF permite a descrição de recursos por meio de propriedades e valores, que são representados por conjuntos de triplas (sujeito + predicado + objeto), o que permite expressar as informações sob a forma de grafos, descrevendo os dados por meio de nós e arestas [Chakkarwar et al. 2016]. O sujeito é identificado por uma URI, que dá acesso a uma definição para um conceito, permitindo a eliminação de ambiguidades. O predicado também é um recurso e é responsável por estabelecer o relacionamento entre o sujeito e o objeto, permitindo relacionar recursos aos dados e relacionamentos entre recursos. Já o objeto pode representar um recurso identificado por uma URI ou ser representado por um literal. Essa forma em que os dados são estruturados e disponibilizados facilita a compreensão e a troca de informações. Um exemplo é ilustrado na Figura 2, onde o sujeito "Tim Berners-Lee" é representado pela URI "http://dbpedia.org/resource/Tim_Berners-Lee", o predicado se refere à uma relação de equivalência entre o sujeito e o objeto, representado pela URI "http://www.w3.org/2002/07/owl\#sameAs". O objeto é representado pela URI "http://www4.wiwiss.fu-berlin.de/dblp/resource/person/100007", que faz referência à mesma pessoa representada pelo sujeito.

Figura 2 - Representação de Tripla (sujeito + predicado + objeto)

Sujeito: <http://dbpedia.org/resource/Tim_Berners-Lee>

Predicado: <http://www.w3.org/2002/07/owl\#sameAs>

Objeto: <http://www4.wiwiss.fu-berlin.de/dblp/resource/person/100007>

Fonte: O Autor (2018).

A Figura 3 apresenta um exemplo de arquivo RDF (serializado em Turtle), em que é possível identificar dois conjuntos de triplas, cada uma delas representa uma determinada entidade. Cada tripla apresenta informações de funcionários. A interligação entre o sujeito e objeto é realizada por meio de predicados referenciados por termos de vocabulários, onde: o id dos funcionários é representado pelo predicado "exOrg:employeeId"; o nome é representado pelo predicado "foaf:name"; e o email pelo "foaf:mbox".

Figura 3 - Exemplo de Documento RDF (Turtle).

```
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
@prefix exOrg:  <http://org.example.com/employees#> .

_:emp1  exOrg:employeeId   "1234";
        foaf:name "Mary Smith";
        foaf:mbox "mary@example.com".

_:emp2  exOrg:employeeId   "5678";
        foaf:name "John Green";
        foaf:mbox "john@example.com".
```

Fonte: O Autor (2018).

O *SPARQL Protocol and RDF Query Language* (SPARQL) é a linguagem de consulta padrão, recomendada pelo W3C, para recuperação de informações contidas em dados nos formatos RDF e OWL. Essa linguagem permite que o usuário tenha acesso a resultados de pesquisa categorizados e personalizados. Para facilitar o processamento das consultas e o acesso aos dados, geralmente são disponibilizados SPARQL endpoints, sendo este um serviço que recebe consultas SPARQL de clientes. A consulta mais utilizada é a "Select", que possui a seguinte estrutura [Prud'hommeaux et al. 2013]:

- PREFIX: fornece declarações de prefixo de *namespace*, para abreviar URIs;
- Select: Especifica uma projeção sobre os dados, com uma ou mais variáveis para armazenar os resultados da consulta;
- From: Declara qual o grafo RDF que será consultado;

- Where: Impõe restrições na consulta, com padrões de triplas a serem localizados no grafo RDF definido.

A consulta SPARQL segue o conceito de BGP (*Basic Graph Pattern*), que é o conjunto de padrões de tripla (ou grafos), sendo estruturado no formato sujeito + predicado + objeto. A Figura 4 mostra um exemplo de consulta SPARQL realizada sobre o RDF representado na Figura 3. A consulta tem como objetivo recuperar o nome e o email do funcionário que corresponde ao id "1234". Essa consulta é composta de três padrões de triplas: o primeiro procura por triplas que possuam o predicado "foaf:name", que especifica o nome dos funcionários. De forma semelhante, o segundo padrão procura por triplas que possuam o predicado "foaf:mbox", que especifica o email dos funcionários. E o terceiro padrão de tripla especifica que o funcionário procurado deve possuir o id "1234". A resposta desta consulta é recebida no formato de uma tabela.

Figura 4 - Consulta *select* SPARQL sobre o RDF representado na Figura 3.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX exOrg:  <http://org.example.com/employees#>

SELECT ?name ?email WHERE {
            ?emp foaf:name ?name;
                 foaf:mbox ?email;
                 exOrg:employeeId "1234"
            }
```

Fonte: O Autor (2018).

2.2 PRODUÇÃO E CONSUMO DE CONJUNTOS DE DADOS NA WEB

Com a crescente geração de informações nos últimos anos, iniciou-se um processo de aumento exponencial na quantidade de dados na Web. Isso resultou em uma grande variedade de conjuntos de dados disponibilizados na Web, o que possibilita um cenário ilimitado de informações, e combinações dessas informações podem trazer descobertas importantes. Um conjunto de dados pode ser definido como uma coleção de dados, publicados ou curados por um agente e disponível para acesso ou download em um ou mais formatos, onde uma distribuição representa uma forma específica de sua disponibilização [Maali et al. 2014].

As diferentes distribuições (e.g., CSV, JSON[15]) facilitam o compartilhamento e reuso dos conjuntos de dados, se estiverem em formato aberto, por grupos de consumidores de

[15] https://www.json.org/

dados, sejam eles humanos ou aplicações [Lóscio et al. 2017]. As distribuições também facilitam o compartilhamento de dados em larga escala, permitindo a utilização de conjuntos de dados por vários grupos de consumidores de dados, independente do propósito.

No ecossistema de produção e consumo de conjuntos de dados existem diferentes atores, podendo ter papéis de publicadores de dados, que estão relacionados à publicação, e/ou de consumidores de dados, que estão relacionados ao seu consumo.

De acordo com o W3C, um publicador pode ser uma pessoa ou um grupo responsável por gerar, manter e compartilhar dados abertamente ou com acesso controlado [Lóscio et al. 2017]. Entre os publicadores podem existir diferentes atores, relacionados à publicação de informações e processos, por exemplo, relacionados à definição de licenças, criação de regras de definição do formato das URIs, escolha dos formatos dos dados e das plataformas para a distribuição das informações, definição do conjunto obrigatório de metadados e documentos, seleção de estratégias para garantir a persistência, a preservação e o arquivamento dos dados [Laufer 2015]. Já os consumidores de dados utilizam os dados publicados, o que requer acesso e meios que possibilitem adquiri-los e/ou filtrá-los. Os consumidores podem estar interessados no consumo direto dos dados, como também podem estar interessados em transformar esses dados e agregar algum valor, podendo resultar na publicação de outro conjunto de dados [Laufer 2015].

Entretanto, normalmente os publicadores e consumidores de dados não se conhecem, tornando-se necessário um meio que possibilite o entendimento comum dos dados. Para isso, os publicadores de dados devem fornecer informações adicionais que descrevem seus conjuntos de dados e distribuições, como sua representação, descrição e forma de disponibilização. Isso facilita o entendimento dos dados, por humanos e pelas máquinas, como também facilita a comunicação entre os publicadores e consumidores de dados, permitindo a compreensão e reutilização dos conjuntos de dados.

Os metadados representam informações adicionais que ajudam os consumidores de dados a entenderem melhor o significado dos dados, sua estrutura e esclarecer outros problemas, como direitos e termos de licença, a organização que gerou os dados, qualidade de dados, métodos de acesso aos dados e o cronograma de atualização de conjuntos de dados [Lóscio et al. 2017]. Vocabulários que possuem uma semântica bem definida também podem ser utilizados para a interpretação dos dados de maneira semelhante entre as diferentes organizações e plataformas envolvidas em links de dados. Os vocabulários tornam possível que o significado pretendido pelo publicador dos dados seja o mesmo que o significado

entendido pelo consumidor dos dados, eliminando ambiguidades em termos utilizados em diferentes conjuntos de dados.

O W3C criou uma recomendação com um conjunto de Melhores Práticas para Publicação e Consumo de Dados na Web[16]. Por meio dela, são abordados diferentes aspectos relacionados à publicação, ao consumo de dados, sua reutilização e possibilidades de acesso, sendo possível obter alguns benefícios, como: compreensão, processabilidade, descoberta, reuso, acesso e interoperabilidade [Lóscio et al. 2017].

2.3 METADADOS

Os conjuntos de dados publicados na Web geralmente não disponibilizam descrições sobre seu conteúdo. Entretanto, para que os consumidores melhor compreendam esses conjuntos de dados, é necessário fornecer algumas informações sobre o conteúdo e suas distribuições, o que também contribui para a sua reutilização. Isso normalmente é realizado por meio de metadados [Clarke et al. 2014].

Com a disponibilização de metadados é possível associar semântica aos dados, fornecendo aos consumidores um melhor entendimento de seu significado, que deve ocorrer tanto pelos usuários quanto pelas aplicações que irão consumi-los. Metadados descrevem conjuntos de dados de forma geral, disponibilizando informações que facilitam seu uso e descoberta [Oliveira et al. 2016]. Os metadados melhoram a compreensão e processamento dos dados por meio da disponibilização de informações descritivas sobre o conteúdo, estrutura, qualidade e outras características dos conjuntos de dados que facilitam sua localização e acesso por parte de mecanismos de busca ou aplicações. Metadados possuem diferentes classificações, sendo duas delas as principais:

- Metadados descritivos: Contêm características identificadoras mais importantes de um recurso e a análise de seu conteúdo para fins de descoberta, identificação, seleção e aquisição [Joudrey et al. 2017].
- Metadados estruturais: Referem-se à composição ou organização interna de um objeto digital, conjunto de dados ou outro recurso que está sendo descrito [Joudrey et al. 2017].

Os metadados devem ser disponibilizados utilizando padrões, que podem ser termos precisos e adequados para a descrição e qualidade da informação, como os termos reutilizados de vocabulários recomendados pelo W3C. Por exemplo, podem ser empregados os

[16] https://www.w3.org/TR/dwbp/

vocabulários Dublin Core Metadata Initiative Metadata Terms[17] (DCTerms) e o Data Catalog Vocabulary (DCAT) para descrever conjuntos de dados e para agregar maior significado aos metadados, aumentando a interoperabilidade, reutilização de dados e evitando redundâncias e ambiguidades.

Entretanto, o fornecimento de metadados que representam corretamente o conteúdo dos conjuntos de dados não é uma prática comum entre os publicadores de dados [Abele 2016]. Em portais de conjuntos de dados abertos, é comum encontrar alguns metadados, entretanto, nem sempre os mesmos são apresentados de forma estruturada com informações suficientes para o seu entendimento e processamento [Oliveira et al. 2016]. Para que se tornem úteis, esses metadados podem ser enriquecidos para que sejam geradas melhores descrições dos conjuntos de dados, como a identificação de domínio e correção dos metadados.

2.4 ENRIQUECIMENTO SEMÂNTICO

O enriquecimento de dados refere-se a um conjunto de processos que podem ser utilizados para aumentar, refinar ou melhorar dados brutos ou processados anteriormente [Lóscio et al. 2017], resultando no melhoramento das descrições sobre um conjunto de dados e em análises de dados mais eficientes. Esse enriquecimento também pode ser realizado em nível de metadados, o que ajuda na atribuição de significado, melhora suas descrições e complementa informações que podem promover a compreensão e processamento dos dados por usuários e aplicações (consumidores).

Enriquecimento semântico de metadados é entendido como o resultado de processamentos para identificar algumas informações adicionais [Parinov 2014]. Durante esse processo também devem ser utilizados vocabulários já existentes, a fim de agregar maior significado a esses metadados, complementando seu significado, e também a representação dos dados em formatos estruturados (e.g. RDF), facilitando seu processamento, compreensão e reuso pelas diferentes organizações e plataformas. A Figura 5 apresenta um exemplo em que os metadados pertencentes a um conjunto de dados são escritos em sintaxe RDF. Primeiramente são especificados os vocabulários utilizados, em que é atribuído um prefixo para cada um deles. O conjunto de dados é representado pelo elemento “:dataset-02123”, o qual compõe uma série de triplas, que retratam seus metadados.

[17] http://dublincore.org/documents/dcmi-terms/

Figura 5 - Exemplo de conjunto de metadados disponibilizados em sintaxe RDF.

```
@prefix dc:    <http://purl.org/dc/terms/> .
@prefix dcat:   <http://www.w3.org/ns/dcat#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .

:dataset-1336
        a dcat:Dataset ;
        dc:title "Football Events" ;
        dcat:keyword ["culture", "sports", "association football"] ;
        dcat:theme "sports" ;
        dc:publisher "Alin Secareanu" ;
        dc:issued "2017-01-25T01:19:19.907Z" ;
        dc:modified "2017-01-25T01:19:19.907" ;
        owl:versionInfo "1" ;
        dcat:landingPage "https://www.kaggle.com/secareanualin/football-events" .
```

Fonte: O Autor (2018)

Os metadados, apresentados no exemplo da Figura 5, são referenciados por termos de vocabulários que possuem significado correspondente, o que permite associar semântica e integrá-los com outros dados e recursos na Web, facilitando também seu o entendimento e manipulação. Entretanto, em muitos casos, esses metadados são representados apenas por literais, tornando necessário encontrar vocabulários que permitam referenciá-los. Por exemplo, o metadado "*keyword*" que, por apresentar significado correspondente, pode ser referenciado pelo termo *dcat:keyword*, pertencente ao vocabulário DCAT.

Segundo Clarke et al. (2014), o enriquecimento semântico pode ser realizado por meio de uma categoria adicional de metadados que melhore ainda mais a utilidade, descoberta e interoperabilidade dos dados. Isso resulta na geração de novos metadados que representam informações úteis para a descrição e acesso aos conjuntos de dados. Também é possível enriquecer dados por meio da criação de ligações semânticas entre metadados de diferentes conjuntos de dados, podendo criar relacionamentos entre eles. No trabalho apresentado por Parinov (2014), foi proposto o enriquecimento semântico por meio da criação de ligações semânticas de metadados provenientes de artigos publicados que estão relacionados entre si, permitindo que os autores desses artigos criem relações científicas a partir de suas próprias publicações.

2.5 QUALIDADE DA INFORMAÇÃO

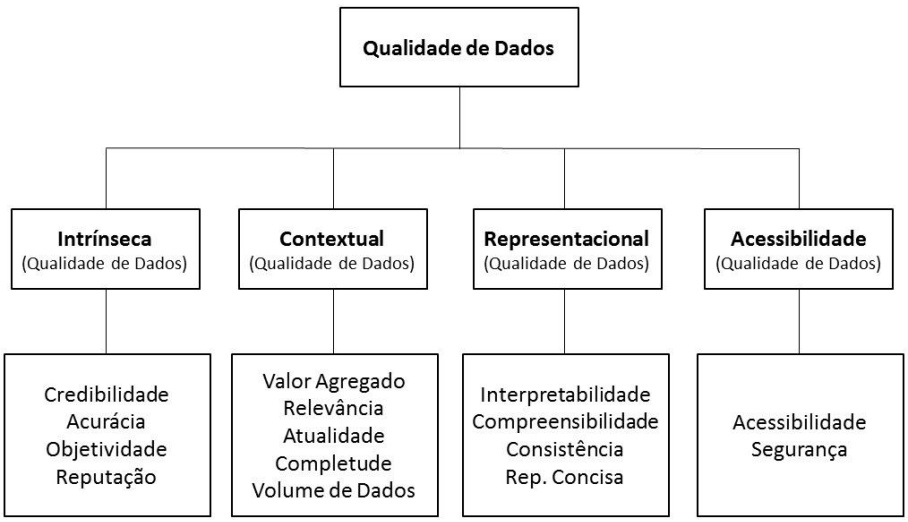
Apesar da grande quantidade de dados disseminados pela Web, encontrar informações úteis não é um trabalho simples. Identificar a qualidade de dados facilita o processo de extração de informações, importante para a tomada de decisões. A Qualidade da Informação

(QI) é considerada um dos aspectos mais importantes para os consumidores e usuários da internet [Naumann et al. 2000]. A QI pode ser definida como um conjunto de critérios ou dimensões utilizados para indicar o grau de qualidade geral de uma informação obtida por um sistema [Batista et al. 2007]. É possível encontrar na literatura uma grande quantidade de trabalhos que abordam a QI, entre estes trabalhos se destacam os trabalhos dos autores Wang et al. (1996), Naumann et al. (2000) e Pipino et al. (2002).

A QI é definida por Wang et al. (1996) com o termo "adequação ao uso" ("*fitness for use*"). Uma informação é considerada valiosa se corresponder a um conjunto de requisitos para uma determinada finalidade. Para isso, são definidos critérios de qualidade que são avaliados por métricas, e cada métrica resulta em um escore de avaliação. Essas métricas são heurísticas desenvolvidas para adequar-se a uma situação de avaliação específica [Pipino et al. 2002]. Muitos trabalhos estabeleceram suas próprias classificações para os critérios de qualidade, mas nenhum deles foi estabelecido como padrão, sendo possível encontrar critérios que possuem a mesma definição, mas são representados por diferentes nomenclaturas.

No trabalho de Wang et al. (1996) foi proposto um framework que avalia a qualidade dos dados por meio de critérios de QI. Foram definidas métricas que mapeiam funções do mundo real para um sistema de informação, sendo considerada uma referência na área de critérios de qualidade, pois foi um dos primeiros trabalhos na área. Na Figura 6 são apresentados os quinze critérios de qualidade definidos, que são agrupados em quatro categorias (intrínseca, contextual, representacional e acessibilidade).

Figura 6 - Categorias e critérios de Qualidade dos Dados.



Fonte: Adaptado de Wang et al. (1996).

Outro trabalho, mais recente, foi desenvolvido por Flemming (2011) com o objetivo de avaliar a qualidade de fontes de dados Linked Data, sendo apresentados critérios que representam a qualidade de uma fonte de dados. A fonte de dados pode ser representada por um conjunto de dados que pode suportar vários meios de acesso. Para avaliar a qualidade de uma fonte de dados, foram estabelecidos onze critérios de qualidade, e cada um deles possui um conjunto de indicadores. Esses indicadores constituem aspectos mensuráveis de um critério, permitindo a avaliação da qualidade de uma fonte de dados. Alguns desses critérios e seus respectivos indicadores podem ser observados no Quadro 1.

Quadro 1 - Categorias, critérios e indicadores de qualidade.

Categoria	Critério	Indicadores
Conteúdo	Consistência	* nenhuma definição de entidades como membros de classes disjuntas; * aplicação válida de propriedades de funcionalidade inversa; * nenhuma redefinição de propriedades existentes; * tipos de dados homogêneos; * nenhuma afirmação de valores inconsistentes para propriedades.
Representação	Compreensibilidade	* marcação legível por humanos de classes, propriedades e entidades, fornecendo um rdfs: label; * descrição legível por humanos de classes, propriedades e entidades, fornecendo um rdfs: comentário; * indicação de metadados sobre um conjunto de dados; * indicação de uma ou mais URIs exemplares; * indicação de uma expressão regular que corresponde às URIs de um conjunto de dados; * indicação de uma consulta SPARQL exemplar; * indicação de alguns dos vocabulários utilizados; * operacionalidade de documentos HTML; * fornecimento de fóruns e listas de discussão.
Uso	Licenciamento	* indicação de uma licença legível por máquina; * indicação de uma licença legível por humanos; * permição para a reprodução de dados; * distribuição permitida de dados; * permição para modificar e redistribuir dados; * nenhuma atribuição necessária; * não é necessário Copyright.
Sistema	Acessibilidade	* acessibilidade do servidor; * URIs referenciáveis; * URIs não contêm informações variáveis; * letra homogênea (minúsculas) de URIs; * redirecionamento usando o código de status 303; * acessibilidade do SPARQL endpoint; * acessibilidade dos RDF dumps.

Fonte: Adaptado de Flemming (2011).

Por exemplo, o critério de qualidade “Acessibilidade”, que se refere ao bom funcionamento de todos os métodos de acesso e pertence à categoria “Sistema”, apresenta um conjunto de indicadores que estão relacionados à acessibilidade do servidor, disponibilização de URIs, redirecionamento de página e acesso a SPARQL endpoint. Cada um dos 11 critérios de qualidade estabelecidos pertence a uma das seguintes categorias:

- Conteúdo - composto por três critérios que influenciam na correção de dados fornecidos por uma fonte de dados, sendo: consistência, pontualidade e veracidade.
- Representação - contém os seguintes critérios relacionados à disponibilização dos dados: uniformidade, versatilidade e compreensão geral.
- Uso - possui critérios que influenciam na usabilidade dos dados, sendo eles: validade dos documentos, quantidade de dados e licenciamento dos dados.
- Sistema - aborda critérios relativos ao sistema relacionado à publicação de dados, sendo: acessibilidade e desempenho.

Ao observar estes e outros trabalhos, é possível identificar uma grande quantidade de critérios de qualidade que podem ser utilizados para representar aspectos de um conjunto de dados, o que permite avaliar sua adequação para uma determinada tarefa. São exemplos os critérios: exatidão, completude, relevância e compreensibilidade [Naumann et al. 2000].

2.6 PERFIL DE CONJUNTO DE DADOS

Um perfil de um conjunto de dados pode ser definido como um grupo de informações descritivas e estatísticas a seu respeito, podendo ser geradas informações gerais sobre os conjuntos de dados, como título, descrição, data de atualização e informações de licença [Abele 2016]. O perfil possibilita a disponibilização de informações associadas aos conjuntos de dados, visando melhorar a descrição e descoberta de dados [Abele 2016; Assaf et al. 2015; Ellefi et al. 2014; Fetahu et al. 2014]. Ellefi et al. (2014) define um perfil de conjunto de dados como um conjunto de características, semânticas e estatísticas, que permite descrever da melhor forma um conjunto de dados. Esses autores se baseiam no conceito de *Dataset Profiling*, que trata do fornecimento de informações semânticas que possam ser associados aos conjuntos de dados, visando melhorar sua descrição e facilitar a descoberta de dados [Ellefi et al. 2014].

A disponibilização de perfis de conjuntos de dados pode facilitar o processo de recuperação de informações mais relevantes [Fetahu et al. 2014]. Prover um perfil também facilita a comunicação entre publicadores e consumidores de dados e o uso integrado dos

conjuntos de dados. Segundo [Ellefi et al. 2014], a criação de um perfil ajuda na identificação de conjuntos de dados, permitindo sua melhor descrição, representação e a criação de links entre os conjuntos de dados.

Os perfis podem ser representados em formato estruturado por metadados que descrevem os conjuntos de dados, tornando possível o acesso a informações adicionais para o entendimento de seu significado e de sua estrutura. Os metadados podem ser gerados em formato RDF e utilizando vocabulários, o que ajuda na atribuição de semântica aos dados [Fetahu et al. 2014]. Como o exemplo apresentado na Figura 5, o qual pode ser considerado um PCD referente ao conjunto de dados "*Football Events*"[18], gerado em formato RDF e composto de metadados descritivos que são referenciados por vocabulários, como o DCAT e o DCTerms.

A geração de um perfil de conjuntos de dados pode incluir etapas relacionadas aos tópicos apresentados anteriormente. Ao considerar os princípios da Web Semântica, é possível aprimorar o processo de produção e consumo de conjuntos de dados na Web por meio da geração de metadados enriquecidos semanticamente, que representem características descritivas, estruturais e de qualidade dos conjuntos de dados.

2.7 CONSIDERAÇÕES

Neste capítulo foram apresentados alguns conceitos considerados relevantes para o desenvolvimento da proposta deste trabalho. Foram apresentados aspectos e conceitos referentes à Web Semântica, incluindo conceitos relacionados a ontologias e vocabulários, e modelo de dados RDF. Também foram introduzidos conceitos acerca da produção e consumo de conjuntos de dados na Web e, em seguida, conceitos relacionados aos metadados, destacando sua importância e principais classificações. Posteriormente foram apresentados conceitos relacionados ao enriquecimento semântico para o melhoramento das descrições sobre um conjunto de dados. Também foram apontadas as principais definições e alguns trabalhos encontrados na literatura que propõem e classificam critérios de Qualidade da Informação. Por fim, foram abordados conceitos e definições relacionadas ao perfil de conjunto de dados. O próximo capítulo apresenta os trabalhos que estão relacionados ao contexto desta dissertação.

[18] https://www.kaggle.com/secareanualin/football-events

3

TRABALHOS RELACIONADOS

A geração de perfil de conjuntos de dados com metadados enriquecidos é uma atividade composta por diferentes etapas quando se vislumbra a geração tanto de metadados descritivos e estruturais quanto de qualidade. Este capítulo tem como objetivo apresentar e discutir alguns trabalhos existentes na literatura que estão relacionados a estas etapas que compõem a proposta desta pesquisa.

Os trabalhos foram divididos em seções da seguinte forma: Seção 3.1 expõe os trabalhos relacionados à geração de perfil de conjuntos de dados; a Seção 3.2 apresenta os trabalhos relacionados ao enriquecimento semântico dos conjuntos de dados e de seus metadados; e a Seção 3.3 mostra os trabalhos relacionados à qualidade. Ao fim de cada uma destas seções são apresentadas algumas discussões comparando as estratégias dos trabalhos à proposta desta dissertação. Na Seção 3.4 são apresentadas algumas considerações.

3.1 GERAÇÃO DE PERFIL DE CONJUNTOS DE DADOS

Prover um perfil de um conjunto de dados facilita a comunicação entre publicadores e consumidores, o uso integrado dos conjuntos de dados, e o processo de identificação de conjuntos de dados.

Nos últimos anos estudos acerca da geração de perfil de conjunto de dados têm sido realizados. Exemplos são as abordagens propostas por Abele (2016), Assaf et al. (2015) e Fetahu et al. (2014), sendo alguns dos trabalhos mais relevantes que são apresentados a seguir.

3.1.1 Linked Data Profiling

O trabalho apresentado por Abele (2016) propõe abordagens para a descrição detalhada de conjuntos de dados conectados. Apesar da crescente quantidade de dados publicados no padrão *Linked Data*, esses dados geralmente não estão conectados uns aos outros, mesmo este sendo um dos princípios para a publicação de dados neste padrão. Essa prática impossibilita a descoberta de conjuntos de dados que apresentam recursos semelhantes, o autor a credita ao fato de que os produtores de dados não tem conhecimento sobre os conjuntos de dados já publicados, ligado ao fato dos conjuntos de dados publicados geralmente não disponibilizarem descrições sobre seu

conteúdo. Com o fornecimento de metadados é possível representar corretamente o conteúdo dos conjuntos de dados e criar ligações entre os conjuntos de dados na nuvem *Linked Open Data*[19] (LOD). Para isso, o autor propõe as seguintes abordagens para a descrição detalhada de conjuntos de dados conectados:

- **Abordagem para descoberta de recursos LOD:** realiza uma análise sobre os conjuntos de dados existentes na nuvem LOD e disponibiliza informações detalhadas dos recursos contidos nesses conjuntos de dados por meio de um conjunto de metadados. Entre estes metadados incluem: descrição, data de atualização, estatísticas sobre frequências e distribuições de sujeitos, predicados e objetos, informações de licença, lista dos tipos de dados usados para literais, e uma lista de vocabulários utilizados.
- **Abordagem para assistência de publicação de dados:** A partir dos metadados gerados, são sugeridos ao publicador os conjuntos de dados que podem ser vinculados ao conjunto de dados em questão.

As abordagens compartilham os seguintes processos: (i) Coleta de estatísticas: fornece estatísticas sobre assuntos distintos, predicados e objetos, lista dos diferentes tipos de dados usados para literais e uma lista de vocabulários usados; (ii) Identificação do domínio: identifica o domínio do conjunto de dados, por meio da análise de literais de um dado conjunto de dados que são vinculados às categorias do DBpedia[20]. O autor escolheu o DBpedia por considerar que abrange grandes domínios e por não encontrar uma situação em que o domínio que descreve um conjunto de dados não esteja presente nas categorias DBpedia.

Para os experimentos foram processados conjuntos de dados no formato RDF, e como resultado são gerados metadados que são salvos em arquivo RDF utilizando o Vocabulário de Conjuntos de Dados Interligados[21] (VoID). O editor pode adicionar esses metadados ao conjunto de dados ou mantê-los como um arquivo de metadados separado. A abordagem foi avaliada por meio da criação de um *baseline* com o objetivo de determinar a eficácia para a identificação de domínios a partir de conjuntos de dados contidos no diagrama da nuvem LOD. Com os experimentos realizados foi possível obter resultados de precisão consideráveis. Entretanto, os experimentos iniciais não foram conclusivos, pois a classificação de domínio existente na nuvem LOD não cobre toda a gama de informações que realmente são representadas na nuvem. Portanto o sistema será reavaliado como um todo, comparando-o com abordagens alternativas.

[19] http://lod-cloud.net/
[20] http://wiki.dbpedia.org/
[21] http://vocab.deri.ie/void

3.1.2 Roomba: Validação Automática, Correção e Geração de Metadados do Conjunto de Dados

Assaf et al. (2015) propõem o Roomba, uma abordagem automática e escalável para extração, validação, correção e geração de perfil de conjuntos de dados, em que os seus campos são corrigidos automaticamente quando possível. O foco do trabalho é o *metadata profiling*, que pode ser utilizado para prover informações gerais sobre o conjunto de dados (e.g. descrição, publicação e data de atualização), informações legais (e.g. informações de licença) e práticas (e.g. pontos de acesso).

A abordagem foi avaliada com conjuntos de dados abertos, dentre eles a nuvem LOD, em que os metadados são extraídos e, então, é possível identificar os recursos associados ao conjunto de dados. Com os metadados extraídos é realizado seu enriquecimento e a validação de formato. Neste processo, cada campo é verificado e é identificado se a informação não está em falta e se o valor atribuído é válido, podendo ser realizada uma correção automática dos dados. Também são normalizadas informações da licença e são gerados metadados adicionais, como o domínio e o mantenedor, a partir das informações da licença de código-fonte aberto. Após esse processo de correção é gerado um relatório sobre as informações ausentes dos conjuntos de dados, podendo ser enviado ao mantenedor do conjunto de dados. Como resultado os perfis aprimorados são gerados em formato JSON utilizando o modelo de dados CKAN[22].

Nos resultados foi demonstrado que a maioria dos conjuntos de dados possui metadados de má qualidade, o que dificulta a pesquisa sobre os conjuntos de dados. Com uso da abordagem proposta é possível gerar um perfil enriquecido em que os problemas encontrados podem ser corrigidos automaticamente, podendo também melhorar o processo de busca e recuperação de conjuntos de dados.

3.1.3 Uma Abordagem Escalável para Gerar Eficientemente Perfis Estruturados sobre Tópicos de Conjuntos de Dados

Fetahu et al. (2014) propõe uma abordagem para criar automaticamente perfis de conjuntos de dados conectados, com o objetivo de facilitar sua busca e reutilização. Um perfil consiste de metadados estruturados que descrevem tópicos e a relevância de um conjunto de dados por meio de um grafo ponderado de categorias selecionadas do DBpedia. O DBpedia é utilizado como conjunto de dados de referência onde suas entidades e categorias representam instâncias de entidades e tópicos. Os perfis de conjuntos de dados são gerados no formato RDF e expostos

[22] https://ckan.org/

como parte de um catálogo de conjuntos de dados estruturados públicos, baseados no VoID e no Vocabulário de Links[23] (VoL).

A abordagem combina técnicas personalizadas para amostragem de conjuntos de dados e extração de tópicos de conjuntos de dados, através de técnicas de Reconhecimento de Entidade Nomeada (*Named Entity Recognition*[24] - NER) que utilizam conjuntos de dados de referência, e ranking de tópicos relevantes. As principais contribuições consistem em: (i) um método escalável para gerar eficientemente perfis estruturados de conjuntos de dados conectados, combinando e configurando métodos adequados para NER, extração de tópicos e classificação como parte de uma configuração otimizada experimentalmente; e (ii) geração de perfis estruturados para conjuntos de dados da nuvem LOD de acordo com os vocabulários estabelecidos para a descrição de conjunto de dados.

Durante a avaliação experimental, os perfis foram gerados para conjuntos de dados da nuvem LOD. Nos resultados foi encontrado um alto nível de precisão e representatividade dos perfis gerados para a os conjuntos de dados, a abordagem também demonstrou desempenho superior às técnicas de modelagem de tópicos estabelecidas.

3.1.4 Análise Comparativa entre os Trabalhos

Os trabalhos apresentados [Abele (2016); Assaf et al. (2015); Fetahu et al. (2014)] se baseiam no conceito de *Dataset Profiling* com o objetivo de fornecer informações semânticas que possam ser associados aos conjuntos de dados para melhorar sua descrição e facilitar a descoberta.

No trabalho realizado por Abele (2016), são propostas duas abordagens para a geração de metadados que representam corretamente o conteúdo dos conjuntos de dados e para a identificação de conexões entre os conjuntos de dados publicados na nuvem LOD. O trabalho considera apenas os conjuntos de dados já no formato RDF, além disso, para a geração dos metadados são considerados apenas aspectos descritivos e estruturais.

Assaf et al. (2015) propõe o Roomba, uma abordagem automática e escalável para extração, validação, correção e geração de perfil de conjuntos de dados. Entretanto, os metadados que compõe o perfil não são referenciados por vocabulários, e também não são gerados metadados estruturais ou de qualidade.

Fetahu et al. (2014) apresentam uma abordagem que utiliza o DBpedia Spotlight[25] para identificar entidades e categorias presentes em conjuntos de dados, com o objetivo de criar perfis compostos de metadados estruturados para a descrição de tópicos que possam vincular os

[23] http://data.linkededucation.org/vol/
[24] https://msdn.microsoft.com/en-us/library/azure/dn905955.aspx
[25] http://www.dbpedia-spotlight.org/

conjuntos de dados. O trabalho propõe um perfil diferente dos trabalhos anteriores, no qual os conjuntos de dados são agrupados em tópicos para facilitar a recuperação de conjuntos de dados mais relevantes. Logo os metadados descritivos disponibilizados são utilizados para estabelecer a relevância de um conjunto de dados, e não para ajudar aos consumidores de dados no entendimento desses conjuntos de dados.

O Quadro 2 apresenta um resumo das principais características identificadas nos trabalhos relacionados.

Quadro 2 - Quadro comparativo entre os trabalhos relacionados à geração de perfil.

	Objetivo	**Uso de vocabulários ou ontologias**	**Nível de automação**	**Experimentos Realizados**	**Tipo de Metadados Gerados**
Abele (2016)	Fornecer descrições com diferentes níveis de granularidade dos conjuntos de dados existentes.	Sim	Semiautomático	Sim	Metadados descritivos e estruturais
Assaf et al. (2015)	Realizar a extração, validação, correção e geração de metadados descritivos para conjunto de dados abertos.	Não	Automático	Sim	Metadados descritivos e de proveniência
Fetahu et al. (2014)	Descrever automaticamente conjuntos de dados vinculados para facilitar sua busca e reutilização.	Sim	Automático	Sim	Metadados descritivos

Fonte: O Autor (2018)

3.2 ENRIQUECIMENTO SEMÂNTICO

O enriquecimento semântico é uma prática que pode ser realizada para várias finalidades e em diferentes níveis, mas, deve resultar no melhoramento das descrições sobre um conjunto de dados, metadados, informações, notícias, ou outros elementos [Diamantini et al. 2006; Mannens et al. 2009; Fileto et al. 2015; Lóscio et al. 2017]. Diante da diversidade apresentada pelos trabalhos, selecionamos trabalhos semelhantes ao enriquecimento semântico proposto neste trabalho.

3.2.1 Recomendação de Vocabulários

O uso e o reuso de vocabulários tornam possível que o significado pretendido pelo publicador dos dados seja o mesmo que o significado entendido pelo consumidor dos dados,

eliminando ambiguidades em termos utilizados nos diferentes conjuntos de dados. Para os conjuntos de dados que não utilizam vocabulários é possível recomendar vocabulários de domínio que possam ajudar na conversão de vários formatos semiestruturados ou estruturados para o formato RDF, promovendo a reutilização de metadados e integração dos dados. A seguir são apresentados alguns dos trabalhos relacionados à recomendação de vocabulários para referenciamento semântico de conjuntos de dados.

3.2.1.1 Datavore: Uma Ferramenta de Recomendação de Vocabulários que Auxilia na Modelagem de Dados Vinculados

O Datavore [Ellefi et al. 2015] é um sistema de recomendação de vocabulários que utiliza o LOV como mecanismo de busca de vocabulários para auxiliar na geração de metadados. A entrada do Datavore é uma lista de termos extraídos da fonte de dados. Mas, na maioria dos casos, é necessária a realização de uma limpeza na cadeia de caracteres extraída, removendo ou modificando os caracteres indesejados. O LOV foi utilizado como motor de busca de vocabulários, pois, de acordo com os autores, é o único criado especificamente na Web com um índice atualizado.

A partir dos termos identificados, o LOV é consultado, e o resultado é uma lista de conceitos para cada termo de origem classificado pela métrica do LOV. Essa lista é gerada de acordo com a popularidade dos termos do vocabulário nos conjuntos de dados LOD e no ecossistema LOV, também são recuperadas as relações entre conceitos de diferentes listas. Os usuários escolhem os termos para a geração dos metadados baseando-se na lista de recomendação de conceitos, deve-se dar preferência aos vocabulários populares, mas que ao mesmo tempo seja o vocabulário mais apropriado.

3.2.1.2 LOVER: Suporte para Modelagem de Dados Usando *Linked Open Vocabularies*

Schaible et al. (2013) apresentam o LOVER, que tem o objetivo de recomendar classes e propriedades de vocabulários existentes mais utilizados na nuvem LOD para a representação dos dados em RDF.

O LOVER possui um mecanismo iterativo que suporta o aumento da reutilização de vocabulários existentes e uma quantidade adequada de diferentes vocabulários. Esse mecanismo iterativo inclui informações contextuais específicas que adaptam e atualizam as recomendações das classes e propriedades de vocabulários, como os vocabulários já utilizados no modelo,

tornando-os melhor ranqueados em relação aos outros vocabulários recomendados. Foi observado que o uso da abordagem resultou em um efeito positivo na busca por vocabulários.

3.2.2 Identificação do domínio/tema

A identificação do domínio ajuda aos usuários no entendimento sobre o conteúdo dos conjuntos de dados, como também nos mecanismos de busca para o retorno de resultados mais relevantes por meio das consultas e de filtros de pesquisa. Foram selecionados os trabalhos que mais se assemelham com a identificação de domínio apresentada neste estudo.

3.2.2.1 Identificação Automática de Domínio para Dados Abertos Conectados

Nesse trabalho, Lalithsena et al. (2013) apresentam uma abordagem que identifica automaticamente os principais tópicos de domínios de conjuntos de dados LOD, o que pode ajudar aos consumidores de dados a identificar os conjuntos de dados relevantes para um determinado propósito. Para a identificação dos tópicos são extraídas entidades dos conjuntos de dados, e são utilizadas fontes de conhecimento existentes, como a Freebase[26], para a identificação de instâncias correspondentes ou diretamente relacionadas para todas as entidades do conjunto de dados.

Com a identificação das instâncias correspondentes, é possível gerar uma hierarquia de categorias a partir da identificação de suas categorias e domínios correspondentes. Para a identificação das categorias mais significativas, são extraídos os domínios que ocorrem com mais frequência. Os autores consideram que quanto maior a frequência de um termo, maior a evidência de que o termo seja um bom descritor para o conjunto de dados. Isso porque mostra que um grande número de instâncias pode ser descrita pelo termo especificado.

Avaliações realizadas demonstraram que a abordagem fornece uma precisão e cobertura significativamente melhores para recuperação de conjuntos de dados LOD, em comparação com outras abordagens. Esses resultados evidenciaram que a abordagem pode ser útil para categorizar sistematicamente os conjuntos de dados e encontrar conjuntos de dados relevantes na nuvem LOD.

3.2.2.2 Identificação de Tema em Grafos RDF

Ouksili et al. (2014) apresentam uma abordagem que possibilita a identificação de temas para um determinado conjunto de dados RDF, permitindo que usuários identifiquem mais

[26] https://developers.google.com/freebase/

rapidamente os conjuntos de dados que possuem informações relevantes para suas necessidades específicas.

A abordagem proposta utiliza um algoritmo com combinação de critérios estruturais e semânticos para o agrupamento dos grafos, de modo que cada agrupamento (*cluster*) corresponde a um tema que melhor defina sua semântica. O algoritmo de agrupamento tentará identificar áreas altamente conectadas, independente da orientação das arestas, para a identificação da existência de alguma relação semântica entre os recursos. Entretanto, apenas a estrutura do grafo não é suficiente para a identificação de *clusters* significativos. Como os usuários podem ter diferentes pontos de vista e estarem interessados em propriedades distintas sobre um mesmo conjunto de dados, o agrupamento leva em consideração essas propriedades como critérios semânticos.

O objetivo é fornecer ao usuário uma visão do conteúdo do agrupamento a partir da extração de um conjunto de rótulos relevantes que descrevem o tema. O conjunto de rótulos é extraído dos nomes de recursos RDF, sendo composto pelas palavras-chave com alto peso no agrupamento.

3.2.3 Análise Comparativa entre os Trabalhos

Os trabalhos apresentados possuem uma grande característica em comum, o enriquecimento visando melhorar a descrição de fontes de dados em diferentes níveis. Os trabalhos de Ellefi et al. (2015) e Schaible et al. (2013) abordam de forma semelhante o enriquecimento semântico pela recomendação de vocabulários e os trabalhos de Lalithsena et al. (2013) e Ouksili et al. (2014) abordam o enriquecimento semântico pela identificação do domínio/tema.

No trabalho de Ellefi et al. (2015) é proposto um sistema de recomendação de vocabulários que disponibiliza uma lista de conceitos para cada termo de uma fonte de dados classificado pela métrica do LOV. Schaible et al. (2013) possibilitam a recomendação de classes e propriedades de vocabulários existentes por meio de um mecanismo iterativo que inclui informações contextuais específicas que adaptam e atualizam as recomendações das classes e propriedades de vocabulários, como os vocabulários já utilizados no modelo que ficam melhor ranqueados em relação aos outros vocabulários recomendados. O trabalho apresentado nesta dissertação a principio também identifica vocabulários para as propriedades do conjunto de dados, mas, ao contrário destes trabalhos, o que é recomendado ao usuário são os vocabulários de domínio e sem a necessidade de interação com o usuário.

Lalithsena et al. (2013) apresentam uma abordagem que identifica automaticamente os principais tópicos de domínios de conjuntos de dados LOD por meio da identificação de

instâncias correspondentes ou diretamente relacionadas para todas as entidades do conjunto de dados. Ouksili et al. (2014) apresentam uma abordagem que possibilita a identificação de temas para um determinado conjunto de dados RDF. A abordagem proposta utiliza um algoritmo com combinação de critérios estruturais e semânticos para o agrupamento dos grafos, de modo que cada agrupamento (*cluster*) corresponde a um tema que melhor defina sua semântica. Os trabalhos apresentados permitem identificar o tema apenas para conjuntos de dados já disponibilizados no formato RDF, podendo até agrupar os grafos de acordo com os temas identificados, e também não são gerados metadados sobre o domínio do conjunto de dados que possam ser disponibilizados aos usuários.

O Quadro 3 apresenta um resumo das principais características identificadas entres esses trabalhos.

Quadro 3 - Quadro comparativo entre os trabalhos relacionados ao enriquecimento semântico.

	Objetivo	**Uso de vocabulários ou ontologias**	**Nível de automação**	**Geração de Metadados**	**Geração de Perfil**
Ellefi et al. (2015)	Fornecer listas de termos de vocabulários para auxiliar na modelagem de metadados.	Sim	Automático	Metadados estruturais	Não
Schaible et al. (2013)	Recomendar a reutilização de classes e propriedades de vocabulários existentes e ativamente usados na nuvem LOD.	Sim	Semiautomático	Não	Não
Lalithsena et al. (2013)	Identificar os domínios de tópicos dos conjuntos de dados, utilizando fontes de conhecimento em outros conjuntos de dados LOD.	Sim	Automático	Não	Não
Ouksili et al. (2014)	Identificar um conjunto de temas e extrair os rótulos ou tags que melhor capturam a sua semântica.	Não	Semiautomático	Não	Não

Fonte: O Autor (2018)

3.3 QUALIDADE

Os critérios de qualidade devem prover a melhor representação dos aspectos dos dados ou conjunto de dados, possibilitando avaliar sua adequação para uma determinada tarefa. Na literatura existem diversos critérios de qualidade que podem ser utilizados para abranger os mais diferentes aspectos de um conjunto de dados, como exatidão, completude, consistência,

interpretação, objetividade, relevância e acessibilidade [Naumann et al. 2000; Pipino et al. 2002]. Nesse contexto, foram selecionados os trabalhos a seguir.

3.3.1 Geração de um Perfil de Qualidade para Fontes de Dados Dinâmicas

Silva Neto et al. (2016) propõe a criação de um Perfil de Qualidade, composto de um conjunto de metadados sobre a qualidade de uma fonte de dados dinâmica. Este perfil é atualizado periodicamente de acordo com os resultados obtidos pela avaliação contínua da qualidade, tornando possível refletir o aspecto dinâmico das fontes. Para isso, foi especificada uma estratégia em que os critérios de Qualidade da Informação (QI) são avaliados de forma contínua, com o objetivo de acompanhar a evolução e avaliar a qualidade de fontes de dados dinâmicas, que são fontes de dados cujo conteúdo pode sofrer modificações com alta frequência.

O Perfil de Qualidade (Figura 7) se apresenta como um facilitador para os produtores e consumidores de dados, uma vez que o seu uso pode ser aproveitado para diversos fins, inclusive no processo de seleção de fontes de dados. Completude e Corretude foram os critérios de QI considerados relevantes para avaliar a qualidade de uma fonte de dados dinâmica. Esses critérios avaliam especificamente o conteúdo da fonte de dados. A Completude é definida como o grau em que uma fonte de dados não possui dados nulos ou faltantes, e pode ser estimada, por meio das métricas de Densidade e Cobertura. Já a Corretude indica o grau em que os dados contidos em uma fonte de dados estão livres de erro. Para sua avaliação foram utilizadas regras de validação, que são definidas de acordo com o domínio dos dados e que podem ser usadas para uma avaliação automática da corretude. O processo de geração do perfil de qualidade consiste de duas etapas:

- **Avaliação da Qualidade:** as fontes de dados são avaliadas de forma contínua com base nos critérios de qualidade definidos;
- **Atualização dos Metadados:** os valores dos critérios de qualidade são calculados e anotados no Perfil de Qualidade que será disponibilizado junto com a fonte de dados.

Figura 7 - Exemplo do Perfil de Qualidade em JSON.

```
{
   "url_fonte": "http://fontededados.com.br/acesso",
   "ultima_atualizacao": "20/05/2016 17:00:01",
   "geracao_perfil": "20/05/2016 17:15:06",
   "volume_dados": 12.395,
   "criterios_qi": [{
        "nome_criterio": "Completude",
        "valor_criterio": 0.89
   }, {
        "nome_criterio": "Corretude",
        "valor_criterio": 0.95
   }],
   "medida_global": 0.92
}
```

Fonte: Silva Neto et al. (2016).

Os experimentos foram realizados utilizando fontes de dados do domínio Meteorológico e demonstraram que a estratégia de avaliação proposta produz resultados satisfatórios. Por meio dos experimentos realizados foi possível confirmar que ao longo do tempo a qualidade das fontes de dados se altera em decorrência das modificações que ocorrem em seu comportamento. Foi observado que a estratégia de avaliação contínua se mostrou mais eficiente, com boa precisão e custo, quando comparada com outras estratégias avaliadas.

3.3.2 Um Framework de Avaliação Objetiva e Ferramenta para Dados Conectados: Enriquecendo Perfis de Conjuntos de Dados com Indicadores de Qualidade

Neste trabalho, Assaf et. al (2016) primeiramente propõem um quadro de avaliação para a qualidade das fontes de dados conectados. Para isso, foram utilizados princípios de qualidade de dados descritos em Assaf et. al (2012) e pesquisados no trabalho de Zaveri et. al (2013), e então foram agrupados atributos para avaliação de qualidade.

Os atributos selecionados resultaram nas seguintes medidas de qualidade: completude, disponibilidade, exatidão, consistência, modificação, proveniência, licenciamento, compreensão, coerência e segurança. Para avaliar essas medidas os autores identificaram um total de 64 indicadores de qualidade, que podem pertencer às seguintes categorias: entidade, conjunto de dados, links e modelos.

Uma grande parte dos indicadores pode ser examinada automaticamente a partir de metadados de conjuntos de dados encontrados em portais de dados. Onde, 30 indicadores de qualidade estão relacionados ao conjunto de dados e qualidade dos links. Os outros 34 indicadores estão relacionados à qualidade das entidades e modelos, e não podem ser verificados através dos metadados.

Os autores também apresentam uma ferramenta que mede a qualidade dos conjuntos de dados, e que ajuda aos proprietários a avaliar seus conjuntos de dados e aos consumidores para escolher fontes de dados. Essa ferramenta de medição de qualidade foi desenvolvida como uma extensão do Roomba [Assaf et al. 2015], que é uma abordagem para avaliar e criar perfis de conjuntos de dados (apresentada na Seção 3.1.2), permitindo medir a qualidade dos conjuntos de dados na nuvem LOD. O Roomba foi ampliado com 7 submódulos que verificam cerca de 46 indicadores de qualidade de dados. Como resultado, a ferramenta gera um relatório sobre a qualidade do conjunto de dados. Roomba é uma abordagem executada em portais de dados baseados em CKAN, logo os autores também construíram a extensão de qualidade baseada no modelo CKAN.

A ferramenta foi avaliada com conjuntos de dados da nuvem LOD. Os resultados demonstram que o estado dos conjuntos de dados da nuvem LOD precisa de mais atenção, pois a maioria deles apresentaram baixos índices de qualidade, proveniência, licenciamento e qualidade de compreensão.

3.3.3 Análise Comparativa entre os Trabalhos

Os trabalhos de Silva Neto et al. (2016) e Assaf et al. (2016) avaliam a qualidade de fontes de dados com diferentes metodologias, buscando facilitar a avaliação dessas fontes para uma determinada finalidade. Silva Neto et al. (2016) propõe um perfil de qualidade que apresenta metadados de qualidade em formato JSON, deixando a critério do usuário a comparação entre as fontes de dados. Entretanto, esses metadados de qualidade se referem aos dados e não a um conjunto de dados, além disso, o foco do trabalho é a geração de metadados de qualidade para fontes de dados dinâmicas e os metadados disponibilizados não são referenciados semanticamente por vocabulários. Já Assaf et al. (2016) desenvolveram um framework para avaliar a qualidade de conjuntos de dados conectados por meio de critérios e indicadores de qualidade. Apesar do framework de qualidade estar relacionado à geração de um perfil, os resultados obtidos em relação à qualidade do conjunto de dados não são inseridos no perfil, eles fazem parte de um relatório que reflete a qualidade do conjunto de dados. O Quadro 4 apresenta um resumo das principais características identificadas entres esses trabalhos.

Quadro 4 - Quadro comparativo entre os trabalhos relacionados à qualidade.

	Objetivo	**Uso de vocabulários ou ontologias**	**Nível de automação**	**Geração de Metadados**	**Geração de Perfil**
Silva Neto et al. (2016)	Geração de um Perfil de Qualidade para fontes de dados dinâmicas, o qual será gerado e atualizado de acordo com uma avaliação contínua da qualidade das fontes de dados.	Não	Automático	Metadados de qualidade	Sim
Assaf et al. (2016)	Avaliar a qualidade de conjuntos de dados conectados para enriquecer o processo de geração de perfil de conjuntos de dados.	Não	Automático	Não	Não

Fonte: O Autor (2018)

3.4 CONSIDERAÇÕES

Este capítulo apresentou os trabalhos relacionados ao tema desta dissertação. Ao final de cada seção foi realizada uma breve discussão entre os trabalhos, com uma análise comparativa entre os trabalhos, agrupando as principais características de cada abordagem em um quadro comparativo. Foram levantados os pontos mais relevantes de cada trabalho com o objetivo de identificar o que já existe relacionado ao tema central desta pesquisa. Nenhum dos trabalhos apresenta uma proposta de um perfil de conjuntos de dados que inclua metadados descritivos, estruturais e de qualidade enriquecidos semanticamente para conjuntos de dados que estejam em qualquer formato. O próximo capítulo apresenta a proposta central desta dissertação.

4

DSPRO+ - UMA ABORDAGEM PARA GERAÇÃO DE PERFIL DE CONJUNTO DE DADOS

Neste capítulo é apresentada a DSPro+ - uma abordagem para geração de Perfil de Conjunto de Dados. Primeiramente, a Seção 4.1 introduz a definição do problema. A Seção 4.2 apresenta algumas definições preliminares acerca dos principais conceitos relacionados à geração do Perfil de Conjunto de Dados. Na Seção 4.3 é descrita uma visão geral das etapas que compõem o processo da abordagem. A extração de informações é descrita na Seção 4.3.1. O processo de geração dos metadados é detalhado na Seção 4.3.2. A Seção 4.3.3 especifica a formação do PCD, onde são descritos os metadados que compõem o perfil. Na Seção 4.4 é apresentado um exemplo para facilitar o entendimento da abordagem DSPro+. Na Seção 4.5 é realizada uma comparação entre este trabalho e os trabalhos relacionados, apresentados no capítulo anterior. Ao fim deste capítulo, na Seção 4.6 são apresentadas algumas considerações.

4.1 DEFINIÇÃO DO PROBLEMA

A Web atualmente representa um ambiente propício à publicação e consumo de conjuntos de dados. Nesse panorama, onde é importante fornecer informações sobre os conjuntos de dados, a disponibilização de metadados pode ajudar. Entretanto, geralmente os conjuntos de dados publicados na Web são disponibilizados sem metadados suficientes para descrever seus dados. Além disso, quando disponíveis, em geral, não estão organizados de forma estruturada e não são referenciados semanticamente por vocabulários, o que torna mais difícil o entendimento do conteúdo do conjunto de dados e seu processamento automático. Diante disso, é relevante gerar metadados que forneçam informações que contribuam para o entendimento e uso dos dados. Esses metadados podem ser enriquecidos semanticamente, o que permite melhorar as descrições do conjunto de dados, complementando informações que facilitem sua compreensão e seu processamento. Nesse contexto, o problema tratado nesta dissertação é definido como segue:

Dado um ecossistema de produção e consumo de conjuntos de dados na Web, como gerar metadados enriquecidos semanticamente que forneçam informações descritivas sobre o

conteúdo, estrutura e qualidade dos conjuntos de dados, a fim de facilitar a compreensão e o processamento dos dados por consumidores?

Como meio para geração de metadados, este trabalho propõe a abordagem denominada DSPro+ (*Enriched DataSet Profile*) para criação de perfil que descreve um determinado conjunto de dados. O perfil é composto por metadados descritivos, estruturais e de qualidade enriquecidos semanticamente. O enriquecimento semântico permite complementar informações sobre os conjuntos de dados, podendo facilitar sua localização e acesso por mecanismos de busca e aplicações. Os metadados são disponibilizados e referenciados a partir de termos de vocabulários existentes, sendo também representados em formato compreensível por máquina, o que resulta na geração de descrições mais significativas e facilita o processamento e manipulação dos dados. Prover um perfil de conjunto de dados contribui para comunicação entre publicadores e consumidores de dados e o uso desses conjuntos de dados. A seguir, são apresentadas algumas definições preliminares, incluindo o conceito de Perfil de Conjunto de Dados.

4.2 DEFINIÇÕES PRELIMINARES

Nesta seção, são fornecidas algumas definições preliminares para auxiliar no entendimento da abordagem proposta.

Definição 1 Conjunto de Dados (ɗ) **-** Um conjunto de dados ɗ representa uma coleção de dados publicados na Web que está disponível para acesso por meio de uma ou mais distribuições.

Definição 2 Indicador de Qualidade ($IQ(ɗ)_n$) - Um indicador de qualidade $IQ(ɗ)_n$ representa uma característica mensurável de um conjunto de dados ɗ que está relacionada à qualidade de seus dados ou metadados.

Definição 3 Critério de Qualidade ($CQ(ɗ)$) - Um critério de qualidade $CQ(ɗ)$ permite avaliar e medir um aspecto específico da qualidade de um conjunto de dados ɗ que está associado à um conjunto de $IQ(ɗ)$**,** tal que $IQ(ɗ)$= $\{IQ(ɗ)_{CQ1}, IQ(ɗ)_{CQ2}, \dots, IQ(ɗ)_{CQn}\}$.

Definição 4 Métrica de Qualidade (**MtQ**(ɗ)) **–** Uma métrica de qualidade MtQ(ɗ) representa o valor associado a um determinado $CQ(ɗ)$, que é estabelecido por meio do cômputo de seu conjunto de $IQ(ɗ)$.

Definição 5 Metadados Descritivos (MD(ɗ)) - Metadados descritivos MD(ɗ) proveem informações sobre características e conteúdo de um conjunto de dados ɗ.

Definição 6 Metadados Estruturais (ME(ɗ)) - Metadados estruturais ME(ɗ) disponibilizam descrições a respeito do esquema estrutural que compõe um conjunto de dados ɗ.

Definição 7 Metadados de Qualidade (MQ(ɗ)) - Metadados de qualidade MQ(ɗ) indicam resultados de MtQ(ɗ) utilizadas para avaliação de $CQ(ɗ)$.

Definição 8 Perfil de Conjunto de Dados (PCD(ɗ)) **-** Um Perfil de Conjunto de Dados PCD(ɗ) é composto por metadados descritivos, estruturais e de qualidade enriquecidos semanticamente e referentes a um determinado conjunto de dados ɗ publicado na Web, definido como PCD(ɗ) = {MD(ɗ) + ME(ɗ) + MQ(ɗ)}.

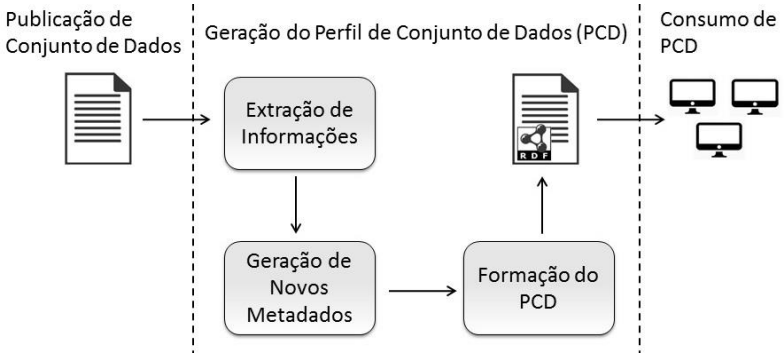
4.3 VISÃO GERAL DA ABORDAGEM DSPRO+

A geração do PCD(ɗ) permite atribuir um melhor entendimento acerca de um conjunto de dados e de seu conteúdo, tornando-o mais acessível e compreensível tanto por pessoas quanto por máquinas. A abordagem proposta é composta de algumas etapas, como a extração de informações, geração de novos metadados descritivos, estruturais e de qualidade, e a formação do PCD. Ao representar os metadados no modelo de dados RDF, é possível atribuir informações e significado, permitindo que esses metadados sejam referenciados semanticamente por termos de vocabulários e por meio de triplas (s + p + o), onde:

- **"s" (sujeito):** faz referência a um recurso, como um conjunto de dados que é representado por um identificador (e.g., "dataset-1336");
- **"p" (predicado):** permite relacionar o sujeito ao predicado. Neste caso, recebe como valor um metadado proveniente ou gerado a partir do conjunto de dados, que deve ser referenciado por um vocabulário. Por exemplo, o metadado "Domínio", que pode ser referenciado pelo termo *dcat:theme*.
- **"o" (objeto):** corresponde ao valor que está associado a um predicado do sujeito, podendo ser um recurso ou um literal. Por exemplo, o predicado *dcat:theme* recebe como valor o literal "esportes".

A Figura 8 ilustra a abordagem proposta para a geração do PCD(ɗ), com as principais etapas que a compõem, até a sua disponibilização para o consumo.

Figura 8 - Abordagem para geração do PCD(ɗ) - DSPro+



Fonte: O Autor (2018)

No ecossistema no qual a abordagem está inserida existem diferentes atores, podendo ter papéis de publicadores, relacionados à publicação, e/ou de consumidores, relacionados ao consumo dos conjuntos de dados publicados. Os publicadores ou consumidores de dados podem criar um novo PCD a partir de um conjunto de dados existente. O processo de geração do perfil recebe como entrada um conjunto de dados publicado na Web, que pode ser acessado por meio de sua URL, permitindo a extração de informações que descrevem o conjunto de dados, por exemplo, informações sobre sua estrutura e metadados existentes.

A extração de informações, descrita na Seção 4.3.1, é uma das etapas mais importantes da geração do perfil, pois permite o acesso às informações necessárias para a composição dos metadados descritivos, estruturais e de qualidade, assim como a identificação de metadados já disponibilizados no conjunto de dados, que serão enriquecidos com termos de vocabulários e disponibilizados no PCD.

Na etapa de geração de novos metadados (Seção 4.3.2) são gerados metadados descritivos (domínio e vocabulário de domínio), estruturais e de qualidade. Após a geração dos metadados, é iniciada a etapa de formação do PCD, descrita na Seção 4.3.3, onde os metadados são referenciados por termos de vocabulários e organizados de forma estruturada em sintaxe RDF, resultando em um PCD composto por metadados enriquecidos semanticamente. A seguir são descritas cada uma dessas etapas.

4.3.1 Extração de Informações

Nesta etapa, a partir da URL do conjunto de dados são extraídas informações que permitem disponibilizar os metadados que compõem os metadados descritivos, estruturais e de qualidade do PCD.

Entre os metadados $MD(d')$ escolhidos para a composição do perfil (Quadro 5), alguns fazem parte do esquema proposto no documento de Melhores Práticas para Publicação e Consumo de Dados na Web disponibilizado pelo W3C [Lóscio et al., 2017]. Sua estrutura e os metadados recomendados são utilizados neste trabalho como modelo para representação de metadados descritivos. Entre onze metadados descritivos recomendados, foram selecionados os metadados mais comuns e de acordo com a proposta deste trabalho: título, descrição, palavras-chave, data da última modificação, data de publicação, publicador, domínio (tema) e distribuição.

Quadro 5 - MD(d)

Metadados Descritivos	Descrição	Tipo de dado
Identificador	Identificador do conjunto de dados	Literal
Título	Título atribuído ao conjunto de dados	Literal
Descrição	Descreve brevemente o conjunto de dados	Texto
Palavras-chave	Termos que podem ser usados para buscas pelo conjunto de dados	Array
Domínio	Tema ou domínio de conhecimento ao qual o conjunto de dados está associado	Array
Vocabulário de domínio	Vocabulário recomendado que pode ajudar na conversão do conjunto de dados.	URI
Endereço	Endereço web em que o conjunto de dados está disponível	URL
Data de modificação	Última data em que o conjunto de dados foi atualizado	Data
Data de publicação	Data em que o conjunto de dados foi disponibilizado para os usuários	Data
Publicador	Nome do responsável por publicar o conjunto de dados	Literal
Versão	Versão atual do conjunto de dados	Numeral
Distribuição	Distribuição do conjunto de dados	Array
Data de criação do PCD	Data em o PCD foi gerado pela abordagem DSPro+	Data

Fonte: O Autor (2018)

Alguns desses metadados descritivos são extraídos do conjunto de dados. Dentre eles, é possível extrair o identificador, título, palavras-chave, descrição, data da última atualização, data de publicação, publicador, versão e distribuição. Quando não são disponibilizadas palavras-chave, e como são utilizadas para a geração de novos metadados descritivos, é possível identificar os termos mais frequentes do conjunto de dados utilizando o TF-IDF (Term Frequency–Inverse Document Frequency). Isso pode ser realizado por meio de ferramentas de indexação, e.g., Apache Solr[27], que disponibilizam uma função correspondente ao TF-IDF, permitindo o cálculo do peso dos termos de um documento.

Outros metadados descritivos também são gerados a partir das informações disponibilizadas pelos conjuntos de dados, que é o caso dos metadados descritivos referentes ao

[27] http://lucene.apache.org/solr/

domínio do conjunto de dados e aos vocabulários de domínio recomendados, apresentados na Seção 4.3.2.1.

Para os metadados estruturais, primeiramente é verificado se eles são disponibilizados pelo conjunto de dados, o que permite sua extração e inclusão no PCD, mas, caso não estejam disponíveis, esses metadados são identificados no esquema estrutural do conjunto de dados, conforme apresentado na Seção 4.3.2.2.

Para os metadados de qualidade, são extraídas informações correspondentes aos indicadores de qualidade que compõem os critérios de qualidade. Com base nesses indicadores os metadados de qualidade são gerados (vide Seção 4.3.2.3).

4.3.2 Geração de Metadados

A etapa de geração de metadados é composta de três fases, onde cada uma delas resulta na geração de um tipo de metadado. São elas: (i) geração de metadados descritivos; (ii) geração de metadados estruturais e (iii) geração de metadados de qualidade.

4.3.2.1 Metadados Descritivos

A geração dos metadados descritivos se refere à identificação do domínio do conjunto de dados e a recomendação de vocabulários de domínio.

Identificação do Domínio

Durante a etapa de identificação do domínio ao qual o conjunto de dados pertence, seu resultado é incluído no PCD por meio do metadado referente ao domínio do conjunto de dados. Sendo assim, a identificação do domínio do conjunto de dados é definida como segue.

Definição 9 Identificação do Domínio (*IdD(*ɗ)) - É uma forma de enriquecimento semântico que resulta na identificação do domínio de conhecimento ao qual o conjunto de dados ɗ pertence.

Para *IdD(*ɗ) deve ser utilizada uma referência semântica. Uma boa opção se refere à Ontologia do DBpedia[28], que disponibiliza informações sobre uma grande quantidade de domínios de conhecimento. A partir das palavras-chave do conjunto de dados são realizadas consultas sobre o SPARQL *endpoint*[29] da ontologia para a identificação das classes ou propriedades que correspondem a cada uma das palavras. Os resultados dessas consultas também consideram a hierarquia das classes retornadas. Entre esses resultados, as classes que apresentam em seu nome, representado pelo metadado *rdfs:label*, a correspondência exata a uma determinada palavra-chave, recebe uma melhor colocação entre os resultados. Como domínio, é

[28] http://wiki.dbpedia.org/services-resources/ontology
[29] https://dbpedia.org/sparql

retornada a classe que apareceu mais vezes entre todos os resultados das consultas. Por exemplo, um conjunto de dados que disponibiliza as palavras-chave "*food"*, "*drink*", "*culture*" e "*alcohol*", retorna como domínio a classe do DBpedia Food[30], que define como alimento qualquer substância comestível ou potável que é normalmente consumida por humanos. Para a etapa de identificação do domínio ao qual o conjunto de dados pertence foi implementado o Algoritmo 1.

Algoritmo 1: Identificação do Domínio

Entrada: palavras-chave de ɗ
Saída: *IdD(*ɗ)
Início
1: **If (***palavras-chave não nulo***) Faça**
2: **While(***temProx(palavras-chave)***) Faça**
3: palavra = *próximo.palavras-chave;*
4: Lista_R*esultados = Resultados _Ontologia(Exec_Consulta(palavra));*
5: **fim While;**
6: **While (***temProx(*Lista_R*esultados)***) Faça**
7: classe_prop = *próximo.*Lista_R*esultados*;
8: Lista_Res_H = R*esultados_Hierarquia(Exec_Cons_Hierarquia(*classe_prop*));*
9: **While (***temProx(*Lista_R*es_H)***) Faça**
10: classe = *próximo.*Lista_R*es_H;*
11: *AddResultados(Map_Dominios,* classe*);*
12: **fim While;**
13: **fim While;**
14: *Map_Domínios.ordenar();*
15: **fim If;**
16: **retorna** *(maior_Resultado(Map_Domínios());*
Fim

O Algoritmo 1 recebe como parâmetro as palavras-chave do conjunto de dados, que são utilizadas em consultas realizadas sobre o SPARQL *endpoint* do DBpedia para a obtenção de informações de domínios de conhecimento (linhas 2-5). Nesta consulta, mostrada na Figura 9, são identificadas as subclasses ou propriedades da ontologia do DBpedia que faz referência a uma dada palavra-chave. Os resultados das consultas de cada palavra-chave são armazenados em uma lista (linha 4).

[30] http://mappings.dbpedia.org/server/ontology/classes/Food

Figura 9 - Consulta SPARQL para recuperação de subclasses e propriedades do DBpedia.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?typeClass ?labelOfMatch ?masterClassOfTypeClass WHERE {
    {
        ?typeClass a owl:Class .
        ?typeClass rdfs:label ?labelOfMatch .
        ?typeClass  rdfs:subClassOf ?masterClassOfTypeClass .
        FILTER(CONTAINS (str(?masterClassOfTypeClass), "dbpedia.org")
                        && LANGMATCHES(lang(?labelOfMatch), "EN"))
        FILTER(CONTAINS (lcase(str(?labelOfMatch)), lcase("word")))
    }UNION{
        ?masterClassOfTypeClass a owl:Class .
        ?masterClassOfTypeClass  rdfs:subClassOf ?masterClassOfTypeClass2 .
        ?typeClass rdfs:domain ?masterClassOfTypeClass .
        ?typeClass rdfs:label ?labelOfMatch .
        FILTER(CONTAINS(str(?masterClassOfTypeClass2), "dbpedia.org")
                        && LANGMATCHES (lang(?labelOfMatch),"EN"))
        FILTER(CONTAINS (lcase(str(?labelOfMatch)), lcase("word")))
    }
}
```

Fonte: O Autor (2018)

Em seguida (linhas 6-13), todos os resultados identificados passam por um processamento que permite percorrer a hierarquia das subclasses ou propriedades encontrada nos resultados. Para isso é executada outra consulta SPARQL (Figura 10) no mesmo *endpoint* utilizado anteriormente. Essa consulta retorna a classe ao qual a subclasse ou propriedade pertence. A classe retornada por essa consulta é inserida em um *array* que contabiliza a quantidade de vezes que cada classe apareceu entre os resultados (linha 11). Nesse caso, as classes em que seu metadado *rdfs:label* corresponde exatamente a uma determinada palavra-chave recebe uma melhor colocação entre os resultados. Como resultado é retornado a classe que apresentou maior número de ocorrências entre os resultados das consultas. Nos casos em que mais de uma classe apresente o valor máximo de ocorrências, pode ser retornada como domínio do conjunto de dados mais de uma classe (linha 14 – 16).

Figura 10 - Consulta SPARQL para recuperação da classe ao qual uma subclasse ou propriedade pertence.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?type ?label ?subpropof {
    ?type a owl:Class .
    ?type rdfs:label ?label .
    ?type  rdfs:subClassOf ?subpropof.
    FILTER(REGEX(str(?type), "classURI") && LANGMATCHES (lang(?label),"EN"))
    FILTER(REGEX(str(?subpropof), "dbpedia.org") || REGEX(str(?subpropof), "owl#"))
}
```

Fonte: O Autor (2018)

Recomendação de Vocabulários

Nesta etapa são recomendados vocabulários de domínio para os metadados de um conjunto de dados de forma geral. Seu resultado é incluído no PCD por meio do metadado referente ao vocabulário de domínio. Sendo assim, a recomendação de vocabulários de domínio é definida como segue.

Definição 10 Recomendação de Vocabulários de Domínio (*RVD(*ɗ)) – É uma forma de enriquecimento semântico que resulta na sugestão de vocabulários de domínio para um conjunto de dados ɗ.

Para gerar *RVD(*ɗ) utiliza-se como fonte de dados um repositório de vocabulários. Uma alternativa de repositório a ser utilizado é o repositório de vocabulários abertos LOV, que possibilita o acesso aos vocabulários e suas descrições semânticas. Utilizando as palavras-chave e propriedades que compõem a estrutura do conjunto de dados, são realizadas consultas no SPARQL *endpoint*[31] do repositório, o que permite a recuperação da URI de vocabulários relacionados a essas palavras ou propriedades. São selecionados os vocabulários que apresentam maior número de ocorrências entre os resultados e que, de preferência, estejam ativos. Os vocabulários ativos são verificados por meio de uma consulta SPARQL no mesmo *endpoint* indicado, retornando um valor booleano referente à disponibilidade do vocabulário. Estes vocabulários recomendados podem ajudar na conversão de um conjunto de dados ɗ de formatos semiestruturados ou estruturados para o modelo RDF.

Para esta etapa de recomendação de vocabulários de domínio foi desenvolvido o Algoritmo 2. Este algoritmo recebe como parâmetros as palavras-chave e propriedades que compõem a estrutura do conjunto de dados. Essas palavras-chave e propriedades são inseridas em uma lista de termos (linha 1). Cada termo é inserido em consultas escritas na linguagem SPARQL para a recuperação de informações contidas no SPARQL *endpoint* do LOV (linha 4-8). Para cada propriedade é utilizada uma consulta, apresentada na Figura 11, que retorna a URI de vocabulários de domínio candidatos que apresentem classes ou propriedades equivalentes às propriedades selecionadas do conjunto de dados. Para cada palavra-chave é utilizada outra consulta (Figura 12), que retorna a URI de vocabulários que estão relacionados à palavra-chave. Cada consulta prioriza em seus resultados os vocabulários que são mais reutilizados por outros vocabulários, sendo ordenados de forma decrescente e retornando até seis vocabulários por consulta.

[31] http://lov.okfn.org/dataset/lov/sparql

Algoritmo 2: Recomendação de Vocabulários

Entrada: palavras-chave e propriedades de ɗ
Saída: *RVD(*ɗ)
Início

```
1:  termos = palavrasChave + propriedades;
2:  If (termos não nulo) Faça
3:     While (temProx(termos)) Faça
4:        If(ÉpalavraChave(proximo.termos)) Faça
5:           consultaSPARQL = Gerar_Consulta_palavarsChave(proximo.termos);
6:        Else
7:           consultaSPARQL = Gerar_Consulta_propriedades(proximo.termos);
8:        fim If;
9:        Resultados = add(Exec_Consulta(consultaSPARQL));
10:       AddResultados(Map_Vocabulários, Resultados);
11:    fim While;
12:    While (temProx(Map_Vocabulários)) Faça
13:       vocabulario = próximo.Map_Vocabulários;
14:       If(Exec_Consulta_VocAtivo(vocabulario)) Faça
15:          valor = Map_Vocabulários.get(vocabulario);
16:          Map_Vocabulários.put(vocabulario, ++valor);
17:       fim If;
18:    fim While;
19:    Map_Vocabulários.ordenar();
20: fim If;
21: retorna (maior_Resultado(Map_Vocabulários);
Fim
```

Figura 11 - Consulta SPARQL para recuperação de vocabulários relacionados às propriedades que compõem a estrutura do conjunto de dados

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

SELECT DISTINCT ?vocab {
    ?x rdfs:isDefinedBy ?vocab.
    {   ?classORprop rdfs:label ?cORpLabel.
        {  SELECT  DISTINCT ?vocab ?reused {
                ?vocab  voaf:reusedByVocabularies ?reused
            }GROUP BY ?vocab ?reused HAVING((MIN(?reused) >= 0))
         }
    } UNION{
        ?classORprop rdfs:label ?cORpLabel
    }
    FILTER (CONTAINS(lcase(str(?cORpLabel)), lcase("property")))
}ORDER BY DESC (?reused) LIMIT 6
```

Fonte: O Autor (2018)

Figura 12 - Consulta SPARQL para recuperação de vocabulários relacionados às palavras-chave do conjunto de dados

```
PREFIX voaf: <http://purl.org/vocommons/voaf#>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT  ?vocab {
    ?vocab  a voaf:Vocabulary.
    {    ?vocab dcterms:title ?vLabel.
    }UNION{
        ?vocab dcterms:title?vLabel.
        {   SELECT DISTINCT  ?vocab ?reused {
                ?vocab voaf:reusedByVocabularies ?reused
            }GROUP BY ?vocab ?reused HAVING((MIN(?reused) >= 0))
        }
    }
    FILTER (CONTAINS(lcase(str(?vLabel)), lcase("word")))
}ORDER BY DESC (?reused) LIMIT 6
```

Fonte: O Autor (2018)

Cada consulta é executada (linha 9) e seu resultado é adicionado em um *array* que contabiliza a quantidade de vezes que cada URI apareceu entre os resultados (linha 10). Entre os vocabulários recomendados são priorizados os vocabulários que possuírem o maior número de ocorrências entre os resultados e que, de preferência, estejam ativos (linha 12 – 18). Os vocabulários ativos são verificados por meio de uma consulta SPARQL (Figura 13) que também é executada no SPARQL *endpoint* do LOV, a qual retorna um valor booleano referente à disponibilidade do vocabulário (se ele está ativo). Caso seja retornado um valor positivo, esse vocabulário recebe um peso maior, que melhora sua colocação entre os resultados (linha 16). Por fim, os resultados são ordenados (linha 19) e são retornados um ou mais vocabulários que apresentem o valor máximo de ocorrências entre os resultados das consultas (linha 21).

Figura 13 - Consulta SPARQL para verificar vocabulários ativos

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

ASK {
    ?ind rdf:type ?class.
    ?prop rdfs:domain ?class.
    ?prop rdfs:isDefinedBy ?vocab.
    FILTER (CONTAINS(str(?vocab), str("vocabURI")))
}
```

Fonte: O Autor (2018)

4.3.2.2 Metadados Estruturais

Os ME(d) permitem uma visão sobre o esquema estrutural dos conjuntos de dados, facilitando o entendimento de seu esquema e o processamento dos dados por aplicações de

consumo. No PCD(d) os ME(ɗ) expressam as propriedades que compõem a estrutura de ɗ e são organizados por meio de um esquema específico. Nesse esquema é indicada a quantidade de propriedades apresentadas no conjunto de dados e para cada propriedade é especificado o seu nome e o tipo de dado que recebe, conforme descrito no Quadro 6.

Quadro 6 - ME(ɗ)

Metadados Estruturais	**Descrição**	**Tipo de dado**
Quantidade de Propriedades	Quantidade de propriedades que compõem a estrutura do conjunto de dados	Numeral
Propriedade	Lista que apresenta cada propriedade, incluído seu Nome e o Tipo	Array
Propriedade – Nome	Nome da propriedade no conjunto de dados	Literal
Propriedade – Tipo	Tipo de dado que a propriedade recebe no conjunto de dados	Literal

Fonte: O Autor (2018)

Os valores recebidos por esses metadados podem ser identificados a partir do conjunto de dados. De onde são extraídas todas as informações necessárias para compor os metadados estruturais, como o nome de cada propriedade que compõe a estrutura interna do conjunto de dados, o tipo de dado que a propriedade recebe e a quantidade de propriedades que o conjunto de dados apresenta.

4.3.2.3 Metadados de Qualidade

A qualidade da informação é um dos aspectos mais importantes para os consumidores e usuários da internet [Naumann et al. 2000]. A qualidade de um conjunto de dados pode ser avaliada por meio de $CQ(ɗ)$, o que permite avaliar a adequação do conjunto de dados para uma determinada tarefa. Para isso são definidos $IQ(ɗ)$, onde cada $IQ(ɗ)$ está relacionado a um $CQ(ɗ)$.

Critérios de Qualidade e Métricas de Avaliação

Existem diversos critérios de qualidade que podem ser utilizados para abranger os mais diferentes aspectos de um conjunto de dados, como acessibilidade, completude, segurança e compreensibilidade [Wang et al. 1996; Naumann et al. 2000; Pipino et al. 2002]. Para identificar critérios de qualidade a serem avaliados para conjuntos de dados, também foram considerados os principais benefícios que podem ser obtidos com a aplicação das Melhores Práticas para Publicação e Consumo de Dados na Web [Lóscio *et al.*, 2017] que são: compreensão,

processabilidade, descoberta, reuso, confiança, capacidade de ligação, acesso, e interoperabilidade.

Ao considerar como base esses critérios de qualidade e as características dos conjuntos de dados, para a composição de MQ(ɗ), foram considerados os critérios Compreensibilidade e Processabilidade. Estes critérios foram selecionados por representarem aspectos relacionados ao consumo do conjunto de dados que pode ocorrer tanto por humanos quanto por aplicações que compartilham, manipulam e reutilizam esses dados.

A disponibilização de metadados de qualidade que representem a Compreensibilidade e a Processabilidade ajudará aos consumidores a selecionar conjuntos de dados que estejam mais apropriados para o consumo, como também pode indicar aos publicadores a necessidade de aprimoramento do conjunto de dados. A seguir, serão descritos os critérios selecionados, especificando as métricas aplicadas para a composição de cada um.

Compreensibilidade

A compreensibilidade considera alguns indicadores de qualidade que facilitam o entendimento do conjunto de dados. Ao seguir as recomendações disponibilizadas pelo documento de Melhores Práticas para Publicação e Consumo de Dados na Web do W3C, como fornecer metadados descritivos e reutilizar vocabulários de preferência recomendados, é possível que os seres humanos tenham uma melhor compreensão sobre a natureza do conjunto de dados, sua estrutura e significado dos dados. Como resultado, surge o critério denominado de Compreensibilidade, definido a seguir.

Definição 11 Compreensibilidade ($C(ɗ)$) - A compreensibilidade do conjunto de dados ɗ, denotada por $C(ɗ)$, é representada pelo grau em que ɗ apresenta informações que promovam ou facilitem seu entendimento por parte de usuários humanos. $C(ɗ)$ considera um conjunto de indicadores de qualidade $IQ(ɗ) = \{IQ(ɗ)_{C1}, IQ(ɗ)_{C2}, \dots, IQ(ɗ)_{C6}\}$.

A compreensibilidade $C(ɗ)$ é avaliada pelo somatório de todos os valores do conjunto $IQ(ɗ) \in C(ɗ)$, dividido pela quantidade de indicadores de qualidade contidos em $IQ(ɗ)$, representado por $\#IQ(ɗ)$, conforme apresentado a seguir:

$$C(ɗ) = \frac{\sum_{n=1}^{6} IQ(ɗ)_{Cn}}{\#IQ(ɗ)}$$

Onde,

$IQ(ɗ)_{Cn}$ é o valor do indicador de qualidade Cn, pertencente ao conjunto $IQ(ɗ) = \{IQ(ɗ)_{C1}, \dots, IQ(ɗ)_{C6}\}$;

$\#IQ(ɗ)$ é a quantidade de indicadores de qualidade associados a $C(ɗ)$.

O critério da compreensibilidade $C(ɗ)$ é composto por um conjunto de seis indicadores $IQ(ɗ)$, apresentados a seguir:

$IQ(ɗ)_{C1}$: Metadados descritivos

$$IQ(ɗ)_{C1} = \frac{\#\mathrm{MD}(ɗ)_D}{\#\mathrm{MD}_D}$$

Onde,

$\#\mathrm{MD}(ɗ)_D$ é a quantidade de metadados descritivos desejáveis (MD_D) encontrados no conjunto de dados ɗ;

$\#\mathrm{MD}_D$ é a quantidade de metadados descritivos desejáveis (MD_D), considerados suficientes para descrever os conjuntos de dados.

Para o conjunto de dados ɗ deve ser disponibilizada uma certa quantidade de metadados descritivos (MD) que permitam descrever suas características gerais, de forma que possa melhorar sua compreensão e descoberta automática. Baseando-se nessa ideia e de acordo com as sugestões fornecidas pelo W3C [Lóscio et al. 2017], para $IQ(ɗ)_{C1}$ os seguintes metadados descritivos são considerados desejáveis para um conjunto de dados ($\mathrm{MD}(ɗ)_D$): título, palavras-chave, endereço do conjunto de dados, data de publicação, data da última atualização, publicador e distribuição do conjunto de dados.

$IQ(ɗ)_{C2}$: Metadados estruturais

$$IQ(ɗ)_{C2} = \frac{\#\mathrm{ME}(ɗ)}{\#ƥ(ɗ)}$$

Onde,

$\#\mathrm{ME}(ɗ)$ é a quantidade de metadados estruturais (ME) disponibilizados pelo conjunto de dados ɗ;

$\#ƥ(ɗ)$ é a quantidade de propriedades encontradas na estrutura do conjunto de dados ɗ.

No conjunto de dados ɗ devem ser disponibilizados metadados estruturais ($\mathrm{ME}(ɗ)$) para descrever as propriedades que o compõem, o que permite uma visão geral do conjunto de dados e entendimento sobre seu conteúdo. Devem ser fornecidas informações sobre todas as propriedades do esquema do conjunto de dados.

$IQ(\text{ɗ})_{C3}$: Metadados Descritivos Referenciados Semanticamente

$$IQ(\text{ɗ})_{C3} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 significa que os metadados descritivos desejáveis (MD_D) disponibilizados pelo conjunto de dados ɗ são referenciados semanticamente por vocabulários recomendados;

0 significa que os metadados descritivos desejáveis (MD_D) disponibilizados pelo conjunto de dados ɗ não são referenciados semanticamente por vocabulários recomendados.

Os metadados descritivos desejáveis (MD_D) disponibilizados pelo conjunto de dados ɗ devem ser referenciados por vocabulários recomendados. Essa semântica fornecida aumenta a descoberta de metadados e a capacidade de consumo. Uma das referências atuais diz respeito ao uso do Data Catalog Vocabulary (DCAT), utilizado para descrever conjuntos de dados, aumentando a capacidade de descoberta e consumo dos metadados. O DCAT reutiliza termos de outros vocabulários, em especial, do *Dublin Core Metadata Initiative* (DCTerms). Esses metadados descritivos devem ser referenciados conforme apresentado no Quadro 7.

Quadro 7 - Vocabulários recomendados para referenciar MD_D

Metadado	**Vocabulário**	**Metadado referenciado**
Título	DCTerms	dcterms:title
Palavras-chave	DCAT	dcat:keyword
Endereço do conjunto de dados	DCAT	dcat:landingPage
Data da última atualização	DCTerms	dcterms:modified
Data de publicação	DCTerms	dcterms:issued
Publicador	DCTerms	dcterms:publisher
Distribuição	DCAT	dcat:distribution

Fonte: O Autor (2018)

$IQ(\text{ɗ})_{C4}$: Conjunto de dados em distribuição com formato RDF

$$IQ(\text{ɗ})_{C4} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 O conjunto de dados ɗ é disponibilizado em distribuição com formato RDF;

0 O conjunto de dados ɗ não é disponibilizado em distribuição com formato RDF.

O conjunto de dados ɗ pode ser disponibilizado em vários formatos de distribuição, mas entre os formatos disponibilizados deve existir um formato com serialização RDF. No RDF cada recurso e vocabulários são identificados por uma URI, sendo este um identificador único, que possibilita a eliminação de ambiguidades. Além disso, os dados em RDF são semanticamente descritos e geralmente têm conexões nomeadas. Essa prática fornece não apenas o significado correto de um recurso, mas também possíveis relacionamentos com outros conjuntos de dados, facilitando sua compreensão e a troca de informações.

$IQ(ɗ)_{C5}$: Metadados em distribuição com formato RDF

$$IQ(ɗ)_{C5} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 Os metadados são disponibilizados em distribuição com formato RDF;

0 Os metadados não são disponibilizados em distribuição com formato RDF.

Os metadados correspondentes ao conjunto de dados ɗ devem estar disponíveis em formato RDF. O modelo RDF é indicado para a representação de metadados, pois permite definir formalmente a semântica dos metadados de acordo com seu significado e também facilita a localização e acesso dos conjuntos de dados por parte de mecanismos de busca ou aplicações.

$IQ(ɗ)_{C6}$: Ponto de contato

$$IQ(ɗ)_{C6} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 É fornecido um ponto de contato entre o publicador e o consumidor de dados;

0 Não é fornecido um ponto de contato entre o publicador e o consumidor de dados.

O publicador de dados deve fornecer meios que facilitem a comunicação com os consumidores de conjuntos de dados. Por exemplo, e-mails ou formulários de contato podem ser fornecidos.

Processabilidade

Para que as máquinas possam processar e manipular automaticamente os dados dentro de um conjunto de dados, é importante que o publicador empregue algumas boas práticas [Lóscio et al., 2017] como: fornecer metadados estruturais; disponibilizar distribuições que apresentam

formatos de arquivos legíveis por máquina; fornecer os dados em várias distribuições; reutilizar vocabulários, de preferência padronizados; e disponibilizar dados por meio de uma API. Como resultado surge o critério denominado de Processabilidade, definido a seguir.

Definição 12 Processabilidade ($P(ɗ)$) - A processabilidade do conjunto de dados ɗ, denotada por $P(ɗ)$, mede o grau em que ɗ é processável por máquinas ou agentes de software. $P(ɗ)$ considera um conjunto de indicadores de qualidade $IQ(ɗ) = \{IQ(ɗ)_{P1}, IQ(ɗ)_{P2}, \dots, IQ(ɗ)_{P5}\}$

A processabilidade $P(ɗ)$ é avaliada pelo somatório de todos os valores do conjunto $IQ(ɗ) \in P(ɗ)$, dividido pela quantidade de indicadores de qualidade contidos em $IQ(ɗ)$, representado por $\#IQ(ɗ)$, conforme apresentado a seguir:

$$P(ɗ) = \frac{\sum_{n=1}^{5} IQ(ɗ)_{Pn}}{\#IQ(ɗ)}$$

Onde,

$IQ(ɗ)_{Pn}$ é o valor do indicador de qualidade *P*n, pertencente ao conjunto $IQ(ɗ) = \{IQ(ɗ)_{P1}, \dots, IQ(ɗ)_{P5}\}$;

$\#IQ(ɗ)$ é a quantidade de indicadores de qualidade de $P(ɗ)$.

O critério da processabilidade $P(ɗ)$ é composto por um conjunto de cinco indicadores $IQ(ɗ)$, apresentados a seguir:

$IQ(ɗ)_{P1}$: API de dados

$$IQ(ɗ)_{P1} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 É disponibilizada uma API para acesso ao conjunto de dados ɗ.

0 Não é disponibilizada uma API para acesso ao conjunto de dados ɗ.

Entre as distribuições do conjunto de dados deve ser disponibilizada uma API, o que permite maior flexibilidade e capacidade de processamento dos conjuntos de dados, possibilitando o uso de dados em tempo real, podendo também ser disponibilizada por meio de um *Web Service* (e.g., SOAP[32], REST[33]).

[32] https://www.w3.org/TR/soap12/
[33] https://www.w3.org/2001/sw/wiki/REST

$IQ(ɗ)_{P2}$: Metadados estruturais

$$IQ(ɗ)_{P2} = \frac{\#\mathrm{ME}(ɗ)}{\#ƥ(ɗ)}$$

Onde,

#ME*(ɗ)* é a quantidade de metadados estruturais (ME(ɗ)) disponibilizados pelo conjunto de dados ɗ*;*

#ƥ*(ɗ)* é a quantidade de propriedades encontradas na estrutura do conjunto de dados ɗ**.**

No conjunto de dados ɗ devem ser disponibilizados metadados estruturais (ME(ɗ)) para descrever as propriedades que o compõem, o que permite uma visão geral do conjunto de dados e entendimento sobre seu conteúdo. Devem ser fornecidas informações sobre todas as propriedades do esquema do conjunto de dados.

$IQ(ɗ)_{P3}$: Distribuições em formatos que possam ser processáveis por máquinas

$$IQ(ɗ)_{P3} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 Os conjuntos de dados ɗ são disponibilizados em distribuições com formatos de arquivos processáveis por máquinas;

0 Os conjuntos de dados ɗ não são disponibilizados em distribuições com formatos de arquivos processáveis por máquinas.

O conjunto de dados ɗ deve ser disponibilizado por meio de distribuições. As distribuições devem disponibilizar os dados em formatos de arquivos que sejam mais facilmente processáveis por máquinas, como XML, RDF, JSON e CSV. Essa boa prática pode ajudar as aplicações a processar os dados.

$IQ(ɗ)_{P4}$: D*ownload* do conjunto de dados

$$IQ(ɗ)_{P4} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 Permite o *download* do conjunto de dados ɗ;

0 Não permite o *download* do conjunto de dados ɗ.

Um conjunto de dados ɗ disponível para download pode facilitar o consumo e acesso por parte dos consumidores, e aumentar as possibilidades de processamento e uso de seus dados.

$IQ(ɗ)_{P5}$: Conjuntos de dados em mais de uma distribuição

$$IQ(ɗ)_{P5} = \begin{cases} 1 \\ 0 \end{cases}$$

Onde,

1 O conjunto de dados ɗ é disponibilizado em mais de um formato de distribuição;

0 O conjunto de dados ɗ não é disponibilizado em mais de um formato de distribuição;

O conjunto de dados ɗ deve ser disponibilizado em mais de uma distribuição. Por exemplo, ele pode ser fornecido em formatos abertos, como XML, RDF, JSON e/ou CSV. Essa prática aumenta as chances de consumo, já que os consumidores de dados podem ter preferências em relação aos formatos de dados.

4.3.3 Formação do PCD

Após a obtenção dos metadados descritivos, estruturais e de qualidade que compõem o PCD, esses metadados são referenciados por termos de vocabulários recomendados, sendo eles: (i) DCAT: utilizado para descrever conjuntos de dados em catálogos de dados; (ii) DCTerms: é um vocabulário simples, utilizado para a descrever recursos. (iii) VOID: utilizado para expressar metadados de conjuntos de dados RDF, como metadados de acesso, metadados estruturais e *links* entre conjuntos de dados; (iv) SKOS[34]: utilizado para compartilhar e vincular sistemas de organização do conhecimento, permite expressar a estrutura e o conteúdo de esquemas; (v) DQV[35]: esse vocabulário é considerado uma extensão do DCAT, utilizado para descrever a qualidade de um conjunto de dados.

No Quadro 8 são apresentados os metadados que compõem o PCD, assim como os termos dos vocabulários utilizados para referenciá-los semanticamente e o seu significado. O PCD, que é composto por metadados enriquecidos semanticamente, é disponibilizado aos consumidores de dados no formato RDF.

[34] https://www.w3.org/2009/08/skos-reference/skos.html
[35] https://www.w3.org/TR/vocab-dqv/

Quadro 8 - PCD(d) = {MD(d) + ME(d) + MQ(d)}

Metadado	**Metadado Referenciado por Vocabulário**	**Significado**
Identificador	*dcterms:identifier*	Representa o identificador do conjunto de dados
Título	*dcterms:title*	Indica o título do conjunto de dados.
Descrição	*dcterms:description*	Representa uma breve descrição do conjunto de dados
Palavras-chave	*dcat:keyword*	Representa as palavras-chave do conjunto de dados.
Domínio	*dcat:theme*	Indica o domínio ao qual o conjunto de dados pertence. Este metadado é composto pelo nome do domínio e a uri desse domínio.
Domínio - Nome	*rdfs:label*	Indica o nome do domínio.
Domínio - URI	*void:uriSpace*	Indica a URI desse domínio.
Vocabulário de domínio	*void:vocabulary*	Recebe o vocabulário de domínio recomendado que pode ajudar na conversão do conjunto de dados.
Endereço	*dcat:landingPage*	Indica a URL do conjunto de dados
Data da última atualização	*dcterms:modified*	Representa a data da última atualização do conjunto de dados
Data de publicação	*dcterms:issued*	Representa a data da criação do conjunto de dados
Publicador	*dcterms:Publisher*	Recebe o nome do publicador do conjunto de dados
Versão	*owl:versionInfo*	Indica a versão do conjunto de dados
Distribuição	*dcat:distribution*	Recebe instâncias de distribuições do conjunto de dados. Cada distribuição é representada por: formato, tamanho, URL de download/acesso à distribuição e o tipo de distribuição.

Metadado	**Metadado Referenciado por Vocabulário**	**Significado**
Distribuição – Formato	*dcterms:format*	Representa o formato que o conjunto de dados está disponível.
Distribuição – Tamanho	*dcat:byteSize*	Indica o tamanho do conjunto de dados.
Distribuição – URL	*dcat:downloadURL*	Representa a URL de download/acesso à distribuição.
Distribuição – Tipo	*dcat:mediaType*	Representa o tipo de distribuição
Esquema do Conjunto de dados	*skos:inSchema*	Representa o esquema das propriedades que compõem a estrutura do conjunto de dados. Apresenta a quantidade de propriedades e cada propriedade presente no conjunto de dados.
Esquema do Conjunto de dados - Quantidade de propriedades	*void:properties*	Representa a quantidade de propriedades presentes no conjunto de dados
Esquema do Conjunto de dados – Propriedade	*void:property*	Representa cada propriedade da estrutura do conjunto de dados, composto por seu nome e o tipo de dado.
Esquema do Conjunto de dados - Propriedade – Nome	*rdfs:label*	Indica o nome da propriedade.
Esquema do Conjunto de dados - Propriedade – Tipo	*dcterms:type*	Indica o tipo de dado que a propriedade recebe.
Critérios de Qualidade	*dqv:hasQualityMetadata*	Representa instâncias de critérios de qualidade. Cada critério é representado pelo seu nome, o valor calculado para esse critério e a definição do critério de qualidade.
Critérios de Qualidade - Nome	*rdfs:label*	Representa o nome do critério de qualidade.

Metadado	Metadado Referenciado por Vocabulário	Significado
Critérios de Qualidade - Valor	*dqv:value*	Representa o valor calculado para esse critério.
Critérios de Qualidade – Definição	*skos:definition*	Representa a definição do critério de qualidade.
Data de criação do PCD	*dcterms:created*	Indica a data de criação do PCD.

Fonte: O Autor (2018)

4.4 EXEMPLO

Esta seção tem o objetivo de exemplificar a abordagem DSPro+ para geração do PCD. Para tal, foi utilizado o conjunto de dados publicado na Web denominado *Rolling Stone's 500 Greatest Albums of All Time*[36], que disponibiliza os 500 melhores álbuns de todos os tempos, baseado em listas publicadas em edições da revista americana Rolling Stone[37]. Este conjunto de dados é proveniente da plataforma Kaggle[38], onde são disponibilizados conjuntos de dados pertencentes a diversos domínios de conhecimento.

As informações e metadados extraídos deste conjunto de dados podem ser acessadas por meio de sua página. Alguns metadados descritivos são representados no formato JSON-LD[39], que utiliza o vocabulário Schema.org. Na página de acesso de alguns conjuntos de dados também é possível encontrar metadados estruturais, mas, algumas vezes, os metadados estruturais disponibilizados não correspondem a todas as propriedades presentes na estrutura do conjunto de dados.

A abordagem recebe como entrada a URL do conjunto de dados, onde é verificado se é uma URL válida para o acesso às informações sobre o conjunto de dados e identificação de metadados, como título e distribuições do conjunto de dados, propriedades que compõem a estrutura do conjunto de dados e outros dados necessários para a geração de metadados enriquecidos semanticamente. Esses dados extraídos são utilizados para a composição dos metadados descritivos, estruturais e de qualidade, e também são utilizados para as etapas de recomendação de vocabulários de domínio e identificação do domínio do conjunto de dados.

[36] https://www.kaggle.com/notgibs/500-greatest-albums-of-all-time-rolling-stone
[37] https://www.rollingstone.com
[38] https://www.kaggle.com/
[39] https://json-ld.org/

Para os metadados descritivos é possível extrair diretamente do conjunto de dados os valores para sua composição. Neste caso também foi identificado o metadado referente às palavras-chave, mas, nos casos em que não estejam disponíveis, é possível identificar os termos mais frequentes.

A geração do metadado descritivo referente à recomendação de vocabulários de domínio é realizada a partir das propriedades e palavras-chave do conjunto de dados, que são utilizadas em consultas SPARQL sobre o *endpoint* de um catálogo de vocabulários, como o *Linked Open Vocabularies* (LOV). Neste caso, o vocabulário que apresentou maior número de ocorrências entre os resultados das consultas SPARQL foi o *The Music Ontology*[40], que fornece conceitos e propriedades principais para descrever músicas, incluindo artistas, álbuns e faixas.

Para geração do metadado descritivo referente à identificação do domínio ao qual o conjunto de dados pertence, as palavras-chave do conjunto de dados também integram consultas SPARQL, que são realizadas sobre o *endpoint* da ontologia utilizada como referência semântica, e.g., Ontologia do DBpedia, para a extração de informações sobre domínios de conhecimento. Neste exemplo, como domínio geral foi identificada apenas a classe do DBpedia *Musical Work*[41], que apresenta as subclasses *Album* e *Song* e obteve o maior número de ocorrências entre os resultados das consultas, mas em alguns casos podem ser encontrados mais de um domínio do conjunto de dados. Desta classe são extraídos o seu nome e URI, o que permite ao usuário o acesso a maiores informações.

A partir de uma visão prévia disponibilizada sobre o conjunto de dados, são identificadas as propriedades que compõem sua estrutura, correspondendo aos metadados estruturais. Neste exemplo foi identificado que a estrutura do conjunto de dados é composta por seis propriedades: *Number* (*numeric*), *Year* (*numeric*), *Album* (*string*), *Artist* (*string*), *Genre* (*string*), S*ubgenre* (*string*). Conforme apresentado na Figura 15, esses metadados estruturais compõem o array de metadados *skos:inSchema*.

Os metadados de qualidade também são calculados a partir das informações coletadas do conjunto de dados e já considerando os benefícios da geração do PCD. Foram utilizados os dois critérios de qualidade apresentados na Seção 4.3.2.3, onde cada critério de qualidade é avaliado de acordo com o seu conjunto de indicadores de qualidade. A compreensibilidade do conjunto de dados $C(d)$ é composta pelo conjunto de indicadores de qualidade $IQ(d) = \{IQ(d)_{C1}, IQ(d)_{C2}, IQ(d)_{C3}, IQ(d)_{C4}, IQ(d)_{C5}, IQ(d)_{C6}\}$, para o qual foram encontrados os seguintes valores:

[40] http://purl.org/ontology/mo/
[41] http://mappings.dbpedia.org/server/ontology/classes/MusicalWork

$\boldsymbol{IQ(d)_{C1}}$ = 1.0, o conjunto de dados disponibilizou todos os metadados descritivos desejáveis;

$\boldsymbol{IQ(d)_{C2}}$ = 1.0, todas as propriedades que compõem a estrutura do conjunto de dados estão representadas pelos metadados estruturais;

$\boldsymbol{IQ(d)_{C3}}$ = 1.0, os metadados descritivos desejáveis disponibilizados são referenciados semanticamente por vocabulários recomendados;

$\boldsymbol{IQ(d)_{C4}}$ = 0.0, o conjunto de dados não é disponibilizado em distribuição com formato RDF;

$\boldsymbol{IQ(d)_{C5}}$ = 1.0, os metadados estão disponíveis no formato RDF;

$\boldsymbol{IQ(d)_{C6}}$ = 1.0, é fornecido um ponto de contato entre o publicador e o consumidor de dados.

$C(d)$ é calculada por meio de sua métrica, sendo obtido um $C(d)$= 0.83.

A processabilidade do conjunto de dados $P(d)$ é composta pelo conjunto de indicadores de qualidade $IQ(d)$ = $\{IQ(d)_{P1}, IQ(d)_{P2}, IQ(d)_{P3}, IQ(d)_{P4}, IQ(d)_{P5}\}$, sendo encontrados os seguintes valores:

$\boldsymbol{IQ(d)_{P1}}$ = 0.0, não é disponibilizada uma API para acesso ao conjunto de dados;

$\boldsymbol{IQ(d)_{P2}}$ = 1.0, todas as propriedades que compõem a estrutura do conjunto de dados estão representadas pelos metadados estruturais;

$\boldsymbol{IQ(d)_{P3}}$ = 1.0, os conjuntos de dados são disponibilizados em distribuições com formatos de arquivos processáveis por máquinas;

$\boldsymbol{IQ(d)_{P4}}$ = 1.0, permite o *download* do conjunto de dados;

$\boldsymbol{IQ(d)_{P5}}$ = 0.0, o conjunto de dados não é disponibilizado em mais de um formato de distribuição.

$P(d)$ é calculada por meio de sua métrica, sendo obtido um $P(d)$= 0.6.

Uma vez que os metadados são obtidos, eles são disponibilizados por meio de um PCD, que utiliza vocabulários recomendados pelo W3C para representar esses metadados. O PCD gerado, dividido na Figura 14 e Figura 15, é disponibilizado em formato RDF (serializado em Turtle). Conforme observado na Figura 14, primeiramente são listados os prefixos dos vocabulários utilizados, e em seguida é possível observar todos os metadados descritivos, que iniciam no metadado *dcterms:identifier* indo até o metadado *dcat:distribution*, onde o domínio e cada distribuição são instanciados e posteriormente especificados. Já na Figura 15 estão os metadados estruturais, representados pelo array *skos:inScheme*, que dispõem as informações de todas as propriedades encontradas na estrutura do conjunto de dados. Os metadados de qualidade apresentam seus critérios instanciados no metadado *dqv:hasQualityMetadata*, onde cada critério é instanciado e posteriormente especificado. Logo depois é exibido o metadado descritivo

dcterms:created que representa a data em que o PCD foi gerado e a especificação de todos os metadados que tiveram instâncias estabelecidas.

Figura 14 - Exemplo de Perfil de Conjunto de Dados – Prefixos + MD(ɗ)

```
@prefix :        <http://example/ns#> .
@prefix void:    <http://rdfs.org/ns/void#> .
@prefix owl:     <http://www.w3.org/2002/07/owl#> .
@prefix dqv:     <http://www.w3.org/ns/dqv#> .
@prefix xsd:     <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms:  <http://purl.org/dc/terms/> .
@prefix skos:    <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:    <http://www.w3.org/ns/dcat#> .

:dataset.ID629.Kaggle
      a       dcat:Dataset ;
      dcterms:identifier "629" ;
      dcterms:title "Rolling Stone's 500 Greatest Albums of All Time" ;
      dcterms:description "From [Wikipedia](https://en.wikipedia.org/wiki/Rolling_Stone's_500_Greatest_Albums_of_All_Time):
      ----------   >'The 500 Greatest Albums of All Time' is a 2003 special issue of American magazine Rolling Stone, and a
      related book published in 2005. The lists presented were compiled based on votes from selected rock musicians, critics,
      and industry figures, and predominantly feature American and British music from the 1960s and the 1970s.  >In 2012,
      Rolling Stone published a revised edition of the list drawing on the original and a later survey of albums in the 2000s.
      It was made available in 'bookazine' format on newsstands in the US from April 27 to July 25. The new list contained 38
      albums not present in the previous one, 16 of them released after 2003.   ----------   I took the albums from
      [MusicBrainz](https://musicbrainz.org/series/8668518f-4a1e-4802-8b0d-81703ced6418) but the genres weren't listed. I
      wrote a Python script to get the genres and subgenres of each album from the [Discogs API](
      https://www.discogs.com/developers/#).  The data collected are:  * Position on the list * Year of release * Album name *
      Artist name * Genre name * Subgenre name  Some of the genres/subgenres may not be entirely correct - Discogs seems to
      not consider some of the smaller genres. Let me know if there are any glaring issues and I'll try to fix them." ;
      dcat:keyword "Music" , "Humanities" , "Performing Arts" , "Critical Theory" , "Culture" ;
      dcat:theme :theme-Musical_Work ;
      void:vocabulary "http://purl.org/ontology/mo/" ;
      dcat:landingPage "https://www.kaggle.com/notgibs/500-greatest-albums-of-all-time-rolling-stone" ;
      dcterms:modified "2017-01-06T02:39:49.18Z"^^xsd:dateTime ;
      dcterms:issued "2017-01-06T02:39:49.18Z"^^xsd:dateTime ;
      dcterms:publisher "Gibs" ;
      owl:versionInfo "1"^^xsd:interger ;
      dcat:distribution :distribution2-DataDownload.csv , :distribution1-DataDownload.zip ;
```

Fonte: O Autor (2018)

Figura 15 - Exemplo de Perfil de Conjunto de Dados - MD(ɗ) + ME(ɗ) + MQ(ɗ)

```
      skos:inScheme
              [ void:properties "6" ;
                void:property
                        [ rdfs:label "Genre"^^xsd:string ;
                          dcterms:type "string"
                        ] ;
                void:property
                        [ rdfs:label "Year"^^xsd:string ;
                          dcterms:type "numeric"
                        ] ;
                void:property
                        [ rdfs:label "Number"^^xsd:string ;
                          dcterms:type "numeric"
                        ] ;
                void:property
                        [ rdfs:label "Artist"^^xsd:string ;
                          dcterms:type "string"
                        ] ;
                void:property
                        [ rdfs:label "Album"^^xsd:string ;
                          dcterms:type "string"
                        ] ;
                void:property
                        [ rdfs:label "Subgenre"^^xsd:string ;
                          dcterms:type "string"
                        ]
              ] ;
      dqv:hasQualityMetadata  :dimensionComprehensibility , :dimensionProcessability ;
      dcterms:created "2018-02-04T00:58:19.965Z"^^xsd:dateTime .

:theme-Musical_Work
      a       dcat:theme ;
      rdfs:label "Musical_Work"@en ;
      void:uriSpace "http://dbpedia.org/ontology/MusicalWork" .

:distribution2-DataDownload.csv
      a       dcat:Distribution ;
      dcterms:format "csv" ;
      dcat:byteSize "37423 bytes" ;
      dcat:downloadURL "https://www.kaggle.com//notgibs/500-greatest-albums-of-all-time-rolling-stone/downloads/albumlist.csv" ;
      dcat:mediaType "DataDownload" .

:distribution1-DataDownload.zip
      a       dcat:Distribution ;
      dcterms:format "zip" ;
      dcat:byteSize "13040 bytes" ;
      dcat:downloadURL "
      https://www.kaggle.com//notgibs/500-greatest-albums-of-all-time-rolling-stone/downloads/rolling-stone-s-500-greatest-album
      s-of-all-time.zip/1" ;
      dcat:mediaType "DataDownload" .

:dimensionComprehensibility
      a       dqv:Dimension ;
      rdfs:label "Comprehensibility"@en ;
      dqv:value "0.83"^^xsd:float ;
      skos:definition "Represented the degree to which a dataset presents information that promotes/facilitates its
      understanding by human users."@en .

:dimensionProcessability
      a       dqv:Dimension ;
      rdfs:label "Processability"@en ;
      dqv:value "0.6"^^xsd:float ;
      skos:definition "Represents the degree to which a dataset is processable by machines or software agents."@en .
```

Fonte: O Autor (2018)

4.5 ANÁLISE COMPARATIVA COM TRABALHOS RELACIONADOS

Nesta seção, por meio do Quadro 9, é apresentada uma comparação entre os trabalhos relacionados e a abordagem DSPro+. Nesta abordagem, é proposto um processo para a geração de perfil que descreve um determinado conjunto de dados. O PCD é gerado utilizando termos de vocabulários recomendados pelo W3C e em formato legível por máquina.

Os trabalhos analisados no Capítulo 3 abordam aspectos relacionados às etapas de geração do Perfil de Conjunto de Dados, sendo trabalhos relacionados ao enriquecimento semântico, qualidade e geração de perfil. Dessa forma, foram considerados os recursos mais relevantes nos trabalhos relacionados e na abordagem proposta, como: os tipos de metadados gerados, a recomendação de vocabulários de domínio, a identificação do domínio do conjunto de dados, metadados referenciados por termos de vocabulários, o nível de automação da abordagem e a geração de um perfil para a representação dos metadados.

Ao comparar os trabalhos relacionados com a abordagem proposta, em geral percebe-se que esses trabalhos consideram no máximo dois tipos de metadados, não disponibilizam metadados em formatos legíveis por máquinas e também não utilizam vocabulários para referenciar esses metadados. Mesmo os trabalhos que realizam a geração de um perfil, como em Abele (2016), Assaf et al. (2015), Fetahu et al. (2014) e Silva Neto et al. (2016), não apresentam uma abordagem que disponibilize informações mais detalhadas acerca dos conjuntos de dados, que incluam metadados descritivos, estruturais e de qualidade. Isso permitiria aos consumidores de dados o acesso a informações mais completas, facilitando também o processo de identificação dos conjuntos de dados mais apropriados para uma determinada tarefa. Além disso, alguns deles não utilizam termos de vocabulários para referenciar os metadados disponibilizados pelo perfil, o que permitiria atribuir um melhor significado e representação dos metadados. Isso evidencia a necessidade de considerar não apenas aspectos descritivos, mas também aspectos relacionados à estrutura e à qualidade dos conjuntos de dados, com descrições completas que permitem a compreensão pelos humanos e pelas máquinas.

Quadro 9 - Comparativo entre trabalhos relacionados e abordagem sugerida

	Tipo de Metadados Gerados	**Recomendação de Vocabulários**	**Identificação de Domínio**	**Metadados Referenciados por Termos de Vocabulários**	**Nível de Automação**	**Geração de Perfil**
Abele (2016)	Descritivos e Estruturais	-	Sim	Sim	Semiautomático	Sim
Assaf et al. (2015)	Descritivos e de Proveniência	-	-	-	Automático	Sim
Fetahu et al. (2014)	Descritivos	-	-	Sim	Automático	Sim
Ellefi et al. (2015)	Estruturais	Sim	-	Sim	Automático	-
Schaible et al. (2013)	-	Sim	-	Sim	Semiautomático	-
Lalithsena et al. (2013)	-	-	Sim	Sim	Automático	-
Ouksili et al. (2014)	-	-	Sim	-	Semiautomático	-
Silva Neto et al. (2016)	Qualidade	-	-	-	Automático	Sim
Assaf et al. (2016)	-	-	-	-	Automático	-
DSPro+	Descritivos, Estruturais e de Qualidade	Sim	Sim	Sim	Automático	Sim

Fonte: O Autor (2018)

4.6 CONSIDERAÇÕES

Neste capítulo foi apresentada a abordagem DSPro+, proposta nesta pesquisa. Inicialmente foi introduzida a definição do problema e algumas definições preliminares. A abordagem proposta foi detalhada, considerando cada uma de suas etapas. Também foram apresentados os metadados descritivos e estruturais que compõem o perfil, assim como os critérios de qualidade e suas respectivas métricas propostas a serem representadas pelos metadados de qualidade. Em seguida, foi apresentado um exemplo que ilustra as etapas da abordagem para geração do PCD. Por último, foi realizada uma análise comparativa entre a abordagem e trabalhos relacionados. No próximo capítulo, será apresentada a implementação de um protótipo que automatiza as etapas da abordagem proposta, assim como os resultados obtidos por meio dos experimentos.

5

IMPLEMENTAÇÃO E EXPERIMENTOS

Neste capítulo serão descritos aspectos relacionados a implementação e aos experimentos realizados para avaliar a abordagem proposta. Na Seção 5.1 é apresentado o protótipo DSPro+, na qual é descrita a arquitetura proposta para a ferramenta e suas funcionalidades. A Seção 5.2 apresenta aspectos relacionados à implementação do protótipo. A Seção 5.3 descreve e discute os experimentos realizados e os resultados obtidos, assim como o cenário em que os experimentos foram realizados. Em seguida, na Seção 5.4 são apresentadas as considerações deste capítulo.

5.1 PROTÓTIPO - DSPRO+

Com o objetivo de avaliar a abordagem para geração do PCD, detalhado no capítulo anterior, foi desenvolvida a ferramenta denominada DSPro+. Esta ferramenta permite que seus usuários gerem, a partir de conjuntos de dados publicados na Web, perfis de conjunto de dados com metadados descritivos, estruturais e de qualidade enriquecidos semanticamente. A seguir são detalhadas a arquitetura e funcionalidades do DSPro+.

5.1.1 Apresentação da Arquitetura

Conforme apresentado na Figura 16, a arquitetura da ferramenta foi dividida em três camadas: (i) camada do usuário, permite a interação do usuário com a ferramenta; (ii) camada lógica, responsável pela execução das funcionalidades do protótipo; e (iii) camada de dados, gerencia o acesso à base de dados e o acesso às fontes de dados utilizados no processo de geração do PCD. A seguir é detalhada cada camada e os componentes da arquitetura.

Camada do Usuário: É formada pela Interface do Usuário, responsável pela interação do usuário com a camada lógica. Permite que o usuário tenha acesso às funcionalidades da ferramenta, detalhadas na Seção 5.1.2.

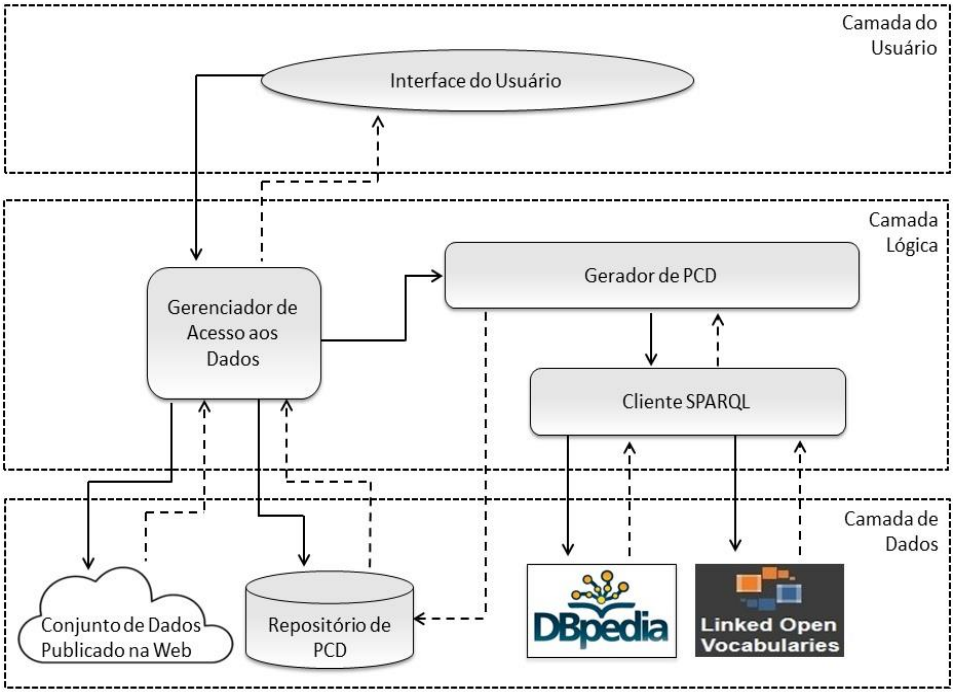
Camada Lógica: Esta camada apresenta os seguintes componentes que permitem executar as funcionalidades do DSPro+:

- **Gerenciador de Acesso aos Dados:** Permite o acesso aos conjuntos de dados utilizados na abordagem. Verifica e intermedeia o acesso aos dados e informações

utilizados durante a etapa de geração do perfil, também retorna aos usuários os resultados obtidos durante o processo de geração do PCD;

- **Gerador de PCD:** É o responsável pela geração do perfil. Realiza a geração de novos metadados e o enriquecimento semântico dos metadados já existentes, resultando em um PCD em formato RDF, composto por metadados enriquecidos semanticamente;
- **Cliente SPARQL:** É uma aplicação responsável por buscar informações complementares em repositórios semânticos específicos, realizando consultas sobre SPARQL *endpoints* e retornando seus resultados para a aplicação.

Figura 16 - Visão geral da arquitetura do DSPro+.



Fonte: O Autor (2018)

Camada de Dados: É composta pelas fontes de dados utilizadas na abordagem, sendo elas:

- **Conjunto de Dados Publicado na Web:** Acesso ao conjunto de dados e suas informações. A partir de sua URL são extraídas informações para compor os metadados descritivos, estruturais e de qualidade;
- **Repositório de PCD:** É responsável por armazenar os resultados obtidos pela geração do PCD e informações sobre o conjunto de dados correspondente, por exemplo, título, URL e palavras-chave, ficando disponíveis para acesso por outros usuários. Para

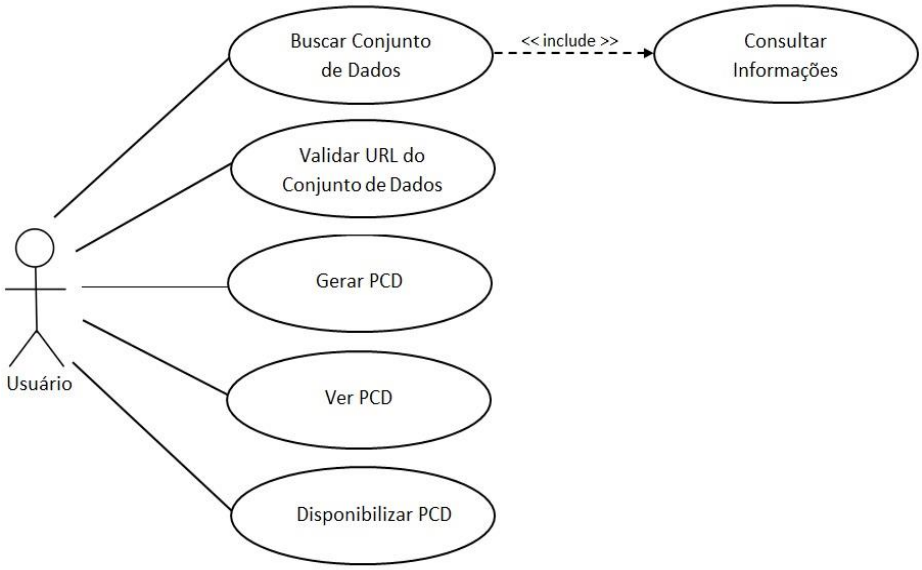
armazenar essas informações foi utilizado o gerenciador de banco de dados objeto relacional PostgreSQL[42];

- **DBpedia:** Disponibiliza informações da Wikipédia[43] de forma estruturada, permite acesso aos seus dados por meio de um SPARQL *endpoint*. Sua ontologia é utilizada como referência semântica durante a identificação do domínio do conjunto de dados;
- ***Linked Open Vocabularies* (LOV):** É um dos maiores catálogos de vocabulários abertos existentes, também permite recuperar informações sobre vocabulários e suas descrições semânticas por meio de um SPARQL *endpoint.* É utilizado durante a etapa de recomendação de vocabulários de domínio.

5.1.2 Funcionalidades do DSPro+

O diagrama de casos de uso da aplicação (Figura 17) ressalta as principais funcionalidades desenvolvidas que o DSPro+ disponibiliza para seus usuários. Cada uma dessas funcionalidades é descrita a seguir.

Figura 17 - Diagrama de Caso de Uso do protótipo DSPro+.



Fonte: O Autor (2018)

Buscar Conjunto de Dados: A partir das palavras-chave, o usuário pode buscar um ou mais conjuntos de dados que tenham passado pelo processo de geração do PCD anteriormente.

Consultar Informações de Conjunto de Dados: Está vinculada à funcionalidade "Buscar Conjunto de Dados". A ferramenta pode recuperar as informações correspondentes ao conjunto de dados selecionado.

[42] https://www.postgresql.org/
[43] https://pt.wikipedia.org/wiki/Wikipédia:Página_principal

Validar URL do Conjunto de Dados: Antes de iniciar o processo de geração do PCD, o usuário deve validar o conjunto de dados. Para isso, o usuário deve disponibilizar a URL do conjunto de dados, e a abordagem verifica se é uma URL válida que permite a extração de informações.

Gerar PCD: Após validar a URL do conjunto de dados, o usuário pode iniciar o processo de geração do PCD. As informações sobre esse conjunto de dados e os resultados obtidos ficarão disponíveis para os usuários. Para isso, pode-se utilizar a funcionalidade "Buscar Conjunto de Dados".

Ver PCD: O usuário pode visualizar o PCD gerado pela abordagem, contendo os metadados descritivos, estruturais e de qualidade.

Disponibilizar PCD: O usuário pode realizar o *download* do PCD em formato RDF, para que possa consumir ou publicar juntamente aos seus conjuntos de dados em portais de Dados Abertos.

5.2 IMPLEMENTAÇÃO DO PROTÓTIPO

A abordagem proposta foi desenvolvida com a tecnologia Java Server Pages[44] (JSP) para plataforma Web, que é baseada na linguagem de programação JAVA[45]. O layout do protótipo foi desenvolvido com o *framework* Foundation[46], utilizando HTML, CSS[47] e Javascript[48]. Para o acesso aos dados e metadados do conjunto de dados foi utilizada a API Selenium[49], que permite navegar entre páginas e extrair informações a partir de sua URL. A integração entre o protótipo e os SPARQL *endpoint*s foi realizada por meio da API Jena[50], que fornece um conjunto de bibliotecas que facilitam o acesso às bases de dados RDF.

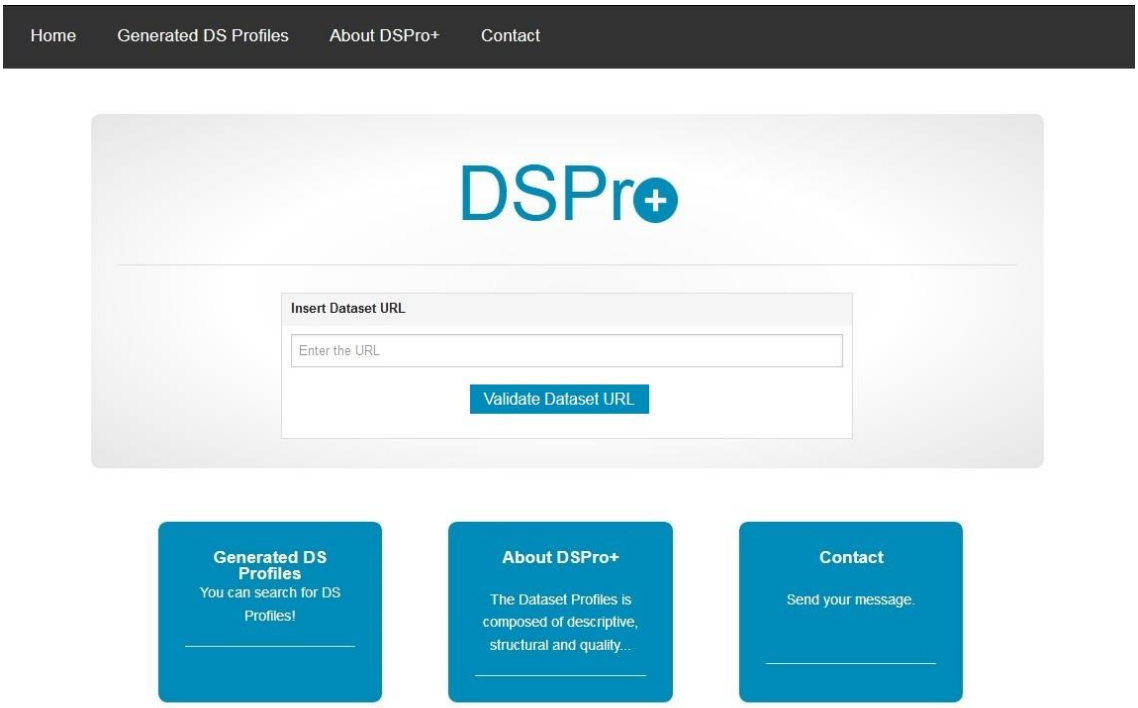
5.2.1 Interface do Protótipo

O protótipo DSPro+ dispõe de uma interface amigável, que permite ao usuário utilizar as funcionalidades da ferramenta. Na página inicial, ilustrada na Figura 18, o usuário deve indicar a URL do conjunto de dados para o qual será gerado o PCD. O protótipo está implementado para aceitar conjuntos de dados que permitem acessar informações e seus metadados a partir de sua URL. Caso o usuário carregue a URL de um conjunto de dados que não os disponibilize, o protótipo apresenta uma mensagem informando que a página não possui o formato desejado.

[44] http://www.oracle.com/technetwork/java/javaee/jsp/index.html
[45] https://www.java.com/pt_BR/
[46] https://foundation.zurb.com/
[47] https://www.w3schools.com/css/
[48] https://www.javascript.com/
[49] https://seleniumhq.github.io/selenium/docs/api/javascript/index.html
[50] https://jena.apache.org/

Nesta versão foram considerados conjuntos de dados que estejam na língua inglesa, principal idioma de publicação de dados na Web. Para a utilização de outros idiomas é necessário configurar as etapas de identificação do domínio e recomendação de vocabulários, que estão relacionadas à geração de metadados descritivos.

Figura 18 - Tela inicial do DSPro+



Fonte: O Autor (2018)

5.3 EXPERIMENTOS

Os experimentos descritos nesta seção têm o objetivo de avaliar a abordagem proposta para a geração de PCD. Primeiramente é apresentado o cenário escolhido para a realização dos experimentos. Em seguida, são apresentados e discutidos os experimentos realizados e os resultados obtidos.

5.3.1 Cenário

O cenário escolhido para a avaliação experimental abrange três grupos temáticos de dados, no qual se deseja o acesso a conjuntos de dados com metadados que contenham informações descritivas, estruturais e de qualidade. A abordagem proposta pode ser utilizada com conjuntos de dados provenientes de qualquer domínio. Para a realização dos experimentos foram selecionados 30 conjuntos de dados escritos na língua inglesa, sendo 10 conjuntos de dados sobre cada um dos seguintes domínios: “videogames”, “automóveis” e “música”. Os conjuntos

de dados selecionados são provenientes da plataforma Kaggle, pois são disponibilizados conjuntos de dados pertencentes a diversos domínios de conhecimento com alguns metadados descritivos de forma estruturada. Estes conjuntos de dados selecionados estão publicados apenas em distribuição com formato CSV, com propriedades estruturais que não são referenciadas semanticamente por meio de vocabulários.

As informações e metadados necessários para a composição do PCD são extraídas a partir da URL do conjunto de dados, que permite o acesso aos dados referentes ao seu conteúdo e estrutura. Os metadados (e.g. nome, descrição, url, versão, palavras-chave, criador e distribuição) são disponibilizados por meio de um JSON-LD que está inserido no HTML. Em alguns casos, também são disponibilizados metadados estruturais, entretanto, muitas vezes os metadados estruturais encontrados não correspondem a todas as propriedades apresentadas na estrutura do conjunto de dados.

5.3.2 Avaliação Experimental

Durante os experimentos foi utilizada uma máquina com processador Intel core i3, 4GB de memória RAM e Sistema Operacional Windows 7. O acesso à Internet foi realizado por meio de uma conexão banda larga de 15MB. Como servidor Web foi utilizado o Web Server Apache Tomcat[51], que é de código aberto e focado nas tecnologias Java Servlets e JSP. Para avaliar a abordagem DSPro+ foram realizados os seguintes experimentos, que objetivaram validar as hipóteses de pesquisa apresentadas no Capítulo 1:

- Experimento 1: Verificar se a geração automática do PCD, com metadados descritivos, estruturais e de qualidade enriquecidos semanticamente, resulta em descrições mais completas sobre o conjunto de dados.
- Experimento 2: Verificar se o domínio de conhecimento identificado está adequado para o conjunto de dados comparado ao *gold standard* gerado pelos especialistas.
- Experimento 3: Verificar se os vocabulários recomendados estão adequados para o conjunto de dados comparado ao *gold standard* gerado pelos especialistas.
- Experimento 4: Verificar se o PCD gerado melhora aspectos relacionados a qualidade do conjunto de dados, em que são agregadas informações que possam melhorar sua compreensão e/ou processamento por parte dos usuários e aplicações consumidoras de dados.

O primeiro experimento verifica se a geração de um PCD resulta em uma representação mais abrangente e rica semanticamente sobre um determinado conjunto de dados. Dessa forma,

[51] http://tomcat.apache.org

os metadados disponibilizados por meio do JSON-LD dos conjuntos de dados são comparados com os metadados disponibilizados pelo DSPro+.

Durante o segundo experimento é realizada uma comparação entre os resultados obtidos por meio da etapa de identificação do domínio de conhecimento do conjunto de dados com o domínio identificado de forma manual por especialistas.

O terceiro experimento realiza uma comparação entre os resultados obtidos por meio da etapa de recomendação de vocabulários de domínio do conjunto de dados com os vocabulários recomendados de forma manual por especialistas.

No quarto experimento são verificados aspectos da qualidade do conjunto de dados, em relação à compreensão e ao processamento, comparando os valores atribuídos às métricas de qualidade antes e depois da geração do PCD.

Para a realização do Experimento 2 e Experimento 3 foram definidos *gold standards* dos resultados esperados, considerados ideais (Apêndice A). Para a geração dos *gold standards* foram convidados dois especialistas que, apesar de não terem acompanhado diretamente o desenvolvimento deste trabalho, possuem conhecimento sobre os conceitos abordados e as tecnologias semânticas, além de estarem familiarizados com vocabulários e ontologias.

Para mensurar os resultados obtidos nos Experimentos 2 e 3 foram calculadas as seguintes métricas [Baeza-Yates et al., 1999]: (i) Precisão (*Precision*), a fração de resultados relevantes que são recuperados; (ii) Cobertura (*Recall*), sendo esta a fração de resultados relevantes que são esperados; e (iii) *F-Measure*, representa a média harmônica entre os valores de Precisão e Cobertura. As fórmulas utilizadas são apresentadas a seguir:

$$Precisão = \frac{\#\text{ResultadosRelevantes}}{\#\text{ResultadosRetornados}}$$

$$Cobertura = \frac{\#\text{ResultadosRelevantes}}{\#\text{ResultadosEsperados}}$$

$$F - Measure = \frac{(2 * Precisão * Cobertura)}{(Precisão + Cobertura)}$$

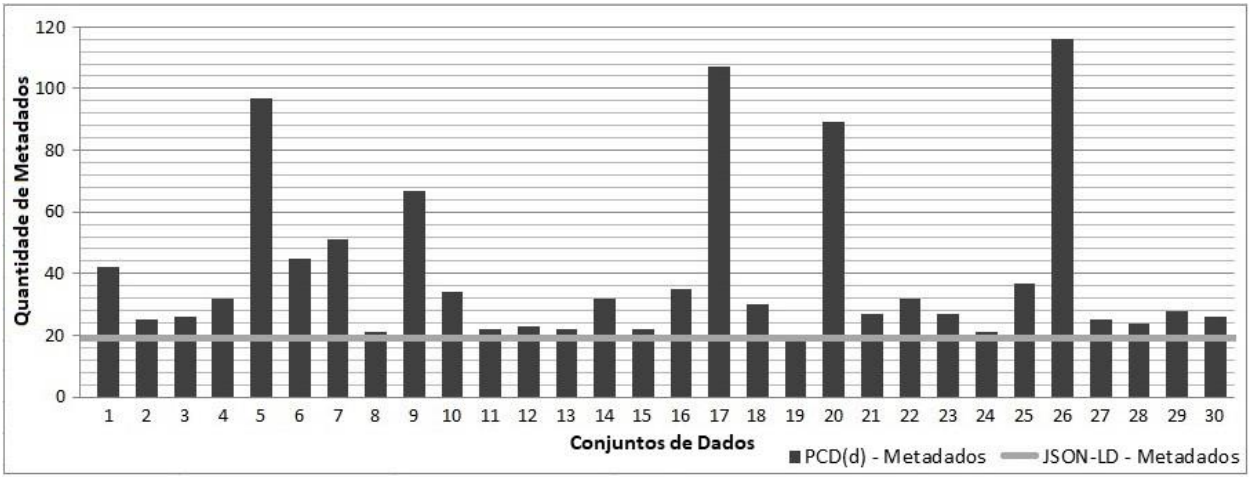
Nas fórmulas apresentadas, #ResultadosRelevantes representam a quantidade de resultados retornados considerados relevantes, #ResultadosEsperados indicam o total de resultados que poderiam ser retornados, e #ResultadosRetornados é o número total de todos os resultados retornados.

5.3.2.1 Experimento 1

Neste experimento, que avalia a hipótese H1, é verificada se a geração de um PCD resulta em uma representação mais abrangente e rica semanticamente sobre um determinado conjunto de dados. Observa-se que é gerado um PCD para cada um dos 30 conjuntos de dados utilizados no experimento, alguns exemplos se encontram no Apêndice B.

Os metadados disponibilizados por meio do JSON-LD de cada conjunto de dados utilizado no experimento são analisados para identificar os metadados em comum que também são disponibilizados pelo PCD gerado. Foi observado que, apesar de alguns conjuntos de dados disponibilizarem metadados estruturais em sua página Web, no conjunto de metadados dispostos no formato JSON-LD não são encontrados metadados estruturais. Também não são disponibilizados metadados de qualidade, nem metadados correspondentes ao domínio do conjunto de dados e a recomendação de vocabulários. Conforme mostrado na Figura 19, os metadados disponibilizados pelo PCD gerado permitem detalhar melhor o conjunto de dados, principalmente em relação à sua estrutura e qualidade. No JSON-LD disponibilizado são fornecidos apenas metadados descritivos, como título, descrição, identificador, versão, e alguns dados estatísticos, como quantidade de comentários e de downloads.

Figura 19 - Experimento 1: Metadados disponibilizados pelos PCDs gerados comparados aos metadados disponibilizados pelo JSON-LD dos conjuntos de dados.



Fonte: O Autor (2018)

Análise do Experimento 1

Com os resultados apresentados, é possível concluir que a abordagem proposta é capaz de gerar os PCDs propostos com todos os metadados descritivos, estruturais e de qualidade enriquecidos semanticamente. Também é possível melhorar as descrições dos conjuntos de dados sem a necessidade de intervenção humana. Como a geração do perfil é um processo

minucioso e depende diretamente de dados provenientes de outras fontes, essa seria uma atividade demorada caso fosse realizada de forma manual. Além disso, poderia ocorrer, em algum caso, das fontes de dados não disponibilizarem as informações necessárias para a geração dos metadados, principalmente durante as etapas de identificação do domínio de conhecimento e durante a recomendação de vocabulários.

Quando comparados os metadados disponibilizados por meio do JSON-LD dos conjuntos de dados com os metadados gerados pelo DSPro+, foi observado que a abordagem propõe uma melhor representação dos dados, em que são abordados mais aspectos de um conjunto de dados por meio da geração de novos metadados descritivos (domínio e vocabulário recomendado), além de metadados estruturais e de qualidade.

Então, ao invés dos publicadores de dados gerarem todos esses metadados de forma manual, a abordagem proposta permite que sejam gerados PCDs de forma automática, o que também complementa os metadados disponibilizados na plataforma de conjunto de dados. Esses resultados demonstraram que há possibilidade de futuramente a abordagem ser integrada a um portal de conjuntos de dados abertos, sendo possível agregar descrições mais completas aos conjuntos de dados por meio dos metadados que compõem o PCD.

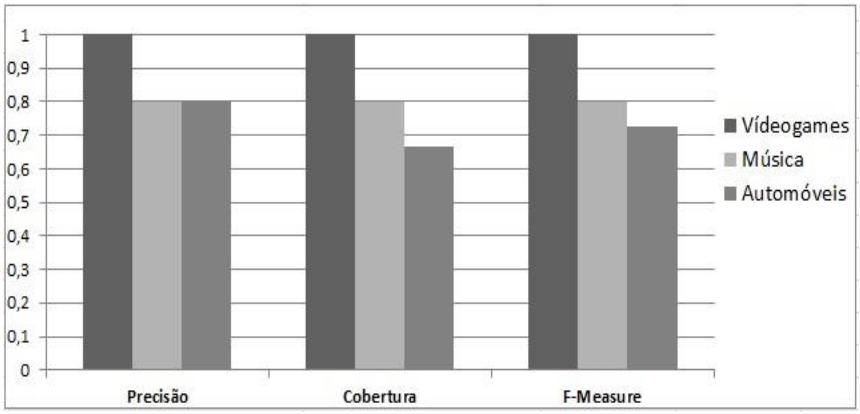
5.3.2.2 Experimento 2

Neste experimento, para avaliar a Hipótese H2, é realizada uma análise comparativa, onde os resultados obtidos pela etapa de identificação de domínio do conjunto de dados são comparados com os *gold standards* gerados pelos especialistas. Durante a análise comparativa, foi constatado que entre os 30 conjuntos de dados utilizados no experimento, apenas quatro deles não receberam a recomendação conforme o domínio recomendado pelos especialistas e em alguns casos foi identificado mais de um domínio para os conjuntos de dados.

Conforme observado na Figura 20, os conjuntos de dados sobre videogames alcançaram melhores resultados. Todos os dez conjuntos de dados desse grupo obtiveram em seus resultados o domínio definido como o *gold standard*. Já entre os conjuntos de dados do grupo de automóveis, dois deles não receberam o domínio definido pelos especialistas. Também não foi identificado o domínio recomendado pelos especialistas para dois conjuntos de dados pertencentes ao grupo de música, e para outros dois conjuntos de dados desse grupo não foram identificados todos os domínios recomendados pelos especialistas. Nos conjuntos de dados do grupo de videogames foi obtido o valor 1 para a Precisão, Cobertura e F-Measure. Para os conjuntos de dados do grupo de automóveis foi obtido valor 0,8 para as métricas de Precisão,

Cobertura e F-Measure, e para os conjuntos de dados do grupo de música foram obtidos os seguintes valores: Precisão = 0,8, Cobertura = 0,67 e F-Measure = 0,73.

Figura 20 - Experimento 2: Identificação do domínio do conjunto de dados.



Fonte: O Autor (2018)

Análise do Experimento 2

No geral, de acordo com os resultados encontrados, a identificação do domínio realizada por meio do protótipo DSPro+ apresentou resultados correspondentes aos *gold standards* definidos pelos especialistas. Nos casos em que os conjuntos de dados não apresentaram o resultado esperado, foi observado que esse fato ocorreu devido às suas palavras-chave. Mesmo para os conjuntos de dados pertencentes a um mesmo grupo temático, suas palavras-chave variam e isso afeta diretamente nos resultados. Por exemplo, o conjunto de dados 11, que disponibiliza as seguintes palavras-chave: *Linguistics, Social Sciences, Music, Performing Arts, Writing, Research Tool, Topic*. Por ser a entrada do algoritmo para a identificação do domínio do conjunto de dados, a variedade de domínios aos quais essas palavras-chave pertencem infere diretamente na obtenção dos resultados esperados.

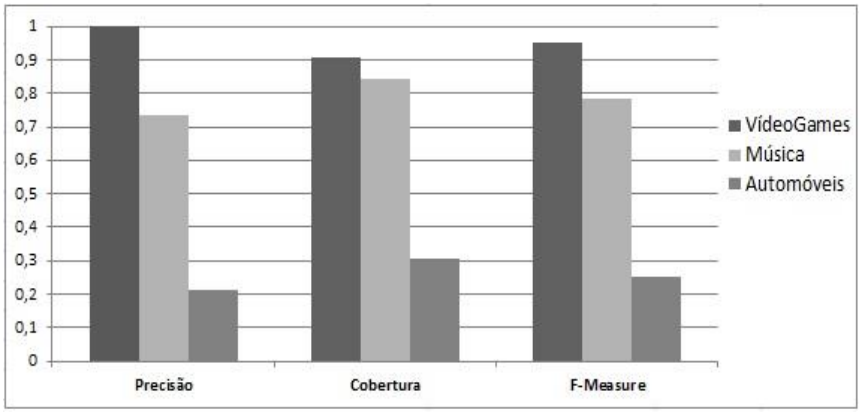
5.3.2.3 Experimento 3

Neste experimento é avaliada a Hipótese H3. Para isso, é realizada uma análise comparativa onde o metadado descritivo referente ao vocabulário de domínio de cada PCD gerado é comparado com o *gold standard* gerado pelos especialistas. Nos resultados foi observado que, em muitos casos, a abordagem recomenda mais de um vocabulário. Isso também ocorreu entre os *gold standards* gerados pelos especialistas. Entretanto, devido à quantidade de vocabulários que estão relacionados a uma mesma área de conhecimento, entre os resultados obtidos para cada conjunto de dados, algumas vezes nem todos os vocabulários recomendados coincidiram com todos os vocabulários do *gold standard*.

Entre os resultados apresentados pelos conjuntos de dados do grupo de videogames, para todos eles foram recomendados vocabulários definidos como o *gold standard*, mas para um conjunto de dados não foram recomendados todos os vocabulários esperados. Em relação aos conjuntos de dados do grupo de música, sete deles obtiveram em sua recomendação vocabulários definidos como o gold standard, entretanto para alguns deles não foram recomendados todos os vocabulários esperados. Já entre os conjuntos de dados do grupo de automóveis, apenas quatro conjuntos de dados apresentaram entre seus resultados algum vocabulário correspondente ao vocabulário recomendado pelos especialistas.

Como apresentado na Figura 21, para os conjuntos de dados do grupo de videogames, foram obtidos valores de Precisão = 1, Cobertura = 0,91 e F-Measure = 0,95. Para os conjuntos de dados do domínio de música, foram obtidos valores de Precisão = 0,73, Cobertura = 0,85 e *F-Measure* = 0,78. E para os conjuntos de dados do grupo de automóveis foram obtidos valores de Precisão = 0,21, Cobertura = 0,31 e *F-Measure* = 0,25.

Figura 21 - Experimento 3: Recomendação de Vocabulários



Fonte: O Autor (2018)

Análise do Experimento 3

Ao analisar os resultados obtidos, foi constatado que o domínio dos conjuntos de dados reflete diretamente nos resultados. Por exemplo, entre os conjuntos de dados sobre videogames, todos apesentaram entre seus resultados um ou mais vocabulários que foram definidos como o *gold standard*. Isso pode ser atribuído às palavras-chave e propriedades que compõem a estrutura do conjunto de dados que contribuem diretamente para os resultados. Quando o conjunto de dados é de uma área mais específica, suas palavras e propriedades estruturais são bastante especializadas, o que torna mais provável alcançar os resultados esperados. Outro fator importante é a variedade de vocabulários disponíveis no repositório de vocabulários LOV para certo domínio, em alguns casos é possível encontrar vários vocabulários que podem ser utilizados para referenciar os mesmos dados.

5.3.2.4 Experimento 4

Para avaliar a Hipótese H4, neste experimento os valores atribuídos aos critérios de qualidade (compreensibilidade e processabilidade), representados pelos metadados de qualidade que compõe o PCD gerado, foram comparados aos valores que seriam atribuídos a esses critérios antes da geração do PCD. Nesta análise foi observada uma melhora significativa entre os valores atribuídos aos critérios de qualidade após a geração do perfil.

Antes da geração do PCD, o maior valor apresentado para a compreensibilidade (Figura 22) foi de 0,5 e para a processabilidade (Figura 23) foi de 0,8. Após a geração dos PCDs, para a compreensibilidade o maior valor obtido foi de 0,83 e a processabilidade continuou com o maior valor de 0,8. Para a compreensibilidade foi observada uma melhoria em relação aos indicadores de qualidade correspondentes à disponibilização de metadados estruturais, reutilização de vocabulários recomendados e a disponibilização dos metadados em formato RDF. Já para a processabilidade, foram observadas melhorias apenas no critério que indica a disponibilização de metadados estruturais.

Figura 22 - Experimento 4: Compreensibilidade antes e após da geração do PCD

Fonte: O Autor (2018)

Figura 23 - Experimento 4: Processabilidade antes e após da geração do PCD

Fonte: O Autor (2018)

Análise do Experimento 4

Para o critério de compreensibilidade foi observada uma melhoria de no mínimo 33% para o valor recebido considerando todos os 30 conjuntos de dados utilizados no experimento. Já para o critério de processabilidade não foram observadas melhorias significativas quando analisados os conjuntos de dados como um todo, pois o maior valor obtido para esse critério continuou

sendo o de 0,8. Isso ocorreu principalmente porque a maior parte dos conjuntos de dados já disponibilizavam todos os metadados estruturais em sua página, fazendo com que ocorresse pouca variação entre os resultados obtidos antes e após a geração do PCD. Mas, quando comparados os conjuntos de dados de forma individual, foram observadas melhorias bastante significativas. Como o conjunto de dados 23, do grupo de videogames, que antes da geração do PCD apresentou uma processabilidade de 0,4 e após o perfil gerado passou a apresentar um valor de 0,6. Isso resulta em uma melhoria de 20% no critério de processabilidade após a geração do PCD. Com esses resultados foi observado que para todos os conjuntos de dados ocorreram melhorias de qualidade, ao menos para o critério de compreensibilidade.

5.4 CONSIDERAÇÕES

Neste capítulo foi apresentada a implementação da abordagem proposta e os resultados dos experimentos realizados para sua avaliação. Primeiramente foi apresentado o protótipo DSPro+, descrevendo sua arquitetura, principais funcionalidades e detalhes da implementação. Em seguida, o cenário dos experimentos foi descrito, também foram detalhados e discutidos os experimentos e resultados obtidos. Os experimentos realizados avaliaram as hipóteses, apresentadas no Capítulo 1, e apresentaram bons resultados para a confirmação dessas hipóteses, indicando a eficácia da abordagem no que diz respeito à geração do PCD e, em especial, ao enriquecimento semântico e melhoria da qualidade dos conjuntos de dados. No próximo capítulo são apresentadas as conclusões, limitações e indicações de trabalhos futuros.

6

CONCLUSÃO

Neste trabalho foi proposta uma abordagem que facilita o processo de geração de metadados para descrever conjuntos de dados na Web por meio da criação de PCDs. Um PCD é composto por metadados descritivos, estruturais e de qualidade e é enriquecido semanticamente. Para avaliar a abordagem proposta foi implementado um protótipo que realiza de forma automática o processo de geração do PCD, podendo ser utilizado por produtores de dados que desejem disponibilizar junto ao conjunto de dados um PCD. Os consumidores de dados, por sua vez, também podem gerar o perfil sem a necessidade de terem conhecimento sobre esses dados.

Os experimentos realizados utilizaram conjuntos de dados provenientes de três domínios de dados diferentes e demonstraram que a estratégia proposta produz bons resultados, permitindo a geração de metadados enriquecidos semanticamente. Os experimentos mostraram também melhorias na qualidade dos conjuntos de dados, em relação a sua compreensão por parte dos usuários humanos e processamento pelas aplicações de consumo de dados. De fato, os resultados obtidos foram satisfatórios para a confirmação das hipóteses de pesquisa levantadas. Desta forma, é possível concluir que o presente trabalho atingiu seus objetivos, disponibilizando uma abordagem para a geração de PCDs com metadados enriquecidos. Na próxima seção são apresentadas as contribuições deste trabalho. Na Seção 6.2 são discutidas algumas limitações e na Seção 6.3 os trabalhos futuros.

6.1 CONTRIBUIÇÕES

As principais contribuições deste trabalho são listadas a seguir:

- **Especificação de uma abordagem para geração de Perfil de Conjunto de Dados com metadados enriquecidos semanticamente:** De acordo com o levantamento realizado na literatura, não foram encontrados trabalhos que realizam a geração automática de um perfil composto por metadados enriquecidos semanticamente e que considere aspectos descritivos, estruturais e de qualidade sobre um conjunto de dados. Dessa forma, neste trabalho foi especificada uma abordagem, denominada DSPro+, para a geração de um Perfil de Conjunto de Dados que disponibiliza metadados descritivos, estruturais e de qualidade enriquecidos semanticamente. Esses perfis

gerados podem ser consumidos ou publicados juntamente aos conjuntos de dados em catálogos ou portais de dados, agregando descrições mais completas aos conjuntos de dados.

- **Identificação do domínio do conjunto de dados:** O resultado obtido na etapa de identificação do domínio de conhecimento ao qual o conjunto de dados está associado corresponde ao metadado descritivo sobre o domínio do conjunto de dados. A identificação do domínio enriquece semanticamente o conjunto de dados, ajudando aos usuários no entendimento sobre o conteúdo dos conjuntos de dados e no retorno de resultados mais relevantes pelos mecanismos de busca.
- **Recomendação de vocabulários de domínio para o conjunto de dados:** Os vocabulários de domínio recomendados nesta etapa correspondem ao valor associado ao metadado descritivo sobre o vocabulário de domínio. Essa recomendação permite identificar vocabulários para os metadados de maneira geral, podendo ajudar na conversão do conjunto de dados em formatos semiestruturados ou estruturados para o formato RDF.
- **Definição de critérios e métricas de qualidade para avaliar conjuntos de dados:** Neste trabalho foram definidos os critérios de Compreensibilidade e Processabilidade. Cada critério é representado por uma métrica que é mensurada a partir de um conjunto de indicadores de qualidade associados a aspectos específicos relativos à publicação e consumo dos conjuntos de dados. .
- **Desenvolvimento do protótipo DSPro+:** Para avaliar a abordagem proposta foi implementado um protótipo que automatiza o processo de geração dos PCDs. Os metadados já existentes são enriquecidos semanticamente e também são gerados novos metadados, resultando em um PCD em formato estruturado que aborda aspectos descritivos, estruturais e de qualidade. Experimentos realizados demonstraram a relevância dos resultados obtidos por meio da ferramenta desenvolvida.

As seguintes publicações foram geradas, até a presente data, a partir dos resultados obtidos durante o desenvolvimento desta dissertação:

- Targino, N., Souza, D., Salgado, A. C. (2017) **Uma Abordagem para Criação e Uso de Perfis de Conjuntos de Dados com Metadados Enriquecidos Semanticamente.** In: VI Workshop de Teses e Dissertações em Banco de Dados - 32nd Simpósio Brasileiro de Bando de Dados (SBBD), 2017, Uberlândia. 32nd SBBD - WTDBD. Porto Alegre: SBC, 2017. p. 115-121.

- Targino, N., Souza, D., Salgado, A. C. (2017). **Uma Proposta de Perfil de Conjuntos de Dados na Web com Enriquecimento Semântico.** In: 32nd Simpósio Brasileiro de Banco de Dados (SBBD), 2017, Uberlândia. 32nd Simpósio Brasileiro de Banco de Dados (SBBD). Porto Alegre: SBC, 2017. p. 172-183.

6.2 LIMITAÇÕES

A seguir são indicadas algumas limitações observadas na abordagem desenvolvida:

- **Critérios de qualidade:** Atualmente a abordagem proposta disponibiliza metadados de qualidade que consideram apenas os critérios de Compreensibilidade e Processabilidade, mas futuramente podem ser incluídos outros critérios para avaliar a qualidade dos conjuntos de dados.
- **Recomendação de vocabulários:** Nesta versão são recomendados vocabulários de domínio para os metadados de maneira geral, com isso, pretende-se futuramente recomendar vocabulários para cada metadado estrutural do conjunto de dados.
- **Identificação do domínio e recomendação de vocabulários relacionadas ao tema do conjunto de dados:** Nos Experimentos 2 e 3, que realizam a identificação do domínio e a recomendação de vocabulários, foi observado que seus resultados estão relacionados ao tema do conjunto de dados. Sendo obtidos melhores resultados quando o conjunto de dados é de um de uma área mais específica.
- **Experimentos:** Os experimentos realizados apresentaram resultados satisfatórios. Entretanto, foi observada a necessidade de realizar novos experimentos com usuários especialistas e conjuntos de dados pertencentes a uma maior variedade de domínios.

6.3 TRABALHOS FUTUROS

A seguir são apresentadas algumas direções que podem ser exploradas em trabalhos futuros para a abordagem desenvolvida:

- ***Feedback/Crowdsourcing* de usuários:** Com o *feedback/crowdsourcing* é possível, a partir do esforço humano, o acesso a informações relevantes acerca da qualidade dos dados, como anomalias e falhas, que podem estar ligadas às etapas de processamento ou até à interpretação dos dados. A ideia é incluir na abordagem proposta o *feedback/crowdsourcing* de usuários, o que permitirá enriquecer os dados e metadados e também viabilizará uma maior interação ente os publicadores e consumidores de dados. Alguns trabalhos encontrados na literatura utilizam o *feedback/crowdsourcing*

para a combinação de esforços de usuários que visem corrigir e aprimorar a qualidade de dados. Um exemplo é o trabalho apresentado por van der Bij et al. (2017), que utiliza uma ferramenta de *feedback* de qualidade de dados para o melhoramento da qualidade da geração do Registro Eletrônico de Saúde (EHR). Outro exemplo é o trabalho de Acosta et al. (2013) que utiliza o *crowdsourcing* com usuários especialistas e não especialistas para aprimorar a qualidade de conjuntos de dados conectados.

- **Incluir outros critérios de qualidade:** Analisar outros critérios de qualidade que possam ser incluídos no PCD para avaliar a qualidade dos conjuntos de dados. Isso enriquecerá mais ainda as informações contidas no PCD e também ajudará aos consumidores na seleção de conjuntos de dados adequados para uma determinada tarefa.
- **Recomendação de vocabulários para cada metadado estrutural do conjunto de dados:** Incluir no PCD a recomendação de vocabulários para cada metadado estrutural do conjunto de dados, promovendo maior facilidade para entendimento e reutilização do conjunto de dados.
- **Acoplamento da abordagem a um catálogo/portal de conjuntos de dados:** Vincular a abordagem proposta nesta dissertação a um catálogo/portal de conjuntos de dados que permitirá a geração automática de PCDs, fazendo com que seus usuários tenham acesso a metadados enriquecidos semanticamente com melhores descrições sobre os conjuntos de dados.
- **Novos experimentos:** Pretende-se realizar novos experimentos com usuários especialistas e conjuntos de dados pertencentes a uma maior variedade de domínios. Podendo também ser avaliada uma nova versão da abordagem que inclua outros critérios de qualidade e, se possível, também considere o *feedback/crowdsourcing* dos usuários. Nesses novos experimentos também poderá ser avaliado se o uso da ferramenta proposta facilita a localização, compreensão, processamento e reuso dos conjuntos de dados.

REFERÊNCIAS

ABELE, A. (2016) Linked Data Profiling: Identifying the Domain of Datasets Based on Data Content and Metadata, In: 25th International Conference Companion on World Wide Web. Canada, p. 287-291.

ACOSTA, M.; ZAVERI, A.; SIMPERL, E.; KONTOKOSTAS, D.; AUER, S.; LEHMANN, J. (2013) Crowdsourcing Linked Data Quality Assessment. In: 12th International Semantic Web Conference (ISWC). Australia, p. 260-276.

ASSAF, A.; SENART, A. (2012) Data Quality Principles in the Semantic Web. In: 2012 IEEE Sixth International Conference on Semantic Computing (ICSC). Italy, p. 226-229.

ASSAF, A.; SENART, A.; TRONCY, R. (2016) An Objective Assessment Framework &Tool for Linked Data: Enriching Dataset Profiles with Quality Indicators. In: International Journal on Semantic Web and Information Systems (IJSWIS), Special Issue on Dataset Profiling and Federated Search for Linked Data, Vol. 12, N° 3, p. 111-133, ISSN: 1552-6283.

ASSAF, A.; TRONCY, R.; SENART, A. (2015) Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In: 24th International Conference on World Wide Web, Italy, p. 159-162.

BATISTA, M. C. M.; SALGADO, A. C. (2007) Information Quality Measurement in Data Integration Schemas. In: Workshop on Quality in Databases (QDB'07), at the VLDB'07 conference, ACM. Austria, p. 61–72.

BAEZA-YATES, R.; RIBEIRO-NETO, B. (1999) Modern Information Retrieval. Addison-Wesley, First Edition.

BERNERS-LEE T.; HENDLER J.; LASSILIA O. (2001) The Semantic Web. Scientific American, Vol. 284, N° 5, p. 34–44. http://dx.doi.org/10.1038/scientificamerican0501-34.

BERNSTEIN, M. S.; LITTLE, G.; MILLER, R. C.; HARTMANN, B.; ACKERMAN, M. S.; KARGER, D. R.; CROWELL, D; PANOVICH, K. (2010) Soylent: a word processor with a crowd inside. In: 23nd annual ACM symposium on User interface software and technology, United States, p. 313–322.

CHAKKARWAR, V. A.; JOSHI, A. (2016) Semantic Web Mining using RDF Data. In: International Journal of Computer Applications. Published by Foundation of Computer Science (FCS). United States, Vol. 133, N° 10, p. 14-19.

CLARKE, M.; HARLEY, P. (2014) How smart is your content? Using semantic enrichment to improve your user experience and your bottom line. Science Editor, Vol. 37, N°. 2, p. 40–44.

DIAMANTINI, C.; BOUDJLIDA, N. (2006) About Semantic Enrichment of Strategic Data Models as Part of Enterprise Models. In: 4th Business Process Management Workshops. Lecture Notes in Computer Science, Springer-Verlag. Austria, Vol. 4103, p. 348-359.

ELLEFI, M. B.; BELLAHSENE, Z.; SCHARFFE, F.; TODOROV, K. (2014) Towards Semantic Dataset Profiling In: International Workshop on Dataset Profiling & Federated Search for Linked Data co-located with the 11th Extended Semantic Web Conference. Greece.

ELLEFI, M. B.; BELLAHSENE, Z.; TODOROV, K. (2015) Datavore: a vocabulary recommender tool assisting Linked Data modeling, In: 14th International Semantic Web Conference, Posters and Demonstrations Track. United States.

FALBO, R.; QUIRINO, G.; NARDI, J.; BARCELLOS, M.; GUIZZARDI, G.; GUARINO, N.; LONGO, A.; LIVIERI, B. (2016). An Ontology Pattern Language for Service Modeling. In: 31st Annual ACM Symposium on Applied Computing, April 04-08, 2016. Italy, p. 321-326

FETAHU, B.; DIETZE, S.; PEREIRA NUNES, B.; ANTONIO CASANOVA, M.; TAIBI, D.; NEJDL, W. (2014) A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. In: Presutti V., d'Amato C., Gandon F., d'Aquin M., Staab S., Tordai A. (eds) The Semantic Web: Trends and Challenges. ESWC 2014. Lecture Notes in Computer Science, Springer, Cham. Vol. 8465, p. 519-534.

FILETO, R.; BOGORNY, V.; MAY, C.; KLEIN, D. (2015) Semantic Enrichment and Analysis of Movement Data: Probably it is just Starting! In: Newsletter SIGSPATIAL Special. United States. Vol. 7. N° 1, p. 11-18.

FLEMMING, A. (2011). Quality Characteristics of Linked Data Publishing Datasources. Master's Thesis, Humboldt-Universität zu Berlin, Institut für Informatik.

GRUBER, T. R. (1993) A Translation Approach to Portable Ontology Specifications. Journal Knowledge Acquisition - Special issue: Current issues in knowledge modeling. Vol. 5, N° 2, p. 199 – 220.

JOUDREY, D. N.; TAYLOR, A. G. (2017). The Organization of Information (Library and Information Science Text), 4th Edition., Libraries Unlimited.

LALITHSENA, S.; HITZLER, P.; SHETH, A. P.; JAIN, P. (2013). Automatic Domain Identification for Linked Open Data. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies. United States, p. 205-212.

LAUFER, C. (2015) Guia de Web Semântica. Centro de Estudos sobre Tecnologia Web – CeWeb. br. Disponível em: http://ceweb.br/guias/web-semantica/ Último Acesso: 10 de dezembro de 2017.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (2017) Data on the Web Best Practices. The World Wide Web Consortium (W3C) recommendation. Versão: https://www.w3.org/TR/2017/REC-dwbp-20170131/ Último Acesso: 10 de dezembro de 2017.

MAALI, F.; ERICKSON, J.; ARCHER, P. (2014). Data Catalog Vocabulary (DCAT). The World Wide Web Consortium (W3C) recommendation. Versão: https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/ Último Acesso: 10 de dezembro de 2017.

MANNENS, E.; TRONCY, R.; BRAECKMAN, K.; VAN DEURSEN, D.; VAN LANCKER, W.; DE SUTTER, R.; VAN DE WALLE, R. (2009) Automatic Information Enrichment in News Production. In: 10th International Workshop on Image Analysis for Multimedia Interactive Services. United Kingdom, p. 61-64.

MENDONÇA, A.; MACIEL, P.; SOUZA, D.; SALGADO, A. (2015). CORE - A Context-based Approach for Rewriting User Queries. In: 17th International Conference on Enterprise Information Systems. Spain, p. 391-398. ISBN 978-989-758-096-3.

NAUMANN, F.; ROLKER, C. (2000) Assessment Methods for Information Quality Criteria. In: 5th Conference on International Quality (IQ). United States, p. 148-162.

NEUMANN, T; WEIKUM, G. (2010) The RDF-3X Engine for Scalable Management of RDF Data. In: The VLDB Journal — The International Journal on Very Large Data Bases archive. Vol. 19, N° 1, February 2010, p. 91-113 .

NICHOLS, D. M.; CHAN, C. H.; BAINBRIDGE, D.; MCKAY, D.; TWIDALE, M. B. (2008). A Light Weight Metadata Quality Tool. In: 8th ACM/IEEE-CS Joint Conference on Digital Libraries. United States, p. 385-388.

OLIVEIRA, M. I. S.; OLIVEIRA, L. A.; LIMA, G. F. B; LÓSCIO, B. F. (2016). Enabling a Unified View of Open Data Catalogs, In: 18th International Conference on Enterprise Information Systems (ICEIS). Italy, p. 230-239.

OUKSILI, H.; KEDAD, Z.; LOPES, S. (2014) Theme Identification in RDF Graphs, In: International Conference on Model and Data Engineering (MEDI). Cyprus, p. 321-329.

PARINOV, S. (2014) Semantic Enrichment of Research Outputs Metadata: New CRIS Facilities for Authors. In: Closs S., Studer R., Garoufallou E., Sicilia MA. (eds) Metadata and Semantics Research. MTSR 2014. Communications in Computer and Information Science, Springer, Cham. Vol. 478, p. 206-217.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. (2002) Data Quality Assessment. In: Communications of the ACM - Supporting community and building social capital, Vol. 45, N°.4, April 2002, p. 211-218.

PRUD'HOMMEAUX, E.; BUIL-ARANDA, C. (2013) SPARQL 1.1 Federated Query. The World Wide Web Consortium (W3C) recommendation. Versão: https://www.w3.org/TR/2013/REC-sparql11-federated-query-20130321/ Último Acesso: 10 de dezembro de 2017.

SCHAIBLE, J.; GOTTRON, T.; SCHEGLMANN, S.; SCHERP, A. (2013) LOVER: Support for Modeling Data Using Linked Open Vocabularies, In: Joint EDBT/ICDT Workshops. Italy, p. 89-92.

SHAHI, D. (2015) Apache Solr: a practical approach to enterprise search. Apress, First Edition. ISBN: 978-1-4842-1071-0

SILVA NETO, E. C.; LÓSCIO, B. F.; SALGADO, A. C. (2016) Geração de um Perfil de Qualidade para Fontes de Dados Dinâmicas. In: 31° Simpósio Brasileiro de Banco de Dados - SBBD. Brasil, p. 52-63.

VAN DER BIJ, S.; KHAN, N.; TEN VEEN, P.; DE BAKKER, D. H.; VERHEIJ, R. A. (2017). Improving the Quality of EHR Recording in Primary Care: a data quality feedback tool. In: Journal of the American Medical Informatics Association, Vol. 24, N° 1, p. 81–87.

WANG, R. Y.; STRONG, D. M. (1996) Beyond Accuracy: What data quality means to data consumers. In: Journal of Management Information Systems. M. E. Sharpe, Inc. United States. Vol. 12, N°. 4, p. 5–33. ISSN 0742-1222.

W3C. (2014) Resource Description Framework (RDF). https://www.w3.org/2001/sw/wiki/RDF Último Acesso: 10 de janeiro de 2018.

ZAVERI, A.; RULA, A.; MAURINO, A.; PIETROBON, R.; LEHMANN, J.; AUER, S.; HITZLER, P. (2013). Quality Assessment Methodologies for Linked Open Data. In: Semantic Web Journal.

APÊNDICE A – *GOLD STANDARDS* DEFINIDOS PELOS ESPECIALISTAS

	Gold Standard Domínio	Gold Standard Vocabulários
Conjunto de Dados 01	MeanOfTransportation/Automobile	Schema (http://schema.org)
Conjunto de Dados 02	MeanOfTransportation/Automobile	Schema (http://schema.org)
Conjunto de Dados 03	MeanOfTransportation/Automobile, Activity/Sales	Used Cars Ontology Metadata (http://purl.org/uco/ns#)
Conjunto de Dados 04	MeanOfTransportation/Automobile	Vehicle Sales Ontology (http://www.heppnetz.de/ontologies/vso/ns);
Conjunto de Dados 05	MeanOfTransportation/Automobile	Vehicle Sales Ontology (http://www.heppnetz.de/ontologies/vso/ns);
Conjunto de Dados 06	MeanOfTransportation/Automobile	UCO(http://ontologies.makolab.com/uco/ns.html); geo(http://www.w3.org/2003/01/geo/wgs84_pos);
Conjunto de Dados 07	MeanOfTransportation/Automobile	Time Ontology(http://www.w3.org/2006/time); Vehicle Sales Ontology (http://www.heppnetz.de/ontologies/vso/ns);
Conjunto de Dados 08	MeanOfTransportation/Automobile	dbpedia - (http://mappings.dbpedia.org/server/ontology/);
Conjunto de Dados 09	MeanOfTransportation/Automobile, Place/PopulatedPlace/Settlement/City	http://data.ordnancesurvey.co.uk/ontology/admingeo/ Vehicle Sales Ontology (http://www.heppnetz.de/ontologies/vso/ns);

	Gold Standard Domínio	Gold Standard Vocabulários
Conjunto de Dados 10	MeanOfTransportation/Automobile	Vehicle Sales Ontology (http://www.heppnetz.de/ontologies/vso/ns);
Conjunto de Dados 11	Musical Work	Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 12	Musical Work	Music Vocabulary - (http://www.kanzaki.com/ns/music); Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 13	Musical Work	Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 14	Musical Work	Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 15	Musical Work	Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 16	Musical Work	Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 17	Musical Work	Music Vocabulary - (http://www.kanzaki.com/ns/music); Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 18	Musical Work	Gnd ontology (https://d-nb.info/standards/elementset/gnd);
Conjunto de Dados 19	Musical Work	Music Vocabulary - (http://www.kanzaki.com/ns/music); Music Ontology - (http://purl.org/ontology/mo/);
Conjunto de Dados 20	Musical Work	dbpedia - (http://mappings.dbpedia.org/server/ontology/);
Conjunto de Dados 21	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntolog);
Conjunto de Dados 22	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);
Conjunto de Dados 23	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);
Conjunto de Dados 24	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);

	Gold Standard Domínio	Gold Standard Vocabulários
Conjunto de Dados 25	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);
Conjunto de Dados 26	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology); The SEAS Statistics ontology - (ttps://w3id.org/seas/StatisticsOntology);
Conjunto de Dados 27	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);
Conjunto de Dados 28	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);
Conjunto de Dados 29	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);
Conjunto de Dados 30	Software/Video Game	The Video Game Ontology - (http://purl.org/net/VideoGameOntology);

APÊNDICE B – EXEMPLOS DE PCDS GERADOS PELO PROTÓTIPO

Exemplo 01 – PCD: Conjunto de Dados "*Automobile Dataset*" (https://www.kaggle.com/toramky/automobile-dataset)

```
@prefix :       <http://example/ns#> .
@prefix void:   <http://rdfs.org/ns/void#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix dqv:    <http://www.w3.org/ns/dqv#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms:  <http://purl.org/dc/terms/> .
@prefix skos:   <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:   <http://www.w3.org/ns/dcat#> .

:dataset.ID1291.Kaggle
    a      dcat:Dataset ;
    dcterms:identifier "1291" ;
    dcterms:title "Automobile Dataset" ;
    dcterms:description "### Context  This dataset consist of data From 1985 Ward's Automotive Yearbook. Here are the sources  Sources:  1) 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.  2) Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038  3) Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037  ### Content  This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process 'symboling'. A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.  The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.  Note: Several of the attributes in the database could be used as a 'class' attribute.  ### Inspiration  Please bring it on whatever inferences you can get it." ;
    dcat:keyword "Automobiles" , "Transport" ;
    dcat:theme :theme-Automobile ;
    void:vocabulary "http://schema.mobivoc.org/" ,
"http://linkeddata.finki.ukim.mk/lod/ontology/tao#" ;
    dcat:landingPage "https://www.kaggle.com/toramky/automobile-dataset" ;
    dcterms:modified "2017-05-24T04:45:13.673Z"^^xsd:dateTime ;
    dcterms:issued "2017-05-20T09:27:56.223Z"^^xsd:dateTime ;
    dcterms:publisher "Ramakrishnan Srinivasan" ;
    owl:versionInfo "2"^^xsd:int ;
    dcat:distribution :distribution2-DataDownload.csv , :distribution1-DataDownload.zip ;
    skos:inScheme
        [ void:properties "26" ;
```

```
void:property
      [ rdfs:label "body-style"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "highway-mpg"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "aspiration"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "peak-rpm"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "compression-ratio"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "height"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "length"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "engine-location"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "engine-type"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "bore"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "make"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "engine-size"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "symboling"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "city-mpg"^^xsd:string ;
```

```
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "drive-wheels"^^xsd:string ;
            dcterms:type "string"
           ] ;
      void:property
           [ rdfs:label "horsepower"^^xsd:string ;
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "num-of-doors"^^xsd:string ;
            dcterms:type "string"
           ] ;
      void:property
           [ rdfs:label "curb-weight"^^xsd:string ;
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "width"^^xsd:string ;
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "price"^^xsd:string ;
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "wheel-base"^^xsd:string ;
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "num-of-cylinders"^^xsd:string ;
            dcterms:type "string"
           ] ;
      void:property
           [ rdfs:label "fuel-type"^^xsd:string ;
            dcterms:type "string"
           ] ;
      void:property
           [ rdfs:label "normalized-losses"^^xsd:string ;
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "stroke"^^xsd:string ;
            dcterms:type "numeric"
           ] ;
      void:property
           [ rdfs:label "fuel-system"^^xsd:string ;
            dcterms:type "string"
           ]
    ] ;
dqv:hasQualityMetadata  :dimensionComprehensibility , :dimensionProcessability ;
dcterms:created "2018-02-04T03:02:17.324Z"^^xsd:dateTime .
```

```
:theme-Automobile
    a      dcat:theme ;
    rdfs:label "Automobile"@en ;
    void:uriSpace "http://dbpedia.org/ontology/Automobile" .

:distribution2-DataDownload.csv
    a      dcat:Distribution ;
    dcterms:format "csv" ;
    dcat:byteSize "25070 bytes" ;
    dcat:downloadURL                    "https://www.kaggle.com//toramky/automobile-
dataset/downloads/Automobile_data.csv" ;
    dcat:mediaType "DataDownload" .

:distribution1-DataDownload.zip
    a      dcat:Distribution ;
    dcterms:format "zip" ;
    dcat:byteSize "5015 bytes" ;
    dcat:downloadURL                    "https://www.kaggle.com//toramky/automobile-
dataset/downloads/automobile-dataset.zip/2" ;
    dcat:mediaType "DataDownload" .

:dimensionComprehensibility
    a      dqv:Dimension ;
    rdfs:label "Comprehensibility"@en ;
    dqv:value "0.83"^^xsd:float ;
    skos:definition "Represented the degree to which a dataset presents information that
promotes/facilitates its understanding by human users."@en .

:dimensionProcessability
    a      dqv:Dimension ;
    rdfs:label "Processability"@en ;
    dqv:value "0.6"^^xsd:float ;
    skos:definition "Represents the degree to which a dataset is processable by machines or software
agents."@en .
```

Exemplo 02 – PCD: Conjunto de Dados "*Vehicle Collisions in NYC, 2015-Present*" (https://www.kaggle.com/nypd/vehicle-collisions)

```
@prefix :       <http://example/ns#> .
@prefix void:   <http://rdfs.org/ns/void#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix dqv:    <http://www.w3.org/ns/dqv#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix skos:   <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:   <http://www.w3.org/ns/dcat#> .

:dataset.ID679.Kaggle
    a      dcat:Dataset ;
    dcterms:identifier "679" ;
    dcterms:title "Vehicle Collisions in NYC, 2015-Present" ;
    dcterms:description "# Content  The motor vehicle collision database includes the date and time, location (as borough, street names, zip code and latitude and longitude coordinates), injuries and fatalities, vehicle number and types, and related factors for all 65,500 collisions in New York City during 2015 and 2016.  # Acknowledgements  The vehicle collision data was collected by the NYPD and published by NYC OpenData." ;
    dcat:keyword "Walking" , "Road Transport" , "Exercise" , "Transport" ;
    dcat:theme :theme-Mean_Of_Transportation ;
    void:vocabulary "http://schema.org/" , "http://linkeddata.finki.ukim.mk/lod/ontology/tao#" ;
    dcat:landingPage "https://www.kaggle.com/nypd/vehicle-collisions" ;
    dcterms:modified "2017-03-09T05:14:51.17Z"^^xsd:dateTime ;
    dcterms:issued "2017-01-18T17:55:44.147Z"^^xsd:dateTime ;
    dcterms:publisher "NYPD" ;
    owl:versionInfo "2"^^xsd:int ;
    dcat:distribution :distribution2-DataDownload.csv , :distribution1-DataDownload.zip ;
    skos:inScheme
          [ void:properties "29" ;
           void:property
                [ rdfs:label "ON STREET NAME"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "OFF STREET NAME"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "ZIP CODE"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "VEHICLE 4 FACTOR"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "LONGITUDE"^^xsd:string ;
                  dcterms:type "string"
                ] ;
```

```
void:property
      [ rdfs:label "PEDESTRIANS KILLED"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "CYCLISTS KILLED"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "PERSONS KILLED"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "UNIQUE KEY"^^xsd:string ;
        dcterms:type "numeric"
      ] ;
void:property
      [ rdfs:label "VEHICLE 5 TYPE"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "VEHICLE 2 FACTOR"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "TIME"^^xsd:string ;
        dcterms:type "dateTime"
      ] ;
void:property
      [ rdfs:label "VEHICLE 1 TYPE"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "VEHICLE 3 TYPE"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "LOCATION"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "CROSS STREET NAME"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "VEHICLE 5 FACTOR"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "BOROUGH"^^xsd:string ;
        dcterms:type "string"
      ] ;
void:property
      [ rdfs:label "LATITUDE"^^xsd:string ;
```

```
                dcterms:type "string"
               ] ;
          void:property
               [ rdfs:label "CYCLISTS INJURED"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "MOTORISTS INJURED"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "PERSONS INJURED"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "PEDESTRIANS INJURED"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "DATE"^^xsd:string ;
                dcterms:type "dateTime"
               ] ;
          void:property
               [ rdfs:label "MOTORISTS KILLED"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "VEHICLE 1 FACTOR"^^xsd:string ;
                dcterms:type "string"
               ] ;
          void:property
               [ rdfs:label "VEHICLE 3 FACTOR"^^xsd:string ;
                dcterms:type "string"
               ] ;
          void:property
               [ rdfs:label "VEHICLE 2 TYPE"^^xsd:string ;
                dcterms:type "string"
               ] ;
          void:property
               [ rdfs:label "VEHICLE 4 TYPE"^^xsd:string ;
                dcterms:type "string"
               ]
         ] ;
    dqv:hasQualityMetadata  :dimensionComprehensibility , :dimensionProcessability ;
    dcterms:created "2018-02-04T03:05:03.855Z "^^xsd:dateTime .

:theme-Mean_Of_Transportation
    a      dcat:theme ;
    rdfs:label "Mean_Of_Transportation"@en ;
    void:uriSpace "http://dbpedia.org/ontology/MeanOfTransportation" .

:distribution2-DataDownload.csv
    a      dcat:Distribution ;
    dcterms:format "csv" ;
```

```
    dcat:byteSize "16583744 bytes" ;
    dcat:downloadURL "https://www.kaggle.com//nypd/vehicle-collisions/downloads/database.csv" ;
    dcat:mediaType "DataDownload" .

:distribution1-DataDownload.zip
    a      dcat:Distribution ;
    dcterms:format "zip" ;
    dcat:byteSize "16583744 bytes" ;
    dcat:downloadURL          "https://www.kaggle.com//nypd/vehicle-collisions/downloads/vehicle-
collisions.zip/2" ;
    dcat:mediaType "DataDownload" .

:dimensionComprehensibility
    a      dqv:Dimension ;
    rdfs:label "Comprehensibility"@en ;
    dqv:value "0.83"^^xsd:float ;
    skos:definition "Represented the degree to which a dataset presents information that
promotes/facilitates its understanding by human users."@en .

:dimensionProcessability
    a      dqv:Dimension ;
    rdfs:label "Processability"@en ;
    dqv:value "0.6"^^xsd:float ;
    skos:definition "Represents the degree to which a dataset is processable by machines or software
agents."@en .
```

Exemplo 03 – PCD: Conjunto de Dados "*Spotify's Worldwide Daily Song Ranking*" (https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking)

```
@prefix :       <http://example/ns#> .
@prefix void:   <http://rdfs.org/ns/void#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix dqv:    <http://www.w3.org/ns/dqv#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix skos:   <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:   <http://www.w3.org/ns/dcat#> .

:dataset.ID2089.Kaggle
    a      dcat:Dataset ;
    dcterms:identifier "2089" ;
    dcterms:title "Spotify's Worldwide Daily Song Ranking" ;
    dcterms:description "### Context  Music streaming is ubiquitous. Currently, Spotify plays an important part on that. This dataset enable us to explore how artists and songs' popularity varies in time. ### Content  This dataset contains the daily ranking of the 200 most listened songs in 53 countries from 2017 and 2018 by Spotify users. It contains more than 2 million rows, which comprises 6629 artists, 18598 songs for a total count of one hundred five billion streams count. The data spans from 1st January 2017 to 9th January 2018 and will be kept up-to-date on following versions. It has been collected from Spotify's regional chart [data][1]. ### Inspiration  - Can you predict what is the rank position or the number of streams a song will have in the future? - How long does songs 'resist' on the top 3, 5, 10, 20 ranking? - What are the signs of a song that gets into the top rank to stay? - Do continents share same top ranking artists or songs? - Are people listening to the very same top ranking songs on countries far away from each other? - How long time does a top ranking song takes to get into the ranking of neighbor countries? ### Example  To start out, you can take a look into a simple Kernel I have made in order to read the data, filter data from a song, plot is temporal tendency per country than make a simple forecast of the its streams count [here][2]. ### Crawler  The crawler used to collect this data can be found [here][3]. [1]: https://spotifycharts.com/regional [2]: https://www.kaggle.com/edumucelli/initial-analysis-and-forecast-example [3]: https://github.com/edumucelli/spotify-worldwide-ranking" ;
    dcat:keyword "Music" , "Performing Art" ;
    dcat:theme :theme-Musical_Work ;
    void:vocabulary "http://www.kanzaki.com/ns/music" , "http://purl.org/ontology/mo/" ;
    dcat:landingPage "https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking" ;
    dcterms:modified "2018-01-12T20:32:57.537Z"^^xsd:dateTime ;
    dcterms:issued "2017-08-20T20:02:54.577Z"^^xsd:dateTime ;
    dcterms:publisher "Eduardo" ;
    owl:versionInfo "3"^^xsd:int ;
    dcat:distribution :distribution2-DataDownload.csv , :distribution1-DataDownload.zip ;
    skos:inScheme
        [ void:properties "7" ;
          void:property
              [ rdfs:label "URL"^^xsd:string ;
                dcterms:type "string"
              ] ;
          void:property
              [ rdfs:label "Track Name"^^xsd:string ;
                dcterms:type "string"
              ] ;
```

```
            void:property
                [ rdfs:label "Position"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
            void:property
                [ rdfs:label "Streams"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
            void:property
                [ rdfs:label "Artist"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "Date"^^xsd:string ;
                  dcterms:type "dateTime"
                ] ;
            void:property
                [ rdfs:label "Region"^^xsd:string ;
                  dcterms:type "string"
                ]
           ] ;
    dqv:hasQualityMetadata  :dimensionComprehensibility , :dimensionProcessability ;
    dcterms:created "2018-02-04T03:09:46.785Z "^^xsd:dateTime .

:theme-Musical_Work
    a      dcat:theme ;
    rdfs:label "Musical_Work"@en ;
    void:uriSpace "http://dbpedia.org/ontology/MusicalWork" .

:distribution2-DataDownload.csv
    a      dcat:Distribution ;
    dcterms:format "csv" ;
    dcat:byteSize "368989292 bytes" ;
    dcat:downloadURL                "https://www.kaggle.com//edumucelli/spotifys-worldwide-daily-song-
ranking/downloads/data.csv" ;
    dcat:mediaType "DataDownload" .

:distribution1-DataDownload.zip
    a      dcat:Distribution ;
    dcterms:format "zip" ;
    dcat:byteSize "45167371 bytes" ;
    dcat:downloadURL                "https://www.kaggle.com//edumucelli/spotifys-worldwide-daily-song-
ranking/downloads/spotifys-worldwide-daily-song-ranking.zip/3" ;
    dcat:mediaType "DataDownload" .

:dimensionComprehensibility
    a      dqv:Dimension ;
    rdfs:label "Comprehensibility"@en ;
    dqv:value "0.83"^^xsd:float ;
    skos:definition "Represented the degree to which a dataset presents information that
promotes/facilitates its understanding by human users."@en .

:dimensionProcessability
    a      dqv:Dimension ;
```

```
    rdfs:label "Processability"@en ;
    dqv:value "0.6"^^xsd:float ;
    skos:definition "Represents the degree to which a dataset is processable by machines or software agents."@en .
```

Exemplo 04 – PCD: Conjunto de Dados "*17 Years of Resident Advisor Reviews*" (https://www.kaggle.com/marcschroeder/17-years-of-resident-advisor-reviews)

```
@prefix :       <http://example/ns#> .
@prefix void:    <http://rdfs.org/ns/void#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix dqv:    <http://www.w3.org/ns/dqv#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms:  <http://purl.org/dc/terms/> .
@prefix skos:    <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:    <http://www.w3.org/ns/dcat#> .

:dataset.ID8482.Kaggle
    a      dcat:Dataset ;
    dcterms:identifier "8482" ;
    dcterms:title "17 Years of Resident Advisor Reviews" ;
    dcterms:description "### Context  Started in 2001, [Resident Advisor][1] (RA) has become the web's
largest resource for information about underground electronic music around the world. The site
maintains a huge database of music and tech reviews, artists, labels, news, podcasts, and events.  ###
Content  Gathered using a web-scraping [script][2], the dataset below is of the site's entire collection of
music reviews from the start of the site through the end of 2017. It contains the following fields:  -
Release Type (album or single)  - Artist Release  - Title  - Label  - Release Month  - Release Year  - Style
(genres of release listed by RA)  - Rating (score out of 5 given by RA)  - Date of Review  - Review Author  -
Body of Review  - Release Tracklist  ### Acknowledgements  Thanks to the RA team for the journalism
over the years, and for (hopefully) being cool with this dataset being published here.  [1]:
http://residentadvisor.com%22Resident%20Advisor%22  [2]:
https://github.com/schroedermarc/RAReviewsData/blob/master/scraper.py" ;
    dcat:keyword "Music" , "Journalism" , "Performing Arts" , "News Agencie" ;
    dcat:theme :theme-Musical_Work ;
    void:vocabulary "http://purl.org/ontology/mo/" ;
    dcat:landingPage "https://www.kaggle.com/marcschroeder/17-years-of-resident-advisor-reviews" ;
    dcterms:modified "2018-01-03T01:19:55.77Z"^^xsd:dateTime ;
    dcterms:issued "2018-01-03T01:19:55.77Z"^^xsd:dateTime ;
    dcterms:publisher "MarcSchroeder" ;
    owl:versionInfo "1"^^xsd:int ;
    dcat:distribution :distribution2-DataDownload.csv , :distribution1-DataDownload.zip ;
    skos:inScheme
          [ void:properties "14" ;
           void:property
                [ rdfs:label "release_year"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
           void:property
                [ rdfs:label "style"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "artist"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
```

```
                [ rdfs:label "release_title"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "label"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "release_month"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "ra_review_id"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
            void:property
                [ rdfs:label "release_type"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "author"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "review_body"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "tracklist"^^xsd:string ;
                  dcterms:type "string"
                ] ;
            void:property
                [ rdfs:label "num_comments"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
            void:property
                [ rdfs:label "rating"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
            void:property
                [ rdfs:label "review_published"^^xsd:string ;
                  dcterms:type "dateTime"
                ]
           ] ;
    dqv:hasQualityMetadata  :dimensionComprehensibility , :dimensionProcessability ;
    dcterms:created "2018-02-04T03:17:02.125Z "^^xsd:dateTime .

:theme-Musical_Work
    a      dcat:theme ;
    rdfs:label "Musical_Work"@en ;
    void:uriSpace "http://dbpedia.org/ontology/MusicalWork" .

:distribution2-DataDownload.csv
    a      dcat:Distribution ;
```

```
    dcterms:format "csv" ;
    dcat:byteSize "36617068 bytes" ;
    dcat:downloadURL "https://www.kaggle.com//marcschroeder/17-years-of-resident-advisor-reviews/downloads/RA_cleaned.csv" ;
    dcat:mediaType "DataDownload" .

:distribution1-DataDownload.zip
    a dcat:Distribution ;
    dcterms:format "zip" ;
    dcat:byteSize "15266902 bytes" ;
    dcat:downloadURL "https://www.kaggle.com//marcschroeder/17-years-of-resident-advisor-reviews/downloads/17-years-of-resident-advisor-reviews.zip/1" ;
    dcat:mediaType "DataDownload" .

:dimensionComprehensibility
    a dqv:Dimension ;
    rdfs:label "Comprehensibility"@en ;
    dqv:value "0.83"^^xsd:float ;
    skos:definition "Represented the degree to which a dataset presents information that promotes/facilitates its understanding by human users."@en .

:dimensionProcessability
    a dqv:Dimension ;
    rdfs:label "Processability"@en ;
    dqv:value "0.6"^^xsd:float ;
    skos:definition "Represents the degree to which a dataset is processable by machines or software agents."@en .
```

Exemplo 05 – PCD: Conjunto de Dados "*Video Game Sales*" (https://www.kaggle.com/gregorut/videogamesales)

```
@prefix :       <http://example/ns#> .
@prefix void:   <http://rdfs.org/ns/void#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix dqv:    <http://www.w3.org/ns/dqv#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms:  <http://purl.org/dc/terms/> .
@prefix skos:   <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:   <http://www.w3.org/ns/dcat#> .

:dataset.ID284.Kaggle
    a      dcat:Dataset ;
    dcterms:identifier "284" ;
    dcterms:title "Video Game Sales" ;
    dcterms:description "This dataset contains a list of video games with sales greater than 100,000 copies.  It was generated by a scrape of [vgchartz.com][1].  Fields include  * Rank - Ranking of overall sales  * Name - The games name  * Platform - Platform of the games release (i.e. PC,PS4, etc.)  * Year - Year of the game's release  * Genre - Genre of the game  * Publisher - Publisher of the game  * NA_Sales - Sales in North America (in millions)  * EU_Sales - Sales in Europe (in millions)  * JP_Sales - Sales in Japan (in millions)  * Other_Sales - Sales in the rest of the world (in millions)  * Global_Sales - Total worldwide sales.  The script to scrape the data is available at https://github.com/GregorUT/vgchartzScrape. It is based on BeautifulSoup using Python. There are 16,598 records.  2 records were dropped due to incomplete information.    [1]: http://www.vgchartz.com/" ;
    dcat:keyword "Video Games" , "Toy" , "Game" ;
    dcat:theme :theme-Software ;
    void:vocabulary "http://purl.org/net/VideoGameOntology" ;
    dcat:landingPage "https://www.kaggle.com/gregorut/videogamesales" ;
    dcterms:modified "2016-10-26T09:10:49.853Z"^^xsd:dateTime ;
    dcterms:issued "2016-10-26T08:17:30.23Z"^^xsd:dateTime ;
    dcterms:publisher "GregorySmith" ;
    owl:versionInfo "2"^^xsd:int ;
    dcat:distribution :distribution2-DataDownload.csv , :distribution1-DataDownload.zip ;
    skos:inScheme
          [ void:properties "11" ;
           void:property
                [ rdfs:label "Platform"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "Year"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
           void:property
                [ rdfs:label "Genre"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "Publisher"^^xsd:string ;
```

```
                dcterms:type "string"
               ] ;
          void:property
               [ rdfs:label "EU_Sales"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "JP_Sales"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "Other_Sales"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "Rank"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "Global_Sales"^^xsd:string ;
                dcterms:type "numeric"
               ] ;
          void:property
               [ rdfs:label "Name"^^xsd:string ;
                dcterms:type "string"
               ] ;
          void:property
               [ rdfs:label "NA_Sales"^^xsd:string ;
                dcterms:type "numeric"
               ]
         ] ;
    dqv:hasQualityMetadata  :dimensionComprehensibility , :dimensionProcessability ;
    dcterms:created "2018-02-04T03:22:08.765Z "^^xsd:dateTime .

:theme-Software
    a      dcat:theme ;
    rdfs:label "Software"@en ;
    void:uriSpace "http://dbpedia.org/ontology/Software" .

:distribution2-DataDownload.csv
    a      dcat:Distribution ;
    dcterms:format "csv" ;
    dcat:byteSize "412254 bytes" ;
    dcat:downloadURL "https://www.kaggle.com//gregorut/videogamesales/downloads/vgsales.csv" ;
    dcat:mediaType "DataDownload" .

:distribution1-DataDownload.zip
    a      dcat:Distribution ;
    dcterms:format "zip" ;
    dcat:byteSize "412254 bytes" ;
    dcat:downloadURL
"https://www.kaggle.com//gregorut/videogamesales/downloads/videogamesales.zip/2" ;
    dcat:mediaType "DataDownload" .
```

```
:dimensionComprehensibility
    a       dqv:Dimension ;
    rdfs:label "Comprehensibility"@en ;
    dqv:value "0.83"^^xsd:float ;
    skos:definition "Represented the degree to which a dataset presents information that promotes/facilitates its understanding by human users."@en .

:dimensionProcessability
    a       dqv:Dimension ;
    rdfs:label "Processability"@en ;
    dqv:value "0.6"^^xsd:float ;
    skos:definition "Represents the degree to which a dataset is processable by machines or software agents."@en .
```

Exemplo 06 – PCD: Conjunto de Dados "*Steam Video Games*" (https://www.kaggle.com/tamber/steam-video-games)

```
@prefix :       <http://example/ns#> .
@prefix void:   <http://rdfs.org/ns/void#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix dqv:    <http://www.w3.org/ns/dqv#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix skos:   <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:   <http://www.w3.org/ns/dcat#> .

:dataset.ID915.Kaggle
    a      dcat:Dataset ;
    dcterms:identifier "915" ;
    dcterms:title "Steam Video Games" ;
    dcterms:description "# Context    Steam is the world's most popular PC Gaming hub, with over 6,000 games and a community of millions of gamers. With a massive collection that includes everything from AAA blockbusters to small indie titles, great discovery tools are a highly valuable asset for Steam. How can we make them better?   # Content  This dataset is a list of user behaviors, with columns: user-id, game-title, behavior-name, value. The behaviors included are 'purchase' and 'play'. The value indicates the degree to which the behavior was performed - in the case of 'purchase' the value is always 1, and in the case of 'play' the value represents the number of hours the user has played the game.   # Acknowledgements  This dataset is generated entirely from public Steam data, so we want to thank Steam for building such an awesome platform and community!   # Inspiration  The dataset is formatted to be compatible with [Tamber][1]. Build a Tamber engine and take it for a spin!   Combine our collaborative filter's results with your favorite Machine Learning techniques with Ensemble Learning, or make Tamber do battle with something else you've built.  Have fun, The Tamber Team      [1]: https://tamber.com" ;
    dcat:keyword "Video Games" , "Toy" , "Game" ;
    dcat:theme :theme-Software ;
    void:vocabulary "http://purl.org/net/VideoGameOntology" ;
    dcat:landingPage "https://www.kaggle.com/tamber/steam-video-games" ;
    dcterms:modified "2017-03-09T03:01:26.933Z"^^xsd:dateTime ;
    dcterms:issued "2017-03-04T01:02:12.53Z"^^xsd:dateTime ;
    dcterms:publisher "Tamber" ;
    owl:versionInfo "3"^^xsd:int ;
    dcat:distribution :distribution2-DataDownload.csv , :distribution1-DataDownload.zip ;
    skos:inScheme
          [ void:properties "5" ;
           void:property
                [ rdfs:label "151603712"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
           void:property
                [ rdfs:label "purchase"^^xsd:string ;
                  dcterms:type "string"
                ] ;
           void:property
                [ rdfs:label "1.0"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
```

```
            void:property
                [ rdfs:label "0"^^xsd:string ;
                  dcterms:type "numeric"
                ] ;
            void:property
                [ rdfs:label "The Elder Scrolls V Skyrim"^^xsd:string ;
                  dcterms:type "string"
                ]
           ] ;
    dqv:hasQualityMetadata  :dimensionComprehensibility , :dimensionProcessability ;
    dcterms:created "2018-02-04T03:29:00.395Z"^^xsd:dateTime .

:theme-Software
    a      dcat:theme ;
    rdfs:label "Software"@en ;
    void:uriSpace "http://dbpedia.org/ontology/Software" .

:distribution2-DataDownload.csv
    a      dcat:Distribution ;
    dcterms:format "csv" ;
    dcat:byteSize "1636145 bytes" ;
    dcat:downloadURL          "https://www.kaggle.com//tamber/steam-video-games/downloads/steam-
200k.csv" ;
    dcat:mediaType "DataDownload" .

:distribution1-DataDownload.zip
    a      dcat:Distribution ;
    dcterms:format "zip" ;
    dcat:byteSize "1636145 bytes" ;
    dcat:downloadURL "https://www.kaggle.com//tamber/steam-video-games/downloads/steam-video-
games.zip/3" ;
    dcat:mediaType "DataDownload" .

:dimensionComprehensibility
    a      dqv:Dimension ;
    rdfs:label "Comprehensibility"@en ;
    dqv:value "0.83"^^xsd:float ;
    skos:definition "Represented the degree to which a dataset presents information that
promotes/facilitates its understanding by human users."@en .

:dimensionProcessability
    a      dqv:Dimension ;
    rdfs:label "Processability"@en ;
    dqv:value "0.6"^^xsd:float ;
    skos:definition "Represents the degree to which a dataset is processable by machines or software
agents."@en .
```