# Uma Proposta de Perfil de Conjuntos de Dados na Web com Enriquecimento Semântico

**Natacha Targino[1], Damires Souza[2], Ana Carolina Salgado[1]**

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – Pernambuco – Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB)
João Pessoa – Paraíba - Brasil

{ntrsb, acs}@cin.ufpe.br, damires@ifpb.edu.br

***Abstract.*** *The lack of metadata to describe datasets published on the Web makes their location and access by search engines or applications more difficult. Providing a dataset profile facilitates communication between publishers and consumers and also the integrated use of datasets. This paper proposes an approach that describes datasets on the Web by the generation of a semantically enriched descriptive and structural metadata profile. The enrichment occurs by means of the knowledge domain identification of the dataset at hand and a vocabulary recommendation in order to semantically reference the data. This work presents some accomplished experiments that indicate the relevance of this enrichment.*

***Resumo.*** *A ausência de metadados para a descrição de conjuntos de dados publicados na Web dificulta sua localização e acesso por parte de mecanismos de busca ou aplicações. Prover um perfil do conjunto de dados facilita a comunicação entre publicadores e consumidores e o uso integrado dos conjuntos de dados. Este trabalho propõe uma abordagem que descreve conjuntos de dados na Web por meio da geração de um perfil de metadados descritivos e estruturais enriquecidos semanticamente. O enriquecimento ocorre por meio da identificação do domínio de conhecimento do conjunto de dados e da recomendação de vocabulários para referenciar semanticamente os dados. O trabalho apresenta alguns experimentos realizados que indicam a relevância deste enriquecimento.*

## 1. Introdução

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 publicados em diferentes distribuições (e.g., CSV[1]) que facilitam seu compartilhamento e reuso, se estiverem em formato aberto, por grupos de consumidores de dados, sejam eles humanos ou aplicações [Lóscio et al. 2017]. Entretanto,

[1] https://www.w3.org/TR/tabular-data-primer/

nem sempre os publicadores de dados e os consumidores de dados se conhecem. Portanto, é necessário fornecer algumas informações sobre os conjuntos de dados e distribuições que possam contribuir para a sua compreensão e reutilização. Isso normalmente é realizado por meio de metadados [Clarke e Harley 2014].

Os conjuntos de dados publicados geralmente não contêm descrição sobre conteúdo, pois ainda não é uma prática comum os publicadores de dados fornecerem metadados que representem corretamente o conteúdo de seus próprios 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. Em portais de dados abertos, é comum encontrar alguns metadados, entretanto, nem sempre são apresentados de forma estruturada com informações suficientes para o seu entendimento e processamento [Oliveira et al., 2016].

Como uma boa prática, o *World Wide Web Consortium* (W3C)[2] recomenda o enriquecimento de dados publicados. O enriquecimento de dados compreende um conjunto de processos que podem ser utilizados para aumentar, refinar ou melhorar dados brutos ou processados anteriormente [Lóscio et al. 2017]. Isso pode ser realizado em nível de metadados também, 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).

Nesse contexto, este artigo apresenta uma abordagem que descreve conjuntos de dados abertos na Web por meio de um perfil composto de metadados descritivos e estruturais que são enriquecidos semanticamente. O enriquecimento semântico dos metadados descritivos é realizado por meio da identificação do domínio de conhecimento ao qual o conjunto de dados pertence (e.g., "saúde", "música"). Para o enriquecimento dos metadados estruturais, este trabalho provê a recomendação de vocabulários de domínio de acordo com as propriedades encontradas no conjunto de dados em questão. Como resultado, um perfil do conjunto de dados é gerado. Para avaliar a abordagem, foi desenvolvida uma aplicação e, por meio dela, alguns experimentos foram realizados com o intuito de verificar a relevância do enriquecimento tanto em termos do domínio identificado quanto do vocabulário recomendado.

Este artigo está organizado da seguinte forma: a Seção 2 introduz alguns conceitos; a Seção 3 propõe a abordagem; a Seção 4 apresenta resultados obtidos por meio da implementação da abordagem e dos experimentos realizados; a Seção 5 discute os trabalhos relacionados, e a Seção 6 expõe as conclusões e aponta trabalhos futuros.

## 2. Conceitos Fundamentais

Segundo a *Open Knowledge Foundation Brasil*[3], os dados estão "abertos" quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los, estando estes sujeitos, no máximo, à exigência de creditar a sua autoria e a compartilhar pela mesma licença. Dados

[2] https://www.w3.org/

[3] http://br.okfn.org/2016/04/13/uma-revolucao-de-dados-para-quem/

Abertos devem ser disseminados publicamente em formatos abertos (e.g., JSON[4]) na Web, de acordo com alguns critérios e aspectos que possibilitem sua reutilização como, por exemplo, a disponibilização de metadados [Lóscio et al. 2017]. Assim, é possível o desenvolvimento de aplicativos que consumam esses dados.

Com a grande quantidade de dados produzidos pelas mais diversas fontes como, por exemplo, organizações governamentais, surgiram iniciativas para a publicação e consumo de dados em formato aberto por meio de catálogos, que fornecem uma interface entre os publicadores e consumidores dos dados [Oliveira et al. 2016]. De acordo com Maali et al. (2014), um catálogo de dados é uma coleção curada de metadados sobre conjuntos de dados. Um conjunto de dados pode ser definido como uma coleção de dados publicados ou curados por um agente, que está disponível para acesso ou *download.*

Para facilitar o efetivo uso dos conjuntos de dados publicados, metadados descritivos (descrevem características gerais) e metadados estruturais (descrevem a estrutura interna) devem ser providos de modo a facilitar seu entendimento. Os metadados devem ser disponibilizados utilizando padrões e vocabulários recomendados, e podem ser enriquecidos sempre que se julgar necessário para a geração de descrições mais significativas. Segundo [Clarke e Harley 2014], o enriquecimento semântico pode ser realizado por meio de uma categoria adicional de metadados que melhore ainda mais a utilidade, descoberta e interoperabilidade de conteúdo.

O W3C propõe um esquema de metadados descritivos a serem empregados aos conjuntos de dados publicados. Dentre eles, *temas* ou *categorias* referem-se a um conjunto de domínios de conhecimento ao qual um conjunto de dados está vinculado [Lóscio et al. 2017]. Exemplos de domínios de conhecimento são "Música", "Dados Bibliográficos". O domínio está associado, por sua vez, aos principais literais (e.g., palavras-chave) identificados em um conjunto de dados [Abele 2016].

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. Segundo [Ellefi et al. 2014], a criação de um perfil ajuda na identificação de conjuntos de dados, podendo ser definido como um conjunto de características, tanto semânticas quanto estatísticas, que permitem sua melhor descrição.

## 3. Abordagem Proposta

Considerando a literatura sobre perfis de conjuntos de dados [Abele 2016; Ellefi et al. 2014; Assaf et al. 2015] e, de acordo com as indicações de boas práticas para publicação de dados na Web do W3C [Lóscio et al. 2017], este trabalho define o conceito de Perfil de Conjunto de Dados como segue.

**Definição. Perfil de Conjunto de Dados (PCD)**. Um Perfil de Conjunto de Dados pode ser compreendido como uma anotação específica que contém metadados descritivos e estruturais referentes a um determinado conjunto de dados publicado na Web.

[4] http://www.json.org/

O PCD é organizado a partir de vocabulários recomendados. São eles: (i) DCAT[5]: descreve conjuntos de dados em catálogos de dados; (ii) VOID[6]: normalmente utilizado para expressar metadados de conjuntos de dados RDF[7] como, por exemplo, a descrição dos *links* entre os conjuntos de dados; (iii) SKOS[8]: utilizado para compartilhar e vincular sistemas de organização do conhecimento utilizando a Web Semântica; (iv) Schema[9]: vocabulário desenvolvido para ajudar mecanismos de busca a encontrar páginas.

Nesta proposta, os metadados descritivos que compõem o PCD são o *título*, *palavras-chave*, *domínio* (tema) e *vocabulário de domínio*. Os metadados estruturais indicam as propriedades identificadas no conjunto de dados e são organizados por meio de um esquema específico. Os metadados do PCD são apresentados na Tabela 1, de acordo com os vocabulários (prefixos) usados para sua formação.

**Tabela 1: Metadados que compõem a estrutura do PCD**

| Metadado | Significado |
|---|---|
| dct:title | Título do conjunto de dados. |
| dcat:keyword | Palavras-chave identificadas no conjunto de dados. |
| dcat:theme | Domínio ou tema ao qual o conjunto de dados pertence. Este metadado é composto de duas propriedades: "rdfs:label", que identifica o nome do domínio, e o "void:uriSpace" que indica a uri desse domínio. |
| void:vocabulary | Vocabulário de domínio recomendado para referenciar as propriedades do conjunto de dados. |
| skos:inSchema | Representa o esquema das propriedades do conjunto de dados. Este metadado contém o "void:properties" que contabiliza o total de propriedades e um "void:property" para cada propriedade do documento. Cada metadado "void:property" possui o "schema:name", que indica seu nome , e o "dct:type", que mostra seu tipo. |

O processo de geração do PCD com enriquecimento semântico é realizado a partir de algumas estratégias: (i) indexação dos conjuntos de dados; (ii) recomendação de vocabulário para metadados estruturais; e (iii) identificação do domínio. Cada uma dessas estratégias será descrita nas seções a seguir.

### 3.1. Indexação de Conjuntos de Dados

Para esta etapa são coletados arquivos referentes a conjuntos de dados. A indexação desses conjuntos de dados possibilita a realização das etapas fundamentais para a geração do PCD. Para a indexação, é utilizada a ferramenta Apache Solr[10]. Com esta ferramenta, calcula-se o peso dos termos por meio da função TF-IDF, que possibilita encontrar os termos mais

[5] https://www.w3.org/TR/vocab-dcat/

[6] https://www.w3.org/TR/void/

[7] https://www.w3.org/RDF/

[8] https://www.w3.org/TR/skos-reference/

[9] http://schema.org/

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

frequentes do arquivo. Conforme apresentado em alguns trabalhos como em Ouksili (2014) e Abele (2016), o TF-IDF pode ser utilizado para identificar os termos mais importantes de um documento e é muito utilizado nos engenhos de busca (ex: Google). Para a obtenção de termos melhorados, esses arquivos passam por um pré-processamento, onde é realizado o processo de (i) *stemming*, em que a palavra é substituída pelo seu radical para associação parcial entre variações de uma mesma palavra; e (ii) retirada de *stopwords*, que são palavras muito frequentes, mas que não possuem semântica associada [Shahi 2015]. Durante a indexação, são obtidas as seguintes informações para cada conjunto de dados: *id*: identificador único, *text*: conteúdo textual e *properties*: lista das propriedades identificadas.

### 3.2. Recomendação de Vocabulários

Após a indexação dos conjuntos de dados, é realizada a recomendação de vocabulários de domínio. Para isso, são selecionadas as propriedades que estão no campo "properties" do arquivo indexado e, para cada uma delas, são retornadas informações de vocabulários de domínio candidatos, como propriedades e classes. Cada propriedade recebe até quatro recomendações de termos de vocabulários que podem ser utilizados para representá-las.

Como fonte de vocabulários foi utilizado o repositório de vocabulários abertos denominado *Linked Open Vocabularies*[11] (LOV), sendo este um dos maiores catálogos aberto existentes. Por meio dele, e é possível obter vocabulários que reusam outros e suas descrições semânticas. O LOV vem sendo utilizado em alguns trabalhos, como apresentado por Ellefi et al. (2015). Após alguns testes com consultas em vários domínios, ele foi considerado suficiente para uma primeira versão da abordagem. Também é realizada a identificação do vocabulário que recebeu mais recomendações para um conjunto de dados específico, o qual é inserido no perfil gerado como o vocabulário de domínio recomendado. O Algoritmo 1 apresenta a estratégia definida.

Inicialmente, a partir do id do conjunto de dados são identificadas suas propriedades (linha 1). Cada propriedade identificada é inserida em consultas SPARQL[12] para a recuperação de informações contidas no SPARQL *endpoint* do LOV (linha 4). O objetivo é encontrar os vocabulários que possuam em suas classes as propriedades que sejam equivalentes às propriedades do conjunto de dados. Para este processo foi definida uma consulta escrita em linguagem SPARQL, que retorna a URI de até quatro propriedades de vocabulários mais populares que o *endpoint* tenha acesso. Considerando que um conjunto de dados geralmente depende de vários vocabulários para descrever seus recursos, a partir desses resultados, é possível fornecer ao usuário um resultado mais completo. O resultado de cada consulta é adicionado em uma lista, assim como os vocabulários recomendados para cada propriedade (linha 5). Estes vocabulários são inseridos em uma lista, cuja ordenação é realizada de acordo com o número de ocorrências dos vocabulários entre os resultados de todas as propriedades do documento (linha 7). São selecionados vocabulários que possuírem maior número de ocorrências entre os resultados e que estejam ativos (linha 8 – 12). Os vocabulários ativos são verificados por meio de uma consulta SPARQL no

[11] http://lov.okfn.org/dataset/lov/

[12] https://www.w3.org/TR/rdf-sparql-query/

mesmo *endpoint* indicado, que retorna um valor booleano referente à disponibilidade do vocabulário (se ele está ativo). São retornados os resultados detalhados de cada propriedade e os vocabulários melhor ranqueados na lista (linha 14).

**Algoritmo 1 Recomendação de Vocabulários**

**Entrada:** Id do conjunto de dados indexado
**Saída:** Lista de vocabulários recomendados para cada propriedade e os vocabulários melhor ranqueados

```
1:  propriedades = getPropriedades(id_Conjunto_Dados);
2:  If (propriedades não nulo) Then
3:     While temProx(propriedades) Then
4:        Resultados_Conjunto_Dados = add(Executar(Consulta_Endpoint(proximo.propriedades)));
5:        Lista_Vocabulários = add(getVocabularios(Resultados_Conjunto_Dados));
6:     end While;
7:     Lista_Vocabulários.ordenar();
8:    While (temProx(Lista_Vocabulários)) Then
9:          If(Executar(Consulta_Endpoint_Disponibilidade(próximo.Lista_Vocabulários))) Then
10:           Vocabulários = add(Vocabulário);
11:         end If;
12:   end While;
13: end If;
14: return (Resultados_Conjunto_Dados, Vocabulários);
```

### 3.3. Identificação do Domínio

Nesta etapa, é identificado de forma automática o domínio ao qual o conjunto de dados pertence. Para isso, é utilizada como referência semântica a Ontologia do DBpedia[13]. Esta ontologia foi escolhida por conter informações sobre uma grande quantidade de domínios de conhecimento de forma bem estruturada, apresentando a taxonomia de suas classes. A estratégia de identificação do domínio é mostrada por meio do Algoritmo 2.

Essa etapa também recebe como parâmetro o id do conjunto de dados. São identificadas as palavras-chave e propriedades, a partir dos termos que possuírem os maiores valores na aplicação da função TF-IDF nos campos "text" e "properties", do conjunto de dados indexado (Linha 1 - 3). Esses termos identificados são utilizados em consultas realizadas sobre o SPARQL *endpoint* do DBpedia para a obtenção de informações de domínios de conhecimento (linha 6). Nesta consulta são identificadas as classes ou propriedades da ontologia do DBpedia, que correspondem a cada um dos termos. Em seguida (linha 9), todas as classes ou propriedades identificadas passam por um processamento onde é executada outra consulta SPARQL que percorre sua hierarquia até a identificação das classes do nó raiz, a classe retornada por essa consulta é inserida em uma lista de domínios que são referentes aos resultados provenientes de cada termo consultado. É retornado como domínio a classe da hierarquia de domínios que apresente maior número de ocorrências entre os resultados das consultas (linha 11 – 13). Na seção seguinte, um exemplo será apresentado.

[13] http://wiki.dbpedia.org/

**Algoritmo 2 Identificação do Domínio**

**Entrada:** Id do conjunto de dados indexado
**Saída:** Os domínios melhor ranqueados

```
1:  palavrasChave = getPalavrasChave(id_Conjunto_Dados);
2:  propriedades = getPropriedades(id_Conjunto_Dados);
3:  termos = palavrasChave + propriedades;
4:  If (termos não nulo) Then
5:     While temProx(termos) Then
6:        Resultados_Ontologia = add(Executar(Consulta_Ontologia(próximo. termos));
7:     end While;
8:     While temProx(Resultados_Ontologia) Then
9:      Lista_Domínios = add(Executar(Consulta_Hierarquia_Onto(próximo.Resultados_Ontologia)));
10:   end While;
11:   Lista_Domínios.ordenar();
12: end If;
13: return (Lista_Domínios.getPrimeiro());
```

## 4. Implementação e Experimentos Realizados

Esta seção apresenta alguns resultados da implementação e dos experimentos realizados.

### 4.1. Implementação

A abordagem proposta foi desenvolvida na linguagem Java como uma aplicação baseada na Web com uma interface amigável, o que facilita a realização das etapas de indexação de conjuntos de dados, recomendação de vocabulários, identificação do domínio e geração do PCD. Nesta versão foram considerados conjuntos de dados que estejam em formato JSON e em inglês, principal idioma de publicação de dados na Web. Para a utilização de outros idiomas, é necessário configurar a etapa de indexação (o *stemming* e a retirada de *stopwords)*. Na recomendação de vocabulários, é necessário o acesso a vocabulários que forneçam o nome de seus recursos na língua desejada. Já na identificação do domínio, foi utilizado o DBpedia que fornece a nomeação dos seus recursos em diversos idiomas, inclusive em português.

Na Figura 1 é exibida a página principal da aplicação, que possibilita a execução das estratégias da abordagem, gerando resultados intermediários e finais. Como exemplo, considere um conjunto de dados sobre música, intitulado "Wanderlust". Primeiramente esse conjunto de dados é indexado. Depois, é realizada a identificação dos vocabulários recomendados. Neste caso, foi recomendado como vocabulário geral o *The Music Ontology* [14]. Durante a execução da identificação do domínio foi identificado como domínio geral a classe do DBpedia *Musical Work*[15]. A partir deste domínio são extraídos o seu nome e a URI, que possibilita ao usuário o acesso a maiores informações.

[14] http://purl.org/ontology/mo/

[15] http://dbpedia.org/ontology/MusicalWork

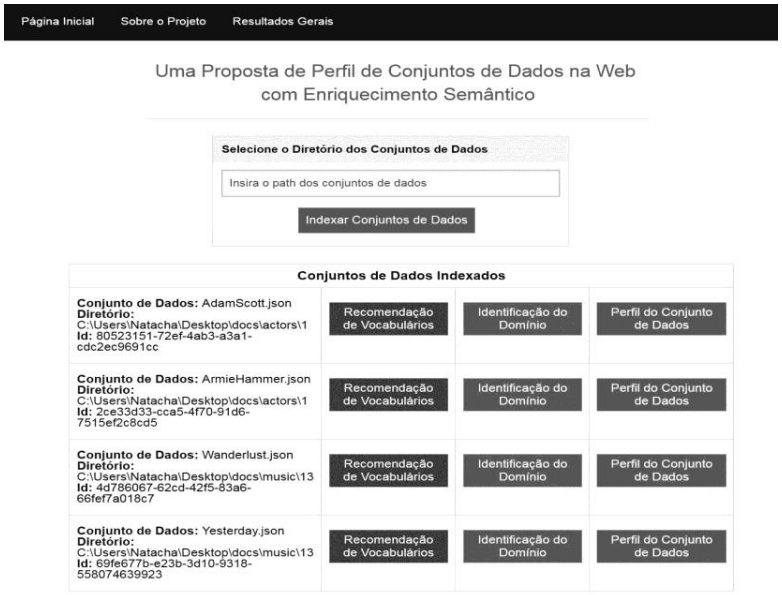



**Figura 1. Página com Etapa inicial de Indexação.**

Para a identificação do título, caso o conjunto de dados não possua alguma propriedade referente, será retornado seu nome de arquivo. No exemplo, foi encontrada a propriedade *title:Wanderlust*. Já as palavras-chave são identificadas com a aplicação da função TF-IDF no campo *text* do conjunto de dados indexado. A descrição do esquema do conjunto de dados é composta dos nomes e tipos das propriedades identificadas.

Na Figura 2 é apresentado o perfil que é gerado no formato RDF, sendo este o formato recomendado pelo W3C para a geração de metadados. São exibidos os prefixos dos vocabulários utilizados e, em seguida, é possível observar os metadados (definidos na Seção 3) que receberam os valores encontrados durante a execução da aplicação.

```
@prefix skos:   <http://www.w3.org/2004/02/skos/core# >.
@prefix void:   <http://rdfs.org/ns/void# >.
@prefix dct:    <http://purl.org/dc/terms/> .
@prefix dcat:   <http://www.w3.org/ns/dcat#:> .
@prefix  rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix  schema: <http://schema.org/>.

[]  a      dcat:Dataset ;
    dct:title "Wanderlust" ;
    dcat:keyword "singer" , "artist" , "name" ;
    dcat:theme
          [ rdfs:label "Musical Work";
            void:uriSpace "http://dbpedia.org/ontology/MusicalWork"
          ] ;
    void:vocabulary <http://purl.org/ontology/mo/> ;
    skos:inScheme
          [ void:properties "3" ;
            void:property
                  [ schema:name "id" ;
                    dct:type "string"
                  ] ;
            void:property
                  [ schema:name "title" ;
                    dct:type "string"
                  ];
            void:property
                  [ schema:name "name" ;
                    dct:type "string"
                  ]
          ] .
```

**Figura 2. Perfil exemplo gerado no formato RDF/Turtle**

### 4.2. Experimentos Realizados

Alguns experimentos foram realizados com o intuito de avaliar a abordagem proposta. Dois objetivos foram identificados: (i) avaliar a recomendação de vocabulários de domínio para o conjunto de dados, e (ii) avaliar o domínio identificado de acordo com o conjunto de dados. Para cada um dos objetivos, foram definidos *gold standards* associados para cada conjunto de dados com os resultados esperados. Foram selecionados 100 conjuntos de dados no formato JSON, escritos em língua inglesa e identificados como pertencentes aos domínios de "biografia de atores" e "música". Os conjuntos de dados são provenientes do Internet Movie Database (IMDB)[16], e da enciclopédia aberta sobre música MusicBrainz[17]. Eles se encontram publicados sem especificações de metadados descritivos e estruturais.

Os resultados obtidos com a geração do PCD foram comparados com os *gold standards* estabelecidos, e então foram calculadas as métricas de Precisão, Cobertura e F-Measure, cujas fórmulas são apresentadas a seguir:

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

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

$$F - Measure = \frac{(2 * \text{Pr}ecisão * Cobertura)}{(\text{Pr}ecisão + Cobertura)}$$

Nas fórmulas apresentadas, #ResultadosRelevantes é a quantidade de resultados retornados considerados relevantes de acordo com o *gold stan*dard, #ResultadosEsperados é o total de resultados que poderiam ser retornados definidos pelos *gold standards*, e #ResultadosRetornados é o número total de todos os resultados retornados.

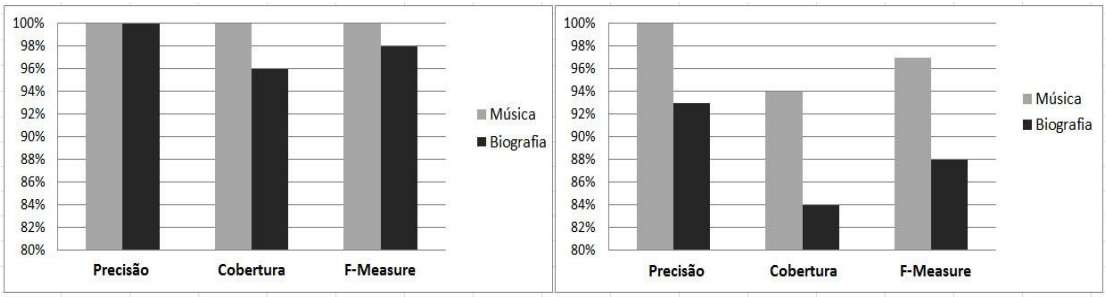
Conforme mostra a Figura 3(a), para a recomendação de vocabulários de domínio, foi observado que os conjuntos de dados dos dois domínios obtiveram resultados semelhantes. Para todos os conjuntos de dados do domínio de música foi recomendado o vocabulário definido como o *gold standard* e, para os conjuntos de dados do domínio de biografia de atores, apenas dois deles não obtiveram em sua recomendação o vocabulário definido como o *gold standard*. Dessa forma, nos conjuntos de dados do domínio de música, foram obtidos valores de precisão, cobertura e F-measure de 100% para cada uma dessas métricas. E para os conjuntos de dados do domínio de biografia, foram obtidos valores de precisão de 100%, cobertura de 96% e F-measure de 98%.

De acordo com a Figura 3(b), na identificação do domínio, os conjuntos de dados sobre música também alcançaram melhores resultados. Apenas três deles não obtiveram em seus resultados o domínio definido como o *gold standard*. Já nos conjuntos de dados sobre a biografia de atores, por possuírem termos mais diversos, em oito deles não foi

[16] http://www.imdb.com/

[17] https://musicbrainz.org/

identificado o domínio definido como *gold standard*. Nos conjuntos de dados do domínio de música, foram obtidos valores de precisão de 100%, uma cobertura de 94% e F-measure de 97%. Já para os conjuntos de dados sobre biografia, foram obtidos valores de precisão de 93%, cobertura de 84% e F-measure de 88%.




(a) Recomendação de Vocabulários (b) Identificação do Domínio

**Figura 3 – Resultados Métricas**

Com os resultados das métricas calculadas foi observado que a especialidade dos conjuntos de dados reflete diretamente nos resultados. Por exemplo, nos conjuntos de dados sobre música, quase todos obtiveram em seus resultados o ideal estabelecido. Isso pode ser atribuído à semântica desses conjuntos de dados, que é utilizada nas palavras-chave e interferem diretamente nos resultados. Quando o conjunto de dados é de uma área mais específica, suas palavras são bastante especializadas, o que torna mais provável alcançar os resultados esperados.

## 5. Trabalhos Relacionados

Considerando as estratégias definidas, alguns trabalhos relacionados foram identificados. Em relação à recomendação de vocabulários, Ellefi et al. (2015) propuseram um sistema de recomendação de vocabulários baseado no repositório LOV. Schaible et al. (2013) apresentaram uma abordagem que identifica as classes e propriedades de vocabulários que são mais utilizadas na nuvem *Linking Open Data* - LOD[18], para a representação dos dados em RDF. Em nosso trabalho, é apresentada a recomendação de vocabulários para cada propriedade do conjunto de dados de forma automática, recomendando apenas os vocabulários que estão ativos, de forma diferente dos trabalhos citados. Com essa recomendação é possível referenciar os metadados estruturais, possibilitando a conversão de conjuntos de dados de vários formatos semiestruturados ou estruturados para o formato RDF, promovendo a reutilização de metadados e integração dos dados.

Com relação à identificação do domínio, o trabalho de Ouksili et al. (2014) apresenta uma abordagem que possibilita a identificação de temas de um determinado conjunto de dados RDF. Nesta abordagem é utilizada uma combinação de critérios estruturais e semânticos para o agrupamento (*clustering*) dos grafos, onde cada agrupamento corresponde a um tema. Em nosso trabalho, as palavras-chave de um conjunto de dados são extraídas para identificar o domínio, por meio de um mecanismo que possui

[18] http://lod-cloud.net/

informações sobre uma grande quantidade de domínios de conhecimento de forma bem estruturada.

Em termos de geração de perfis, como exemplo, o trabalho apresentado por Assaf et al. (2015) propõe o Roomba, uma abordagem automática e escalável para extração, validação, correção e geração de perfil de conjuntos de dados. A abordagem proposta foi avaliada diante de um conjunto de portais de dados abertos. Outro trabalho foi apresentado por Abele (2016), onde é proposta uma abordagem para a descrição detalhada dos conjuntos de dados, utilizando metadados para prover informações gerais sobre os mesmos como, por exemplo, descrição, data de atualização e informações de licença. Diferentemente desses trabalhos, propomos a geração de um perfil com metadados descritivos e estruturais enriquecidos para conjuntos de dados que estejam em qualquer formato semiestruturado ou estruturado, e onde também não há necessidade da existência prévia de metadados descritivos e/ou estruturais. Além do perfil de conjunto de dados gerado, o usuário tem acesso à recomendação de vocabulários do domínio para cada propriedade existente no conjunto de dados em questão.

## 6. Conclusões e Trabalhos Futuros

Este trabalho apresentou uma abordagem para criação de perfil de conjunto de dados na Web com enriquecimento semântico. A abordagem proposta inclui as seguintes etapas: (i) indexação do conjunto de dados; (ii) recomendação de vocabulários de domínio para os metadados estruturais; (iii) identificação do domínio de conhecimento ao qual o conjunto de dados pertence e (iv) geração do perfil do conjunto de dados composto por metadados descritivos e estruturais. Com a implementação da abordagem, foi gerada uma aplicação Web que pode ser utilizada por produtores que desejem disponibilizar junto ao conjunto de dados um perfil com metadados descritivos e estruturais. Também é possível ter acesso a vocabulários para a conversão de seus dados para formato RDF. Consumidores de dados podem gerar de forma automática o perfil do conjunto de dados sem a necessidade de terem conhecimento sobre esses dados.

Experimentos realizados demonstraram que a abordagem é promissora no que diz respeito à geração do perfil do conjunto de dados e, em especial, ao enriquecimento semântico. Esse enriquecimento é obtido por meio da identificação do domínio de conhecimento do conjunto de dados e da recomendação de vocabulários para referenciar semanticamente os dados. Nos experimentos foram selecionados conjuntos de dados provenientes de domínios de dados distintos. Os resultados apresentaram valores de precisão e cobertura significativos, mas também foi observada a necessidade da realização de experimentos mais aprofundados, com usuários especialistas e conjuntos de dados pertencentes a uma maior variedade de domínios.

Como trabalhos futuros, além de novos experimentos, pretende-se identificar um único vocabulário para cada propriedade do conjunto de dados que possa ser incluído nos metadados estruturais do perfil gerado, promovendo maior facilidade para entendimento e reutilização do conjunto de dados. Também é pretendido incluir outros catálogos de vocabulários como fontes adicionais para a etapa de recomendação de vocabulários e permitir o uso de outros formatos de conjunto de dados, além do JSON.

## 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.

Assaf, A., Troncy, R. and Senart, A. (2015) "Roomba: An extensible framework to validate and build dataset profiles", In: 24th International Conference on World Wide Web, Italy, p. 159-162.

Clarke, M. and Harley, P. (2014) "How smart is your content? Using semantic enrichment to improve your user experience and your bottom line", Science Editor, v. 37, n. 2, p. 40–44.

Ellefi, M. B., Bellahsene, Z., Scharffe, F. and 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. and Todorov, K. (2015) "Datavore: a vocabulary recommender tool assisting Linked Data modeling", In: 14th International Semantic Web Conference Posters & Demonstrations Track a Track. United States.

Lóscio, B. F., Burle, C., Calegari, N. (2017) "Data on the web best practices. The World Wide Web Consortium", https://www.w3.org/TR/dwbp/ Último Acesso: 20 de maio de 2017.

Maali, F., Erickson, J., and Archer, P. (2014). "Data catalog vocabulary (DCAT). W3C recommendation, The World Wide Web Consortium", https://www.w3.org/TR/vocab-dcat/ Último Acesso: 20 de maio de 2017.

Oliveira, M. I. S., Oliveira, L. A., Lima, G. F. B. and 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. and Lopes, S. (2014) "Theme Identification in RDF Graphs", In: 4th International Conference on Model and Data Engineering (MEDI). Cyprus, p. 321-329

Schaible, J., Gottron, T., Scheglmann, S. and Scherp, A. (2013) "LOVER: support for modeling data using linked open vocabularies", In: EDBT/ICDT 2013 Joint Conference. Italy, p. 89–92.

Shahi, D. (2015) Apache Solr: a practical approach to enterprise search. Apress, Primeira Edição, p. 82–85. ISBN: 978-1-4842-1071-0