A LIGAÇÃO DE ENTIDADES COMO UMA ETAPA PARA A RECUPERAÇÃO SEMÂNTICA DA INFORMAÇÃO

ENTITY LINKING AS A STEP FOR SEMANTIC INFORMATION RETRIEVAL

Diego Andres Salcedo[1]
Vinícius Cabral Accioly Bezerra[2]
Renato Fernandes Corrêa[3]

RESUMO
Discorre sobre a ligação de entidades como uma etapa primordial para a recuperação semântica da informação, por meio da análise da aplicação do software DBpedia Spotlight em textos descritivos de selos postais. A metodologia consiste em pesquisa exploratória quanto aos objetivos, bibliográfica e estudo de caso quanto aos meios. O estudo de caso explora a aplicação do software em textos descritivos em linguagem natural para reconhecimento de entidades nomeadas e ligação de entidades à base de conhecimento da DBpedia. A pesquisa revela a potencialidade do uso do software para criação de bases de conhecimento especializadas e interconectadas por meio dos dados originados da Wikipédia. Conclui-se que a ligação de entidades é uma etapa importante para a recuperação semântica de objetos que apresentam em sua descrição textual entidades retratadas na enciclopédia.

Palavras-chave: Reconhecimento de entidades. Ligação de entidades. DBpedia. DBpedia Spotlight.

ABSTRACT
This article is about the entity linking as a primordial step for the semantic information retrieval, it's use analyzing the application of the DBpedia Spotlight software in descriptive texts of postage stamps. The methodology consists in an exploratory research on the objectives, bibliographic and case study on the means. The case study explores the application of software in natural language descriptive texts for entity identification and entity linking to the DBpedia knowledge base. The research reveals the software's ability to create specialized and interconnected knowledge bases through data originated from Wikipedia. It is concluded that the entity linking is an important step for the semantic retrieval of objects that present in their textual description entities portrayed in the encyclopedia.

Keywords: Entity identification. Entity linking. DBpedia. DBpedia Spotlight.

Artigo submetido em 06/08/2019 e aceito para publicação em 24/04/2020

1 Universidade Federal de Pernambuco, Brasil. ORCID: http://orcid.org/0000-0002-5936-279X. E-mail: salcedo.da@gmail.com
2 Mestrando no Programa de Pós-Graduação em Ciência da Informação. Universidade Federal de Pernambuco, Brasil. ORCID: http://orcid.org/0000-0002-2649-3232. E-mail: vviniciuscabral@gmail.com
3 Universidade Federal de Pernambuco, Brasil. ORCID: http://orcid.org/0000-0002-9880-8678. E-mail: fc_renato@yahoo.com.br

1 INTRODUÇÃO

As plataformas digitais on-line potencializam a interação entre pessoas ao redor do mundo e consequentemente a produção de conteúdo informacional derivado dessa interação. A capacidade de criação e compartilhamento produzida pelos usuários de sistemas colaborativos possibilita a existência de uma Web Social que conecta em rede muitos-para-muitos, por meio de canais de compartilhamento de conteúdo como redes de relacionamento sociais e wikis, promovendo conectividade, compartilhamento, colaboração, coprodução e consumo de dados e informações.

Toda produção de informação mediada pelo ambiente Web supracitado é ao mesmo tempo um desafio, quanto a organização e recuperação do conteúdo produzido, como uma fonte de informação ou base de conhecimento coletivo (MAIA; BAX, 2016). A Wikipédia é um dos maiores expoentes dessa dinâmica de produção coletiva, colaborativa e imediatista, tendo atualmente 43 milhões de artigos (1.010.682 em português) escritos de forma conjunta por diversos voluntários ao redor do mundo (dados atualizados em 26 de fevereiro de 2019).

Por meio desse abrangente quantitativo de conhecimento socialmente produzido na Wikipédia são gerados verbetes a partir desses artigos que possibilitam a identificação de conceitos e ideias expostas que podem ser acessadas facilmente por URIs (Identificador único do recurso na Web), públicas do Web Site. Nesse contexto, os verbetes podem ser poderosos atributos para a identificação de entidades em texto em linguagem natural, uma vez que entidades são entendidas como qualquer coisa que possuem um nome, que remetem a um conceito compartilhado, sendo referenciadas por meio de termos encontrados em textos. Por exemplo, se pesquisado na Wikipédia o verbete “Pernambuco” pode ser encontrado pelo link “https://pt.wikipedia.org/wiki/Pernambuco” contendo toda informação colaborativamente construída sobre o estado do nordeste brasileiro.

Nesse ínterim é possível identificar que a análise de texto em linguagem natural para a criação de um repositório de dados a partir de uma base de conhecimento existente pode ser dividido em duas tarefas: a identificação de unidades básicas de informação textual ou seja o Reconhecimento de Entidades Nomeadas (REN); e a correlação das entidades encontradas com sua definição em uma base de conhecimento existente, no caso em questão a Wikipédia. Tal atividade é conhecida como Ligação de Entidades (LE). Embora trabalhadas conceitualmente separadas as duas atividades anteriormente expostas são complementares e na prática podem ser trabalhadas juntas para maximizar resultados (MAIA; BAX, 2016).

A DBpedia surge como uma base de conhecimento enciclopédico criada por meio da extração de informação da Wikipédia a estruturação dessas informações em um formato de base de dados RDF (modelo padrão para intercâmbio de dados na Web: <https://www.w3.org/RDF/>). Essa base armazena o conhecimento em um formato legível por máquina e fornece meios para que as informações sejam coletadas, organizadas, compartilhadas, pesquisadas e utilizadas, dando acesso ao grafo de conhecimento aberto disponibilizado em formato de *Linked Data* (padrão de publicação de dados relacionados entre si: <https://www.w3.org/wiki/LinkedData>), dados ligados semanticamente.

Visando o aproveitamento da DBpedia foi desenvolvido o DBpedia Spotlight que "é um projeto de código aberto que desenvolve um sistema para anotação automática de entidades DBpedia em texto em linguagem natural" (DAIBER et al, 2013). Esse software permite identificação de entidades nomeadas em textos submetidos à ferramenta. Dessa forma, a ferramenta a partir da base de conhecimento utilizada e a apresentação das entidades filtradas juntamente com seus hiperlinks na DBpedia e consequentemente seu verbete na Wikipédia apresenta as soluções de: análise textual, encontrabilidade de entidades, seleção das entidades, desambiguação.

Imbuído na problemática do Reconhecimento e da Ligação de Entidades em textos em linguagem natural, visando a construção de um repositório de conhecimento a partir da base de conhecimento existente na DBpedia, este trabalho analisa a ferramenta do DBpedia Spotlight como uma solução prática para as atividades de REN e LE, permitindo posterior criação de uma base de dados não relacional populada por entidades e hiperlinks extraídos de textos em linguagem natural que descrevem itens informacionais.

Assim, o objetivo da pesquisa é a utilização da ferramenta DBpedia Spotlight para as tarefas de REN e LE e posterior criação de uma base de conhecimento para inferências semânticas, tendo como análise inicial textos em linguagem natural sobre itens informacionais. Por sua vez, para o estudo de caso foi escolhido um selo postal por ser um exemplo de item informacional que alude, na forma de documento, à conexão entre as pessoas e as instituições, por meio de processos, fluxos, lembranças e esquecimentos. Por fim, esta pesquisa integra o projeto do Repositório Filatélico Brasileiro (REFIBRA), que contribui para uma nova abordagem de se trabalhar itens informacionais na web pautada, basicamente, com aplicações de tecnologias computacionais articuladas ao campo das Humanidades Digitais. O projeto constitui para das ações do Grupo de Pesquisa Imago e Humanidades Digitais, da Universidade Federal de Pernambuco (SALCEDO e BEZERRA, 2018).

2 RECONHECIMENTO DE ENTIDADES NOMEADAS E LIGAÇÃO DE ENTIDADES

Atualmente, a Data Science desponta como uma área promissora, uma vez que tem como objeto o tratamento informacional dos enormes volumes de dados (*big data*) disponibilizados na Web. Como abordado por Maia e Bax (2016) tal campo científico tem como um de seus fundamentos basilares as atividades de identificação de entidades nomeadas em textos e sua ligação com uma base de conhecimento existente para que assim possa se expandir conhecimento por meio de informações já existentes.

Interpretar e reconhecer conceitos por meio da ação da leitura nem sempre é tarefa fácil, até mesmo para seres humanos. Automatizar tal tarefa é um problema relacionado ao Processamento de Linguagem Natural (PLN), área de pesquisa que busca de forma computacional extrair do discurso escrito ou falado informações relevantes e permitir inferências automatizadas sobre tais informações.

Para que uma entidade seja automaticamente relacionada com uma base de conhecimento existente, primeiramente ela deve ser identificada por meio da atividade de REN, processo de identificação de uma palavra ou frase que remeta a um determinado conceito ou ideia. Por exemplo, na frase "Recife é a capital do estado de Pernambuco" entidades possíveis seriam: Recife, capital, estado e Pernambuco.

Uma vez que as entidades nomeadas identificadas estão isoladas é preciso que elas sejam ligadas com alguma base de conhecimento existente para que assim possa haver uma inter-relação semântica entre os conceitos encontrados, e a posterior criação de novas base de conhecimentos intimamente relacionados. Nesse sentido a atividade de Ligação de Entidades (LE) é este processo posterior ao de Reconhecimento de Entidades Nomeadas (REN) para ligação da entidade. No exemplo exposto anteriormente a entidade nomeada "Recife" pode ser ligada com o verbete "https://pt.wikipedia.org/wiki/Recife" do Wikipédia. Contudo a operação de LE não é tarefa trivial e remete a problemas linguísticos como ambiguidade e polissemia, desafios já lançados nas pesquisas de PLN.

Para Maia e Bax (2016) as duas atividades anteriormente expostas são sobremaneira eficientes para possibilitar a Web Semântica no sentido de permitir que computadores entendam o conteúdo dos documentos existentes e sejam capazes de decidir a sua relevância como resposta à pergunta formulada pelo usuário. Por sua vez, Shen, Wang e Han (2015) apontam que base de dados construídas automaticamente por meio de um processo de LE são eficientes para a consolidação de conceitos como o *Linked Data*, que prima manter integridade semântica entre os diversos conceitos espalhados em bases de conhecimento na Web. Os autores supracitados, assim como Maia e Bax (2016) apontam a DBpedia como uma referência de base de conhecimento, uma vez que possui dados extraído da Wikipedia.

2.1 Wikipédia, DBpedia e DBpedia Spotlight

As atuais ferramentas de colaboração na Web possibilitam o desenvolvimento de projetos enciclopédicos mundiais. O maior exemplo é a Wikipédia (2018),

> [...] um projeto de enciclopédia multilíngue de licença livre, baseado na web e escrito de maneira colaborativa; encontra-se, atualmente, sob administração da Fundação Wikimedia, uma organização sem fins lucrativos cuja missão é 'empoderar e engajar pessoas pelo mundo para coletar e desenvolver conteúdo educacional sob uma licença livre ou no domínio público, e para disseminá-lo efetivamente e globalmente'. Integrando um dos vários projetos mantidos pela Wikimedia, os mais de 43 milhões de artigos (1 009 025 em português, até 31 de outubro de 2018) hoje encontrados na Wikipédia foram escritos de forma conjunta por diversos voluntários ao redor do mundo.

Dessa forma a Wikipedia é atualmente considerada um dos maiores repositórios de conhecimento socialmente construído em formato de páginas de enciclopédia on-line majoritariamente compostas por texto em linguagem natural escrita em diversas línguas diferentes. Visando aproveitar todo conhecimento disponível na Wikipédia foi desenvolvido um projeto comunitário chamado DBpedia. Tal esforço consiste na automação da extração de conhecimento disponível em recursos da Wikipédia para uma forma mais acessível e recuperável, seguindo conceitos de *Linked Data* e utilizando o framework RDF.

Segundo Bizer et al. (2009, p. 2) a DBpedia é um projeto que contribui diretamente para a potencialização da Web Semântica, que nada mais é que "um movimento colaborativo para organizar a informação de maneira legível para computadores e máquinas através de padrões de formatação de dados como o RDF." Dessa forma os autores apontam a contribuição do DBpedia no desenvolvimento de uma estrutura de extração de informações que converte o conteúdo da Wikipédia em uma rica base de conhecimento de vários domínios.

Para extrair informações das páginas da Wikipédia o DBpedia se apropria de práticas e ferramentas que enriquecem a qualidade dos dados coletados, uma dessas ferramentas é a Ontologia da DBpedia que também é colaborativamente construída. Tal tecnologia é o principal instrumento que possibilita homogeneização dos dados em diferentes línguas e formas para conceitos centralizados e recuperáveis (MENDES et al, 2012).

Extrair informações das páginas da Wikipédia envolve a tarefa de identificação de entidades nomeadas e consequentemente seus desafios como a ambiguidade. Para contornar esses desafios a DBpedia usa estratégias e recursos disponíveis na própria Wikipédia, como os títulos das páginas e suas relações com o conceito abordado no verbete, as páginas de redirecionamento quando as formas dos conceitos se apresentam de forma (escrita) diferente e as páginas de desambiguação. Nesta última

também é utilizada técnicas de análise contextual dos conceitos, levando em consideração a aparição por proximidade de termos relacionados (MENDES et al, 2012).

Aproveitando o resultado do trabalho realizado para a criação do DBpedia como um repositório de dados, a ferramenta DBpedia Spotlight se constitui em um sistema adaptável para a desambiguação de termos em línguas naturais a partir de recursos DBpedia (MENDES et al, 2011. p. 1). Assim, esses autores discorrem sobre a criação dessa ferramenta com o intuito de ser uma potencializadora para a Web Semântica, uma vez que tem com função a identificação, em texto em linguagem natural fornecido pelo usuário, de entidades mapeadas no DBpedia, bem como seus hiperlinks de acesso. O DBpedia Spotlight se vale de uma interface configurável a partir de classes e categorias disponíveis na Ontologia DBpedia bem como o nível de desambiguação e de qualidade das entidades encontradas no texto fornecido pelo usuário.

Como resultado das etapas de identificação de entidades, seleção e desambiguação, a ferramenta produz entidades destacadas e suas respectivas páginas do DBpedia (MENDES et al, 2011). A partir das etapas supracitadas abordadas por esses autores pode ser percebido que o DBpedia Spotlight consiste em uma ferramenta para identificação de entidades nomeadas e sua ligação com a base de conhecimento da DBpedia, buscando resolver de forma transparente os desafios dessas atividades.

3 METODOLOGIA

Esse artigo consiste em uma pesquisa caracterizada pela busca do aprimoramento de ideias o que segundo Gil (2009) a define, quanto ao objetivo, como exploratória. Ainda segundo Gil (2009) essa pesquisa também pode ser classificada quanto aos meios como bibliográfica, uma vez que se baseia principalmente em livros e artigos como fonte de informação. Também pode ser caracterizada como estudo de caso uma vez que desenvolve o experimento prático de utilização de uma ferramenta tecnológica.

Dessa forma esse trabalho pode ser dividido em três (3) etapas:

1) Pesquisa bibliográfica

Nessa etapa inicial foram buscadas informações sobre Reconhecimento e Ligação de entidades principalmente utilizando a Brapci e SciElo como fonte de pesquisa. Posteriormente a pesquisa foi expandida para artigos recomendados no site oficial da ferramenta DBpedia Spotlight.

2) Utilização da ferramenta DBpedia Spotlight

Uma vez compreendido as etapas realizadas pelo software DBpedia Spotlight foi possível iniciar o estudo de caso com a utilização da ferramenta, por meio da interface gráfica disponibilizada no site oficial do projeto. Nesse momento foi escolhido um texto sobre um Selo Postal encontrado no livro "Pernambuco nos Selos postais: fragmentos verbovisuais de pernambucanidades" de Salcedo (2011) para ser utilizado como experimento.

Assim como outros documentos históricos, o selo postal é objeto que merece apreciação do profissional que se dedica ao estudo dos fenômenos informacionais, uma vez este item e sua desconstrução como monumento histórico pode levar o pesquisador à diversas interligações semânticas com outros conhecimentos, segundo Salcedo (2010, p. 73),

> [...] o selo postal oferece a oportunidade para que possamos, se olharmos atentamente, perceber as transformações pelas quais temos passado, como conduzimos o desenvolvimento tecnológico, como nos distanciamos ou aproximamos do Outro, como lidamos com as diferenças e as semelhanças, como continuamos contando a nossa própria história e a da Natureza, como dizemos ou silenciamos nossos discursos e como os Estados ramificam os seus.

Por conseguinte, uma prática possível é a utilização do DBpedia Spotlight como ferramenta de análise e identificação de entidades nos textos encontrados no livro supracitado, a relação dessas entidades para a criação de uma base de dados onde seja possível inferências semânticas.

3) Análise de corretude

Após a submissão do texto escolhido para o processamento do DBpedia Spotlight foi realizada uma análise de corretude dos resultados encontrados pela ferramenta. Nesse sentido foi feita uma validação a respeito do resultado final do processo de LE, ou seja, foi verificado se as entidades encontradas no texto foram corretamente ligadas com seus verbetes na base de conhecimento do DBpedia. Assim sendo possível verificar a assertividade da ferramenta por uma média aritmética simples: divisão de entidades corretamente ligadas dividido pela quantidade total de itens encontrados.

4 ANÁLISE DE RESULTADOS

Como abordado anteriormente, o DBpedia Spotlight é uma ferramenta que pode ser utilizada para a identificação de entidades nomeadas em textos em linguagem natural, assim como apresentar a ligação dessas entidades com a base de conhecimento da DBpedia.

Para validar o uso da ferramenta para tal prática foi escolhida uma descrição de um selo postal encontrado no livro intitulado "Pernambuco nos selos postais: fragmentos verbovisuais de pernambucanidades", de Salcedo (2011). Nesse livro o autor escreveu sobre 32 imagens em selos postais que indicam características do Estado de Pernambuco, bem como criou quadros com metadados para cada imagem. Assim, para o estudo de casos deste artigo, por uma questão discricionária dos autores, o selo postal escolhido foi o representado na Figura 1. Ele retrata a imagem de uma bandeira da Revolução Republicana em Pernambuco.

Salcedo (2011, p. 12) disserta sobre o selo como indicado abaixo:

> Este selo comemora os cem anos da Revolução Republicana ou "dos Padres", ocorrida no ano de 1817 em Pernambuco. Foi um movimento contrário à estada da família real portuguesa no Brasil. Essa ação tratou, ao mesmo tempo, de "cortar relações com o Império estabelecido no Rio de Janeiro" e acarretar indagações sobre a realidade social vigente. O vitorioso governo provisório deliberou sobre a "criação de uma bandeira azul e branca, repartida horizontalmente com um desenho simbólico". Emitir um selo em que a imagem principal é a bandeira resultante de um movimento revolucionário e, além disso, inscrever a sugestiva expressão Revolução Republicana, pode representar a necessidade de renascimento do passado nacional, com o objetivo de enfatizar "manifestações de patriotismo". Sugere, também, coroar uma trajetória de liberdade dessa neófita nação. A imagem tem duas colunas jônicas que sustentam o "BRAZIL", simbolizando assim uma "eterna estabilidade". Acima da coluna da esquerda está o Brasão do Estado de Pernambuco, oficializado pelo Governador General Alexandre José Barbosa Lima (1892-1931), em 1895. Um detalhe interessante na imagem trata sobre as possíveis Flores de Lis 'deitadas', cercando ou protegendo o 'BRAZIL'. A Flor de Lis não existe na natureza. É criação simbólica humana, também conhecida como "flor Heráldica [...] símbolo real desde a Alta Antiguidade". Ao centro tem-se a imagem principal do selo: a bandeira do Estado de Pernambuco, presa ao mastro pelo lado esquerdo, reprodução fiel que prossegue até hoje.

Figura 1 - Selo comemorativo do Centenário da Revolução Republicana em Pernambuco.

Fonte: Salcedo (2011)

Definido o texto a ser analisado, no caso a descrição do selo postal, é possível a utilização da ferramenta DBpedia Spotlight. Utilizando a interface on-line (https://www.dbpedia-spotlight.org/demo) do sistema, o usuário pode simplesmente colar o conteúdo que deseja analisar, definir a linguagem em que está escrita, na opção *confidence* escolher o grau de confiabilidade de acerto da ligação da entidade com seu verbete no Wikipedia e clicar em *annotate* para que a ferramenta encontre e destaque as entidades nomeadas identificadas, como pode ser observado na Figura 2.

Figura 2 - Utilização do Spotlight com a descrição do Selo Postal selecionado

Fonte: dados da pesquisa (2019)

Após a submissão e processamento do texto, o usuário recebe como resultado o destaque das entidades nomeadas encontradas (destaque azul em formato de hiperlink), bem como seus links com os verbetes no repositório do DBpedia, como pode ser observado na Figura 2. Com a análise do resultado encontrado com a utilização do DBpedia Spotlight posterior ao processamento da descrição do selo postal (Figura 1) descrito por Salcedo (2011) foi possível encontrar como resultado os valores apresentados na Figura 3, a seguir:

Figura 3 - Entidades encontradas e seus links para o repositório do DBpedia

Entidades	**Ligação com o DBPedia em português**
Revolução Republicana	http://pt.dbpedia.org/page/Implantação_da_República_Portuguesa
Pernambuco	http://pt.dbpedia.org/page/Pernambuco
Família Real Portuguesa	http://pt.dbpedia.org/page/Monarquia_de_Portugal
Brasil	http://pt.dbpedia.org/page/Reino_Unido_de_Portugal,_Brasil_e_Algarves
Rio de Janeiro	http://pt.dbpedia.org/page/Rio_de_Janeiro
Bandeira	http://pt.dbpedia.org/page/Bandeira
Patriotismo	http://pt.dbpedia.org/page/Patriotismo
Jônica	http://pt.dbpedia.org/page/Ordem_jônica
Alexandre José Barbosa Lima	http://pt.dbpedia.org/page/Barbosa_Lima
Flor de Lis	http://pt.dbpedia.org/page/Flor_de_Lis
Antiguidade	http://pt.dbpedia.org/page/Idade_Antiga

Fonte: dados da pesquisa (2019)

Analisando a Figura 3 juntamente com a leitura do texto submetido para análise na ferramenta pode ser calculado um percentual de corretude e acerto das entidades reconhecidas e ligadas. Uma vez que foram encontradas onze (11) entidade nomeadas, de modo que o percentual de acerto pode ser encontrado por média aritmética onde cada ligação correta acarreta em uma em uma fração de 1/11. Para fins de análise e visualização, na Figura 2 pode ser observado a descrição de como foi analisado se houve acertou ou não.

Por conseguinte, após a análise dos resultados disponibilizados pode ser calculado um percentual de acerto de 81,81% para o texto analisado, observando uma taxa de *confidence* de 0.5 configurado como parâmetro na ferramenta. Ou seja, para o DBpedia Spotlight todas as entidades encontradas e ligadas têm uma probabilidade igual ou maior do que 50% de estar corretamente associada ao verbete.

Figura 4 - Análise percentual de acerto das entidades ligadas pelo DBpedia Spotlight.

Ligação com o DBpedia em português	Análise	Acerto
http://pt.dbpedia.org/page/Implantação_da_República_Portuguesa	Entidade reconhecida corretamente, porém ligada com o verbete incorreto. Essa entidade se refere, no texto, a Revolução Pernambucana de 1817, no entanto o verbete ligado remete a revolução Republicana ocorrida em Portugal no ano de 1910.	Não
http://pt.dbpedia.org/page/Pernambuco	Entidade corretamente reconhecida e ligada com o verbete Pernambuco	Sim
http://pt.dbpedia.org/page/Monarquia_de_Portugal	Entidade corretamente reconhecida e ligada com o verbete sobre a monarquia portuguesa, no texto se referindo a vinda da família real ao Brasil no ano de 1808.	Sim
http://pt.dbpedia.org/page/Reino_Unido_de_Portugal,_Brasil_e_Algarves	Entidade corretamente reconhecida e ligada ao verbete que remete ao Brasil de 1817, conhecido como Reino Unido de Portugal, Brasil e Algarves.	Sim
http://pt.dbpedia.org/page/Rio_de_Janeiro	Entidade corretamente reconhecida e ligada com o verbete Rio de janeiro.	Sim
http://pt.dbpedia.org/page/Bandeira	Entidade corretamente reconhecida e ligada com o verbete Bandeira.	Sim
http://pt.dbpedia.org/page/Patriotismo	Entidade corretamente reconhecida e ligada com o verbete Patriotismo.	Sim
http://pt.dbpedia.org/page/Ordem_jônica	Entidade corretamente reconhecida e ligada com o verbete ligado a arquitetura Jônica.	Sim
http://pt.dbpedia.org/page/Barbosa_Lima	Entidade corretamente reconhecida e ligada com o verbete a respeito do político brasileiro Alexandre José Barbosa Lima.	Sim
http://pt.dbpedia.org/page/Flor_de_Lis	Entidade corretamente encontrada, porém ligada ao verbete sobre a canção do artista brasileiro Djavan. Na verdade, no texto, essa entidade se refere ao um símbolo utilizado pela monarquia francesa e outros povos ao longo do tempo.	Não
http://pt.dbpedia.org/page/Idade_Antiga	Entidade corretamente reconhecida e ligada com o verbete referente ao conceito de um período histórico remoto.	Sim
Taxa de acerto		81,81%

Fonte: dados da pesquisa (2019)

O resultado disponibilizado pela ferramenta e elencado (Figura 3) pode ser utilizado como base para a criação de um repositório passível de inferências semânticas, uma base de dados utilizando **Resource Description Framework** (RDF). O RDF é um modelo de representação de dados largamente utilizado para troca de informações semanticamente relacionadas por sistemas na web, baseado na ideia de que todo item pode ser descrito e relacionado por meio de um conceito de triplas: *subject* -> o recurso em questão; o *predicate* -> propriedade que faz relação com outros recursos; *object* -> o valor da propriedade (LIMA; CARVALHO, 2005). Com a incorporação do valor de URIs tanto para o *subject* como o *object*, o RDF permite a criação de um mapeamento relacional entre os itens, de modo a permitir uma navegabilidade e relacionamento compreensível por humanos e máquinas. Nesse contexto o objeto selo postal ganha uma identificação e passa a ser o sujeito (*subject*), o predicado(*predicate*) pode ser assumido como sendo o "tem relação" e o objeto (*object*) é o link para o DBpedia.

Na medida que esse repositório for populado com o resultado da descrição de outros selos postais ele é carregado com informações em RDF que permite a formação de grafos e/ou inferências semânticas com a utilização uma linguagem de consulta eficiente para recuperação de dados em banco de dados RDF e recomendada pela World Wide Web Consortium (W3C) chamada SPARQL (acrônimo para *SPARQL Protocol and RDF Query Language,* linguagem de consulta e manipulação de dados em RDF). Dessa forma sendo possível relacionar os itens informacionais por meio de suas ligações entre as Entidades Nomeadas encontradas nos textos em linguagem natural processados. Por exemplo, o verbete <http://pt.dbpedia.org/page/Pernambuco>, pode ser usado com inferência de relacionamento entre dois Selos Postais que tiverem essa entidade encontradas nas suas respectivas descrições.

5 CONSIDERAÇÕES FINAIS

O presente artigo analisou a utilização da DBpedia Spotlight na tarefa de ligação de entidades em descrições textuais em português de selos postais, permitindo a construção de uma base de conhecimento especializada nos moldes da Web Semântica. Nesse sentido foi constatado que a ferramenta utilizada para automatizar a ligação de entidades é um facilitador para o processamento de um grande volume de informação em um espaço de tempo reduzido, de modo que uma base de conhecimento seja rapidamente gerada.

Com o intuito de analisar a aplicação do DBpedia Spotlight para a tarefa proposta foi desenvolvido um estudo de caso com o uso da descrição de um selo postal descrito no livro “Pernambuco nos Selos postais: fragmentos verbovisuais de pernambucanidades” de Salcedo (2011). A análise dos resultados desse estudo comprovou uma porcentagem de acerto superior a 81% para a atividade de ligação de entidades nomeadas utilizando o DBpedia Spotlight. Foi descrito também como resultados encontrados pelo software podem ser utilizados para criação de uma base de dados RDF para futura inferência semântica para o conhecimento encontrado.

O presente artigo não tem como escopo cobrir completamente a pesquisa proposta, uma vez que é sabido a necessidade da análise de um maior número dos resultados de processamento textuais utilizando o DBpedia Spotlight, porém tem como objetivo validar a possibilidade de utilização da ferramenta para o fim de automatização da criação de um repositório semântico a partir de tarefa de Ligação de Entidades com a uma base de conhecimento existente, no caso, a DBpedia.

Por conseguinte, esse trabalho se apresenta como o princípio de uma jornada de investigação científica em prol da utilização da informação de maneira semântica, utilizando de técnicas e ferramentas que capacitem a criação de ambiente com dados linkados semanticamente, para que este exista uma recuperação da informação mais contextual, plural e potencializadora da descoberta de novidades.

REFERÊNCIAS

BIZER, C. et al. DBpedia: a crystallization point for the web of data. **Journal of Web Semantics**: science, services and agents on the World Wide Web, v. 7, p. 154–165, 2009. Disponível em: https://bit.ly/2MHeF3k. Acesso em 17, fev. 2019.

DAIBER, J.; JAKOB, M.; HOKAMP, C.; MENDES, P. Improving efficiency and accuracy in multilingual entity extraction. **ACM International Conference Proceeding Series**. Graz [Áustria], 2013. p.121-124. Disponível em: http://jodaiber.de/doc/entity.pdf. Acesso em 17, fev. 2019.

MAIA, E. H. B.; BAX, M. P. Um estudo bibliográfico sobre ligação de entidades. **Inf. Inf.**, Londrina, v. 21, n. 2, p. 245 - 291, maio/ago., 2016. Disponível: http:www.uel.br/revistas/informacao. Acesso em 26 de set. 2018.

MENDES, P. N., JAKOB, M.; BIZER, C. DBpedia for NLP: a multilingual cross-domain knowledge base. INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 2012. **Anais**… Istanbul [Turquia], 21-27 maio, 2012. Disponível em: https://bit.ly/2GOCMce. Acesso em 13, fev. 2019.

MENDES, P. N.; JAKOB, M.; GARCÍA-SILVA, A.; BIZER, C. DBpedia Spotlight: shedding light on the web of documents. INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 7, 2011. **Anais**… Graz [Áustria], 7-9 set., 2011. Disponível em: https://bit.ly/2Tb6Jbx. Acesso em 13, fev. 2019.

SALCEDO, D. A. **A ciência nos selos postais comemorativos brasileiros: 1900-2000**. Recife: EDUFPE, 2010.

SALCEDO, D. A. **Pernambuco nos selos postais**: fragmentos verbo-visuais de pernambucanidades. Recife: Néctar-Liber, 2011.

SALCEDO, D. A.; BEZERRA, Vinícius C. A. A gênese do repositório filatélico brasileiro: uma experiência interdisciplinar nas Humanidades Digitais. **Inf. & Soc.:Est.**, João Pessoa, v.28, n.3, p. 69-80, set./dez. 2018. Disponível em: https://bit.ly/2Kt4kFo. Acesso em: 19 fev. 2019.

SHEN, W.; WANG, J.; HAN, J. **Entity Linking with a Knowledge Base:** issues, techniques, and solutions. 2015. Disponível em: https://bit.ly/2T5G5ko. Acesso em: 17 fev. 2019

WIKIPÉDIA. Disponível em: https://pt.wikipedia.org/wiki/Wikipédia. Acesso em: 02 nov. 2018.