ARTIGO

RDBCI Revista Digital Biblioteconomia e Ciência da Informação | RDBCI Digital Journal of Library and Information Science

doi: 10.20396/rdbci.v17i0.8651791

TECNOLOGIAS DA WEB SEMÂNTICA PARA A RECUPERAÇÃO DA INFORMAÇÃO NO WIKIDATA

SEMANTIC WEB TECHNOLOGIES FOR THE INFORMATION RETRIEVAL ON WIKIDATA

[1]Larissa Pavarini da Luz
[1]Caio Saraiva Coneglian
[1]José Eduardo Santrem Segundo
Universidade da Amazônia[1]

Correspondência

Larissa Pavarini da Luz
Universidade Estadual Paulista
São Paulo, SP – Brasil.
E-mail: larissapavarinidaluz@gmail.com

Submetido em: 27/02/2018
Aceito em: 26/10//2018
Publicado em: 05/11/2018

Checagem Antiplagiarismo
turnitin

JITA: IL. Semantic Web

e-Location ID: 019003

RESUMO

A Recuperação da Informação é responsável pelo armazenamento e pela recuperação automática de informação, podendo estes documentos ser constituídos por textos, páginas Web, áudio, vídeo, imagens, gráficos e figuras. Técnicas de Recuperação de Informação ganharam importância com o crescimento da Web, pois a quantidade ilimitada de informação pode expressar as mais diversas formas e níveis de qualidade ao que se espera. Pensando nisso o presente trabalho estuda métodos e tecnologias capazes de recuperar essas informações, dando enfoque a buscar em bases de dados estruturadas chamadas Linked Data, mas especificamente no Wikidata, uma base de dados estruturada utilizando conceitos da Web Semântica, que reúne conhecimentos da Wikipédia. Buscando compreender como é feita essa recuperação da informação no projeto Wikidata, esta pesquisa tem como objetivo apresentar os meios que o Wikidata fornece para a RI e como eles usam os princípios da Web Semântica. A metodologia utilizada foi um estudo exploratório com embasamento para a pesquisa e aplicada, uma vez que testes foram feitos na base de dados do Wikidata. Como resultados, identificou-se características das diversas formas de acesso e de recuperação dos dados, traçando correlações existentes entre cada uma destas formas, com o arcabouço teórico da Web Semântica e da Recuperação da Informação. Concluiu-se que o Wikidata se coloca como uma base de dados sólida, com um grande volume de conteúdo que possui uma série de mecanismos de recuperação, capazes de atender às mais diversas aplicações existentes na Web, devido a estes mecanismos serem construídos com distintas tecnologias e configurações.

PALAVRAS-CHAVE

Web semântica. Recuperação da informação. Linked data. Wikidata.

ABSTRACT

Information Retrieval is responsible for the storage and automatic retrieval of information, and these documents may consist of texts, web pages, audio, video, images, graphics and figures. Information Retrieval techniques have gained importance with the growth of the Web, because the unlimited amount of information can express the most diverse forms and levels of quality to what is expected. With this in mind, the present work studies methods and technologies capable of retrieving this information, focusing on searching structured databases called Linked Data, but specifically on the Wikidata project, a database structured using Semantic Web concepts, which brings together the knowledge from Wikipedia. Seeking to understand how this information retrieval is done in the Wikidata project, this research has the objective of presenting the media that Wikidata provides to RI and how they use the principles of the Semantic Web. The methodology used was an exploratory study based on the research and applied, since tests were done in the database of Wikidata. As a result, the characteristics of the various forms of data access and retrieval were identified, tracing the correlations between each of these forms, with the theoretical framework of the Semantic Web and Information Retrieval. It was concluded that Wikidata stands as a solid database, with a large volume of contents, quite current, that has a series of recovery mechanisms, capable of serving the most diverse applications on the Web, because these mechanisms are built with different technologies and configurations.

KEYWORDS

Semantic web. Information retrieval. Linked data. Wikidata.

1 Introdução

No âmbito da Web, alguns serviços possuem um papel de reunir grandes quantidade de informações dos mais diversos domínios, como por exemplo a Wikipédia. Com a consolidação da Web Semântica, por meio de iniciativas como o *Linked Data*, estes serviços tratando de domínios gerais estão sendo convertidos em bases de dados seguindo os princípios dos dados ligados. Um expoente desse contexto é a Wikidata, que reúne em sua maioria, informações oriundas da Wikipédia.

A Wikidata foi criada com o intuito de reunir informações estruturadas contemplando técnicas capazes de extrair os dados de páginas Web automaticamente. Essa característica fez com que a quantidade de informações presentes da Wikidata crescesse muito rapidamente, sendo atualmente, uma das principais fontes de dados estruturadas da Web.

Desta forma, a quantidade dos dados que uma base como Wikidata reúne, é extremamente volumosa, trazendo grandes desafios para a Recuperação da Informação nestes ambientes. Tal aspecto impeliu a Wikidata a criar uma série de mecanismos para que suas informações fossem extraídas e localizadas. Visando compreender tais mecanismos, o presente trabalho questiona: Quais são os meios que o Wikidata fornece para a Recuperação da Informação, bem como a forma como tais meios utilizam os princípios da Web Semântica?

Visando responder tal problemática, o presente trabalho tem como objetivo analisar os modos de Recuperação da informação fornecidos pela Wikidata, identificando como ocorre o uso dos princípios da Web Semântica dentro desse serviço, além de analisar como tais princípios tornam a Recuperação da Informação mais contextualizada de acordo com os conteúdos desta base de dados. Assim, busca-se analisar uma das principais bases de dados no âmbito da Web e da Web Semântica, verificando se a Wikidata segue os princípios discutidos pelos criadores da Web Semântica, Berners-Lee, Hendler e Lassila (2001), em que existe uma necessidade dos agentes computacionais serem capazes de compreender com maior precisão o domínio em que os conteúdos estão inseridos.

Como metodologia utilizada, realizou-se um estudo exploratório, buscando na literatura um embasamento para a pesquisa, e que aplicado, foram feitos testes explorando a base de dados da Wikidata. Com este trabalho, identificou-se que a Wikidata é uma base rica de dados, que está estruturada em formatos compatíveis com a Web Semântica, e além disso, foi possível verificar que as formas de recuperar os dados permitem com que os usuários possam localizar as relações com uma ampla gama de opções.

Dessa forma, é possível definir os conteúdos metodológicos da seguinte forma: A pesquisa bibliográfica se deu por meio de levantamentos bibliográficos em bases de dados

nacionais e internacionais, explorando as temáticas: Recuperação da Informação, Web Semântica, *Linked Data* e Wikidata.

A ação aplicada foi realizada a partir da observação direta e participativa no *site* da Wikidata, realizando provas de conceitos a partir das informações expostas pelo serviço e das formas de acesso aos dados.

Nessa perspectiva, este trabalho analisou os modos de Recuperação da Informação fornecidos pela Wikidata, visando identificar como ocorre o uso dos princípios da Web Semântica dentro desse serviço, e como tais princípios tornam a Recuperação da Informação mais contextualizada de acordo com os conteúdos contidos na base de dados. Com a identificação das formas de acessos aos dados, bem como com a realizações de ações visando identificar o funcionamento de tais formas, foi possível traçar as correlações existentes entre cada forma de acesso, com as teorias acerca da Recuperação da Informação e da Web Semântica.

2 Web Semântica e Linked Data

A evolução nas Tecnologias da Informação e Comunicação vem proporcionando mudanças na Web, trocando os paradigmas seguidos por esse ambiente informacional. Desde a sua concepção em 1989, a Web passa constantemente por modificações, que está especialmente vinculado a uma evolução tanto das tecnologias, quanto da utilidade que os usuários encontraram de suas ferramentas.

Nesta perspectiva, um dos principais proveitos da Web, trata do conhecimento que os dados podem fornecer, que perpassam por uma visão mais aprimorada das pessoas e das instituições. Diante do exposto, o surgimento da Web Semântica ocorreu naturalmente, pela necessidade de ter uma compreensão melhor dessas informações, que a princípio eram incompreensíveis para os mecanismos computacionais.

Esse cenário é exposto por Berners-Lee, Hendler e Lassila (2001, p.2, tradução nossa), que afirmam que a Web Semântica é uma: “[...] extensão da Web atual, em que a informação possui um significado claro e bem definido, possibilitando uma melhor interação entre computadores e pessoas.”

Assim, a Web Semântica surgiu como uma solução para permitir uma compreensão melhor das informações disponíveis na Web pelos mecanismos, visando fornecer bases para a criação de aplicações mais inteligentes e que tenham uma maior compreensão do domínio em que dados se encontram. Berners-Lee, Hendler e Lassila (2001) discorrem sobre essa questão,

ao apontar um cenário futuro àquela época, mas em que agentes computacionais seriam capazes de facilitar a vida do usuário, ao conseguirem extrair, interoperar e compreender os dados de diversas fontes.

A proposta da Web Semântica vem evoluindo ano a ano, criando conceitos e tecnologias mais consistentes, capazes de embasarem aplicações que conseguem explorar com precisão bases de dados, compreendendo o contexto de tais informações. Essa fase em que aplicações de uso são construídas e utilizadas em massa, está sendo chamada de materialização da Web Semântica, processo que tem o *Linked Data* como principal aplicação.

O *Linked Data* tem em seu princípio inserir em uma estrutura compatível com a Web Semântica, significado aos dados. Essa iniciativa foi proposta de 2006 por Berners-Lee, apresentando algumas diretrizes para a criação de bases de dados seguindo normas que facilitam a localização e a inserção de significado nestes dados. A partir dessa proposta, foram criadas diversas bases de dados ligadas e abertas, que seguiam a proposta do *Linked Data*, utilizando as tecnologias recomendadas pela Web Semântica, em destaque o DBpedia[1], que reunia informações principalmente da Wikipédia.

As consequências dessa evolução da Web, especialmente no que tange a Web Semântica e ao *Linked Data*, promoveram mudanças profundas dentro de diversos campos de estudos, em especial a Recuperação da Informação. Buscando apontar essas mudanças, bem como as correlações que o arcabouço teórico desse contexto possui com os conceitos e as tecnologias da Web Semântica, a seguir são apresentados conceitos acerca das Tecnologias da Web Semântica na Recuperação da Informação.

3 Tecnologias da Web Semântica na Recuperação da Informação

Devido ao aumento considerável e exponencial de informações armazenadas e disponíveis para o acesso, Sant'Ana (2008, p.145) afirmou que a adoção de tecnologias de informação e comunicação para a transmissão dessas informações ao usuário é de extrema relevância, sobretudo no que se refere aos processos de Recuperação da Informação (RI).

Assim, a RI classifica-se como uma solução para o problema da explosão informacional identificada por Bush em 1945, como o irreprimível crescimento na quantidade de dados gerados. Com o passar das décadas, esta explosão informacional se potencializou, em especial pela massificação do uso das tecnologias.

[1] A DBpedia é uma iniciativa que buscou estruturar os dados da Wikipédia, as disponibilizando na Web. Disponível em: <http://wiki.dbpedia.org/>. Acesso em: 04 fev. 2018.

Tal fenômeno pode ser percebido, ao verificarmos que os sistemas de informação e de comunicação se tornaram centrais na maioria das atividades humanas, que direcionou para a criação e uso dos dados em formatos digitais. Consequentemente, tornou-se mais difícil recuperar as informações nestes ambientes digitais, principalmente após a criação e a multiplicação da Web, que possibilitou a todos serem criadores e consumidores de informações, conectando a maioria da população em ambientes digitais quase que em tempo integral, tornando a geração de dados mais expressiva.

As dificuldades para se localizar e se recuperar os dados dentro deste imenso oceano de informações, a Web, começaram a serem amenizadas com a criação da Web Semântica, que de acordo com Santarém Segundo, Souza e Coneglian (2015) visava ajudar e melhorar a vida do usuário, permitindo que todas as informações a serem buscadas estivessem interligadas e com significado explícito

Neste sentido, vale destacar que a Web Semântica não está somente relacionada ao formato do conteúdo de um recurso, mas também à forma como este conteúdo será disponibilizado e interagirá com outros recursos na Web, ou seja, além do usuário humano ser capaz de buscar por uma informação, as máquinas estariam aptas a proporcionar o retorno correto do sentido que tais dados possuem.

Para tornar implementável e real a Web Semântica, e consequentemente aprimorar a RI na Web, uma série de tecnologias foram desenvolvidas, tais como: o XML (*eXtensible Markup Language*), RDF (*Resource Description Framework)*, RDF Scheme, OWL (*Web Ontology Language*), SPARQL (SPARQL *Protocol and RDF Query Language*) entre outros conceitos que são descritos pelo W3C (W3C, 2017), que serão descritos na sequência.

Criada em 1998, a linguagem *eXtensible Markup Language* (XML) é uma linguagem de marcação extensível, e recomendada pela W3C (W3C, 2017) para descrever os metadados que um documento contém, e seu maior objetivo é criar uma infraestrutura única para diversas linguagens.

Berners-Lee, Hendler e Lassila (2001) fizeram toda a proposição da arquitetura da Web Semântica baseada no XML, visando gerar um padrão de assimilação na troca de documentos eletrônicos, de forma textual, simples, estruturada, extensível, flexível, semanticamente rica e com uma segurança adequada. Assim, a linguagem XML permitiu e permite a criação de estruturas concisas nos documentos contidos na Web.

Vale destacar, que a Web Semântica possui uma estrutura de funcionamento, sintaxe, que utiliza de *Universal Resource Identifier* (URI) para representar os dados. A identificação

dos seus recursos é feita de forma única, assim como seus relacionamentos, utilizando as URIs para nomeá-los.

Outra tecnologia central para a Web Semântica é o *Resource Description Framework* (RDF), uma linguagem declarativa, que se tornou um padrão recomendado pelo W3C em 2004. O RDF representa os dados na forma de sentenças sobre propriedades e relacionamentos entre recursos, que podem ser virtualmente qualquer objeto existente no mundo real.

As características de RDF se concentram em ser independente de domínio e é composto por uma tripla: sujeito, predicado e objeto. A formação de uma tripla se deve pela combinação de um recurso, uma propriedade e um valor para a propriedade de um recurso. (DZIEKANIAK; KIRINUS, 2004).

Outras características relevantes referentes ao RDF podem ser elencadas como: buscar uma representação primitiva com vistas a uma criação maior, tem como finalidade o enriquecimento semântico, além de fornecer a capacidade de se comunicar de forma transparente, ou o mais próximo disso, em relação a semântica para metadados, facilitando a recuperação da informação na web. Apesar de possibilitar a descrição dos recursos com semântica formal, o RDF não fornece subsídios necessários para ser considerada uma linguagem de ontologias, como a linguagem *Web Ontology Language* (OWL)

Definida, segundo W3C (2012) como uma revisão da Linguagem *DARPA agent markup language* (DAML) + *Ontology Inference Layer* (OIL), a OWL possui maior flexibilidade ao expressar significados e semânticas comparadas ao XML/RDF e RDF *Schema*, idealizada para o uso em aplicações que necessitem processar o conteúdo de informações, ao invés de somente apresentar a visualização destas informações.

A OWL é considerada o núcleo da Semântica (W3C, 2015), sendo uma linguagem Web baseada em RDF, e definida como uma linguagem para instanciação de ontologias. Esta linguagem pode formalizar um domínio pela definição de classes e de propriedades, além de definir indivíduos e afirmações sobre elas, capaz de especificar como se deriva sequências lógicas, isto é, fatos que não estão presentes na ontologia, mas que são vinculados pela semântica.

A recuperação dos dados em RDF e em OWL, ocorrem no âmbito da Web Semântica, por meio do protocolo *SPARQL Protocol and RDF Query Language* (SPARQL), que segundo a W3C (2008) constitui a primeira camada da Web Semântica, é classificada como sendo uma linguagem de consulta e um protocolo capaz de recuperar e manipular dados no formato RDF, que permite recuperar valores de dados estruturado e semiestruturados, explorar dados ao

consultar relações desconhecidas, realizando uniões complexas de conjuntos de dados diferentes em uma única e simples consulta.

Partindo dos conceitos apresentados, na seção seguinte são expostos os resultados e as discussões do presente trabalho.

4 Resultados e Discussões

As bases de dados que seguem os princípios do *Linked Data* costumam ser uma fonte de informação bastante abrangente, contendo um elevado nível de contextualização das informações ali contidas. Desta forma, a utilização dessas bases de dados em diversas outras aplicações podem conduzir a um aprimoramento do nível de compreensão que os agentes computacionais possuem do conteúdo disponível na Web.

Esse cenário vai ao encontro dos ideais apresentados por Berners-Lee, Hendler e Lassila, em 2001, quando propuseram a Web Semântica. Todavia, a criação de interfaces capazes de relacionar e de recuperar os dados desses *datasets*[2] é essencial para a popularização do uso dessas ferramentas, especialmente no que tange a aplicação de forma prática e usual.

Ao argumentar tais questões, é necessário que as bases teóricas referentes à Recuperação da Informação sejam requeridas e discutidas, visando identificar como, a partir da necessidade de acessar os dados desses *datasets* pode-se apresentar uma Recuperação da Informação eficiente, seguindo os princípios da Web Semântica.

No âmbito do *Linked Data*, o Wikidata está ganhando notoriedade pela quantidade de informações que ele apresenta, bem como pelos processos que esse ambiente possui, visando ter um crescimento sustentável de seus dados, que busca identificar dados incorretos, permitindo a correção. Outro ponto de destaque do Wikidata é a relação existente entre este serviço com a Wikipédia, possibilitando a extração automática de informações deste segundo ambiente.

Assim, sob a ótica da Recuperação da Informação identificar e analisar as formas de buscas existentes dentro do Wikidata é fundamental para compreender a extensão que esta plataforma pode obter dentro da Web como um todo. Vale destacar ainda, que a base da Web

[2] Segundo Silerchtz, Korth e Sadarshan (2012), *dataset* é um objeto, de qualquer linguagem de programação que faz conexão com banco de dados, que por sua vez, tem como objetivo ser um repositório de dados deste objeto. Possui tabelas, colunas, linhas, etc. O objeto *dataset* também tem recursos para manipulação dos dados nele contidos (inclusão, exclusão e alteração).

Semântica do Wikidata permite uma nova perspectiva de comunicação com outros serviços da Web, mas que deve ser analisada se os princípios da Web Semântica estão sendo utilizados.

O Wikidata aponta as formas que as informações podem ser buscadas, dividindo em dois grupos principais (WIKIDATA, 2016): Acesso a dados por itens e acesso aos despejos. A divisão nestas duas classificações é feita pelo princípio de acesso aos dados, em que uma ocorre centrado na localização de itens específicos, e a segunda permite com que o usuário acesse a todo o conjunto de dados realizados.

A forma nomeada de acesso a dados por itens, possui diversas subdivisões, devido a forma como cada uma destas permite para se recuperar e se acessar os dados. Assim, na sequência busca-se traçar as correlações existentes entre cada uma das formas existentes com o arcabouço teórico da Recuperação da Informação e da Web Semântica. São estas as divisões: a interface dos dados vinculados, a API MediaWiki, a consulta Wikidata, o SPARQL *Endpoint* e os Robôs.

A interface dos dados vinculados é uma página Web em que o usuário pode visualizar as informações e as relações que um determinado item possui. Essa interface pode ser acessada pela própria URI de um recurso, apresentado em uma interface visual e compreensível para as pessoas os conteúdos expressos na base de dados, daquele item específico.

Ao se acessar a URI, o próprio sistema redireciona para a página de Wiki, permitindo modificações e inserções nos conteúdos disponibilizados. As possibilidades dessa página contemplam um dos princípios do Wikidata, que é ser uma base de dados ligados colaborativo, que permite com que os usuários possam realizar alterações no conteúdo. Vale destacar que diferentemente do que ocorre ao se realizar uma alteração no Wikipédia, em que o usuário altera a página Web, no Wikidata o usuário seguirá as estruturas da Web Semântica para realizar as alterações, alterando as próprias relações existentes, tendo a possibilidade de corrigir eventuais erros e inconformidades nas ligações existentes.

Sob a ótica da Web Semântica, essa interface permite com que usuários humanos sejam capazes de visualizar de uma forma mais simples as informações contidas na base de dados. Outro ponto a se destacar que traz benefícios, do ponto de vista da semântica dos dados, é a possibilidade de alterações nos dados, como relatado anteriormente. Tal questão permite uma atualização mais constante dos dados, evitando com que o Wikidata se torne obsoleto. Essa problemática recorrente em outras bases de dados do *Linked Data*, é resolvida neste ponto, juntamente com outras medidas, como o uso de robôs, que será abordado no decorrer do texto.

Para a Recuperação da Informação, essa interface visual contempla uma forma visual clara e representativa, sendo uma tela final de recuperação de informação satisfatória para indicar os dados de um recurso.

A figura 1 apresenta uma tela de interface dos dados vinculados, contendo como recurso Tim Berners-Lee, cientista da computação e fundador da Web e da Web Semântica.

Figura 1. Tela do recurso do Tim Berners-Lee
Fonte: Elaborado pelos autores a partir da consulta a página http://www.wikidata.org/entity/Q80 (Acesso em: 23 jan. 2018)

Nesta figura é possível identificar as informações que o recurso possui, além de demonstrar como os usuários realizam as alterações nos conteúdos, estando condicionados a inserirem ou a alterarem os dados seguindo o padrão do RDF, de triplas. Tais relações podem ser visualizadas pelos destaques realizados em vermelho na figura 1.

Uma segunda forma de acessar os dados é por meio da API MediaWiki, que permite que aplicações externas possam consultar os dados acessando um servidor do Wikidata. Na definição desta API, há uma série de parâmetros que possibilitam a consulta, definindo os padrões que devem ser utilizados na busca.

A busca por meio da API MediaWiki tem um caráter importante para a Recuperação da Informação, uma vez que é destinado a aplicações computacionais que desejam construir padrões próprios para consultar os dados estruturados do Wikidata. Uma análise aprofundada

dos parâmetros desta API, indica uma consistência frente aos principais modelos existentes de busca dentro da Web.

Vale destacar que a API permite uma gama de opções que vão desde *login* e *logout* no sistema, passando por questões relativas aos formatos dos dados, chegando até opções relativas a busca em si. No âmbito desta pesquisa, centra-se a análise sobre as possibilidades oferecidas relativas a busca, mais especificamente, a busca sobre entidades, que reflete com precisão a estrutura semântica em que os dados são apresentados.

O quadro 1 apresenta os principais parâmetros para a realização de uma busca por entidades dentro do Wikidata, utilizando a API MediaWiki.

Quadro 1. Parâmetros básicos para a definição de uma busca API MediaWiki

Parâmetro	**Valores possíveis**	**Comentário**
Action	Todas as ações permitidas, no caso para a busca de entidade, utilizam-se o valor "wbsearchentities"	Por meio desse parâmetro define-se qual a ação será tomada dentro da API.
Search	Deve-se buscar por um valor textual (string) contendo os elementos que se deseja buscar.	Esse parâmetro deve ser utilizado, quando a ação for "wbsearchentities", retornando as entidades que possivelmente atendam a busca realizada.
language	Utilizar código de línguas, como pt-br (para português brasileiro) ou en (para inglês)	Definirá a língua que os resultados obtidos fornecem
format	json, jsonfm, none, php, phpfm, rawfm, xml, xmlfm	Auxilia para definir o formato mais adequado para cada aplicação

Fonte: Elaborado pelos autores a partir de consultas a API

A partir desses parâmetros indicados no quadro 1, é possível realizar uma consulta, recuperando uma série de entidades. Desta forma, construiu-se uma busca, seguindo os parâmetros apontados, que obteve a URL para realizar a consulta, o seguinte conteúdo: "https://www.wikidata.org/w/api.php?action=wbsearchentities&search=Tim%20Berners-Lee&language=en". Vale destacar que o único parâmetro não utilizado dos listados no quadro, foi o *format*, pois buscou-se obter somente uma resposta visual de fácil compreensão, além do

mais, poderiam ser utilizados outros parâmetros, para dar um maior nível de detalhamento nos resultados.

Esta consulta construída, busca por entidades referentes a "Tim Berners-Lee", na língua inglesa. Os resultados obtidos são apresentados na figura 2, em que são listadas as entidades que possuem em seu nome o texto informado como busca.


```
{
    "searchinfo": {
        "search": "Tim Berners-Lee"
    },
    "search": [
        {
            "id": "Q80",
            "concepturi": "http://www.wikidata.org/entity/Q80",
            "url": "//www.wikidata.org/wiki/Q80",
            "title": "Q80",
            "pageid": 139,
            "label": "Tim Berners-Lee",
            "description": "British computer scientist",
            "match": {
                "type": "label",
                "language": "en",
                "text": "Tim Berners-Lee"
            }
        },
        {
            "id": "Q22991023",
            "concepturi": "http://www.wikidata.org/entity/Q22991023",
            "url": "//www.wikidata.org/wiki/Q22991023",
            "title": "Q22991023",
            "pageid": 25007543,
            "label": "Tim Berners-Lee: A Magna Carta for the web",
            "description": "TED2014",
            "match": {
                "type": "label",
                "language": "en",
                "text": "Tim Berners-Lee: A Magna Carta for the web"
            }
        },
```


Figura 2. Fragmento obtido a partir da consulta a "Tim Berners-Lee" no API MediaWiki
Fonte: Elaborado pelos autores, a partir da consulta em:
<https://www.wikidata.org/w/api.php?action=wbsearchentities&search=Tim%20Berners-Lee&language=en>.
Acesso em: 25 jan. 2017.

Nos resultados obtidos com a consulta, apresentados na figura 2, é possível identificar que duas entidades foram recuperadas, todas contendo "Tim Berners-Lee" no nome, inserindo o cientista da computação na primeira posição e um livro na sequência. Tal fato se mostra interessante, pois permite que cada aplicação, ao utilizar a API MediaWiki, possa delimitar a forma como será apresentado os resultados aos usuários, ou do uso que será feito de cada entidade.

Outro ponto a se destacar, é que esse tipo de busca, retorna somente um indicativo da entidade localizada, sem apresentar as relações existentes entre as entidades e as propriedades destas. Quando analisado da perspectiva da Web Semântica, esta API contempla elementos fundamentais que tornam a interoperabilidade e a comunicação entre agentes computacionais e conteúdos criados por pessoas, especialmente ao delinear variáveis de identificação consistentes como a URI, que aponta um endereço e um nome para aquela entidade específica.

O terceiro modo apresentado para realizar acesso aos dados é por meio da Consulta Wikidata. Essa ferramenta tem como proposta permitir com que sejam construídas consultas com um alto nível de complexidade, visando relacionar os dados e explorar as entidades do Wikidata.

Essa forma de acesso se mostra a mais convergente com as principais teorias da Recuperação da Informação, uma vez que propicia a criação de consultas utilizando modelos deste campo de estudos. Dentre eles, destaca-se a possibilidade de uso de operadores lógicos como o *AND* e o *OR*, permitindo a construção de consultas booleanas para recuperar as informações mais aderentes as necessidades informacionais.

Outro destaque desta ferramenta fica por conta da interface que permite buscas utilizando as estruturas de propriedades das triplas RDF, além de dar a possibilidade de buscar sobre as instâncias e as subclasses que organizam as informações dentro dessa base de dados. As funções permitidas por esse mecanismo, explicitam uma relação direta com os conceitos da Web Semântica, pois está aderente as funcionalidades desta proposta, especialmente ao que condiz das triplas RDF.

A construção de consultas neste mecanismo apresenta algumas dificuldades, pela complexidade possibilitada pela ferramenta, contudo, é capaz de atender com precisão o objetivo de consultar os dados do Wikidata, inserindo uma série de elementos que tornam mais aprimorado tal processo. As opções disponíveis para a construção das consultas, podem ser analisadas na página da WikidataQuery API (2017) (https://wdq.wmflabs.org/api_documentation.html)

Visando demonstrar como funciona essa forma de consulta, a figura 3 representa a tela padrão para a construção das consultas, que juntamente com as recomendações para uso desta ferramenta, permite a construção de consultas no Wikidata.

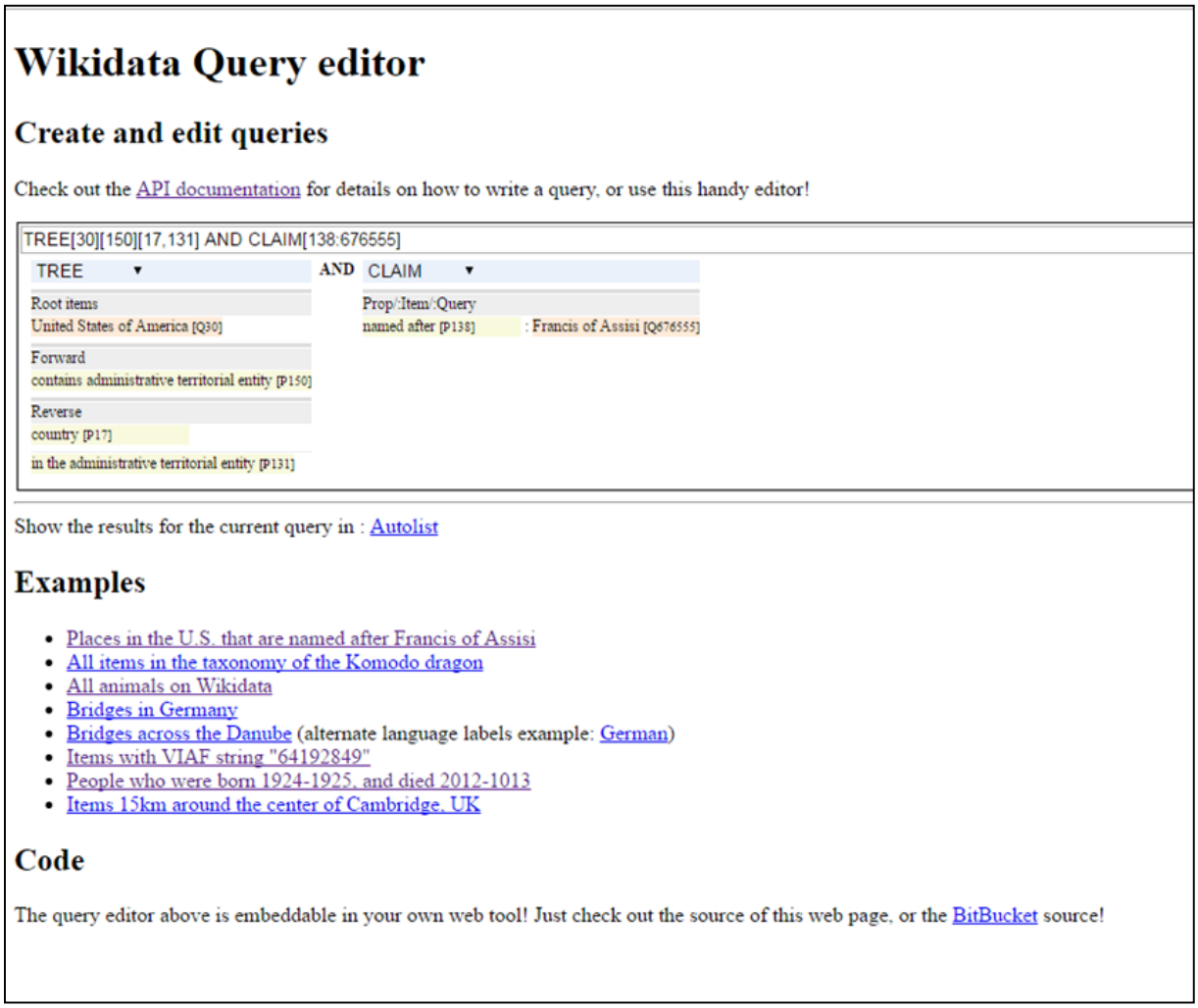

Figura 3. Exemplo de utilização da consulta Wikidata
Fonte: Elaborado pelos autores

O exemplo apresentado na figura 3, contempla uma busca de exemplo dada pelo própria Consulta Wikidata, em que são apresentados os locais dos Estados Unidos cujo o nome contém Francisco de Assis. Neste exemplo é possível identificar a ocorrência da lógica booleana *AND*, ao relacionar duas informações da busca, bem como se verifica o uso das propriedades do RDF para buscar as informações, no caso se utilizou a propriedade "*named after*" para traçar o relacionamento.

Os resultados obtidos com essa consulta são listados apresentando as entidades que correspondem a busca realizada, de forma semelhante ao demonstrado com o uso da API MediaWiki.

Um outro meio de se acessar os dados do Wikidata ocorre com a utilização do SPARQL *Endpoint*. Esta forma é a mais integrada com as propostas da Web Semântica, pois está centrada

em uma das tecnologias de base proposta, o SPARQL, responsável por ser a linguagem de consultas dos dados em RDF e em OWL. O SPARQL *Endpoint* é um ambiente aberto para consultas aos dados de uma base de dados, que por meio do SPARQL, permite aos usuários, humanos ou não, a acessarem a tais dados.

O SPARQL *Endpoint* do Wikidata representa a forma de consulta aos dados que mais aproveita os potenciais da Web Semântica, por ser nativamente integrada às demais tecnologias da Web Semântica e do *Linked Data*. Desta forma, torna-se possível explorar com mais precisão o contexto que os dados se encontram, ao mesmo tempo que permite a construção de inferências e de axiomas capazes de expandir a busca realizada.

Além disso, o SPARQL favorece a integração com outras plataformas que também possuem o SPARQL *Endpoint*. Desta forma, a linguagem RDF juntamente com o SPARQL favorece a interoperabilidade entre as informações de dados disponíveis em formatos compatíveis com o *Linked Data*, favorecendo o reuso e a troca de dados.

Dentro da perspectiva da Recuperação da Informação, o SPARQL é capaz de fornecer um ferramental que explora os dados, sem seguir os paradigmas clássicos desse campo de estudos. Contudo, identifica-se que o SPARQL atua em uma etapa da Recuperação da Informação, estando vinculada aos processos de localização dos dados e visando o relacionamento entre as informações.

Como demonstração e exemplo de uso realizou-se uma busca através do SPARQL *Endpoint*, que mostra todas as pessoas que nasceram na cidade de Marília, Brasil.

Tal busca expressa uma das possibilidades que o trabalho com SPARQL permite, em que é possível traçar relacionamentos, buscando por informações de um nível de especificidade bem elevado. Em suma, localizar todas as personalidades que nasceram na cidade de Marília, caso não se utilize uma ferramenta com essa, não é simples de chegar no resultado, pois por vezes é necessário percorrer diversos sites, localizando em alguns espaços onde tal informações pode ser encontrada. O SPARQL torna esse processo mais simples, uma vez que utiliza das relações existentes para promover essa busca.

A figura 4 demonstra os resultados alcançados com a busca pelas pessoas nascidas em Marília, no próprio ambiente do SPARQL *Endpoint*. Vale destacar que a busca foi limitada as cinco primeiras pessoas recuperadas, além de que a propriedade P19 tem como característica apresentar o local de nascimento de uma pessoa.

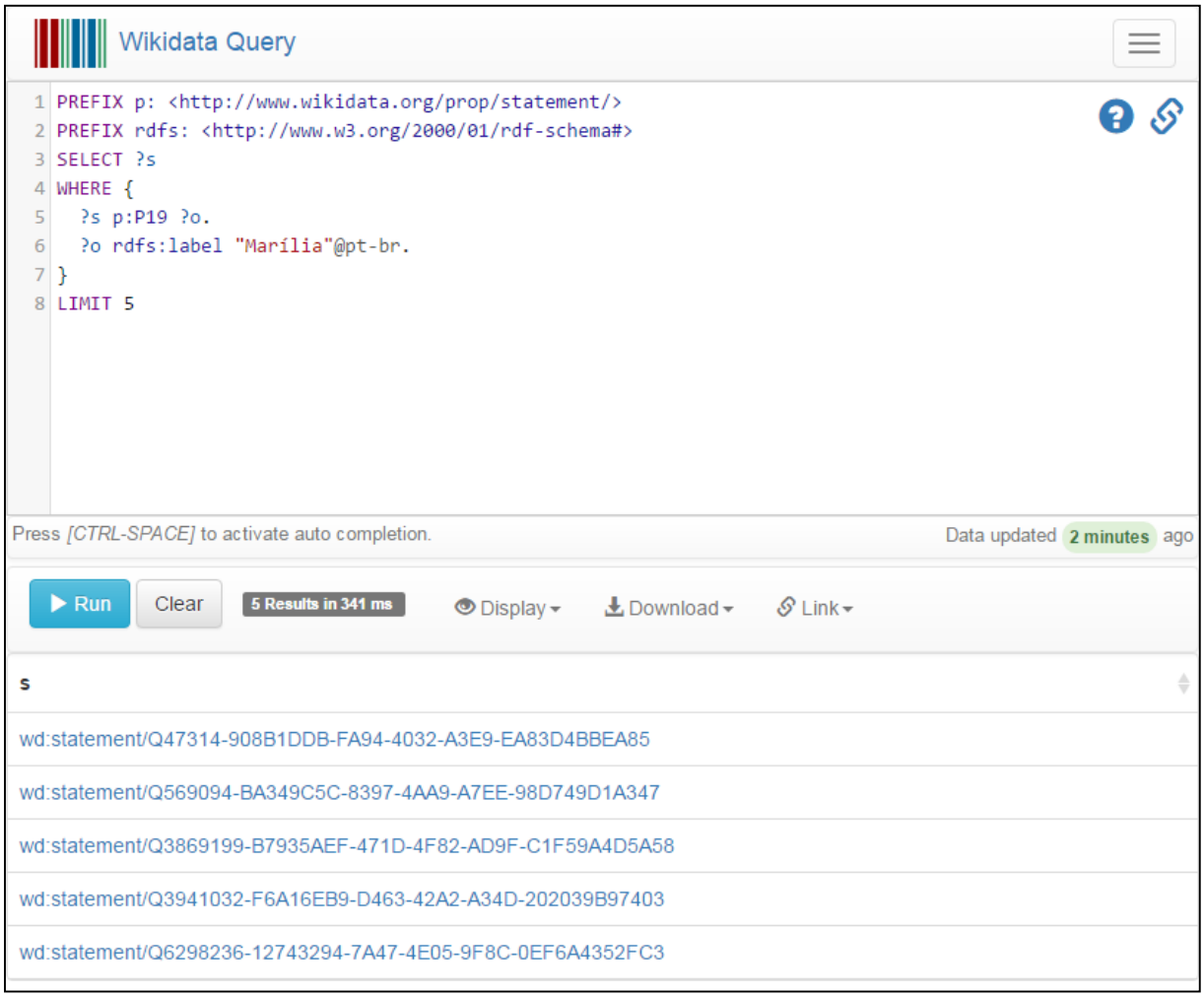

Figura 4. Busca por pessoas nascidos na cidade de Marília
Fonte: Elaborado pelos autores a partir de consulta no SPARQL *Endpoint* do Wikidata.

Os resultados obtidos com a consulta apresentada, são entidades que apresentam uma série de outras propriedades. Tal característica permite com que sejam explorados tais dados a fundo somente com a utilização da linguagem SPARQL, permitindo assim uma gama de opções bem ampla, que pode ser utilizada para relacionar determinadas informações, bem como em programas computacionais que visam utilizar dados na Web, como exemplo de sistemas de tomada de decisões, entre diversos outros cenários em que pode ser útil o uso desses dados.

A última forma de acesso aos dados por item, são os chamados robôs, que tem como função realizar alterações e inserções de informações de forma automática. Tais robôs são construídos pelos próprios usuários que visam colaborar com a proposta, no entanto, a facilidade e a velocidade com que estes robôs são capazes de alterar os conteúdos do Wikidata,

conduziu a comunidade a traçar uma política para a construção dos robôs. (WIKIDATA, 2017b).

Em síntese, essa política definirá questões fundamentais sobre alterações e inserções de conteúdo, inserindo juntamente um processo de aprovação dessas modificações por meio de alguns robôs chamados de robôs administradores. Além disso, há uma série de requisitos que os robôs devem possuir para acessarem aos dados do Wikidata.

O uso desses robôs é um dos principais destaques do Wikidata, frente a outras bases de dados do *Linked Data*, devido a velocidade com que os dados podem ser inseridos ou alterados, mantendo-se sempre atualizados.

Por fim, a outra forma de acesso aos dados é o Acesso aos despejos, ou acesso aos dados brutos (*database dumps*). Essa possibilidade vai ao encontro das diretrizes do *Linked Data* que propaga que os dados devem ser disponibilizados abertamente, possibilitando com que cada usuário faça o uso da melhor maneira possível para suas condições.

O *download* desses dados brutos pode ocorrer em diversos formatos, como em JSON (*JavaScript Object Notatio*), XML e RDF, permitindo com que o usuário identifique o formato que melhor atende às suas necessidades. Como relatado, sob a ótica da Web Semântica, esse acesso é bastante interessante, para que novas aplicações possam ser construídas, e caso nenhuma das opções apresentadas de acesso aos dados seja útil a uma aplicação, esta poderá criar formas mais adequadas de acesso a partir dos dados brutos.

Dentro da concepção da Recuperação da Informação, a opção do *download* aos dados brutos deixa nas mãos da aplicação a melhor maneira de construir os processos para recuperar os dados. Porém, ao permitir diversas opções de disponibilização dos dados, facilita a construção de formas eficientes de recuperação, especialmente pela versatilidade dessa opção.

Todas as opções apresentadas estão centradas na Web Semântica, bem como em possibilitar formas de aprimorar a Recuperação da Informação em um contexto de alterações, tanto da forma como as informações são inseridas na Web, quanto em relação ao modo como as aplicações estão consumindo os dados. As tecnologias da Web Semântica e do *Linked Data* estão auxiliando a uma nova forma de organizar e de recuperar os dados dentro da Web, visando inserir mais informações referentes ao contexto que tais dados possuem.

Desta forma, o Wikidata apresenta uma série de opções que estão embasados fortemente nas teorias e nas tecnologias tanto da Web Semântica, quanto da Recuperação da Informação, sendo uma plataforma bastante atualizada, estando adequada as necessidades atuais da Web.

Os cabeçalhos das seções/subdivisões devem ser breves e claros. O texto do artigo deve ser estruturado preferencialmente contemplando os seguintes itens: introdução, método, resultados e considerações finais. Acrônimos e abreviações devem estar entre parênteses e serem precedidos de seu significado completo quando do primeiro uso no texto.

5 Considerações Finais

A influência dos conceitos e das tecnologias da Web Semântica está ocorrendo em diversas vertentes e campos de estudos, em especial a todas aquelas que estão direta ou indiretamente relacionadas a Web. Essa influência tem como princípio tornar a relação das pessoas com os mecanismos computacionais mais natural, uma vez que permite uma integração e uma compreensão dos conteúdos disponibilizados na Web por estes dois atores.

Dentro desta perspectiva, a criação de conteúdos pelos próprios usuários, que possui na Wikipédia seu principal expoente, provocou uma revolução na Web, uma vez que a quantidade de informações disponibilizadas para consultas cresceu a medida que o número de usuários aumentava. No entanto, tais informações estavam compreensíveis, em sua maioria, somente pelas pessoas, tendo os agentes computacionais, sistemas de Recuperação da Informação e motores de busca um acesso apenas sintático aos conteúdos destas plataformas.

Visando solucionar tal questão, a Web Semântica foi proposta tendo por objetivo acessar de maneira "inteligente" dados heterogêneos, trazendo um salto de eficiência na forma em que as informações são buscadas e exibidas. Essa eficiência se dá pelo fato de que as informações passam a terem sentido, permitindo a compreensão dos conteúdos pelos seres humanos, mas também pelos sistemas computacionais.

Para que todos estes conceitos possam realmente ser útil para as pessoas, usufrutuárias inerentes das tecnologias, os sistemas de Recuperação da Informação possuem um papel fundamental neste contexto, pois são tais sistemas que tornam as informações tangíveis ao usuário, seja promovendo a busca feita em documentos, seja ainda na busca por metadados. Desta forma, é necessário aproximar a Recuperação da Informação e a Web Semântica, quando se analisa as iniciativas que são inspiradas pela Web Semântica, que tem como propósito reunir dados dos mais diversos domínios.

Neste sentido, o presente estudo aprofundou a compreensão acerca do Wikidata, analisando como esta aplicação baseada nos princípios da Web Semântica e do *Linked Data*, promove a Recuperação da Informação de seus dados.

Os resultados obtidos apontaram diversas maneiras para que usuários, tanto humanos quanto não humanos, possam acessar e utilizar os dados ligados do Wikidata. As análises realizadas em cada uma das formas de se acessarem os dados, demonstrou particularidades, tanto quanto ao uso dos conceitos da Web Semântica, quanto dos métodos de Recuperação da Informação promovido.

Essas diferenciações demonstram que o Wikidata está propenso a abarcar um extenso número de aplicações que desejam utilizar sua base de dados, como uma fonte informacional de *Linked Data*, para os mais distintos propósitos. Outro ponto de destaque, ocorre pela solidez que as soluções de busca dos dados apresentam, em que existe uma atenção em manter a estrutura dos dados nos formatos de origem, no caso o RDF.

Todas estas formas diferenciam os modos de se recuperar os dados, sem perder a origem baseada nos formatos da Web Semântica e do *Linked Data*. Vale ressalvar, que o SPARQL se mostra como a linguagem mais preparada para recuperar os dados, porém a opção de acessar com soluções mais comuns na Web, como APIs de consultas, tornam o Wikidata mais abrangente.

Portanto, o Wikidata pode ser visto como uma solução que é capaz de reunir conteúdos que a princípio era compreensível somente por pessoas, originários em sua maioria da Wikipédia, em formatos entendíveis por máquinas, em especial o RDF e o OWL. O trabalho realizado demonstra ainda que essa base de dados, o Wikidata, pode ser de bastante destaque e importância na Web, por possuir mecanismos de busca e de Recuperação da Informação eficientes e acessíveis por aplicações construídas de diferentes configurações.

Referências

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

DZIEKANIAK, G.,V.; KIRINUS, J. B. Web Semântica. Semantic Web. **Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf**., Florianópolis, n.18, 2º sem. 2004. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2004v9n18p20. Acesso em: 31 jan. 2018.

SANTAREM SEGUNDO, J. E.; SOUZA, J. O.; CONEGLIAN, C. S. Web semântica: introdução a recursos de visualização de dados em formatos gráficos. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., João Pessoa, PB. 2015. **Anais eletrônicos...** João Pessoa, PB: ANCIB, 2015. Disponível em:

http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/view/2780. Acesso em: 2 fev. 2018.

SANT'ANA, R. C. G. A importância do papel do profissional da ciência da informação nos processos de recuperação de conteúdos digitais estruturados. *In*: GUIMARÃES, José Augusto Chaves; FUJITA, Mariângela Spotti Lopes (Org.). **Ensino e pesquisa em biblioteconomia no Brasil**: a emergência de um novo olhar. Marília: Cultura acadêmica, 2008. p. 145-154.

SILBERCHATZ, A.; KORTH, H. F.; SADARSHAN S. **Sistema de banco de dados**. 6° edição traduzida. Editora Elservier. 2012.

WIKIDATA QUERY API. **Query construction**. 2017. Disponível em: https://wdq.wmflabs.org/api_documentation.html. Acesso em: 27 jan. 2018

WIKIDATA. **Wikidata**: robôs. 2017b. Disponível em: https://www.wikidata.org/wiki/Wikidata:Bots/pt-br. Acesso em: 27 jan. 2018.

WIKIDATA. **Wikidata**: acesso aos dados. 2016. Disponível em: https://www.wikidata.org/wiki/Wikidata:Data_access/pt-br. Acesso em: 25 jan. 2018.

W3C. **Web semântica**. 2015. Disponível em: http://www.w3c.br/Padroes/WebSemantica. Acesso em: 31 jan. 2018.

W3C. **SPARQL query language for RDF**. 2008. Disponível em: https://www.w3.org/TR/rdf-sparql-query/. Acesso em: 31 jan. 2018

W3C. **OWL**: web ontology language (OWL). 2012. Disponível em: https://www.w3.org/OWL/. Acesso em: 31 mar. 2017

W3C. **Extensible Markup Language**. 2015. Disponível em: https://www.w3.org/XML/. Acesso em: 30 jan. 2018.