Uma Abordagem sobre a Estrutura *Geonames* e suas Contribuições para o *Linking Open Data*

José Eduardo Santarem Segundo
Universidade de São Paulo – USP, Email: santarem@usp.br

Ana Carolina Simionato
Universidade Federal de São Carlos – UFSCar, Email: acsimionato@ufscar.br

RESUMO

A temática do ENANCIB 2016 faz refletir a composição da estrutura informacional dos dados produzidos por inúmeras formas, sejam por humanos ou por máquinas. No entanto, os dados oferecem um nível semântico que outras estruturas mais consolidadas não permitem. Dessa forma, a relação intrínseca entre a Ciência da Informação e a Ciência da Computação é cada vez mais próxima principalmente relacionada aos processos metodológicos para agregar semântica nas ligações entre os dados. Além disso, a amplitude dessa relação transcendem os diferentes domínios do conhecimento, incidindo em diversas preocupações e melhorias sobre as especificidades de cada campo, como as tentativas de ligação de dados de geolocalização do projeto *Geonames*. Nesse contexto, o problema caracteriza-se em como o projeto *Geonames* tem auxiliado na estruturação de dados de geolocalização para reutilização de outros *datasets* a partir das ligações de dados na Web Semântica? Assim, o objetivo desse trabalho é apresentar as contribuições do projeto *Geonames* para o desenvolvimento da *Web* Semântica e do *Linking Open Data*. Em relação aos procedimentos metodológicos foi adotada a pesquisa de natureza teórica aplicada e qualitativa, objetivando analisar as contribuições do *Geonames* e do *Linking Open Data*. Em relação aos objetivos do trabalho essa pesquisa é classificada como exploratória. Como resultados, apresentou a estrutura do projeto *Geonames*, suas possibilidades de contribuição no *Linking Open Data*, além de algumas consultas com protocolo SPARQL para verificar suas funcionalidades. Portanto é possível considerar que as melhores práticas de *Linked Data* estão em expansão e já passíveis de utilização, entretanto ainda carecem de pesquisas e recursos tanto estruturais como tecnológicos.

Palavras-chave: Geonames. Linked data. Web Semântica. Geolocalização.

---

1 INTRODUÇÃO

"Descobrimentos da Ciência da Informação: desafios da Multi, Inter e Transdisciplinaridade", tema do XVII Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação (ENANCIB), faz refletir a composição da estrutura informacional dos dados produzidos por inúmeras formas, sejam por humanos ou por máquinas. São as partículas sem significação direta e em grande parte, nem sempre estão estruturadas de forma coerente para serem disponíveis aos usuários, como o que acontece com o fenômeno *Big Data*, cominado pelos fatores de velocidade, volume, variedade e complexidade dos dados.

No entanto, os dados oferecem um nível semântico que outras estruturas mais consolidadas não permitem. Um exemplo disso, são os dados de geolocalização, que não passam de estruturas numéricas. No entanto, conjunto a uma simbologia adequada, os dados podem representar atributos de importância imensurável sobre as temáticas de geolocalização, climatologia e até mesmo, privacidade de dados.

As tecnologias semânticas e a ubiquidade dos dados são campos de atuação e de pesquisa ligados aos conceitos da *Web* Semântica, provenientes da Ciência da Computação. Entretanto, a interdisciplinaridade com a Ciência da Informação e com outras áreas do conhecimento, oferecem diálogos necessários para o desenvolvimento e o aprofundamento de diversos temas, proporcionando mais agilidade aos processos de busca e recuperação da informação e dados, em diferentes domínios, entre eles os dados de geolocalização.

Nesse contexto, a relação intrínseca entre as duas áreas é cada vez mais próxima, em processos metodológicos para agregar semântica nas ligações entre os dados, antepostos aos conceitos de *Web* Semântica e princípios *Linked Data* e *Linked Open Data*. Além disso, a amplitude dessa relação transcendem os diferentes domínios do conhecimento, incidindo em diversas preocupações e melhorias sobre as especificidades de cada campo, como as tentativas de ligação de dados de geolocalização, apresentados pelo projeto *Geonames*.

Dessa forma, o problema caracteriza-se em como o *Geonames* tem auxiliado na estruturação de dados de geolocalização para reutilização de outros *datasets* a partir das ligações de dados na *Web* Semântica? Assim, o objetivo desse trabalho é apresentar as contribuições do projeto *Geonames* para o desenvolvimento da *Web* Semântica e do *Linking Open Data*, tendo especificidades do objetivo, identificar a estrutura funcional dos elementos geográficos do *Geonames*; analisar o mapeamento de elementos utilizado pelo *Geonames*; analisar como os *datasets* se relacionam com o *Geonames* e como esse contribui de forma decisiva para a evolução da *Web* Semântica e *Linking Open Data*; analisar a capacidade de contribuição do *Geonames* na recuperação de dados interligados com outros *datasets*, por meio do protocolo SPARQL.

Em relação aos procedimentos metodológicos para o desenvolvimento da pesquisa, foi adotada a pesquisa de natureza teórica e aplicada, de caráter qualitativo, objetivando analisar as contribuições dos Geonames e do Linking Open Data. Em relação aos objetivos do trabalho, para o uso dos *Geonames* no desenvolvimento da *Web* Semântica e *Linking Open Data* essa pesquisa é

classificada como exploratória, visto que contempla atividade de uso dos recursos do *dataset Geonames* por meio do protocolo SPARQL para analisar os recursos que o *Geonames* tem a oferecer.

2 *WEB SEMÂNTICA E LINKING OPEN DATA*

A estruturação semântica e sintática dos dados proposta em 2001 por Benners-Lee, Hendler e Lassila, ficou conhecida como *Web* Semântica. A ideia era que os dados fossem utilizados para realizar descobertas, ampliar integração entre sistemas e reusados por meio de várias aplicações. A *Web* Semântica é caracterizada por melhorias no tratamento informacional do conteúdo digital, bem como proporcionar ao usuário uma busca, localização e recuperação mais eficiente e simplificada do conteúdo que procuram.

Como foi também denominada, a *Web* 3.0 é vista como um composto de tecnologias que se complementam e atribuem um valor a cada relacionamento. A proposta estrutural relacionada ao conteúdo significativo de páginas *Web*, cria um ambiente onde agentes possam realizar tarefas facilmente de uma página para outra e cada vez mais sofisticadas para os usuários. (BERNERS-LEE; HENDLER; LASSILA, 2001).

Para isso, Benners-Lee, Hendler e Lassila (2001) descrevem que os sistemas de conhecimento de representação dos dados normalmente são centralizados. Dessa forma, são exigidos um compartilhamento com a mesma definição de conceitos comuns, podendo ser definidos e ligados de uma forma legível pelas máquinas.

Sendo considerada uma extensão da própria *Web*, a *Web* Semântica deve-se preocupar com os usuários, a descentralização e abertura dos dados, para que as tecnologias atreladas sejam eficientes e a interação seja um meio universal para o intercâmbio de dados, de informações e de conhecimento.

A *Web* Semântica foi projetada em camadas, desde o desenvolvimento estrutural, sintático, semântico, ontológico e lógico. Cada uma dessas camadas apresentam tecnologias apropriadas para atender às necessidades de compartilhamento entre os agentes em diferentes aplicações de domínio, propiciando a interpretação dos conteúdos disponíveis na *Web* e com resultados mais eficientes. (BERNERS-LEE; HENDLER; LASSILA, 2001, SOUZA; ALVARENGA, 2004, RAMALHO, 2006; CATARINO; SOUZA, 2012, MARCONDES, 2015).

Por exemplo, os buscadores identificarão se a expressão de busca 'rosa' refere a coloração 'rosa', ou a botânica '*Rosoideae*', ou ainda, ao nome próprio. Nesse sentido, motores de busca como o *Wolfram Alpha*[1] revitalizam competências entre máquinas e usuários.

Santarém Segundo (2014, p. 3864) afirma que as tecnologias da *Web* Semântica "[...] estão diretamente relacionadas ao processo de construção da informação e armazenamento das mesmas, constituindo assim ambientes que possam ter conjunto de dados ligados semanticamente."

A camada ontológica da *Web* Semântica é caracterizada por estabelecer o vínculo semântico dos dados. Entretanto, para que isso seja possível, a sua composição deve ser estruturada pela *Web Ontology Language* (OWL) e serve de apoio as arquiteturas *Resource Description Framework* (RDF), além das recomendações feitas pela *World Wide Web Consortium* (W3C). Em suma, o RDF fornece uma infraestrutura que possibilita a troca de informação na *Web* e a linguagem OWL é utilizada na publicação e compartilhamento dos termos denominados como ontologias, apoiando em pesquisas avançadas na *Web*, pelos agentes de software e gestão do conhecimento.

Nessa divisão semântica, a estruturação das ontologias tem como propósito facilitar o compartilhamento do conhecimento em áreas distintas para o reuso dos dados. Ontologia define os termos usados para representar uma área do conhecimento por meio de definições e conceitos básicos, de diferentes domínios e legíveis por máquina.

Santarem Segundo (2015, p. 226) afirma que:

> Utilizar ontologias é uma das maneiras de se construir uma relação organizada entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação.

Se conceitualmente as ontologias fortalecem a estrutura organizacional dos conjuntos de dados que se baseiam nas tecnologias da *Web* Semântica, as linguagens computacionais capazes de transformá-las em artefatos computacionais, permitem que se produza a grande massa de dados ligados como vem acontecendo com o *Linking Open Data*.

[1] *Wolfram Alpha:* https://www.wolframalpha.com/

Diante da estruturação proposta pela *Web* Semântica, uma forma para redução de barreiras e que possibilita o relacionamento e interoperabilidade de dados, é a ligação dos mesmos, denominada como *Linked data* por Berners-Lee no ano de 2006. A W3C (2013, não paginado, tradução nossa) define *Linked Data* como uma "[...] coleção de conjunto de dados inter-relacionados na *Web* [...]". Heath e Bizer (2011, não paginado, tradução nossa) explica que o *Linked Data* refere-se ao "[...] uso da Web para conectar dados relacionados que não foram previamente ligados, ou usando a Web para reduzir as barreiras de ligação de dados atualmente ligados através de outros métodos."

Bizer, Heath e Berners-Lee (2009, p. 206, tradução nossa) explicam que

> O termo *Linked Data* refere a um conjunto de melhores práticas para a publicação e conexão de dados estruturados na Web. Essas melhores práticas foram adotadas por um crescente número de provedores de dados ao longo dos últimos três anos, levando à criação de um espaço global de dados, que contém bilhões de afirmações - a Web de Dados. (BIZER; HEATH; BERNERS-LEE, 2009, tradução nossa).

O princípio *Linked data* é encontrado e definido por vários autores, Catarino e Souza (2012, p. 79) apontam o *Linked data* como coleções de dados relacionados na *Web*. Sendo considerado, como uma forma de utilizar a *Web* como conjuntos de dados, interligando entre si, criando novas formas e mais específicas aos usuários. O *Linked data* também é entendido como um "[...] estilo de publicar e interligar dados estruturados na Web. *Linked Data* não representa propriamente uma nova tecnologia, mas sim um conjunto de melhores práticas para publicação e interligação de dados estruturados na Web." (PIZZOL; TODESO; TODESCO, 2016, p. 93). Pizzol, Todeso e Todesco (2016) ainda destacam que a arquitetura da *Web* deve ser utilizada para o compartilhamento de dados estruturados em *links hyperdata*, onde as informações dos recursos podem ser ligadas.

Para o *Linked data* é necessário que os dados estejam em formato *Resource Description Framework* (RDF), os objetos precisam ser identificados por *Universal Resource Identifier* (URIs), os dados devem estar acessíveis via *Hypertext Transfer Protocol* (HTTP); e os objetos devem ser referenciados por meio de suas URIs. (BERNERS-LEE, 2006).

Para que os documentos sejam publicados na *Web* é necessário que eles obedeçam a um esquema de requisitos para publicação, o *Five-star open data*. Em suma, os requisitos devem

conter uma licença aberta (1); os dados serem estruturados de forma legível por máquina e serem editáveis (2); publicados em formatos não proprietários, por exemplo, a extensão *.csv* (3); publicados em padrões abertos utilizando de URIs para identificação e direcionamento do item (4), por fim, os dados podem estar vinculados a outros dados, contribuindo para outro conjunto de dados (5). (BERNERS-LEE, 2006).

Um dos principais difusores do *Linked Data* é o projeto *Linking Open Data* (LOD), mantido pelo W3C desde sua criação em 2007. O objetivo é a prática a *Web* de Dados com os princípios da *Web* Semântica e com as regras de publicação do *Linked data*, identificando os conjuntos de dados (*datasets*) em formato abertos e baseados em RDF.

Ressalta-se que o formato de dado aberto condiz à sua forma de utilização, ou seja, o usuário "[...] pode livremente usá-los, reutilizá-lo e redistribuí-los, estando sujeito a, no máximo, a exigência de creditar a sua autoria e compartilhar pela mesma licença.", segundo o *Open Knowledge Foundation* (201-, não paginado).

Desse modo, o projeto *Linking Open Data* apontam maneiras para a ligação e publicação dos dados abertos estruturados na *Web*, permitindo a conexão entre diversas fontes de dados. Santarém Segundo (2015, p. 225) afirma que o LOD retrata como um "[...] conjunto de normas a serem seguidas, que usa os mesmos princípios de ligação semântica da *Web* de Dados, entretanto tem particularidades específicas, indicando um grau de exigência maior na constituição de sua rede de interligações."

Santarém Segundo (2015, p. 224) afirma ainda que

> Estruturar dados abertos de forma semântica não é apenas uma das formas de estabelecer a ligação entre o conceito de Dados Abertos e de Web Semântica, mas sim de estabelecer um modelo de estrutura de dados que favoreça o atendimento ao quinto princípio de dados abertos e também ao inciso da Lei de Acesso a Informação, que indicam a possibilidade dos dados serem processados por máquina, além da ligação entre informações de bases diferentes através de relacionamentos semânticos.

No projeto LOD, os *datasets* são integrados para o compartilhamento global, de maneira crescente desde a sua criação em 2007. O LOD é apresentado em um diagrama e sua última atualização entre as ligações está sua divisão entre temáticas, relacionadas à dados de publicação; ciências da vida; domínio geral (*cross-domain*); dados geográficos; dados governamentais; mídia; dados de uso geral; dados de redes sociais e linguística.

A publicação de dados no projeto LOD possibilita uma maior flexibilidade e integração de inúmeras fontes, possibilitando a universalização do uso e a estruturação em RDF. O uso de URIs como identificadores universal também permite que os *hyperlinks* sejam definidos por entidades diversas, a exemplo dos dois principais conjuntos de dados que mais se relacionam no LOD, que são o *DBpedia* e o *Geonames*.

3 *GEONAMES*

A facilidade oferecida pela estruturação dos dados e pelos relacionamentos semanticamente ricos pode viabilizar a combinação de diferentes domínios. Ao associar um nome ao lugar pode não ser uma tarefa simples quando envolve diferentes contextos, e seus resultados podem ser ambíguos, 'São Paulo' por exemplo, pode referir-se a um santo católico, um logradouro, uma cidade, além de um estado brasileiro.

O *Geonames* é uma base de dados disponível pela licença *Creative Commons* para *download* gratuito referente ao domínio de geolocalização. Segundo o *Geonames* (2016, não paginado), seus registros possuem mais de 10 milhões de nomes geográficos, que consiste em mais de 9 milhões de recursos exclusivos, onde 2,8 milhões são de lugares povoados e 5,5 milhões nomes alternativos.

Os dados do *Geonames* são consumidos por uma grande quantidade de entidades e organizações, algumas conhecidas e reconhecidas internacionalmente como produtoras de conteúdo: *New York Times*, BBC, *Digital Globe*, além de outros tipos de organizações como *Nike* e *Adidas*. Há ainda uma infinidade de outras organizações menos conhecidas mas que utilizam-se constantemente da base de dados do *Geonames*. O *Geonames* atende a mais de 150 milhões de requisições *Web* por dia.

O consumo dos dados do *Geonames* é realizado por meio de uma série de *webservices*, além de permitir consulta diretamente por humanos e também *download* de seus conjuntos de dados, para uso e navegação *offline*.

O *Geonames* conta com uma ampla fonte de dados para alimentar sua base, recebendo dados de agências cartográficas nacionais do mundo todo, como o *National Geospatial Intelligence Agency's* (NGA), Instituto Brasileiro de Geografia e Estatística (IBGE) e outras fontes como os serviços *hotels.com* e *alpharooms.com*.

Os pontos fortes do *Geonames* são: a integração dos dados geográficos, tais como nome de países, cidades, divisões administrativas e elementos geográficos, entre rios, montanhas e planícies, do mundo todo e a incorporação de coordenadas de latitude e longitude, sendo editadas, adicionadas e corrigidas, manualmente ou utilizando uma *wiki* com uma interface amigável. (GEONAMES, 2016).

Os recursos do *Geonames* estão divididos em nove classes e posteriormente subcategorizados em 645 categorias, cada uma recebendo um tipo de código diferente. As nove classes principais são:

- área com fronteira administrativa (países, estados, subregiões administrativas, área eclesiástica, etc.);
- áreas com características hidrográficas (rios, oceanos, mares, lagos, canais, etc.);
- áreas delimitadas (parques, áreas militares, reservas ambientais, reservas agrícolas, portos, regiões econômicas, regiões culturais, etc.);
- áreas povoadas (cidades, vilas, vilarejos, aldeias, capitais, etc.);
- estradas (rodovias, ferrovias, gasodutos, trilhas, túneis, etc.);
- áreas identificadas como construções (aeroportos, casinos, castelos, hotéis, refinarias, estádios, estações de rádio, universidades, zoológicos, etc.);
- áreas com características hipsográficas - representação gráfica das variações de altitude da crosta terrestre (montanhas, praias, deltas, desertos, ilhas, vulcões, penínsulas, picos, etc.);
- áreas submersas (bacias, gargantas, aterros, recifes, etc.); e
- áreas de vegetação (florestas, tundras, áreas de cultivo, pomares, vinhas, cerrados...).

Por meio destas classes e suas subcategorias os recursos estão disponíveis para serem consumidos e utilizados.

Cada um dos aproximadamente 10 milhões de registros geográficos do *Geonames* é representado com um conjunto básico de elementos, sendo eles: número ID ou código *Geonames*, nome geográfico, nomes alternativos, latitude, longitude, código de classe (classes listadas anteriormente), código de categorização, código de país (baseado na ISO-3166), código

alternativo de país, até quatro códigos administrativos (regiões e subregiões), população, elevação (em metros), área no fuso horário e ultima data de modificação.

Como pode ser visto na figura 1, o acesso para usuários comuns, via *browser*, permite identificar algumas dessas informações. Utilizou-se neste caso a cidade de Salvador, capital baiana e sede do XVII ENANCIB em 2016.

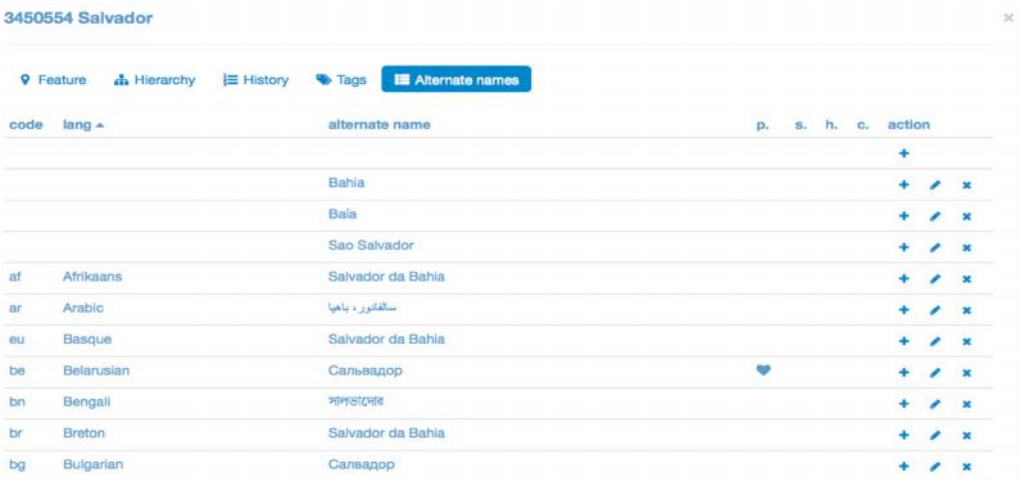
Figura 1 - Informações de Salvador no *Geonames*

Fonte: Elaborado pelos autores.

Algumas das informações disponíveis também podem ser acessadas com telas e opções alternativas a imagem apresentada.

Cada registro, conforme pode ser observado na figura 2, pode receber uma infinidade de nomes alternativos, que são descritos de acordo com os seguintes elementos: nome alternativo (*alternate names*), código do idioma, descrição do idioma, nome preferido, nome curto (*short name*), nome coloquial e nome histórico.

Figura 2 - Nomes alternativos para Salvador no *Geonames*



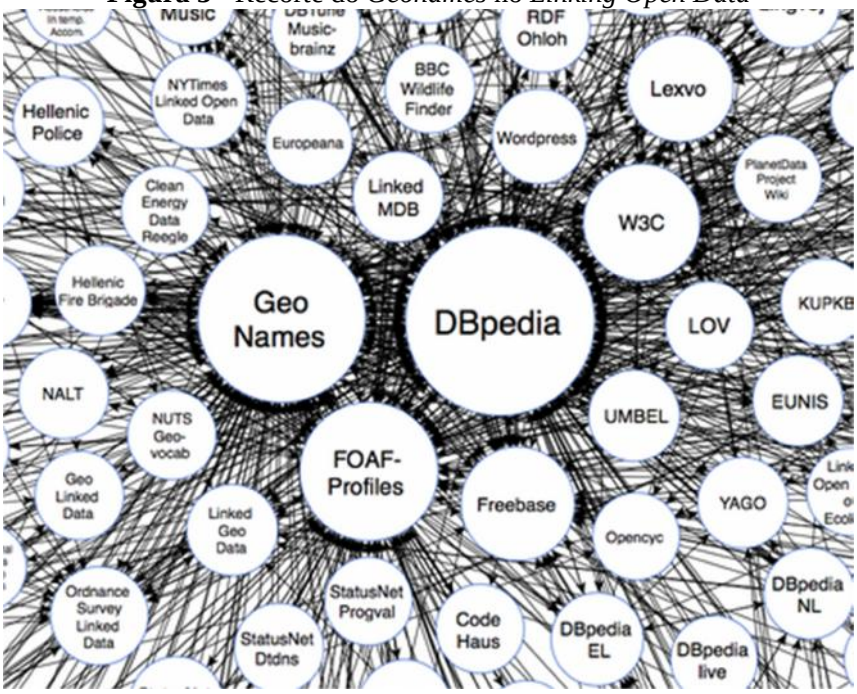
Fonte: Elaborado pelos autores.

As informações como apresentadas nos trazem uma ideia da quantidade de dados que podem ser aproveitadas ao acessar a base do *Geonames*, entretanto ainda não nos caracteriza sua particular e expressiva participação como *dataset* do *Linking Open Data*, conforme será tratado a seguir.

4.1 *GEONAMES NO LINKING OPEN DATA*

O *dataset* do *Geonames* é tido como uma dos mais importantes e também um dos mais representativos dentro do grande grafo do *Linking Open Data*. Como pode ser observado na figura 3, baseada no diagrama apresentado em 2014, pelo grupo dos pesquisadores Richard Cyganiak e Anja Jentzsch. O *dataset* do *Geonames* é representado como uma das circunferências mais expressivas, dada principalmente a quantidade de registros que compartilha e também a quantidade de *datasets* que apontam para o *dataset* do *Geonames*.

Figura 3 - Recorte do *Geonames* no *Linking Open Data*



Fonte: Elaborado pelos autores.

Segundo o *State of the LOD Cloud 2014*, o *Geonames* é o segundo *dataset* que mais recebe conexões de outros *datasets* no *Linking Open Data*, sendo ao todo 141 *datasets* do total de

570 representados no grande grafo. O *Geonames* fica atrás apenas do *DBPedia*, que recebe conexões de 207 outros *datasets*.

O *Geonames* dispõe de uma ontologia disponível em linguagem OWL, que permite que sejam utilizadas informações geoespaciais em toda *Web*. Cada um de seus registros geográficos possui uma *Uniform Resource Locator* (URL) única para representá-los, assim como um arquivo RDF passível de acesso.

Pela capacidade de apresentar e representar localizações geográficas com tão rica quantidade de informações e pela sua regular atualização de dados, o *Geonames* tem se tornado a principal referência para recursos sobre informações geográficas.

Os *datasets* criados ao redor do mundo e que precisam identificar os dados geográficos, mas que em grande parte das vezes não tem esse dado como ponto fundamental em sua estrutura, simplesmente apontam seus recursos para o *Geonames*. Desse modo, os *datasets* ao utilizar informações geoespaciais da *Web*, permitem que agentes computacionais possam realizar consultas que agreguem informações geográficas advindas de uma busca que inicia-se no *dataset* principal e segue posteriormente até o *dataset* do *Geonames* para ser enriquecida com informações geográficas, assim, caracteriza-se a ênfase de uso do *Geonames*.

Para atender as requisições baseadas em *Linked Data* e principalmente em oferecer os conceitos que a *Web* Semântica necessita, o *Geonames* está pronto para fazer a distinção entre conceito e documento. Essa técnica, indicada como redirecionamento 303 nos princípios do *Linked Data*, é de fundamental importância para atender os requisitos de agentes computacionais e principalmente os de *browsers* semânticos e prevê que um servidor ao receber uma requisição em sua URI principal possa redirecionar o pedido para um documento (do tipo RDF) para atender a solicitação de um agente computacional semântico.

A estratégia utilizada pelo *Geonames* prevê que cada recurso tenha duas URIs, como apresentada a seguir, onde utiliza-se novamente o exemplo da cidade de Salvador.

1. http://sws.geonames.org/3450554/
2. http://sws.geonames.org/3450554/about.rdf

A primeira URI refere-se a cidade de Salvador na Bahia, e deve ser utilizada quando há a necessidade de fazer uma referência a cidade. A segunda URI é um documento, em RDF, que contém informações sobre a cidade de Salvador no *Geonames*. Ressalta-se que o servidor

Geonames está preparado para redirecionar pedidos do URI 1 para a URI 2, informado aos agentes computacionais semânticos que há mais informações do que simplesmente a identificação do nome da cidade.

O RDF a ser consumido pelos agentes computacionais semânticos contém o conjunto de dados já descritos anteriormente. A figura 3 apresenta o conjunto de dados que o arquivo RDF contempla, entretanto, para evitar que a figura ficasse muito extensa foram retirados aproximadamente 50 identificações de nomes alternativos que o arquivo contempla para ser apresentado aqui.

Figura 4 - Nomes alternativos para Salvador no *Geonames*

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<rdf:RDF xmlns:cc="http://creativecommons.org/ns#" xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:gn="http://www.geonames.org/ontology#"
xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:wgs84_pos="http://www.w3.org/2003/01/geo/wgs84_pos#">
<gn:Feature rdf:about="http://sws.geonames.org/3450554/">
<rdfs:isDefinedBy rdf:resource="http://sws.geonames.org/3450554/about.rdf"/>
<gn:name>Salvador</gn:name>
<gn:alternateName xml:lang="ja">サルバドール</gn:alternateName>
<gn:alternateName>Bahia</gn:alternateName>
<gn:alternateName>Baía</gn:alternateName>
<gn:alternateName xml:lang="ca">Salvador</gn:alternateName>
<gn:officialName xml:lang="en">Salvador</gn:officialName>
<gn:alternateName xml:lang="es">Salvador de Bahía</gn:alternateName>
<gn:alternateName>Sao Salvador</gn:alternateName>
<gn:alternateName xml:lang="ar">ساڵڤادۆر، باهیا</gn:alternateName>
<gn:officialName xml:lang="ru">Сальвадор</gn:officialName>
<gn:featureClass rdf:resource="http://www.geonames.org/ontology#P"/>
<gn:featureCode rdf:resource="http://www.geonames.org/ontology#P.PPLA"/>
<gn:countryCode>BR</gn:countryCode>
<gn:population>2711840</gn:population>
<wgs84_pos:lat>-12.97111</wgs84_pos:lat>
<wgs84_pos:long>-38.51083</wgs84_pos:long>
<gn:parentFeature rdf:resource="http://sws.geonames.org/6321026/"/>
<gn:parentCountry rdf:resource="http://sws.geonames.org/3469034/"/>
<gn:parentADM1 rdf:resource="http://sws.geonames.org/3471168/"/>
<gn:parentADM2 rdf:resource="http://sws.geonames.org/6321026/"/>
<gn:nearbyFeatures rdf:resource="http://sws.geonames.org/3450554/nearby.rdf"/>
<gn:locationMap rdf:resource="http://www.geonames.org/3450554/salvador.html"/>
<gn:wikipediaArticle rdf:resource="https://pt.wikipedia.org/wiki/Salvador_%28Bahia%29"/>
</gn:Feature>
<foaf:Document rdf:about="http://sws.geonames.org/3450554/about.rdf">
<foaf:primaryTopic rdf:resource="http://sws.geonames.org/3450554/"/>
<cc:license rdf:resource="http://creativecommons.org/licenses/by/3.0/"/>
<cc:attributionURL rdf:resource="http://sws.geonames.org/3450554/"/>
<cc:attributionName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">GeoNames</cc:attributionName>
<dcterms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2006-01-15</dcterms:created>
<dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2014-05-11</dcterms:modified>
</foaf:Document>
</rdf:RDF>
```

Fonte: Elaborado pelos autores.

É importante ressaltar que o arquivo RDF correspondente a cada um dos recursos que o *Geonames* apresenta, pode variar dependendo do tipo de recurso a ser recuperado. Como Salvador é uma cidade, é disponibilizado o dado de sua população, entretanto alguns dos dados como longitude, latitude e em qual país o recurso está posicionado fisicamente estarão disponíveis em praticamente todos os recursos que o *Geonames* disponibiliza.

Pensando no modelo de ligações do *Linked Data*, os recursos do *Geonames* estão interligados entre eles mesmos, além das URIs já citadas anteriormente para o mesmo recurso, é possível ainda trabalhar com outros conjuntos de dados disponíveis para cada um dos recursos.

O *Geonames* frequentemente disponibiliza mais três URIs, com documentos RDF, para cada recurso disponibilizado. Esses documentos, devido as suas características, não atendem a todos os tipos de recursos que o *Geonames* oferece, mas estão disponíveis em grande parte das vezes. Essas URIs adicionais são constituídas a partir dos nomes dos documentos a URI principal do recurso, como segue (ainda usando o exemplo de Salvador):

- O documento *contains.rdf* pode ser adicionado a URI de qualquer recurso tendo como resposta um documento que contem toda a estrutura dos recursos que estão ligados ao recurso principal, ou seja, se o recurso principal é um país, apresenta-se a lista de estados. Se o recurso principal é uma cidade, o documento possui a suas regiões e assim sucessivamente. Para o recurso que apresenta a cidade de Salvador temos a seguinte URI: http://sws.geonames.org/3450554/contains.rdf;
- O documento *nearby.rdf* permite ter acesso a recursos que estão fisicamente próximos ao recurso principal. Esse documento constitui uma ligação semântica muito rica entre os recursos, baseado no posicionamento geográfico de cada um deles. Para a cidade de Salvador tem a seguinte URI: http://sws.geonames.org/3450554/nearby.rdf;
- O terceiro documento, contempla apenas os recursos que identificam países e trata-se do documento *neighbours.rdf*, utilizado para apresentar os países que são vizinhos (fazem fronteira) com o recurso principal. Como a cidade de Salvador não tem a disponibilidade desse recurso, então a URI utilizado como exemplo tem como recurso principal o Brasil. Na figura 4, é possível identificar parte da resposta que o servidor *Geonames* oferece aos agentes quando se utiliza a URI com o documento *neighbours.pdf* tendo o Brasil como recurso principal dessa forma:

  http://www.geonames.org/3469034/neighbours.rdf

Figura 5 - Países que fazem fronteira ao Brasil no *Geonames*

Fonte: Elaborado pelos autores.

Esses documentos tornam os dados do *Geonames* ainda mais aderentes a proposta do *Linked Data* pensada por Tim Berners-Lee e sua equipe, caracterizando o enriquecimento dos dados. Por meio da grande quantidade de dados que o *Geonames* disponibiliza para cada um dos recursos apresentados é possível imaginar a possibilidade da serendipidade tão desejada quando pensamos em ambientes semânticos.

4.2 RECUPERAÇÃO DA INFORMAÇÃO COM SPARQL

Segundo Tim Berners-Lee (2006) tentar utilizar o potencial da *Web* Semântica sem SPARQL é o mesmo que tentar utilizar um banco de dados relacional sem usar a linguagem SQL (SANTAREM SEGUNDO, 2014).

O SPARQL é um conjunto de especificações que fornecem linguagens e protocolos para consultar e manipular o conteúdo publicado em RDF na *Web*. O padrão compreende as seguintes especificações: uma linguagem de consulta para RDF; uma especificação que define uma extensão do SPARQL *Query Language* para executar consultas distribuídas em diferentes terminais SPARQL; uma especificação que define a semântica de consultas SPARQL sob regimes de vinculação, como RDF *Schema*, OWL, ou RIF; um protocolo que define os meios para a transmissão de consultas SPARQL arbitrárias e solicitações de atualização para um serviço

de SPARQL; uma especificação que define um método busca e descoberta e um vocabulário para descrever serviços SPARQL e um conjunto de testes, para avaliação da especificação SPARQL 1.1 (SPARQL, 2013).

O protocolo SPARQL 1.1 é uma versão com muitas alterações e substanciais evoluções em relação a primeira versão SPARQL 1.0, publicada pelo W3C como linguagem de consulta em janeiro de 2008. Atualmente chamado de protocolo, a versão 1.1 é uma recomendação W3C desde março de 2013.

De forma a mostrar o poder de exploração que os dados do *Geonames* disponíveis no *Linking Open Data* proporcionam, apresentam-se duas consultas realizadas utilizando-se recursos do *Dataset DBPedia* integrado ao *dataset Geonames*.

As consultas SPARQL que operam entre vários *datasets* são chamadas de consultas federadas, elas proporcionam que, por meio de apenas uma consulta, dados de diferentes *datasets* tenham como resposta apenas um resultado final.

A primeira consulta, apresentada por meio do código apresentado no quadro 1, é um código SPARQL, representado por meio de uma consulta federada que tem como objetivo recuperar os seguintes dados: nome, nativo (nome atribuído a quem mora no local), código do país, latitude, longitude e população.

Os dados relativos ao 'nome' e 'nativo' estão disponíveis no *dataset DBPedia*, os outros dados foram recuperados por meio do *dataset Geonames*. O valor utilizado como base para o início da consulta foi *dbpedia:Bahia*, mas poderia ter sido utilizado qualquer outro recurso de qualquer outro *dataset* que fizesse referência ao *dataset Geonames*.

Quadro 1 - Código de consulta SPARQL para *dbpedia:Bahia* do *dataset Geonames*

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
prefix wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix gn: <http://www.geonames.org/ontology#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbp: <http://dbpedia.org/property/>

select ?nome ?nativo ?pais ?latitude ?longitude ?populacao

from <http://sws.geonames.org/3471168/about.rdf>

where
```

```
{

  ?geonames     wgs84_pos:lat     ?latitude;
                        wgs84_pos:long  ?longitude;
                        gn:countryCode  ?pais;
                        gn:population    ?populacao;
                        rdfs:seeAlso     dbpedia:Bahia.

  SERVICE <http://DBpedia.org/sparql>
  { SELECT  distinct ?nome ?nativo
   WHERE
   {
   dbpedia:Bahia foaf:name ?nome;
                                dbp:populationDemonym ?nativo
   }limit 1
   }

}
```

Fonte: Elaborado pelos autores.

O resultado da consulta pode ser visto por meio da figura 5:

Figura 6 - Resultado da consulta Sparql apresentada no Quadro 1.

nome	nativo	pais	latitude	longitude	populacao
"Bahia"@en	"Baiano"@en	"BR"	"-12"	"-42"	"14175341"

Fonte: Elaborado pelos autores.

A segunda consulta apresentada por meio do código, no quadro 2, utiliza-se da URI baseada no documento *neighbours.rdf* que o *Geonames* apresenta para efetivar também uma busca federada que a partir de um recurso do *DBPedia*, no caso aqui o recurso *DBPedia:Brazil*, recorre ao *Geonames* para apresentar os países fronteiriços assim como suas respectivas populações.

Essa segunda consulta utiliza um recurso do protocolo SPARQL que é realizar uma busca recursiva no *Geonames*, pois a partir da busca dos vizinhos do Brasil é realizada uma nova consulta em cada um dos recursos originais de cada país no *Geonames* para que se possa obter suas respectivas populações

Quadro 2 - Consulta SPARQL baseada no recurso *DBPedia:Brazil* para dados de países vizinhos ao Brasil com suas respectivas populações

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix gn: <http://www.geonames.org/ontology#>
PREFIX dbpedia: <http://dbpedia.org/resource/>

select ?pais ?vizinho ?populacaoVizinho

from <http://sws.geonames.org/3469034/about.rdf>
from <http://www.geonames.org/3469034/neighbours.rdf>
from <http://sws.geonames.org/3437598/about.rdf>
from <http://sws.geonames.org/3439705/about.rdf>
from <http://sws.geonames.org/3686110/about.rdf>
from <http://sws.geonames.org/3923057/about.rdf>
from <http://sws.geonames.org/3932488/about.rdf>
from <http://sws.geonames.org/3865483/about.rdf>
from <http://sws.geonames.org/3378535/about.rdf>
from <http://sws.geonames.org/3382998/about.rdf>
from <http://sws.geonames.org/3381670/about.rdf>
from <http://sws.geonames.org/3625428/about.rdf>

where
{
  ?geonames     wgs84_pos:lat     ?latitude;
                        gn:population    ?populacao;
                        gn:name ?pais;
                        rdfs:seeAlso     dbpedia:Brazil.

   ?geonames2     gn:neighbour    ?geonames.
   ?geonames2     gn:name ?vizinho;
                                gn:population     ?populacaoVizinho;
 }
```

Fonte: Elaborado pelos autores.

O resultado da consulta pode ser visto por meio da figura 7.

Figura 7 - Resultado da consulta SPARQL apresentada no Quadro 2.

pais	vizinho	populacaoVizinho
"Brazil"	"Suriname"	"492829"
"Brazil"	"Guyana"	"748486"
"Brazil"	"Bolivia"	"9947418"
"Brazil"	"Paraguay"	"6375830"
"Brazil"	"Peru"	"29907003"
"Brazil"	"Argentina"	"41343201"
"Brazil"	"French Guiana"	"195506"
"Brazil"	"Colombia"	"47790000"
"Brazil"	"Venezuela"	"27223228"
"Brazil"	"Uruguay"	"3477000"

Fonte: Elaborado pelos autores.

Como pode ser observado, por meio do *DBPedia* que faz referência ao *Geonames*, as consultas entre *datasets* diferentes tornam-se possíveis gerando um grande banco de dados de informações abertas para serem consumidas.

Para que os agentes computacionais possam realizar as consultas de forma mais automática e inteligente é de fundamental que as ontologias estejam disponíveis para que seja possível identificar a estrutura funcional dos conjuntos de dados que os *datasets* podem oferecer.

O *Geonames* ainda não dispõe de um serviço de recuperação baseado em SPARQL *Endpoint*, dificultando o processo de uso do SPARQL. Entretanto, a maneira que oferece seus dados ainda torna possíveis as consultas, mas é de fundamental importância, para a ligação dos dados e melhoria nos processos de recuperação que esse importante *dataset* ofereça em breve esse tipo de ferramenta.

5 CONSIDERAÇÕES FINAIS

A interdisciplinaridade apresentada pelo diálogo entre o *Linking Open Data*, *Geonames* e Ciência da Informação estão presentes neste trabalho, ratificando sobre a estruturação de dados de geolocalização para reutilização de outros *datasets* a partir das ligações de dados na *Web* Semântica.

A pesquisa apresentou o *dataset Geonames* e sua grande representatividade no *Linking Open Data*, destacando-se não apenas pelo tamanho do círculo no diagrama apresentado pela equipe dos pesquisadores Cyganiak e Jentzsch, mas também pela capacidade de oferecer informações fidedignas e ricas em larga escala para serem utilizadas de forma aberta.

Durante a exploração do *dataset Geonames* foi possível identificar que há uma grande quantidade de informações disponíveis para serem utilizadas, isso implica que a construção de novos *datasets* e disponibilização de novos recursos para o *Linked Data* podem utilizar-se dos dados geográficos já disponíveis no *Geonames*, portanto na construção de um novo *dataset* que tenha algum recurso relativo a áreas geográficas, a ligação deste recurso no novo *dataset* com o recurso disponível no *Geonames* pode agregar rapidamente um enriquecimento semântico as novas informações disponibilizadas.

Além disso, notou-se que as consultas federadas utilizando o protocolo SPARQL, que percorrem os *datasets*, são recursos de extrema relevância no processo de recuperação de dados

no *Linking Open Data*, pois permitem que a integração de dados proposta pela estrutura organizacional construída pelas ontologias possam ser exploradas no processo de recuperação da informação.

Destaca-se que o *Geonames* não dispõe do recurso computacional de SPARQL *Endpoint*, e que essa tecnologia é importante para que os agentes computacionais possam trabalhar de forma mais eficiente, mostrando que mesmo os *datasets* mais conhecidos apresentados no *Linking Open Data* ainda carecem de evolução.

Portanto é possível considerar que as melhores práticas dos princípios *Linked Data* idealizadas por Tim Berners-Lee em 2006 estão em franca expansão e já passíveis de utilização. Entretanto, ainda carecem de pesquisas e recursos tanto estruturais como tecnológicos, mas é possível perceber que parte da *Web* caminha para ser um grande banco de dados aberto, com expressividade semântica possível.

An Approach to the Structure of Geonames and their Contributions to the Linking Open Data

ABSTRACT

ENANCIB 2016 theme does reflect the composition of the informational structure of the data produced by numerous ways, whether by humans or by machines. However, the data provide a semantic level than more consolidated structures do not. The intrinsic relationship between the Information Science and Computer Science is ever closer, mainly related to methodological processes to add semantics on the links between the data. In addition, the breadth of that relationship beyond the different areas of knowledge, on various concerns and improvements in the specifics of each field, such as attempts to link data geo-sensing the Geonames project. The problem is characterized in how the Geonames project has helped in structuring geolocation data for reuse in other datasets from the data connections in the Semantic Web? The objective of this work is to present the contributions from the Geonames project for the development of the Semantic Web and the Linking Open Data. Regarding the methodological procedures was research theoretical, applied and qualitative nature, aiming to analyze the contributions of Geonames and the Linking Open Data. Regarding the work goals this research is classified as exploratory. The results, presented the structure of the Geonames project, its potential contribution to the Linking Open Data, and some consultations with SPARQL protocol to verify its functionality. Therefore it is possible to consider that the best practices of Linked Data are expanding and can already be used, but still lack of research and both structural features such as technology.

Keywords: Geonames. Linked data. Semantic Web. Geolocation.

REFERÊNCIAS

BERNERS-LEE, T. **Linked data**: design issues. [S.l.]: W3C, 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 25 jul. 2016.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, 2001, p. 29-37.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: the story so far. I**nternational Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1-22, 2009. Disponível em: <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linkeddata.pdf>. Acesso em: 25 jul. 2016.

CATARINO, M. E.; SOUZA, T. B. de. A representação descritiva no contexto da web semântica. **Transinformação**, Campinas, v. 24, n. 2, maio/ago. 2012. p. 77-90. Disponível em: <http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/766/746>. Acesso em: Acesso em: 25 jul. 2016.

GEONAMES. Disponível em <http://www.geonames.org/>. Acesso em: 25 jul. 2016.

HEATH, T.;BIZER, C. Linked data: evolving the Web into a global data space. **Synthesis Lectures on the Semantic Web**: theory and technology. [S.l]: Morgan & Claypool, 2011. Disponível em: <http://linkeddatabook.com/editions/1.0/#htoc8>. Acesso em: 20 jul. 2016.

LINKING OPEN DATA. Disponível em: <http://lod-cloud.net/>. Acesso em: 20 jul. 2016.

OPEN KNOWLEDGE. Disponível em: < http://opendefinition.org/>. Acesso em: 20 jul. 2016.

MARCONDES, C. H. “Linked data” dados interligados - e interoperabilidade entre arquivos, bibliotecas e museus na web. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, v. 17, n. 34, p. 171-192. 23 jun. 2012. ISSN 1518-2924. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2012v17n34p171>. Acesso em: 20 jul. 2016.

PIZZOL, L. D.; TODESO, J. L.; TODESCO, B. P. R. Como a Web de dados pode apoiar o processo de inteligência competitiva. **Perspectivas em Gestão & Conhecimento**, João Pessoa, v. 5, Número Especial, p. 87-102, jan. 2016. Disponível em: <http://periodicos.ufpb.br/ojs/index.php/pgc/article/view/27384>. Acesso em: 26 jul. 2016.

RAMALHO, R. A. S. **Web Semântica**: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação. 2006. 120 f. Dissertação (Mestrado em Ciência da Informação)- Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2006.

SANTAREM SEGUNDO, J. E. Web Semântica: introdução a recuperação de dados usando SPARQL. In: **Encontro Nacional de Pesquisa em Ciência da Informação**: além das nuvens, expandindo as fronteiras da Ciência da Informação, 15., 2014. Belo Horizonte, MG. *Anais*... Belo Horizonte, MG: ECI, UFMG, 2014. p. 3863-3882. Disponível em: <http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt8>. Acesso em: 20 jul. 2016.

SANTARÉM SEGUNDO, J. E. Web semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 8, n. 2, p. GT8-2, 2015. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/v/19443>. Acesso em: 20 jul. 2016.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a Ciência da Informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004. Disponível em: <http://dx.doi.org/10.1590/S0100-19652004000100016>. Acesso em: 20 jul. 2016.

SPARQL. Disponível em: <https://www.w3.org/TR/sparql11-query/> .Acesso em: 20 jul. 2016.

WORLD WIDE WEB CONSORTIUM. Semantic Web. [S.l.], 2013. Disponível em: <http://www.w3.org/standards/semanticweb/>. Acesso em: 20 jul. 2016.