Triplificação de dados de notícias sobre a Zika
Alternative Title: Data triplification of Zika news

Luís Fernando Monsores Passos Maia
Graduate Program on Informatics
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
luisfmpm@ufrj.br

Marcela Mayumi Mauricio Yagui
Graduate Program on Informatics
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
marcelayagui@ufrj.br

RESUMO
O objetivo deste trabalho foi construir uma base de dados temática de notícias sobre a Zika, reunidas de fontes on-line como jornais oficiais, blogs e fóruns. Após explicar o processo de coleta e processamento dos dados, é pretendido mostrar como aplicar conceitos da web semântica de modo a triplificar, conectar e realizar consultas nessa base. A motivação deste trabalho é relacionar diferentes tipos de dados oriundos de fontes heterogêneas, como autores, localidades e notícias para compreender os impactos e a repercussão da doença em mídias sociais.

Palavras-Chave
RDF; Dados abertos ligados; mineração de dados; Zika; mídias sociais.

ABSTRACT
The objective of this work was to build a thematic database about Zika news, gathered from on-line sources such as newspapers, blogs and forums. After explaining the collecting and gathering process it is intended to show how to apply semantic web concepts in order to triplify, connect and perform queries on that database. The motivation of this work is to relate different data, such as authors, locations and news to understand the impacts of this disease in social media.

CCS Concepts
• Information systems→Graph-based database models • Information systems→Resource Description Framework (RDF) • Human-centered computing→Social media.

Keywords
RDF; Linked data; data mining; Zika; social media.

1. INTRODUÇÃO
O vírus Zika é um arbovírus da família *Flaviviridae*. Sua identificação ocorreu em 1947 em Uganda. Contudo, ocorrências de infecção humana foram relatadas de forma isolada e esporádica em um grande número de indivíduos [8].

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

Seu potencial epidêmico tornou-se evidente em 2007, quando houve um grande surto na Micronésia. Em 2013, no entanto, uma mudança de padrão pôde ser observada em novos surtos que ocorreram principalmente na Polinésia Francesa, com elevadas taxas de ocorrência e onde estima-se que houve uma incidência de 11% de contaminação pelo vírus na população local [12].

No Brasil, desde novembro de 2014 e principalmente no início de 2015, os estados da região Nordeste relataram um grande surto de uma nova doença exantemática. A condição clínica, caracterizada por pouca ou nenhuma febre, acompanhada por artralgia, edema articular e conjuntivite, levou o infectologista Kleber Luz a considerar a hipótese de que o surto fosse causado pelo Zika vírus. Isto foi confirmado em Abril 2015 através de *Polymerase Chain Reaction* (PCR) realizada em oito das 25 amostras de sangue de casos suspeitos na Bahia e, posteriormente, em oito dos 21 casos no Rio Grande do Norte [11]. O surto apresentou uma alta taxa de ocorrências, afetando milhares de pessoas e causando superlotação dos serviços de emergência públicos e privados, embora não tenha sido medido por um sistema de notificação oficial. Uma vez que esta não era uma doença de notificação obrigatória, mesmo em casos suspeitos de Zika, os médicos foram orientados pela vigilância da saúde a notificar esses casos como Dengue. Após a eclosão da doença em sua forma clássica, casos neurológicos e de microcefalia começaram a surgir [3, 16].

A ocorrência destes eventos gerou grande comoção nos noticiários e, desde então, tornou-se crescente o interesse dos pesquisadores no sentido de coletar mais dados sobre a doença para responder perguntas que são do interesse da população, tais como, "Em que lugares a doença tem causado mais impactos?"; "Fatores geossociais afetam os padrões de disseminação do vírus?"; "Como foi o histórico de evolução da doença ao longo dos anos?"; "Quais foram os tópicos/assuntos sobre a doença com maior repercussão?". Uma forma de responder a essas e a outras perguntas é através da coleta de notícias relacionadas à doença em veículos oficiais (G1, blogs, sites de universidades/instituições de pesquisa) e de dados compartilhados nas mídias sociais (Twitter, Facebook, Youtube, Google+, etc) para construção de bancos de dados que permitam a realização de consultas direcionadas, tornando possível a extração de informações importantes que permitam aos pesquisadores compreender melhor a doença e seus desdobramentos.

Um problema nesta abordagem, no entanto, reside no fato de que a maioria dos bancos de dados disponibilizados atualmente são relacionais, com colunas de informação relacionadas umas às outras. A semântica ou significado dos dados está presente nos bancos de dados relacionais através dos relacionamentos existentes entre essas colunas. Todavia, os dados presentes na *web* e, consequentemente compartilhados nas mídias sociais não estão organizados de forma estruturada, de modo que, extrair esses

dados para atribuir-lhes novos significados torna-se uma tarefa praticamente impossível utilizando os bancos de dados relacionais convencionais. Uma solução para isso é a utilização do *framework* para descrição de recursos (RDF) para triplificar[1] essas bases de dados não relacionais permitindo relacioná-las com outras bases de dados já existentes, como a DBpedia[2] [14]. Deste modo, a partir de dados dispersos na internet e com o auxílio de linguagens de consulta específicas como o SPARQL[3], torna-se possível extrair valiosas informações das bases de dados criadas para que os cientistas de dados possam realizar suas análises e tirar conclusões acerca da doença, que é o cenário da aplicação, agregando valor semântico a dados semiestruturados, além de disponibilizá-los na internet para que outros pesquisadores possam utilizá-los e realizar suas próprias análises.

Este trabalho mostra a construção de uma base de dados sobre notícias e publicações relacionadas à doença Zika mineradas a partir de mídias sociais e veículos oficiais, sua triplificação, construção de vocabulários e atribuição de significado semântico aos dados semiestruturados, além de sua integração com a base de dados DBpedia. A partir disso, será mostrado um estudo de caso com a finalidade de mostrar como as práticas da *web* semântica viabilizam, entre outras possibilidades, consultas integradas em bases de dados construídas a partir de fontes heterogêneas para identificar padrões em publicações de mídias sociais e redes sociais on-line (RSO).

2. TRABALHOS RELACIONADOS

No contexto do jornalismo, alguns estudos vêm sendo realizados com o intuito de aplicar a *web* semântica e conceitos correlatos em seus produtos, a fim de permitir a interoperabilidade entre sistemas e ampliar sua base de conhecimento.

A BBC utiliza a *web* semântica e os recursos de *linked data* em seus sistemas, como no *BBC Music* e *BBC Programmes* e a publicação de dados nesses dois domínios. No trabalho de Kobilarov et al [11] os autores também demostraram a utilização de *web* semântica em um sistema de categorização chamado CIS que permite a integração com diferentes bases de dados, inclusive com a DBpedia. Foi demonstrado no artigo como a utilização de *linked data* pode enriquecer o conteúdo do site.

Outro estudo relacionado diz respeito ao desenvolvimento do NEWS (*News Engine Web Services*), desenvolvido pelo EU IST. Este projeto tem como objetivo o desenvolvimento de uma ontologia no domínio do jornalismo, mais especificamente para notícias. Para isso, os pesquisadores alinham os atuais padrões jornalísticos com outras ontologias existentes e com outros padrões de metadados. A ontologia conta com os módulos estruturais (5 classes, 8 propriedades e 6 regras), categorização (1300 classes e 3 propriedades), empacotamento (52 classes, 96 propriedades, 21 regras e mais de 500 instâncias) e conteúdo (248 classes, 116 propriedades, 30 regras e milhares de instâncias). Além disso, os autores planejam ampliar a ontologia de modo a integrá-la ao *Subject Reference System* do *International Press Telecommunications Council*[4], a um módulo de perfis e um módulo de tempo com suporte a eventos e notícias em tempo real [6, 7].

García, Perdrix e Gil [10] criaram uma ontologia e uma arquitetura para dar suporte ao gerenciamento de jornais on-line e seus recursos de multimídia, ao permitir que dados jornalísticos sejam facilmente recuperados e integrados. Para isso, os pesquisadores criaram mapeamentos XML2RDF automáticos de padrões codificados em XML, como o MPEG-7 e os principais *schemas* do ITPC (NITF, NewsML e NewsCodes) para ontologias OWL. Este projeto foi aplicado na *Segre Media Group*, uma empresa espanhola que elabora conteúdo de jornais, rádio e televisão.

Com base nos estudos apontados, ontologias no domínio do jornalismo são implementadas para suprir necessidades de jornalistas e de grupos de pesquisa da área. Contudo, eles não permitem a integração de forma personalizada com dados de notícias de um determinado assunto ou área de interesse e de dados de mídias sociais. Este trabalho tem como finalidade permitir a integração de notícias de um determinado domínio (Zika), proveniente de fontes heterogêneas, como jornais, blogs e fóruns com dados de RSO e de outras fontes (DBpedia) para enriquecimento dos dados do domínio escolhido.

3. VOCABULÁRIOS UTILIZADOS

DBPEDIA - A DBpedia disponibiliza informações estruturadas da Wikipedia, permitindo que consultas avançadas sejam realizadas nesses dados e também que sejam integrados com quaisquer outras bases de dados triplificadas. A versão disponibilizada no idioma Inglês da base de conhecimentos da DBpedia atualmente descreve 4,58 milhões de coisas e, adicionalmente, 23,8 milhões de termos são descritos em mais 125 diferentes idiomas mas que também existem no Inglês, totalizando de 38,3 milhões de coisas descritas somando todos os idiomas [1].

Neste trabalho, a DBpedia é utilizada para integração de dados, como também para descrever o domínio do trabalho e o país de publicação da notícia.

DUBLIN CORE - *Dublin Core* é um vocabulário para descrever recursos na *Web*, como páginas e a sua relação de autoria, a descrição de seu conteúdo (texto, imagens, vídeos, entre outros) e demais itens relevantes para permitir que essas páginas sejam mais visíveis por motores de busca e mecanismos de recuperação da informação [5]. O Dublin Core é o vocabulário que descreve mais elementos neste trabalho, expressando relações, como autor, título da notícia, linguagem e data de publicação.

FOAF - *Friend of a Friend* é um vocabulário que descreve termos para expressar pessoas e suas relações sociais de forma semântica. É composto por diferentes tipos de redes: redes sociais humanas de colaboração, amizade e associação; redes que expressam relações factuais; além de redes de informação que usam *Web-based linking* para compartilhar vocabulários com descrições semânticas publicados de modo independente [2, 9]. O FOAF está sendo utilizado neste trabalho apenas para expressar a relação de uma notícia com a sua respectiva página principal de publicação.

[1] Descrição dos dados que é realizada através de triplas, seguindo a sintaxe sujeito, predicado e objeto, onde o sujeito é uma URI, o objeto pode ser uma URI ou um literal e o predicado é uma URI que define como sujeito e predicado estão relacionados. Essa relação que o predicado estabelece entre o sujeito e o objeto dá significado aos dados.

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

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

[4] Metadados sobre fotos.

4. METODOLOGIA

Para que seja efetiva, a triplificação de dados deve ser acompanhada de metodologias que auxiliem no processo de extração, transformação e carga dos dados. O processo de *Extraction, Transformation and Loading* (ETL) consiste no mecanismo de (i) extração de dados de uma ou múltiplas bases externas; (ii) transformação dos dados para que atendam à demandas específicas de uma área de negócios e (iii) carga desses dados em ambientes de armazenamento, como *Data Warehouses* ou *Data Marts*, de maneira organizada e estruturada [4]. O processo de triplificação e construção da base de dados de notícias sobre a Zika foi realizado a partir da metodologia ETL, que se dividiu em três etapas: Extração, Transformação e Carga dos dados.

EXTRAÇÃO – Para a coleta de notícias sobre a Zika, foram testadas 8 ferramentas que coletam dados de mídias sociais, como jornais, revistas, blogs e fóruns. Dentre as ferramentas analisadas, a que mais se adequou foi a *Webhose*[5], por permitir que pudessem ser coletados gratuitamente dados de notícias publicadas de até um mês passado, permitir definir parâmetros avançados na *string* de busca, permitir definir critérios como idioma, país, o tipo de mídia (fóruns, notícias e blogs) e até possibilitar a coleta de notícia com um sentimento associado (positivo, negativo ou neutro) [15].

Foram retornadas todas as notícias de fóruns, notícias e blogs correspondentes ao período de 30 dias (29/07 a 29/08), utilizando-se a palavra-chave "Zika". As notícias foram filtradas de modo a permitir apenas dados na linguagem Português, porém sem restrição de país de publicação. Também não houve restrição quanto à categoria ou site de publicação da notícia.

Ao final deste processo foram coletados os dados de 5633 registros de notícias no formato JSON. O armazenamento inicial dos dados foi realizado no banco de dados NoSQL MongoDB[6].

TRANSFORMAÇÃO - Nesta etapa, definiu-se o modelo do grafo RDF, os vocabulários, a criação de novos termos que não foram encontrados nos vocabulários existentes e, finalmente, as notícias foram limpas e triplificadas.

Inicialmente, os registros foram limpos para que restassem apenas dados que corroborassem com a pesquisa. Foi utilizada a ferramenta *Talend Open Studio for Big Data*[7] para gerar um novo JSON contendo os seguintes termos: `URI`, `Author`, `Title`, `Country`, `Homepage`, `Language`, `Published`, Contagem de compartilhamentos na RSO Facebook e Contagem de compartilhamentos na RSO Google+.

A partir disso, foram mapeados quais termos já existem em vocabulários controlados disponibilizados na *Web*. A Tabela 1 mostra a correspondência feita entre os termos das notícias coletadas e termos da DBpedia, Dublin Core e FOAF:

Tabela 1. Correspondência entre termos mapeados com vocabulários existentes

Dados das notícias	Termo correspondente no vocabulário
DBpedia	
[Tema das notícias]	Zika_Virus
Country	Country
Dublin Core	
Published	Date
Language	Language
Title	Title
Author	Creator
FOAF	
Homepage	Homepage

Contudo, apesar do extenso número de termos já disponibilizados na *Web*, foi necessário criar novos termos em função de nenhum ter sido encontrado para expressar semanticamente o domínio de notícias sobre a Zika. Deste modo, foram criados quatro novos termos na ferramenta Protégé[8], versão 5.0.0, conforme é mostrado na Tabela 2.

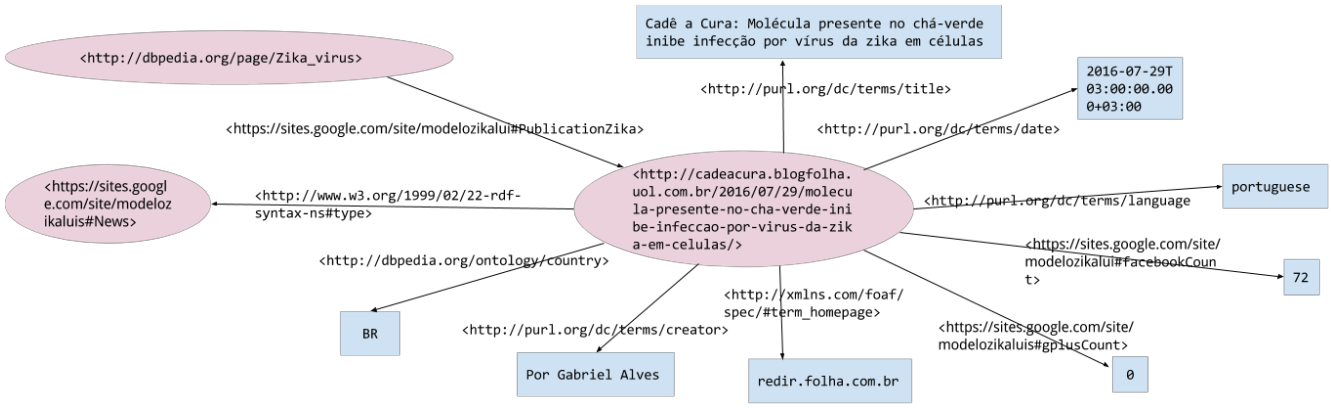


Figura 1. Modelo do grafo RDF

[5] https://webhose.io/

[6] https://www.mongodb.com/

[7] https://www.talend.com/

[8] http://protege.stanford.edu/

Tabela 2. Termos criados para o domínio de notícias sobre a Zika

Termo criado	Descrição
Classe	
News	Um registro de notícia.
Object Properties	
PublicationZika	Relacionamento entre a doença Zika e uma notícia.
Datatype Properties	
facebookCount	Compartilhamentos na RSO Facebook.
gplusCount	Compartilhamentos na RSO Google+.

Os novos termos foram conectados a outros termos e vocabulários com a utilização de propriedades do RDF *Schema* como rdfs:range e rdfs:domain. E a fim de criar URIs para cada novo termo, criou-se um site para registro temporário, disponível em: <https://sites.google.com/site/modelozikaluis>.

A Figura 1 mostra como foi modelado o grafo RDF para uma instância da base de dados.

Após a definição dos termos, os dados foram triplificados com o complemento Talend4sw[9] da ferramenta *Talend Open Studio for Big Data*. A Figura 2 demonstra a primeira etapa do processo de triplificação: a entrada dos dados limpos no formato JSON inicia o fluxo e eles são transportados para um componente que transforma os dados para o formato N-Triple, sendo necessário configurar o componente para que ocorra a formação da tripla com os campos corretos, como também para informar prefixos e outros detalhes adicionais. A saída dos dados ocorre no último componente do processo.

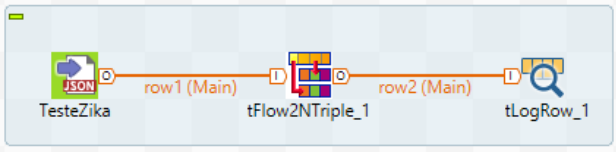


Figura 2. Primeira etapa da triplificação

Para a integração de todas as triplas geradas e a exportação de todas as triplas em outros formatos disponibilizados pela ferramenta (RDF/XML, JSON-LD, N-TRIPLES, *TURTLE*, N3, TRIG e RDF/JSON), foi necessário formular um segundo processo, como pode ser observado na Figura 3. O processo inicia com o carregamento dos dados gerados na primeira etapa seguida de sua associação com o formato que se pretende transformar. A partir disso, o processo faz iterações para que os dados sejam de fato convertidos no componente 'tRDF2RDF'. Por fim, ocorre a saída de um arquivo no último componente do processo, onde acontece a exibição dos dados exportados para os formatos *Turtle* e XML.

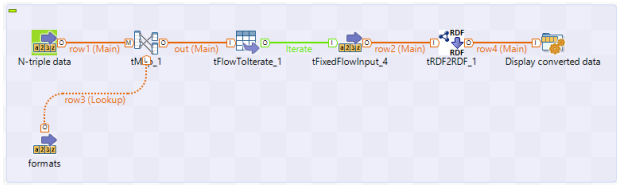


Figura 3. Segunda etapa da triplificação

CARGA – Nesta etapa as triplas foram armazenadas no serviço de banco de dados GraphDB[10]. O GraphDB foi escolhido como banco de triplas por permitir que dados sejam facilmente integrados, com a identificação por URIs; por ser compatível com os padrões da W3C; por suportar formas de classificação de metadados, ontologias, tesauros, taxonomias e classes; por suportar a proveniência de dados; e por ser compatível com *linked data* [13].

Uma base de dados chamada ZikaNews foi criada e os dados foram importados no formato *Turtle*. Após o armazenamento, as triplas ficaram disponíveis para a realização de consultas na linguagem SPARQL. Na Figura 4, pode-se visualizar os dados triplificados juntamente com dados extraídos da DBpedia.

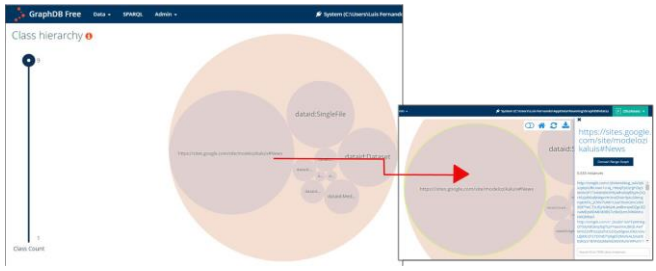


Figura 4. Dados armazenados no GraphDB

5. RESULTADOS E DISCUSSÃO

Com o intuito de extrair informações relacionadas ao noticiário sobre Zika ao longo do mês de Agosto, quatro consultas foram formuladas a fim de responder algumas das perguntas propostas, entre elas:

"Quais foram os tópicos/assuntos sobre a doença com maior repercussão em redes sociais?"

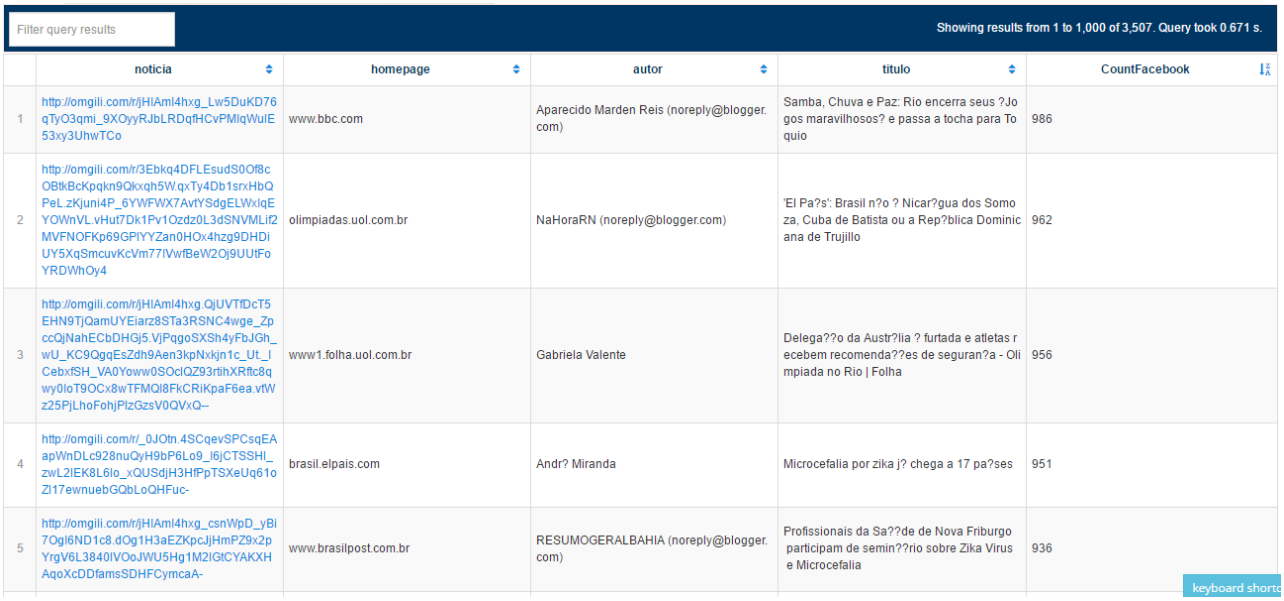
"Quais foram os principais autores durante o período analisado?"

"Quais países mais publicaram na língua portuguesa?"

O código SPARQL a seguir demonstra a consulta elaborada a fim de responder a primeira questão, enfatizando a quantidade de compartilhamentos na RSO Facebook:

[9] https://github.com/fbelleau/talend4sw

[10] http://graphdb.ontotext.com/

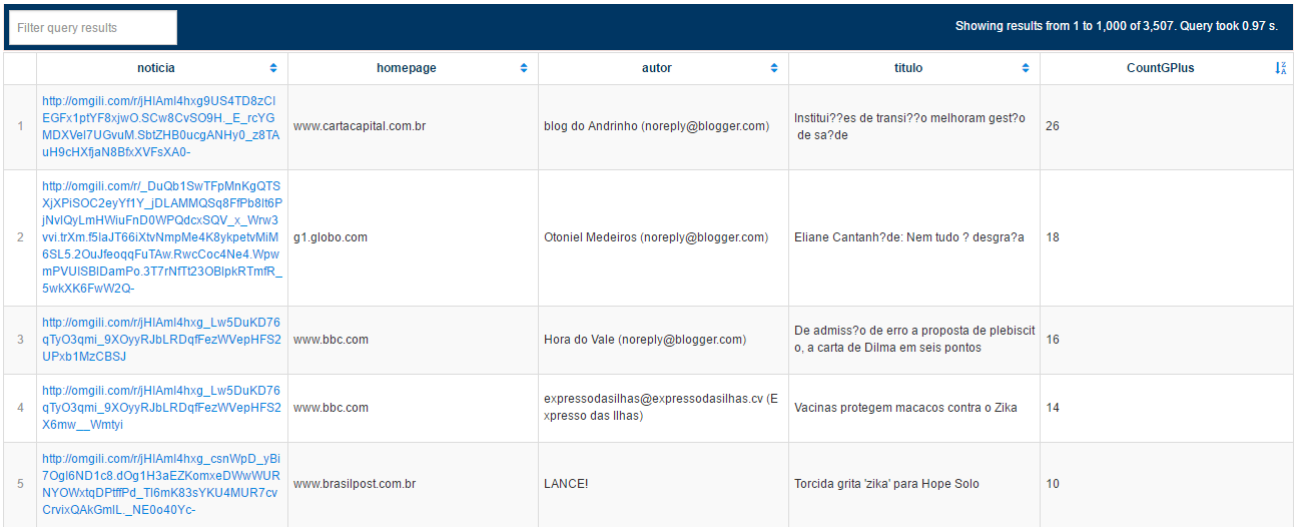


Filter query results — Showing results from 1 to 1,000 of 3,507. Query took 0.671 s.

	noticia	homepage	autor	titulo	CountFacebook
1	http://omgili.com/r/jHIAml4hxg_Lw5DuKD76qTyO3qmi_9XOyyRJbLRDqfHCvPMlqWuIE53xy3UhwTCo	www.bbc.com	Aparecido Marden Reis (noreply@blogger.com)	Samba, Chuva e Paz: Rio encerra seus ?Jogos maravilhosos? e passa a tocha para Toquio	986
2	http://omgili.com/r/3Ebkq4DFLEsudS0Of8cOBtkBcKpqkn9Qkxqh5W.qxTy4Db1srxHbQPeL.zKjuni4P_6YWFWX7AvtYSdgELWxlqEYOWnVL.vHut7Dk1Pv1Ozdz0L3dSNVMLif2MVFNOFKp69GPlYYZan0HOx4hzg9DHDiUY5XqSmcuvKcVm77IVwfBeW2Oj9UUtFoYRDWhOy4	olimpiadas.uol.com.br	NaHoraRN (noreply@blogger.com)	'El Pa?s': Brasil n?o ? Nicar?gua dos Somoza, Cuba de Batista ou a Rep?blica Dominicana de Trujillo	962
3	http://omgili.com/r/jHIAml4hxg.QjUVTfDcT5EHN9TjQamUYEiarz8STa3RSNC4wge_ZpccQjNahECbDHGj5.VjPqgoSXSh4yFbJGh_wU_KC9QgqEsZdh9Aen3kpNxkjn1c_Ut_.ICebxfSH_VA0Yoww0SOclQZ93rtihXRftc8qwy0loT9OCx8wTFMQl8FkCRiKpaF6ea.vtWz25PjLhoFohjPlzGzsV0QVxQ--	www1.folha.uol.com.br	Gabriela Valente	Delega??o da Austr?lia ? furtada e atletas recebem recomenda??es de seguran?a - Olimpiada no Rio | Folha	956
4	http://omgili.com/r/_0JOtn.4SCqevSPCsqEAapWnDLc928nuQyH9bP6Lo9_I6jCTSSHl_zwL2IEK8L6lo_xQUSdjH3HfPpTSXeUq61oZI17ewnuebGQbLoQHFuc-	brasil.elpais.com	Andr? Miranda	Microcefalia por zika j? chega a 17 pa?ses	951
5	http://omgili.com/r/jHIAml4hxg_csnWpD_yBi7Ogl6ND1c8.dOg1H3aEZKpcJjHmPZ9x2pYrgV6L3840lVOoJWU5Hg1M2IGtCYAKXHAqoXcDDfamsSDHFCymcaA-	www.brasilpost.com.br	RESUMOGERALBAHIA (noreply@blogger.com)	Profissionais da Sa??de de Nova Friburgo participam de semin??rio sobre Zika Virus e Microcefalia	936

keyboard shortc

Figura 5. Repercussão das notícias no Facebook



Filter query results — Showing results from 1 to 1,000 of 3,507. Query took 0.97 s.

	noticia	homepage	autor	titulo	CountGPlus
1	http://omgili.com/r/jHIAml4hxg9US4TD8zCIEGFx1ptYF8xjwO.SCw8CvSO9H._E_rcYGMDXVel7UGvuM.SbtZHB0ucgANHy0_z8TAuH9cHXfjaN8BfxXVFsXA0-	www.cartacapital.com.br	blog do Andrinho (noreply@blogger.com)	Institui??es de transi??o melhoram gest?o de sa?de	26
2	http://omgili.com/r/_DuQb1SwTFpMnKgQTSXjXPiSOC2eyYf1Y_jDLAMMQSq8FfPb8lt6PjNvIQyLmHWiuFnD0WPQdcxSQV_x_Wrw3vvi.trXm.f5IaJT66iXtvNmpMe4K8ykpetvMiM6SL5.2OuJfeoqqFuTAw.RwcCoc4Ne4.WpwmPVUISBIDamPo.3T7rNfTt23OBlpkRTmfR_5wkXK6FwW2Q-	g1.globo.com	Otoniel Medeiros (noreply@blogger.com)	Eliane Cantanh?de: Nem tudo ? desgra?a	18
3	http://omgili.com/r/jHIAml4hxg_Lw5DuKD76qTyO3qmi_9XOyyRJbLRDqfFezWVepHFS2UPxb1MzCBSJ	www.bbc.com	Hora do Vale (noreply@blogger.com)	De admiss?o de erro a proposta de plebiscito, a carta de Dilma em seis pontos	16
4	http://omgili.com/r/jHIAml4hxg_Lw5DuKD76qTyO3qmi_9XOyyRJbLRDqfFezWVepHFS2X6mw__Wmtyi	www.bbc.com	expressodasilhas@expressodasilhas.cv (Expresso das Ilhas)	Vacinas protegem macacos contra o Zika	14
5	http://omgili.com/r/jHIAml4hxg_csnWpD_yBi7Ogl6ND1c8.dOg1H3aEZKomxeDWwWURNYOWxtqDPtffPd_.Tl6mK83sYKU4MUR7cvCrvixQAkGmlL._NE0o40Yc-	www.brasilpost.com.br	LANCE!	Torcida grita 'zika' para Hope Solo	10

Figura 6. Repercussão das notícias no Google+

```
  SELECT DISTINCT ?noticias ?homepage ?autor ?titulo
?CountFacebook
WHERE {
?noticias a
<https://sites.google.com/site/modelozikaluis#News> .
?noticias  <http://xmlns.com/foaf/spec/#term_homepage>
?homepage .
  ?noticias <http://purl.org/dc/terms/creator> ?autor
.
    ?noticias         <http://purl.org/dc/terms/title>
?titulo .
    ?noticias                                        <
https://sites.google.com/site/modelozikaluis#facebook
Count> ?CountFacebook .
}
ORDER BY desc (?CountFacebook)
```

Executado a consulta anterior no *Endpoint* do GraphBD foram obtidos os resultados mostrados pela Figura 5.

Observando a Figura 5 (que está limitada aos 5 primeiros resultados), pode-se verificar que os assuntos que mais repercutiram no Facebook estavam relacionados com o tema Olimpíadas e as principais consequências que um surto do Zika vírus poderia trazer para atletas e membros do comitê olímpico, como também quais seriam os principais cuidados para evitar a doença. As outras principais notícias estavam relacionadas com dados estatísticos do contágio da doença.

Uma consulta similar foi realizada para observar os dados envolvendo a RSO Google+:

```
SELECT DISTINCT ?noticias ?homepage ?autor ?titulo
?CountGPlus
WHERE {
        ?noticias a
<https://sites.google.com/site/modelozikaluis#News> .
?noticias  <http://xmlns.com/foaf/spec/#term_homepage>
?homepage .
    ?noticias       <http://purl.org/dc/terms/creator>
?autor .
    ?noticias         <http://purl.org/dc/terms/title>
?titulo .
    ?noticias                                          <
https://sites.google.com/site/modelozikaluis#gplusCou
nt> ?CountGPlus .
}
ORDER BY desc (?CountGPlus)
```

Executado a consulta anterior no *Endpoint* do GraphBD foram obtidos os resultados mostrados pela Figura 6.

Na Figura 6 (que está limitada aos 5 primeiros resultados), observa-se que os assuntos que mais repercutiram no Google+ estavam relacionados a questões políticas, aos impactos e medidas de precaução, previsões quanto a disseminação da doença nos Jogos Olímpicos e a divulgação de avanços de pesquisas científicas.

Pode-se observar também que alguns dos principais sites de notícias, como Carta Capital, G1, BBC e Brasil Post, tiveram um baixo número de compartilhamentos, quando comparados com o Facebook. Dentre essas notícias, a publicação de "blog do Andrinho", do Carta Capital, foi a que teve maior destaque no período analisado.

Para a segunda questão levantada, foi criada a próxima consulta SPARQL, a fim de verificar quais foram os principais autores, e consequentemente, quais tiveram mais publicações e destaque no meio jornalístico.

```
SELECT (COUNT(?noticia) as ?n) ?autor
WHERE {
        ?noticia a
<https://sites.google.com/site/modelozikaluis #News>
.
    ?noticia        <http://purl.org/dc/terms/creator>
?autor .
}
GROUP BY (?autor)
ORDER BY DESC(?n)
LIMIT 10
```

A Tabela 3 expressa os resultados retornado para a consulta anterior (limitada aos 10 primeiros resultados, restringidos pela cláusula *limit* da consulta):

Tabela 3. Principais autores

Autor	Resultados
folha.com.br	71
Portal Saúde Business	61
Redação Folha Vitória	57
Jose Siqueira	51
Fábio Fraga	31
EFE	30
Brasil Soberano e Livre	26
Antônio Carlos Ribeiro	25
Agência Estado	24
Redação	23

De acordo com os resultados retornados, podemos observar que o "autor" com mais destaque ou maior número de publicações foi o site "folha.com.br". Isso se justifica pelo fato de o site não informar explicitamente o nome do autor ou mesmo por não dispor de recursos para descrever semanticamente o nome do criador da publicação, o que impossibilita a identificação deste campo pela ferramenta utilizada durante a fase de coleta dos dados. Pode-se observar também que, dentre os principais autores, suas publicações tiveram pouca repercussão em redes sociais.

A Figura 7 expressa os resultados obtidos por meio de um gráfico de barras.

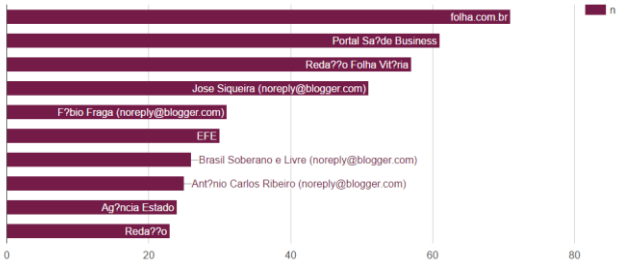


Figura 7. Gráfico contendo os principais autores

Por fim, a quarta consulta tem como finalidade responder a terceira questão, relacionando a quantidade de publicações com os seus respectivos países, lembrando que, foi utilizado um filtro na ferramenta de coleta para retornar apenas resultados na língua portuguesa.

```
SELECT (COUNT(?noticia) as ?n) ?pais
WHERE {
        ?noticia a <
https://sites.google.com/site/modelozikaluis#News> .
    ?noticia     <http://dbpedia.org/ontology/country>
?pais .
}
GROUP BY (?pais)
ORDER BY DESC(?n)
```

A Tabela 4 mostra os 10 primeiros resultados retornados para a quarta consulta.

Tabela 4. Publicações por país

País	Notícias publicadas
Brasil	3998
Estados Unidos	905
Portugal	177
Espanha	50
Angola	44
Reino Unido	41
França	37
Itália	18
Canadá	12
Suíça	10

Observando os resultados, nota-se que não apenas países de língua portuguesa publicaram notícias. A relação de países que publicaram pode ser melhor visualizada no mapa da Figura 10, onde as cores mais escuras indicam um número maior de ocorrências.

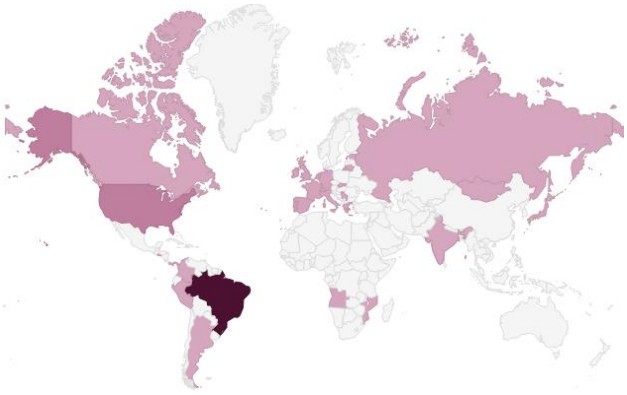


Figura 8. Mapa contendo as publicações por país

Por conta da realização das Olimpíadas no Brasil, houve uma grande repercussão mundial alertando sobre o risco de contrair a doença durante os jogos, e por conta disso, houve uma motivação da imprensa internacional para publicar notícias relacionadas à Zika, conforme foi ilustrado na Figura 8.

6. CONSIDERAÇÕES FINAIS

Com a aplicação da *web* semântica em uma base de dados de notícias sobre Zika coletada de fontes heterogêneas como jornais, revistas, blogs e fóruns, foi possível avaliar (i) qual a repercussão que publicações sobre o tema causaram nas mídias sociais; (ii) quais foram os principais autores e a relação de autoria nas redes sociais; (iii) e quais lugares mais contribuíram com publicações em língua portuguesa.

Com base nos resultados obtidos foi possível observar que o tema que mais inspirou a publicação de notícias sobre a Zika no período analisado foram os Jogos Olímpicos. Pode-se observar impactos no conteúdo das notícias, nas tendências de compartilhamento nas redes sociais e no volume de publicações a nível global.

Outros assuntos comentados se relacionavam a dados do contágio da doença e impactos e medidas de precaução, a questões políticas e a divulgação de avanços de pesquisas científicas. Além disso, as notícias mais compartilhadas pertenciam aos principais sites de notícias no cenário brasileiro, tendo o número de compartilhamentos no Facebook superior quando comparado ao Google+.

Já com relação aos principais autores, a maioria estava associada a perfis jornalísticos, como "folha.com.br". Isso ocorreu devido à não publicação do nome do autor ou por este nome não ter sido semanticamente descrito na página.

Deste modo, o trabalho demostrou, através de um estudo de caso, como as práticas da *web* semântica, como a utilização do *Framework* RDF e de consultas SPARQL, pode gerar aplicações que possibilitam a análise de dados em bases integradas para identificar padrões em publicações de mídias sociais e em RSO.

Como trabalhos futuros, pretende-se ampliar o modelo RDF criado para permitir a conexão com mais termos, como por exemplo, multimídia e outros aspectos que representem o domínio de notícias, além de gerar mais triplas como o texto da notícia para que consultas mais elaboradas sejam executadas. Também é pretendido integrar dados de outras mídias sociais e outras bases de dados abertos ligados para ampliarmos o escopo das análises e para que investigações mais complexas possam ser realizadas acerca de doenças como a Zika, Dengue e Chikungunya.

7. AGRADECIMENTOS

Nossos agradecimentos à Capes, ao CNPQ e ao PPGI/UFRJ pelas bolsas concedidas.

8. REFERÊNCIAS

[1] About | DBpedia: 2016. *http://wiki.dbpedia.org/about*. Accessed: 2016-10-09.

[2] Brickley, D. and Miller, L. 2012. FOAF vocabulary specification 0.98. *Namespace document*. 9, (2012).

[3] Campos, G.S., Bandeira, A.C. and Sardi, S.I. 2015. Zika virus outbreak, Bahia, Brazil. *Emerging infectious diseases*. 21, 10 (2015), 1885.

[4] Campos, S.R. 2013. *Validação de dados em sistemas de data warehouse através de índice de similaridade no processo de ETL e mapeamento de trilhas de auditoria utilizando indexação ontológica*. Universidade de Brasília.

[5] DCMI Home: Dublin Core Metadata Initiative (DCMI): 2016. *http://dublincore.org/*. Accessed: 2016-09-28.

[6] Fernández-García, N., Sánchez-Fernández, L., Blázquez-del-Toro, J.M. and Villamor-Lugo, J. 2007. The News Ontology for Professional Journalism Applications. *Ontologies*. Springer. 887–919.

[7] Fernández-Garcıa, N. and Sánchez-Fernández, L. 2004. Building an ontology for NEWS applications. *Poster Sesssion of the 3rd International Semantic Web Conference, ISWC* (2004).

[8] Fiocruz: 2016. *http://portal.fiocruz.br/pt-br*. Accessed: 2016-10-09.

[9] FOAF Vocabulary Specification: 2013. *http://xmlns.com/foaf/spec/*. Accessed: 2016-09-28.

[10] García, R., Perdrix, F. and Gil, R. 2006. Ontological infrastructure for a semantic newspaper. *Semantic Web Annotations for Multimedia Workshop, SWAMM* (2006).

[11] Luz, K.G., Santos, G.I.V. dos and Vieira, R. de M. 2015. Zika Virus Fever. *Epidemiologia e Serviços de Saúde*. 24, 4 (2015), 785–788.

[12] Martins, M. de F.M. 2016. Análise bibliométrica de artigos científicos sobre o vírus Zika. *Revista Eletrônica de*

Comunicação, Informação & Inovação em Saúde. 10, 1 (2016).

[13] Ontotext 2016. Graph Database Free Downolad - Ontotext Graph DB ™. *Ontotext*.

[14] SILVA, G.C. da and LIMA, T. de S. 2002. RDF e RDFS na Infra-estrutura de Suporte à Web Semântica. *Revista Eletrônica de Iniciação Científica da Sociedade Brasileira de Computação. Porto Alegre, Ano II*. 2, (2002).

[15] Webhose.io: *https://webhose.io/documentation*. Accessed: 2017-02-12.

[16] Zanluca, C., Melo, V.C.A. de, Mosimann, A.L.P., Santos, G.I.V. dos, Santos, C.N.D. dos, Luz, K., Zanluca, C., Melo, V.C.A. de, Mosimann, A.L.P., Santos, G.I.V. dos, Santos, C.N.D. dos and Luz, K. 2015. First report of autochthonous transmission of Zika virus in Brazil. *Memórias do Instituto Oswaldo Cruz*. 110, 4 (Jun. 2015), 569–572.