APDIS
Associação Portuguesa de Documentação e Informação de Saúde

# De *blogs* a dados abertos: um estudo de caso na disseminação de informação em saúde

## From blogs to open data: a case study in the dissemination of health information

**Ivan Luiz M. RICARTE**. Faculdade de Tecnologia, Universidade Estadual de Campinas, Limeira (SP), Brasil. (ricarte@unicamp.br)

**Karina S. HAGIWARA**. Faculdade de Tecnologia, Universidade Estadual de Campinas, Limeira (SP), Brasil. (karina.hagiwara@gmail.com)

**Maria Cristiane B. GALVÃO**. Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto (SP), Brasil. (mgalvao@usp.br)

**Resumo**

**Introdução**: A atual transição da Web para a Web Semântica, com informação disponibilizada também no formato de dados abertos e conectados, permite a integração da informação de distintas fontes, possibilitando o seu uso em novas aplicações e expandindo horizontes de conhecimento. A informação em saúde disponibilizada na Web não pode estar alheia a essa transição. **Objetivo**: Avaliar o esforço necessário e os potenciais benefícios envolvidos na tradução de informação em saúde disseminada em um *blog* para o formato de dados abertos. **Método**: Foi realizado um estudo de caso com um *blog* de disseminação de informação em saúde. Para as publicações desse *blog* foram desenvolvidos um modelo de representação em formato de dados abertos e uma aplicação de *software* para traduzir as informações para esse modelo. O resultado foi avaliado em termos de qualidade da informação, possibilidades de acesso à informação e conexões com outras fontes de dados. **Resultados**: O *blog* selecionado foi o Fale com o Dr. Risadinha, que dissemina informação sobre a saúde de crianças e adolescentes para o público leigo. No momento da realização deste estudo de caso, o *blog* continha 479 publicações, todas em português. Um modelo da informação do *blog* foi desenvolvido em *Unified Modeling Language*, contemplando aspetos genéricos, comuns a qualquer *blog*, e específicos, presentes nas publicações do *blog* selecionado. Os elementos desse modelo foram expressos usando RDF (*Resource Description Framework*), o arcabouço para a representação de dados abertos na Web Semântica. Nesse modelo RDF foram utilizados, quando possível, vocabulários padronizados como *Semantically-Interlinked Online Communities*, *Dublin Core* e *Friend of a friend*. A aplicação de *software* para traduzir as publicações para esse modelo de dados abertos foi desenvolvida em Java. Das 479 publicações foram derivadas 15.812 triplas RDF, às quais foram agregadas 605 conexões a recursos da DBpedia. Todos esses factos puderam ser consultados e analisados usando a linguagem de consulta SPARQL, bem como aplicações desenvolvidas em qualquer linguagem de programação com recursos para manipular dados em RDF, como Java, Python e R. **Discussão e conclusões**: Além dos benefícios já propagados e associados à disseminação de dados abertos, como a integração com dados provenientes de outras fontes, este estudo mostrou que há benefícios para os editores do *blog*, na forma de avaliações sobre a consistência e a qualidade da informação. Por outro lado, ficaram evidentes os limites do processamento automático. Nesse sentido, a atuação de um profissional da informação como mediador nesse processo é

essencial para explorar devidamente o potencial dos dados abertos conectados. Tal profissional deve ter clara compreensão dos modelos de informação usados para a representação de dados abertos e conhecer os seus principais vocabulários, ontologias e conjuntos de dados disponíveis.

**Palavras-chave**

Armazenamento e recuperação da informação; Disseminação de informação; Web semântica

**Abstract**

**Introduction**: The current transition from the Web to the Semantic Web, with information also available as linked open data, allows the integration of information from different sources, enabling reuse and expanding knowledge horizons. Health information available on the Web must also make this transition. **Objective**: Evaluate the required effort and potential benefits involved in translating health information disseminated on a blog into open data format. **Method**: A case study was conducted with a blog focused on disseminating health information. A conceptual model was developed for the posts in this blog, as well as a software application to translate the information from the posts for the linked open data format. The outcome was assessed in terms of information quality, information access possibilities and connections to other data sources. **Results:** The blog selected was Ask Dr. Giggle, which disseminates information about the health of children and adolescents to the lay public. At the time of this case study, the blog contained 479 posts, all in Portuguese. A blog information model was developed in Unified Modeling Language, covering generic aspects, common to any blog, and specific aspects, present at posts from the selected blog. The elements of this model were expressed using Resource Description Framework (RDF), the framework for representing open data in the Semantic Web. In this RDF model, standardized vocabularies such as Semantically-Interlinked Online Communities, Dublin Core, and Friend of a friend were used as much as possible. The software application to translate posts for this model was developed in Java. From the 479 posts, 15812 RDF triples were derived, and 605 connections to DBpedia resources were added. These facts could then be queried and parsed using the SPARQL query language, as well as by applications developed in any other programming language that handles RDF data, as Java, Python, and R. **Discussion and conclusions**: In addition to the benefits already propagated and associated with the dissemination of open data, such as integration with data from other sources, this study showed that there are benefits for blog publishers in assessing information consistency and quality. On the other hand, the limits of automatic processing became evident. In this sense, the role of an information professional as a mediator in this process is essential to properly explore the potential of open data. Such a professional should have a clear understanding of the information models used to represent open data and know their main vocabularies, ontologies and available datasets.

**Keywords**

Information storage and retrieval; Information dissemination; Semantic web

## Introdução

A Web evoluiu rapidamente de um modelo no qual conteúdos eram disponibilizados por detentores de acesso a servidores Web, em sua concepção inicial, para um modelo no qual qualquer usuário pode ser um provedor de informação, com aplicações em redes sociais, *blogs*

e fóruns – um modelo que se tornou conhecido como a Web 2.0[1]. Nessa Web, a informação está disponível a usuários humanos para ser consumida principalmente por meio de navegadores e outras aplicações, que apresentam conteúdos semiestruturados em *Hypertext Markup Language* (HTML). Nos últimos anos, uma transição em outra direção de evolução está em andamento. É a evolução para a Web Semântica[2], voltada primariamente não para os usuários humanos, mas para disponibilizar a infraestrutura de informação necessária para o desenvolvimento de novas classes de aplicações, que possibilitem explorar o amplo espectro de dados da Web.

A disponibilização de dados por meio da Web tornou-se uma prática corrente. Diversas instituições, das mais diversas naturezas, apresentam publicamente dados que coletam ou produzem, em variados formatos. São usuais os portais Web de dados abertos, como da Organização Mundial da Saúde (https://www.who.int/data/gho), do Banco Mundial (https://data.worldbank.org/) e de governos como dos Estados Unidos (https://www.data.gov/), de Portugal (https://dados.gov.pt/pt/) e do Brasil (http://dados.gov.br/).

A diversidade de formatos nos quais esses dados são oferecidos motivou a criação de uma classificação de dados abertos proposta por Berners-Lee[3]. Nessa classificação, recebem 1 estrela os dados que estão disponíveis na Web, em qualquer formato, desde que com alguma licença aberta (para que sejam considerados dados abertos). Dados abertos que estejam em um formato estruturado legível por máquinas (e.g., dados em formato Microsoft Excel) recebem a classificação de 2 estrelas. Os dados abertos recebem a classificação de 3 estrelas se o formato legível por máquinas não for proprietário, como arquivos em formato CSV (valores separados por vírgulas) ou JSON (*JavaScript Object Notation*).

Para melhorar a classificação dos dados abertos é preciso utilizar a infraestrutura de representação de dados da Web Semântica. Na Web Semântica, a informação da Web tradicional é disponibilizada também em um formato mais apropriado para facilitar o acesso por agentes (ou aplicações) de *software* e permitir a conexão entre informação de diferentes fontes[4]. Os dados são representados com o modelo do *Resource Description Framework*[5] (RDF), o arcabouço tecnológico da Web Semântica. Nesse modelo, cada item de informação tem associado um identificador lógico denominado URI (*Universal Resource Identifier*), formado por sequências de caracteres com estrutura similar a um endereço Web sem que, no entanto, precise representar um recurso disponível em algum servidor. Dados abertos representados nesse formato recebem, na classificação de Berners-Lee, 4 estrelas.

No nível máximo dessa classificação, 5 estrelas, explora-se a conexão entre os dados abertos. De modo geral, dados podem estar conectados, no sentido de um conjunto de dados usar ou referenciar parte de outro conjunto de dados. Por exemplo, os dados da Organização Mundial da Saúde sobre incidência de meningite assumem, para um dos atributos, valores de classificação de renda definidos em conjuntos de dados do Banco Mundial. Em dados abertos 5 estrelas, essas conexões são diretas, por meio de URI. Desse modo, os conjuntos de dados se integram, sem necessidade de repetir a informação de um conjunto de dados em outro e oferecem contexto aos dados utilizados.

Para que essa integração efetivamente ocorra, é essencial que os vocabulários utilizados para descrever os dados sejam previamente acordados por instituições ou comunidades de usuários. Nesse sentido, é usual nas representações de dados na Web Semântica a adoção de vocabulários padronizados como os definidos em *Dublin Core*[6] para a descrição de documentos e recursos eletrónicos; em FOAF[7] (*Friend of a Friend*) para a descrição de relações

interpessoais; e em SIOC[8] (*Semantically-Interlinked Online Communities*) para a descrição de comunidades virtuais.

A informação de saúde difundida por meio da Web não deve estar à parte dessa evolução. Uma preocupação inicial em relação à informação em saúde na Web esteve relacionada à credibilidade da informação[9], levando a iniciativas como HON[10] (*Health On the Net Code*) e preocupações sobre a informação disseminada em redes sociais[11] e em conteúdos produzidos pelos usuários[12]. É de se esperar que a informação em saúde confiável, disponibilizada inicialmente em HTML, possa ser também disponibilizada na Web Semântica e, assim, ser utilizada e conectada a outras fontes de informação. No entanto, a conversão de dados apresentados inicialmente em páginas HTML para o formato RDF, exceto no caso de dados tabulares, é um processo manual e, consequentemente, moroso.

Este trabalho tem por objetivo avaliar o esforço necessário para realizar a tradução de conteúdos em HTML para o formato apropriado para a Web Semântica, com dados abertos em formato RDF com a classificação de 5 estrelas. Particularmente, a fonte de dados será um *blog* de disseminação de informação em saúde confiável. Serão avaliadas também as limitações envolvidas nesse processo de tradução, bem como os potenciais benefícios envolvidos na utilização do formato de dados abertos com a classificação de 5 estrelas.

## Método

O método de estudo de caso foi adotado para explorar detalhadamente como pode ser realizada a tradução de publicações em HTML de um *blog* de disseminação de informação em saúde para dados no formato RDF, bem como os benefícios e limitações envolvidos nesse processo.

Para tanto, a primeira etapa do estudo de caso, realizada por programadores de *software*, é o desenvolvimento de uma aplicação de *software* para extrair informação das páginas HTML e produzir dados em RDF. Realizar esse desenvolvimento permite explorar as possibilidades dessa tradução, com estratégias para explorar o conhecimento sobre a organização da informação, registar as etapas realizadas internamente pela aplicação (com a produção de arquivos de *log*, que registam eventos notáveis no processo de tradução) e avaliar os resultados produzidos. Os arquivos de *log* são artefactos que podem ser inspecionados por programadores de *software* para compreender as causas e propor soluções para problemas ou deficiências nos resultados produzidos ou simplesmente para acompanhar as etapas para a realização de uma tarefa.

O resultado produzido pela aplicação de tradução é a criação de um repositório de dados em formato RDF, ou seja, dados 4 estrelas na classificação de Berners-Lee. A partir do repositório, os dados podem ser diretamente consultados ou exportados para arquivos de texto em diferentes formatos. A segunda etapa do estudo é a avaliação desses dados por inspeção de arquivos textuais ou pelo uso dos dados em outras aplicações, por analistas de dados com experiência em Web Semântica.

Uma vez que dados em formato RDF sejam corretamente produzidos a partir da tradução da informação nas publicações, a terceira etapa do estudo envolve avaliar como elevar a classificação desses dados produzidos de 4 para 5 estrelas, com o estabelecimento de conexões dos dados em RDF no repositório à informação em outros conjuntos de dados. O responsável por esta etapa do estudo deve conhecer bem os conjuntos de dados abertos disponibilizados na Web Semântica.

Por fim, os dados produzidos devem ser analisados em termos de qualidade da informação[13], além de ser disponibilizados a outras aplicações que evidenciem as novas possibilidades de uso da informação apresentada nesse formato. Essa etapa deve ser realizada por cientistas da informação com experiência em avaliação da qualidade da informação e em Web Semântica.

## Resultados

Para este estudo foi selecionado como fonte de informação o *blog* Fale com o Dr. Risadinha (http://www.drrisadinha.org.br/), que dissemina informação baseada em evidência sobre a saúde de crianças e adolescentes para o público leigo. No momento da realização deste estudo de caso, o *blog* continha 479 publicações, todas em português.

A aplicação de *software* para este estudo de caso foi desenvolvida com o uso da linguagem de programação Java (https://www.oracle.com/java/). Essa foi uma opção de conveniência, por familiaridade dos autores com os recursos oferecidos por essa plataforma de desenvolvimento. No entanto, todas as funcionalidades descritas poderiam ter sido implementadas em outras linguagens, pois em geral há recursos equivalentes para a obtenção de dados pela Internet e para manipulação de dados em RDF em diferentes plataformas.

A apresentação dos resultados está assim organizada: modelo de informação do *blog*; extração de dados do *blog*; representação dos dados do *blog* em RDF e o processo de conversão; conexões com outras fontes de dados; e consultas e relatórios.

### *Modelo de informação do blog*

Para desenvolver a aplicação de produção de dados abertos a partir das publicações do *blog*, o primeiro passo foi analisar a estrutura da informação dessas publicações. Alguns elementos de informação são inerentes a qualquer *blog*, como o seu nome e endereço na Web. As publicações de todos os *blogs* também têm elementos comuns, como o título da publicação, a data de publicação, o responsável pela publicação e os marcadores (etiquetas) de assunto. O conteúdo das publicações tem formato livre, podendo conter textos e imagens.

O conteúdo das publicações do *blog* Fale com o Dr. Risadinha, especificamente, adotam uma estrutura visual uniforme. A Figura 1 apresenta os elementos presentes em todos os *blogs*, como a data de publicação (no espaço delimitado pelo retângulo 1), o título (2), o responsável pela publicação (9) e os marcadores (10). Já os elementos específicos desse *blog*, dispostos no conteúdo da publicação, são: uma imagem representativa para o conteúdo da mensagem (3); a mensagem curta (4), que sintetiza a informação apresentada na publicação; a mensagem longa (5), o texto detalhado elaborado a partir de evidências em saúde; as referências utilizadas (6); o autor ou autores do texto (7); e o revisor ou revisores do texto (8).

A partir da análise desses elementos, um modelo da informação do *blog* foi elaborado e representado por meio do diagrama de classes de UML (*Unified Modeling Language*). Esse modelo, apresentado na Figura 2, representa que um *blog* (*Blog*) tem propriedades como nome (*name*) e endereço (*url*) e é composto por diversas publicações (*Post*). Cada publicação, por sua vez, tem propriedades como título (*title*), data de publicação (*published*) e endereço (*url*). Além disso, cada publicação tem vários marcadores (*label*), um publicador (*Publisher*) e conteúdo (*Content*). Esses elementos presentes em todos os *blogs* recebem, no modelo, o rótulo (estereótipo) <<blog>>.

domingo, 9 de junho de 2019 1

## Como tratar sangramentos no nariz? 2

3

Para-se o sangramento no nariz sentando-se com as costas retas e inclinando-se o corpo para frente. Além disso, deve-se apertar as narinas uma contra a outra. 4

***

Geralmente, o sangramento no nariz, também chamado de epistaxe, não indica nenhuma doença, embora algumas vezes possa ser um problema de saúde sério, necessitando de atendimento médico. Algumas medidas simples podem ajudar a parar o sangramento na maior parte dos casos. Sentar-se deixando as costas retas e inclinar-se para frente pode ajudar, pois ficando reta, a pessoa diminui a pressão de sangue nas veias do nariz, o que vai levar o sangramento a parar mais rápido. Inclinar-se para frente previne ainda que o sangue que está saindo do nariz seja engolido, o que pode ser irritativo para o estômago. Outra coisa que pode ajudar a parar o sangramento no nariz é fechar as narinas e apertá-las uma contra a outra, usando os dedos das mãos, por cerca de 10 a 15 minutos. Ao apertar o nariz para parar o sangramento, deve-se respirar pela boca. Caso não melhore, pode ser feita a mesma coisa por mais 10 a 15 minutos. Contudo, se, na segunda vez, o nariz não parar de sangrar, deve-se procurar atendimento médico de emergência. Além disso, caso o nariz volte a sangrar, deve-se assoá-lo para tirar pedaços de sangue coagulado que possam ter ficado lá dentro e é importante buscar atendimento médico. Caso alguém costume ter sangramento no nariz mais de uma vez por semana deve-se buscar consulta em um otorrinolaringologista, médico especialista em problemas do nariz, garganta e ouvido. Em pessoas que têm frequentes sangramentos no nariz, pode ser necessária a cauterização de vasos sanguíneos, que significa queimar pequenos vasos sanguíneos do nariz com eletricidade, laser ou uma substância conhecida como nitrato de prata. Assim, busca-se fazer com que esses vasos parem de sangrar. Caso você ou alguém de sua família tenha sangramentos frequentes, procurem um médico e ele indicará qual o melhor tratamento. 5

Referências: 6
MayoClinic [Internet]. Nosebleeds. Informação atualizada em maio de 2018. Disponível em: https://www.mayoclinic.org/symptoms/nosebleeds/basics/when-to-see-doctor/sym-20050914. Acesso em: 16 Abr. 2019.
MayoClinic [Internet]. Nosebleeds: first aid. Informação atualizada em setembro de 2017. Disponível em: https://www.mayoclinic.org/first-aid/first-aid-nosebleeds/basics/art-20056683. Acesso em: 16 Abr. 2019.

Autor do resumo: 7
Jéssica Nara Targino Cavalcante

Revisores do resumo: 8
Profa. Dra. Maria Cristiane Barbosa Galvão, Gabriella Neves Cury

Você achou esta informação útil? Clique AQUI para dar a sua opinião!

Postado por Gabriella Neves às 17:59:00 9

Marcadores: Cauterização, Epistaxe, Laser, Nariz, Nitrato de Prata, Sangramento no nariz, Sangue, Vasos sanguíneos. 10

**Figura 1.** Estrutura de uma publicação do *blog* Fale com o Dr. Risadinha.

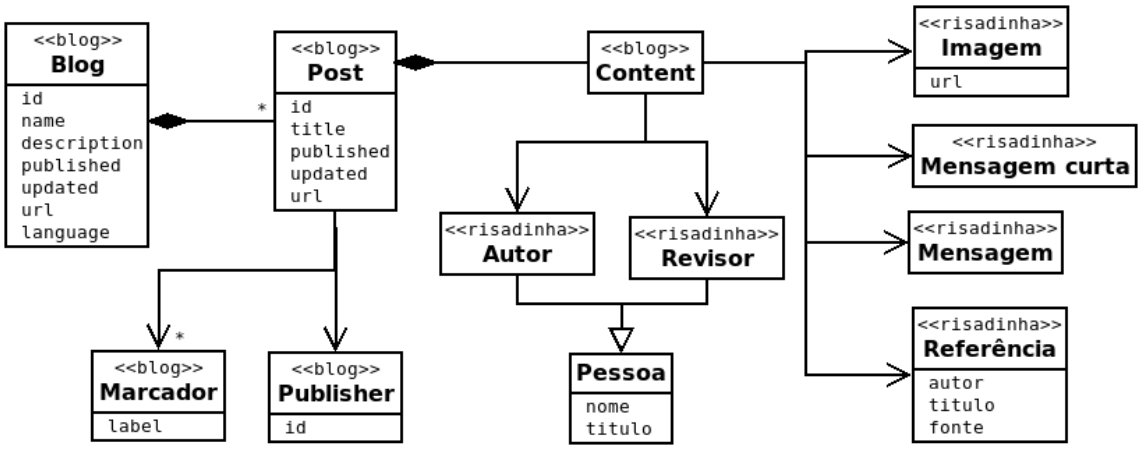


**Figura 2.** Modelo de informação para o *blog* Fale com o Dr. Risadinha.

Nesse modelo da Figura 2, os elementos específicos do *blog* Fale com o Dr. Risadinha são representados com o estereótipo <<risadinha>> e estão todos contidos no elemento *Content* do *blog*. Autores e Revisores são Pessoas, que tem nome e título (de tratamento).

Para a aplicação desenvolvida para este estudo, os elementos desse modelo de informação foram representados internamente como classes e atributos da linguagem de programação Java.

*Extração de dados do blog*

Embora as informações disponibilizadas no *blog* sejam públicas e possam ser extraídas diretamente das páginas em HTML, o modo mais prático de obtê-las para um aplicativo de *software* é por meio de uma interface de programação (API, *Application Programming Interface*). Em geral, as diferentes plataformas de *blogs* e redes sociais oferecem essa opção aos programadores e, para o Fale com o Dr. Risadinha, a opção disponível é a *Blogger API* (https://developers.google.com/blogger). Essa interface de programação oferece funções para obter os dados sobre um *blog* e suas publicações. Para obter dados públicos do *blog*, a aplicação que requisita os dados precisa apenas ser identificada por meio de uma chave de acesso (*API key*). Para o desenvolvimento da aplicação deste estudo de caso, duas solicitações são realizadas: uma para obter os dados gerais do *blog* e outra para obter uma lista com o conteúdo de todas as publicações desse *blog*.

Por intermédio da *Blogger API* são obtidos todos os elementos de informação marcados com o estereótipo <<blog>> no modelo da Figura 2. Cada requisição é realizada como uma solicitação a um servidor Web, identificando no endereço o *blog*, a chave de acesso e a informação desejada. A resposta do servidor é um conjunto de dados no formato JSON. Para traduzir essas respostas para o formato interno de objetos na linguagem de programação Java foi utilizada a biblioteca Gson (https://github.com/google/gson), disponibilizada pela equipe de desenvolvimento do Google.

Já a extração dos elementos de informação marcados no modelo da Figura 2 com o estereótipo <<risadinha>> a partir do elemento de conteúdo da publicação não é tão trivial. O conteúdo é representado internamente como um fragmento de uma página HTML e, embora as publicações do *blog* Fale com o Dr. Risadinha adotem uma estrutura visual uniforme, há nas

publicações variações na representação interna que impossibilitam a localização direta do elemento de informação dentro do conteúdo. Além disso, há algumas poucas publicações que não respeitaram a estrutura adotada, como algumas divulgações de eventos e revisões de publicações anteriores. A solução para esse problema foi combinar a extração direta de dados com os recursos de um *HTML Parser*, uma biblioteca com funções que processam conteúdos HTML. No caso desta aplicação, desenvolvida em Java, a biblioteca utilizada para esse fim foi Jsoup (https://jsoup.org/).

Com essas funcionalidades desenvolvidas, a aplicação de *software* do estudo de caso pode obter todas as informações do *blog* e representá-las como objetos da linguagem de programação Java.

*Representação de dados do blog em RDF e o processo de conversão*

Para traduzir os dados do *blog* do formato interno da linguagem de programação para o formato da Web Semântica, o primeiro passo é estabelecer o modelo e os vocabulários utilizados para a representação dos dados em RDF.

Em RDF, a informação é representada por um conjunto de factos (*statements*), cada facto envolvendo três elementos: um sujeito, uma propriedade e um objeto. Por este motivo, esses factos são usualmente denominados triplas. O objeto de uma tripla, por sua vez, pode ser um valor literal ou uma referência a um sujeito de outra tripla. Por meio desse modelo simples, é possível representar conjuntos de dados arbitrariamente complexos. Por exemplo, supondo que a publicação ilustrada na Figura 1 esteja associada a um identificador interno #25, parte de seus dados pode ser representada por triplas, como:

| *Sujeito* | *Propriedade* | *Objeto* |
|---|---|---|
| #25 | Post.published | 2019-06-09 |
| #25 | Post.title | “Como tratar sangramentos no nariz?” |
| #25 | Revisor | #P12 |
| #P12 | Pessoa.nome | “Maria Cristiane Barbosa Galvão” |
| #P12 | Pessoa.titulo | “Profa. Dra.” |

Nesse exemplo, um identificador interno de outra tripla (#P12) representa uma pessoa que aparece como o sujeito de algumas triplas e como o objeto de outras.

Obviamente, quando o objetivo é disponibilizar tal informação como dados abertos, é fundamental que os usuários que compartilham esses dados adotem uma representação comum para a descrição das triplas. Para tanto, deve-se avaliar quais elementos de informação (apresentados na Figura 2) têm correspondência a termos já definidos em vocabulários existentes. Na Web Semântica, esses vocabulários, denominados *namespaces*, são identificados por URI. No domínio de descrição envolvido neste estudo de caso foram utilizados termos dos vocabulários RDF (identificado pelo URI http://www.w3.org/1999/02/22-rdf-syntax-ns#), *Dublin Core* (DCTerms, http://purl.org/dc/terms/), FOAF (http://xmlns.com/foaf/0.1/) e SIOC *Core Ontology* (http://rdfs.org/sioc/ns#).

Os elementos de informação para os quais foram identificados termos correspondentes nos vocabulários padrões da Web Semântica são apresentados na Tabela 1 para as propriedades do

*blog* e de suas publicações. O termo *type* do vocabulário RDF foi utilizado como a propriedade para representar que autores e revisores são pessoas (objetos da classe *Person* de FOAF).

**Tabela 1.** Mapeamento de elementos de informação para vocabulários padronizados

| *Elemento do modelo de informação* | *Vocabulário* | *Termo* |
|---|---|---|
| Blog | SIOC | Forum |
| Blog.id | DCTerms | identifier |
| Blog.name | DCTerms | title |
| Blog.description | DCTerms | description |
| Blog.published | DCTerms | issued |
| Blog.updated | DCTerms | modified |
| Blog.url | DCTerms | source |
| Blog.language | DCTerms | language |
| Post | SIOC | Post |
| Post.id | DCTerms | identifier |
| Post.title | DCTerms | title |
| Post.published | DCTerms | issued |
| Post.updated | DCTerms | modified |
| Post.url | DCTerms | source |
| Mensagem curta | DCTerms | abstract |
| Mensagem | SIOC | content |
| Imagem | FOAF | depiction |
| Autor | DCTerms | creator |
| Revisor | DCTerms | contributor |
| Pessoa | FOAF | Person |
| Pessoa.nome | FOAF | name |
| Pessoa.titulo | FOAF | title |
| Referência | DCTerms | references |
| Marcador | DCTerms | subject |

Os dados extraídos do *blog*, armazenados internamente como objetos Java na aplicação desenvolvida neste estudo, foram traduzidos em factos no formato RDF com o uso da biblioteca Apache Jena[14] (http://jena.apache.org/), que oferece funcionalidades para criar, manipular e armazenar dados em RDF. Ao conjunto de dados criado por esse processo de conversão foi adicionada uma tripla para especificar a licença de dados abertos que foi adotada (CC-BY 4.0), tendo como propriedade o termo *license* de DCTerms.

Uma das características de dados abertos é que seus itens de dados são identificáveis por meio de um URI. Aos itens de dados que foram produzidos foi atribuído um *namespace* local (http://drrisadinha.org). Por exemplo, a publicação exemplificada na Figura 1 à qual foi atribuída, na plataforma do *blog*, o identificador 250773238602800406, é representada nesse conjunto de dados pelo URI http://drrisadinha.org/250773238602800406.

Nesse processo de conversão foi possível observar, por meio da análise de registos de *log*, uma potencial fonte de inconsistência na identificação de colaboradores (autores e revisores), devido a diferenças na grafia de nomes ou títulos entre publicações distintas. Por exemplo, em algumas publicações o nome completo do colaborador foi utilizado, em outras apenas o primeiro nome e o último sobrenome; em alguns casos, o nome do meio foi grafado por extenso, em outros foi abreviado; nomes ora com, ora sem acento; erros de digitação, com uma letra trocada; e, ainda, houve inconsistências na grafia dos títulos de tratamento como, no caso de um colaborador, terem sido utilizadas três formas diferentes: “Enf.”, “Enfa.” e “Enfa Ms.”.

A solução para esse problema foi a criação manual de um «banco de colaboradores», em RDF, que adotou a forma preferencial dos nomes e títulos dos colaboradores, definida a partir da lista de nomes de membros da equipe e colaboradores apresentada na página inicial do *blog*. Assim, no processo de conversão dos dados para RDF, ao tratar o nome de um autor ou revisor, inicialmente é feita uma busca por similaridade com os nomes presentes nesse banco de colaboradores. Se encontrado um nome muito próximo a um existente, esse elemento RDF é utilizado. Colaboradores eventuais, que não fazem parte da equipe do *blog*, não são encontrados nesse processo e, nesse caso, novas triplas RDF são criadas. Para realizar a busca por similaridade, foi utilizada a funcionalidade de similaridade nebulosa (*fuzzy*) de textos da biblioteca *Apache Commons Text* (https://commons.apache.org/proper/commons-text).

*Conexões com outras fontes de dados*

A execução das etapas anteriormente descritas produz dados representados por um conjunto de triplas RDF, com itens de informação referenciáveis por URI e ao qual foi atribuída uma licença aberta. Assim, atende aos requisitos necessários para receber a classificação de 4 estrelas na escala de dados abertos de Berners-Lee. Para elevar a classificação desses dados ao nível máximo de 5 estrelas, é necessário conectar a informação produzida a itens de dados em outros conjuntos de dados abertos também representados em RDF.

Provedores de dados abertos estão gradualmente oferecendo dados em RDF, embora ainda em quantidade bem inferior a formatos de dados com classificação de até três estrelas. Por exemplo, em fevereiro de 2020 o portal governamental de dados abertos dos Estados Unidos disponibilizava 253.005 conjuntos de dados, dos quais 11.818 (4,7%) estavam em formato RDF. Um diretório de dados abertos disponíveis com a classificação de 5 estrelas pode ser encontrado em *The Linked Open Data Cloud* (https://lod-cloud.net/) que, nessa mesma época, contemplava 1.421 conjuntos de dados, incluindo classificações da Organização Mundial da Saúde (ICD10, ICF, ICPC-2), terminologias em saúde (*Medical Subject Headings*, *Logical Observation Identifier Names and Codes*, SNOMED *Clinical Terms*) e a DBpedia[15], resultado de um esforço colaborativo para disponibilizar o conteúdo da Wikipedia em RDF (http://dbpedia.org).

Para ilustrar como os dados produzidos neste estudo de caso podem ser conectados a outras fontes de dados optou-se por realizar, durante o processo de conversão dos dados, uma consulta à DBpedia para cada marcador (assunto) de cada publicação. A consulta é expressa com a linguagem de consulta criada para dados em RDF, SPARQL[16], sendo realizada ao ponto de acesso remoto (*endpoint*) disponibilizado pelo projeto DBpedia (http://dbpedia.org/sparql). Como os marcadores estão em português, foi buscada correspondência com termos identificados na DBpedia com o rótulo em português, por meio do prefixo http://pt.dbpedia.org/resource. Ao localizar um termo correspondente, uma tripla tendo como assunto o recurso na DBpedia era adicionada ao conjunto de dados, conectando assim a informação local aos dados já apresentados na Wikipedia. Por exemplo, o marcador «alergia» em uma publicação resulta, por meio desse processo, em uma conexão adicional com o recurso «Allergy» da DBpedia, a partir do qual conexões adicionais informam que o código CID-10 para esse termo é T78.4 e o identificador MeSH é D006967, além de outras informações.

*Consultas e relatórios*

No momento da realização deste estudo de caso, o *blog* Fale com o Dr. Risadinha disponibilizava 479 publicações. O processo de conversão para RDF criou um repositório com 15.812 triplas, das quais 1.444 eram triplas que definiam marcadores de assuntos. A partir dessas triplas com marcadores de assuntos e com a consulta ao ponto de acesso da DBpedia, 605 novas triplas estabelecendo conexões a recursos da DBpedia foram adicionadas ao conjunto de dados. O processo de conversão completo, incluindo a obtenção de todas as publicações do *blog*, a conversão para RDF e armazenamento das triplas no repositório, a busca por recursos na DBpedia, o registo dos passos executados em arquivos de *log* e a criação de um arquivo consolidado com as triplas em um formato textual para RDF (*turtle*) levou, em um computador pessoal conectado à Internet, menos de 30 minutos.

Os dados produzidos podem ser manipulados diretamente no repositório por meio de consultas em SPARQL ou ser utilizados em outras aplicações de *software* em qualquer linguagem de programação que ofereça facilidades para manipulação de dados em RDF. Nesse estudo de caso, por exemplo, o arquivo consolidado foi interpretado e processado por meio de *scripts* na linguagem R (https://www.r-project.org/), permitindo derivar informações como uma relação de publicações e seus autores, a quantidade de autores que colaboraram para as publicações (17), quais autores produziram mais publicações (91), quais revisores foram mais atuantes, quantos marcadores distintos foram atribuídos às publicações (1.444), entre muitas outras possibilidades.

A título de exemplo, uma lista de autores e títulos das publicações que eles produziram pode ser obtida pela consulta em SPARQL sobre o repositório criado:

```
SELECT ?autor ?titulo
WHERE {
  ?r <http://purl.org/dc/terms/creator> ?a  .
  ?r <http://purl.org/dc/terms/title> ?titulo  .
  ?a <http://xmlns.com/foaf/0.1/name> ?autor  .
}
```

Nessa consulta, a primeira linha indica quais são os dados desejados (expressos pelos rótulos “?autor” e “?título”) e, na cláusula WHERE, como esses dados podem ser extraídos do repositório: procure todas as triplas que expressem que um recurso correspondente a uma publicação (identificado nessa consulta por “?r”) tem um autor (propriedade *creator* de DCTerms) com uma identificação interna representada nessa consulta por “?a” e, para esses recursos, retorne o título do recurso (propriedade *title* de DCTerms) e o nome do autor (propriedade *name* de FOAF). A partir do resultado dessa consulta, e com as funcionalidades de análise de dados da linguagem R, é possível produzir relatórios como ilustrado na Figura 3.

| autor | posts |
|---|---|
| Claudio Vinicius de Assis Rondado | 91 |
| Nivaldo Sena da Silva | 79 |
| Lenisa de Mello e Souza | 53 |

**Figura 3.** Fragmento da lista de autores e respetiva quantidade de publicações produzida.

## Discussão e Conclusões

Este estudo de caso ilustrou como é possível criar um conjunto de dados abertos e conectados na Web Semântica a partir de informação em saúde disponibilizada em um *blog*. Além dos benefícios já associados à disseminação de dados abertos e conectados, como a integração com dados provenientes de outras fontes, este estudo mostrou que há benefícios para os produtores da informação (no caso, os editores do *blog*), na forma de avaliações sobre a qualidade da informação, como relatórios de análise das produções e recomendações para a correção de erros e para uniformização na grafia de nome de colaboradores e de rótulos.

O estudo realizado é inovador, pois a produção de dados RDF a partir de textos é um processo usualmente manual e laborioso. A produção automática é usualmente restrita a fontes de dados em formato tabular, tarefa para a qual há até grupos de trabalho no Consórcio W3 (https://www.w3.org/TR/csv2rdf/). Neste estudo de caso, essa produção automática foi possibilitada pela estrutura uniforme adotada para as publicações do *blog* Fale com o Dr. Risadinha.

Por outro lado, o estudo evidenciou que a qualidade das conexões de dados produzidas com o processamento automático é limitada. Nesse sentido, um processo semiautomático e com a intermediação de um profissional da informação poderia explorar de modo mais efetivo o potencial dos dados abertos conectados. Para tanto, esse profissional deve ter clara compreensão dos modelos de informação usados para a representação de dados abertos conectados e conhecer os seus principais vocabulários, ontologias e conjuntos de dados disponíveis. Assim, é fundamental que esse conhecimento seja incorporado à formação desses novos profissionais.

O código produzido como parte deste estudo de caso, desenvolvido na linguagem de programação Java, está disponível em https://github.com/IvanRicarte/DrRisadinha. A partir desse mesmo endereço é possível obter o arquivo RDF consolidado em formato textual (*turtle*).

## Agradecimentos

À toda equipe do projeto Fale com o Dr. Risadinha, que produz informação de qualidade em linguagem simples e disponível a todos. KSH agradece também ao Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq) pelo suporte financeiro oferecido por meio do Programa PIBIC/CNPq/Unicamp.

## Referências bibliográficas

1. Newman R, Chang V, Walters RJ, Wills GB. Web 2.0: the past and the future. Int J Inf Manage. 2016;36(4):591-8.
2. Shadbolt N, Hall W, Berners-Lee T. The semantic Web revisited. IEEE Intell Syst. 2006;21(3):96-101.
3. Berners-Lee T. Linked data [Internet]. W3C; 2006. Available from: https://www.w3.org/DesignIssues/LinkedData.html
4. Bizer C. The emerging web of linked data. IEEE Intell Syst. 2009;24(5):87-92.

5. McBride B. The resource description framework (RDF) and its vocabulary description language RDFS. In: Staab S, Studer R, editors. Handbook on ontologies. Berlin: Springer; 2004. p. 51-65.
6. Baker T. Libraries, languages of description, and linked data: a Dublin Core perspective. Libr Hi Tech. 2012;30(1):116-33.
7. Graves M, Constabaris A, Brickley D. FOAF: connecting people on the semantic Web. Cat Classif Q. 2007;43(3/4):191-202.
8. Breslin JG, Decker S, Harth A, Bojars U. SIOC: an approach to connect web-based communities. Int J Web Based Communities. 2006;2(2):133-42.
9. Gaudinat A, Cruchet S, Boyer C, Chrawdhry P. Enriching the trustworthiness of health-related web pages. Health Informatics J. 2011;17(2):116-26.
10. Boyer C, Gaudinat A, Hanbury A, Appel RD, Ball MJ, Carpentier M, et al. Accessing reliable health information on the web: a review of the HON approach. Stud Health Technol Inform. 2017;245:1004-8.
11. Lin WY, Zhang X, Song H, Omori K. Health information seeking in the Web 2.0 age: trust in social media, uncertainty reduction, and self-disclosure. Comput Human Behav. 2016;56:289-94.
12. Ma T, Atkin D. User generated content and credibility evaluation of online health information: a meta analytic study. Telemat Informatics. 2017;34(5):472-86.
13. Hadhiatma A. Improving data quality in the linked open data: a survey. J Phys Conf Ser. 2018;978(1).
14. Yu L. Jena: a framework for development on the Semantic Web. In: A developer's guide to the semantic Web. Berlin: Springer; 2011. p. 491-532.
15. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, et al. DBpedia: a large-scale, multilingual knowledge base extracted from Wikipedia. Semant Web. 2014;1:1-5.
16. Arenas M, Pérez J. Querying semantic web data with SPARQL. In: Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Athens, Greece; 2011. p. 305-16.

**Notas biográficas**

**Ivan Luiz Marques RICARTE.** Professor Titular da Faculdade de Tecnologia da Universidade Estadual de Campinas (Unicamp), atuando principalmente em sistemas de informação em saúde, aprendizagem colaborativa e aplicações da Web Semântica em saúde e educação. Currículo disponível em http://lattes.cnpq.br/4372943322993518.

**Karina Sayuri HAGIWARA.** Bacharel em Sistemas de Informação pela Faculdade de Tecnologia da Universidade Estadual de Campinas (Unicamp). Currículo disponível em http://lattes.cnpq.br/3575796298731824.

**Maria Cristiane Barbosa GALVÃO.** Professora da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (USP), com atuação na área de informação em saúde. Currículo disponível em http://lattes.cnpq.br/9163421021115381.