# Um método para publicação semântica *Linked Data* de bases de dados convencionais e um estudo de caso real de artigos acadêmicos

# *A method for Linked Data semantic publishing of conventional databases and a real case study of academic papers*

Adriano de Oliveira GONÇALVES[1] 0000-0002-5215-8251

Mark Douglas de Azevedo JACYNTHO[2] 0000-0003-3910-1442

## Resumo

Este trabalho propõe, por meio de uma metodologia de pesquisa quali-quantitativa, aplicada e experimental, um método para mapeamento e publicação sistemática de uma base relacional existente, segundo os princípios *Linked Data*, a partir de um estudo de caso de artigos acadêmicos da conferência interna Semana de Integração Acadêmica de uma universidade pública federal brasileira. O método proposto é resultado de mapeamento do domínio de conhecimento estudado em ontologias *Linked Data* de referência (*Schema.org, Friend of a Friend, Bibliographic Ontology, Semantic Web Conference Ontology*, entre outras). O referido método foi aplicado ao banco de dados relacional da conferência, a fim de disponibilizá-lo em formato inteligível por máquinas na *Web*, estabelecendo-se ainda *links* semânticos com a famosa fonte de dados DBpedia, por meio de um processo de *mashup* automatizado. Os resultados obtidos com o método foram bastante satisfatórios, atingindo-se plenamente o objetivo de se publicar uma visão *Linked Data* sobre os dados relacionais, sem alterá-los. Espera-se, com este trabalho, fomentar a disponibilização de dados semânticos na *Web*, em consonância com os princípios *Linked Data*. Assim, contribui-se para a ampla divulgação de conhecimento, impulsionada pela capacidade que a *Web* Semântica provê às máquinas de interligar, compreender e descobrir informações.

**Palavras-chave**: Bases de dados. Ontologia. *Resource Description Framework*. *Web* semântica.

## Abstract

*Using a quali-quantitative, applied, and experimental research methodology, this paper proposes a method for systematic mapping and publication of an existing relational database according to the Linked Data principles and based on a case study of academic papers of the internal conference* Semana de Integração Acadêmica *(Academic Integration Week), carried out in a Brazilian federal public university. The proposed method results from mapping the knowledge domain studied in reputed Linked Data ontologies (Schema.org,*

[1] Universidade Federal do Rio de Janeiro, Superintendência de Tecnologia da Informação e Comunicação, Diretoria de Macaé. Macaé, RJ, Brasil.

[2] Instituto Federal Fluminense, Programa de Pós-Graduação em Sistemas Aplicados à Engenharia e Gestão, Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão. R. Dr. Siqueira, n. 273, Parque Dom Bosco, 28030-130, Campos dos Goytacazes, RJ, Brasil. Correspondência para/*Correspondence to*: M. D. A. JACYNTHO. *E-mail*: <mjacyntho@iff.edu.br>.

Recebido em: 28 de junho de 2018, reapresentado em 31 de maio de 2019 e aprovado em 28 de agosto de 2019.

Como citar este artigo/*How to cite this article*

Gonçalves, A. O.; Jacyntho, M. D. A. Um método para publicação semântica *Linked Data* de bases de dados convencionais e um estudo de caso real de artigos acadêmicos. *Transinformação*, v. 32, e180051, 2020. http://dx.doi.org/10.1590/1678-9865202032e180051

CC BY

*Friend of a Friend, Bibliographic Ontology, Semantic Web Conference Ontology, etc.) and it was applied to the conference's relational database in order to make it available in machine readable format on the Web, establishing, in addition, semantic links with the famous DBpedia dataset through an automated mashup process. The results obtained with the method were quite satisfactory. The goal of publishing a Linked Data view over the relational data was fully achieved without changing it. With this work, we hope to foster making semantic data available on the web in accordance with Linked Data principles, thus contributing to a wide dissemination of knowledge, boosted by the ability the Semantic Web provides to machines for interconnecting, understanding, and discovering information.*

***Keywords***: *Databases. Ontology. Resource Description Framework. Semantic Web.*

## Introdução

Instituições acadêmicas (*e.g.*, universidades) são, por definição, valiosos repositórios de conhecimento. Diversas pesquisas, publicações e eventos são, recorrentemente, gerenciados por sistemas de informação heterogêneos e isolados, que não compartilham conhecimento entre si devido à ausência de um formato de dados padrão, inteligível por máquina, para publicação e intercâmbio de informações. Na sociedade moderna, a *Web* se apresenta, sem dúvida, como o ambiente mais abrangente para a publicação desse conhecimento científico. Não obstante, a busca por publicações científicas geralmente se dá pelo próprio ser humano, o qual tem que pesquisar em vários sítios *Web*, analisando documentos *Hypertext Markup Language* (HTML) ou *Portable Document Format* (PDF), perfazendo suas próprias buscas, associações, interpretações e integrações entre conceitos e documentos, em um processo manual limitado, demorado e sujeito a erros.

Para auxiliar o ser humano na busca, integração e processamento de informação, surge a *Web* Semântica, inicialmente proposta por Tim Berners-Lee, em Berners-Lee *et al.* (2001), e difundida pelo consórcio W3C (W3C, 2018). A *Web* Semântica é uma extensão da *Web* original e consiste em um conjunto de tecnologias e padrões que viabiliza o compartilhamento e o reuso de conhecimento, em escala global. Isso é feito por meio de formatos de dados inteligíveis por máquinas, de forma que estas possam executar automaticamente, em larga escala, as tarefas que têm sido executadas de modo manual até então (Azevedo; Jacyntho, 2014). A utilização desses padrões tem se tornado uma tendência mundial, com considerável adesão nos meios científicos.

Como subconjunto mais pragmático da *Web* Semântica, proposto em Berners-Lee (2006), destaca-se a noção de *Linked Data* (Dados Ligados, em português). A ideia consiste, basicamente, na utilização da *Web* para publicar e interligar dados de forma direta através de *links* semânticos para que pessoas e máquinas possam explorá-los, tornando mais fácil a descoberta de dados relacionados. Esse conceito dá origem à chamada *Web* de Dados Ligados, um gigantesco grafo de dados global (Abele *et al.*, 2017), compreensível por agentes de software, possibilitando buscas por informação mais precisas e consistentes (Heath; Bizer, 2011).

Um dos aspectos mais importantes da publicação de dados na *Web* diz respeito à forma como eles são representados (Azevedo; Jacyntho, 2014). Não basta diversas fontes de dados publicarem seus dados em um formato comum. Para que as máquinas consigam, efetivamente, intercambiar dados e compreendê-los sem ambiguidade, elas precisam compartilhar um vocabulário semântico comum para a área de conhecimento em questão. A *Web* Semântica provê mecanismos para criar, publicar e reutilizar esses vocabulários. De acordo com Jacyntho (2012), vocabulário é um conjunto de termos que denotam elementos de modelos formais de representação do conhecimento chamados de ontologias. Sendo assim, é imperativa a escolha de ontologias adequadas para descrever o domínio de conhecimento pretendido, buscando sempre reusar aquelas globalmente consagradas. À medida que fontes de dados publiquem suas informações reusando as mesmas ontologias ou vocabulários, estabelece-se um entendimento comum, viabilizando a interpretação desses dados por máquinas de fato, independentemente da origem e local da fonte de dados.

Este trabalho procura responder à seguinte questão de pesquisa: como publicar bases de dados relacionais de acordo com os princípios *Linked Data*, sem, no entanto, alterar os dados originais? Desse modo, esta pesquisa

tem o objetivo de propor um método para mapear e publicar, de forma sistemática, dados relacionais como um grafo *Resource Description Framework* (RDF), apenas de leitura, em consonância com os princípios *Linked Data*. Para tal, foi empregada uma metodologia de pesquisa quali-quantitativa, aplicada e experimental, culminando na aplicação do método proposto a um estudo de caso real da conferência acadêmica interna Semana de Integração Acadêmica de uma universidade pública federal brasileira. O estudo de caso consiste no mapeamento do banco de dados relacional convencional de gerenciamento do evento para o modelo de dados padrão da *Web* Semântica, o modelo RDF, contemplando: emprego de um conjunto de ontologias consagradas; estabelecimento automático de dados ligados com uma fonte de dados semânticos bem conhecida/reusada; seleção de *softwares* necessários para a publicação desses dados na *Web* de Dados Ligados.

Além do método proposto, que configura o objetivo deste trabalho, esta obra traz ainda, como contribuições, sugestões de ferramentas para cada um dos seus passos e um estudo de caso real como exemplo da aplicação do método. Adicionalmente, concebe um modelo ontológico para o domínio de conhecimento do estudo de caso.

O modelo ontológico apresentado pode ser sistematicamente empregado para a publicação e interligação de produções acadêmicas na *Web* de Dados. Espera-se que a universidade promotora do evento em estudo possa disponibilizar, na *Web*, os dados referentes a produções científicas de seus eventos internos, em formato inteligível por máquina, ligados a recursos de outras fontes de dados presentes na *Web*. Espera-se também que este modelo se torne uma referência para instituições acadêmicas que pretendam publicar suas produções, contribuindo, assim, para o crescimento do conhecimento científico, fomentado pela capacidade que a *Web* Semântica oferece às máquinas de interligar, compreender e descobrir informações.

## Fundamentos

### *Web* Semântica

O termo *Web* Semântica designa uma extensão da Web atual, de forma a prover tecnologias e padrões que permitam a adição de significado formal explícito à informação publicada, o qual possa ser entendido por máquinas, viabilizando, assim, a execução automática de tarefas até então executadas manualmente em larga escala (Yu, 2011). De acordo com Berners-Lee *et al.* (2001), a *Web* Semântica não é uma *Web* separada, e sim uma extensão da atual, na qual a informação recebe um significado bem definido, de forma a permitir que computadores e pessoas trabalhem cooperativamente.

Conforme Jacyntho e Azevedo (2015, p. 3.):

> A *Web* Semântica baseia-se na simples ideia de reusar a arquitetura da *Web* original para compartilhar e interligar metadados (ou dados diretamente, no caso de *Linked Data*), assim como os recursos (entidades) descritos por eles, de forma homogênea e padrão, abstraindo das idiossincrasias tecnológicas internas de cada fonte (servidor) de dados.

Na *Web* Semântica, também conhecida como *Web* de Dados, entidades do mundo real (pessoas, locais, objetos, publicações, conceitos abstratos, *etc*.) são representadas por meio de recursos, os quais são identificados por um endereço *Web* único, um *Uniform Resource Identifier* (URI). Ao acessar (dereferenciar) esse endereço, retorna-se um documento contendo uma descrição do recurso solicitado, utilizando uma linguagem estruturada inteligível por máquina. Nesse documento, o recurso é descrito utilizando-se *links Web* que o conectam, relacionando-o semanticamente, com representações de outros recursos, geograficamente distribuídos, identificados por outros URIs, usando, pois, a *Web* como um grafo de dados global que cresce a cada dia (Heath; Bizer, 2011; Azevedo; Jacyntho, 2014).

Segundo Antoniou *et al.* (2012), toda linguagem para intercâmbio de dados entre máquinas é definida por três pilares, a saber: modelo de dados, semântica e sintaxe. A seguir, são apresentados esses três componentes da *Web* Semântica.

## *Resource Description Framework* (RDF): modelo de dados

*Resource Description Framework* (RDF) é o modelo de dados padrão da *Web* Semântica. Trata-se de um modelo em grafo que permite a descrição de recursos por meio de triplas [recurso – propriedade – valor] (Jacyntho, 2012). Dessa forma, o modelo RDF adiciona semântica explícita à estrutura de *links* da *Web* ao nomear o *link* com o URI da correspondente propriedade de uma ontologia, assim como seu nó de origem (URI do recurso) e seu nó de destino (URI de outro recurso ou valor literal). A utilização desse simples modelo permite que dados (semi)estruturados sejam combinados, publicados e compartilhados entre diferentes aplicações (RDF Working Group, 2014). Utilizando os autores desse artigo, poder-se-ia afirmar que "Adriano conhece Mark" criando, por exemplo, a seguinte tripla RDF:

*<http://www.example.com/mestrado/saeg/alunos/adrianooliveiragoncalves>*

*<http://xmlns.com/foaf/0.1/knows>*

*<http://www.example.com/mestrado/saeg/professores/markdouglasjacyntho>*.

Nesse exemplo de tripla, o nó de origem é o URI que identifica o aluno "Adriano de Oliveira Gonçalves", a aresta é o URI da propriedade "conhece" da ontologia para descrição de pessoas *Friend of a Friend* (FOAF) (Brickley; Miller, 2014), e o nó de destino é o URI que identifica o professor "*Mark Douglas Jacyntho*". Um conjunto de triplas forma um grafo RDF.

Segundo Jacyntho e Schwabe (2016):

> Em essência, um modelo de dados é apenas uma maneira de visualizar os dados. O modelo relacional estabelecido visualiza os dados por meio de relações e tuplas. O modelo em grafo RDF, baseado em triplas, é uma representação natural para vários tipos de aplicações (por exemplo, *Facebook, Twitter*, sistemas de recomendação, *etc.*), nas quais as entidades estão fortemente conectadas umas com as outras. Em contraste com o modelo relacional, essas aplicações consideram propriedades multivaloradas tão desejáveis na modelagem de dados reais, que trabalham com propriedades multivaloradas por padrão. Consultas de propriedades com valores múltiplos ou valor único são feitas exatamente do mesmo modo, sem preocupações com a necessidade de se associar a uma terceira tabela para modelar um relacionamento n-para-n. Além disso, o modelo RDF é mais conveniente se a aplicação tiver alta heterogeneidade em seu esquema ou necessidade frequente de adaptação de esquema. Os bancos de dados RDF simplificam o desenvolvimento de aplicações de dados ligados e também se alinham muito bem com numerosos algoritmos e técnicas estatísticas desenvolvidas para grafos (Jacyntho; Schwabe, 2016, p.1 tradução nossa)[3].

É importante destacar que não há necessidade de os diferentes recursos envolvidos em ligações semânticas estarem publicados no mesmo servidor. Este é o cerne da *Web* de Dados: fontes de dados distintas com recursos interligados (*mashup* semântico), promovendo o reuso das informações e permitindo à máquina navegar de um

[3] No original: *"In essence, a data model is just a way to view the data. The established relational model views the data through relations and tuples. The RDF graph model, based on triples, is a natural representation for various types of applications (e.g., Facebook, Twitter, recommender systems, etc.), where entities are strongly connected with each other. In contrast with legacy RDBMS, these applications consider multi-valued properties to be so desirable in modeling real-life data that they support multi-valued properties by default. Querying for multi-valued and single-valued properties is done in exactly the same way, without concerns about the need to join with a third table to model an n-to-n relationship. Furthermore, the RDF model is more convenient if the application has high heterogeneity in its schema or frequent need for schema adaptation. RDF stores simplify the development of linked data applications, and also align very well with numerous algorithms and statistical techniques developed for graphs".*



 http://dx.doi.org/10.1590/1678-9865202032e180051

recurso para outro, e de uma fonte para outra, extraindo mais informações, independentemente da localização dos dados (Azevedo; Jacyntho, 2014).

O modelo RDF oferece uma linguagem de consulta padrão chamada de SPARQL *Protocol* and RDF *Query Language* (SPARQL Working Group, 2013). Ela desempenha um papel análogo ao da linguagem de consulta *Structured Query Language* (SQL) nos bancos de dados relacionais.

### *Ontologias*: *semântica*

Um dos componentes fundamentais da *Web* Semântica são modelos formais de representação de conhecimento, denominados ontologias (Berners-Lee *et al.*, 2001), as quais, no campo da Ciência da Computação, possuem a capacidade de promover o compartilhamento e a reutilização do conhecimento (Camilo; Silva, 2009). De acordo com Gruber (1995), uma ontologia é a especificação de uma conceitualização. Conforme W3C OWL Working Group (2012), ontologias são vocabulários formalizados de termos, geralmente abrangendo um domínio de conhecimento específico e compartilhados por uma comunidade de usuários.

A utilização de ontologias traz uma série de vantagens, como, por exemplo, o fato de, diferentemente de como acontece com a representação textual dos documentos na *Web*, estas serem definidas em linguagem formal, não deixando espaço para as limitações semânticas e ambiguidades de entendimento da linguagem natural. Além disso, ontologias podem ser mapeadas entre diferentes linguagens computacionais e especializadas para domínios de conhecimento mais específicos (Souza, 2016).

Na *Web* Semântica, as ontologias são criadas por meio das linguagens (metaontologias) *Web Ontology Language* (OWL) (Horrocks *et al.*, 2012) e RDF *Schema* (RDFS) (Brickley *et al.*, 2014), linguagens declarativas baseadas em lógica descritiva. Como ontologias OWL e RDFS são documentos RDF, as classes e propriedades que compõem seu vocabulário também são identificadas por endereços *Web* (URIs), consistindo, dessa forma, em recursos. Esses recursos podem igualmente ser dereferenciados pelas aplicações para que, através das suas descrições RDF, a máquina compreenda os tipos dos recursos (classes), assim como os relacionamentos entre eles, inclusive sendo capaz de inferir novas triplas com base nas regras (axiomas) descritas na ontologia (Azevedo; Jacyntho, 2014).

Para que haja melhor comunicação entre aplicações, é fortemente recomendado o reuso de ontologias existentes para a descrição dos dados. Essa prática maximiza a probabilidade de os dados serem consumidos entre aplicações sem que seja necessário aplicar alguma modificação ou pré-processamento (Heath; Bizer, 2011).

### *Sintaxes para publicação de arquivos RDF*

Para que os dados do grafo abstrato RDF sejam efetivamente publicados na *Web*, foram criadas diversas sintaxes-padrão para arquivos RDF, entre as quais: RDF/XML (Gandon; Schreiber, 2014), Turtle (Beckett *et al.*, 2014) e JSON-LD (Sporny *et al.*, 2014).

### *Linked Data*

Um paradigma muito importante pertencente à *Web* Semântica é o conceito de *Linked Data* (Dados Ligados, em português). Trata-se de um conjunto de princípios para publicar e interligar dados estruturados na *Web*, provendo um caminho mais genérico e flexível para que os consumidores de dados possam descobrir e integrar dados de diferentes fontes de dados (Heath; Bizer, 2011). A aplicação desse paradigma viabiliza a chamada "*Web* de Dados Ligados" (*Web of Linked Data*, em inglês): uma *Web* de dados estruturados, totalmente compreensível por máquinas, tornando a busca por informações mais precisa e consistente (Jacyntho; Azevedo, 2015). A topologia da *Web* de Dados consiste em um enorme grafo global, interligando diversas fontes de dados abertos. O coração

da *Web* de Dados é a fonte de dados DBpedia (Bizer *et al*., 2009; DBpedia, 2017), que é uma versão *Linked Data* da Wikipédia.

Em Berners-Lee (2009), são apresentados os quatro princípios *Linked Data* que devem nortear a publicação de dados semânticos na *Web* de Dados, a saber: (1) Use URI para nomear recursos; (2) Use URI Hypertext Transfer Protocol (HTTP), de forma que se possa acessar esses recursos; (3) Faça todos os URIs dereferenciáveis. Em outras palavras, quando um URI for acessado, retorne informações úteis, utilizando os padrões (RDF, SPARQL); (4) Inclua *links* para outros URIs, para que se possa descobrir (navegar para) mais recursos (*Linked Data mashup*).

O termo *mashup* denota técnicas utilizadas na *Web* que permitem combinar dados de múltiplas fontes em uma única aplicação, como, por exemplo, reusar serviços de dados proprietários disponibilizados por gigantes da *Web*, como *Google, Yahoo* e *eBay*. *Linked Data mashup* significa utilizar as tecnologias da *Web* Semântica para integrar, de forma padronizada, dados estruturados de diferentes fontes, permitindo, assim, acesso através de clientes genéricos, como navegadores RDF, máquinas de busca RDF e agentes de consulta na *Web*. Supera-se, desse modo, as limitações dos *mashups* tradicionais (Bizer; Cyganiak; Gauß, 2007). Esse tipo de *mashup* semântico permite que as máquinas naveguem de um recurso para outro, obtendo mais informações, independentemente da localização dos dados (Azevedo; Jacyntho, 2014).

### *O modelo cinco estrelas para publicação de dados abertos*

Com o objetivo de encorajar e orientar as pessoas na produção e publicação de dados abertos ligados, Tim Berners-Lee (2009) criou um modelo de classificação de maturidade de dados abertos em cinco níveis, ou cinco estrelas, descrito a seguir: ★ Disponível na *Web* (em qualquer formato, *e. g.*, PDF), mas com uma licença aberta, para serem dados abertos; ★★ A regra anterior, mais: dados estruturados inteligíveis por máquina (*e. g.*., uma planilha Excel em vez de uma imagem digitalizada de uma tabela); ★★★ Todas as regras anteriores, mais: formato não proprietário (*e. g.*, CSV[4], em vez de Excel); ★★★★Todas as regras anteriores, mais: usar os padrões do W3C (RDF e SPARQL) para descrever os recursos (entidades), de forma que seja possível referenciá-los para reuso; ★★★★★Todas as regras anteriores, mais: ligar os dados publicados a dados de outras pessoas para prover contexto (*Linked Data mashup*).

## Método proposto para publicação de dados relacionais, como *Linked Data*

Nesta seção, é apresentado o método proposto para a publicação de bases de dados relacionais, como *Linked Data*, retratado no fluxograma da Figura 1, bem como, para cada um dos passos, a trajetória metodológica percorrida para sua definição a partir do domínio do estudo de caso considerado neste trabalho. Ainda, são sugeridas ferramentas que foram avaliadas e selecionadas em cada um dos passos. Complementando a explanação do método, a seção subsequente a esta é reservada à descrição detalhada da aplicação do método proposto ao estudo de caso real, visando corroborar sua eficácia.

A publicação, na *Web* de Dados, dos trabalhos acadêmicos da conferência realizou-se em um processo iterativo e incremental composto pelos seguintes passos:

(1) *Definir informações a publicar*: a partir do banco de dados relacional do sistema utilizado, foram definidas quais informações seriam pertinentes à publicação *Linked Data*, considerando perguntas como: (a) Quais informações são essenciais para um catálogo de publicações científicas, expressando os principais dados dos trabalhos, seus autores, instituições e informações relacionadas? (b) Quais propriedades possuem ontologias conhecidas para modelar o domínio pretendido? (c) Quais informações, por serem sensíveis ou privadas, não devem constar na publicação semântica a ser disponibilizada abertamente na *Web*?

[4] *Comma Separated Values* (CSV): arquivo texto que armazena dados tabulares, no qual cada linha é um registro e cada registro consiste de um ou mais campos separados por vírgulas.


http://dx.doi.org/10.1590/1678-9865202032e180051

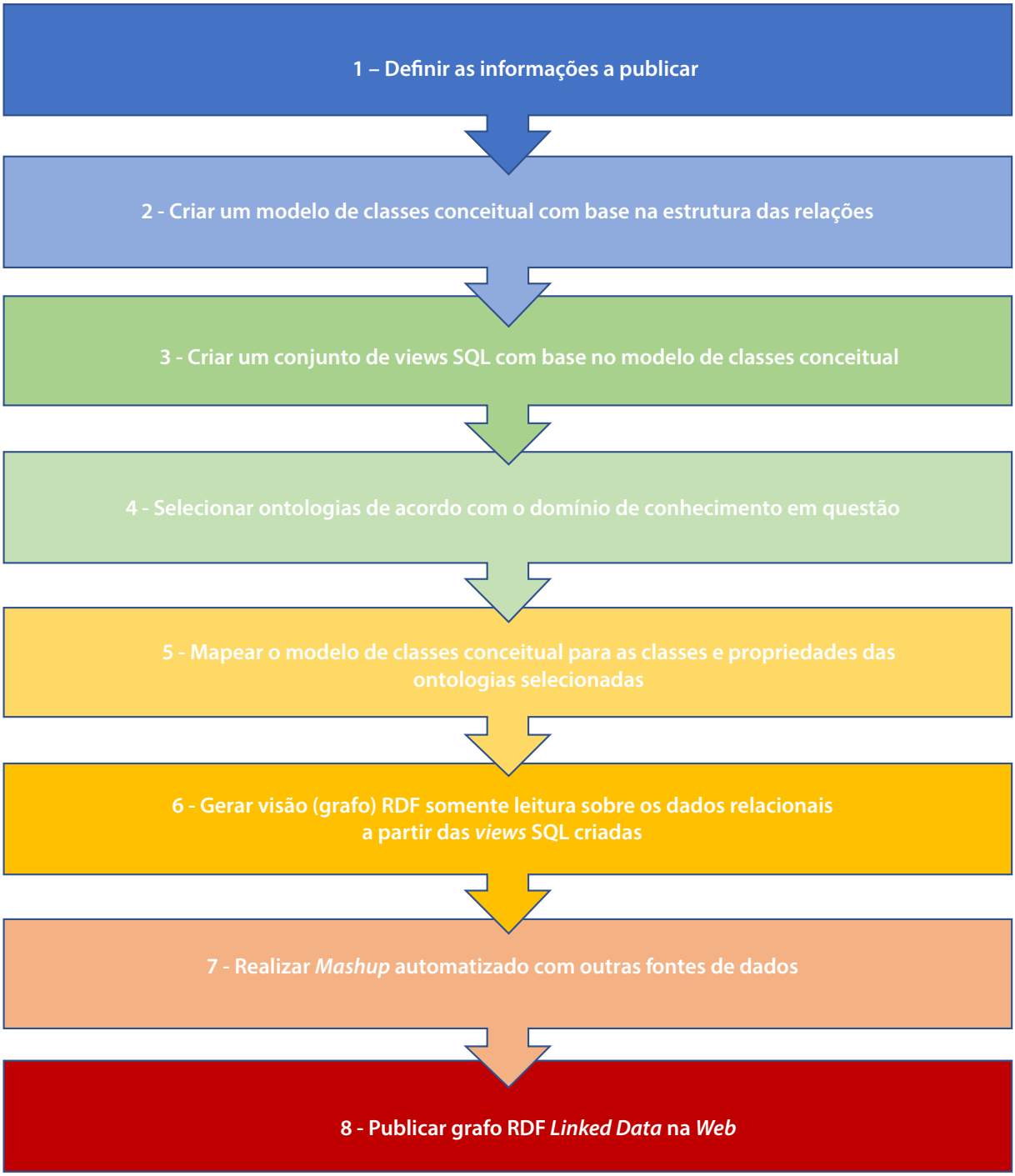


**Figura 1**. Metódo proposto.
Fonte: Elaborada pelos autores (2018).

(2) *Criar um modelo de classes conceitual com base na estrutura das relações*: com o objetivo de visualizar de forma mais simples e clara as informações escolhidas para serem publicadas, fez-se necessária a criação de um modelo de classes conceitual de mais alto nível de abstração, baseado nas relações e seus atributos, utilizando uma linguagem de modelagem adequada, como, por exemplo, *Unified Model Language* (UML) (Object Management

Group, c1997). Esta foi selecionada para este projeto por ser um padrão largamente conhecido e adotado na área de computação.

(3) *Criar um conjunto de views SQL com base no modelo de classes conceitual*: com o intuito de agregar as informações a serem publicadas, bem como não interferir na base de dados original do sistema da conferência, garantindo acesso seguro às informações pretendidas, o mapeamento semântico se deu sobre *views* SQL, ou seja, consultas SQL pré-definidas, armazenadas na estrutura de um banco de dados auxiliar separado, expressando os dados a serem mapeados. Basicamente, deve ser criada uma *view* SQL para cada classe conceitual. Para esse banco de dados auxiliar de *views* SQL, foi utilizado o MySQL (Oracle Corporation, [201-]), apenas por ser o mesmo gerenciador de banco de dados usado na conferência.

(4) *Selecionar ontologias de acordo com o domínio de conhecimento em questão*: uma vez definidos os dados a publicar, foi conduzida uma pesquisa minuciosa, visando encontrar ontologias capazes de descrever o domínio de conhecimento em questão, partindo-se das mais conhecidas e conceituadas até outras mais específicas. A pesquisa foi feita a partir do catálogo de ontologias *Linked Open Vocabularies* (Vandenbussche *et al.*, 2016), o qual se tornou uma referência para pesquisas dentro da área de *Web* Semântica, em sítios de busca e artigos acadêmicos. Com base nos resultados encontrados, foi definido o mapeamento do modelo conceitual da conferência para as classes e propriedades das ontologias selecionadas.

(5) *Mapear o modelo de classes conceitual para as classes e propriedades das ontologias selecionadas*: esta etapa compreende o mapeamento das classes e propriedades (atributos e relacionamentos) do modelo de classes conceitual para classes e propriedades das ontologias, o que pode ser documentado usando a mesma linguagem de modelagem escolhida no passo 2. A seção seguinte, sobre o estudo de caso real da conferência acadêmica, ilustra, por meio de estereótipos UML, como executar este passo mais detalhadamente.

(6) *Gerar visão (grafo) RDF somente leitura sobre os dados relacionais a partir das views SQL criadas*: esta etapa diz respeito à implementação desse mapeamento no acesso à base de dados relacional, a partir das *views* SQL criadas no passo 3 do método. Para o mapeamento em tempo real do banco de dados relacional da conferência em uma visão RDF, foi utilizada a plataforma D2RQ (Bizer *et al.*, 2012). D2RQ permite, por meio da especificação do mapeamento relacional-RDF, expressar, em uma linguagem declarativa própria (D2RQ *Mapping Language*), a disponibilização das tuplas de um banco de dados relacional em triplas RDF, em tempo de execução (*on the fly*), sem a necessidade de armazená-las em um banco de dados RDF. Além disso, oferece um SPARQL endpoint que converte, de forma transparente, *on the fly*, consultas SPARQL em consultas SQL, bem como o retorno SQL no formato definido pelo protocolo SPARQL. Um mapeamento D2RQ é em si um documento RDF, escrito na sintaxe Turtle. Já a linguagem de mapeamento – D2RQ *Mapping Language* –, é uma ontologia RDF Schema simples. Portanto, o mapeamento é expresso usando termos (classes e propriedades) dessa ontologia.

A facilidade de se configurar declarativamente o mapeamento relacional-RDF, aliado a outros recursos, assim como a capacidade de permitir acesso *Linked Data* com negociação de conteúdo HTTP para humanos (HTML) e máquinas (RDF), e a capacidade de disponibilizar um SPARQL endpoint para consultas *ad hoc*, de forma transparente, tornaram o D2RQ a escolha adequada para a implementação deste projeto. Além do D2RQ, foi avaliada também a ferramenta Triplify, que trata o mapeamento por meio de comandos SQL e não fornece SPARQL endpoint (Jaenicke *et al.*, 2010).

O mapeamento foi especificado, então, na linguagem de mapeamento do D2RQ (Cyganiak *et al.*, 2012) e experimentado através de navegação nas páginas *Web* geradas dinamicamente pela plataforma D2RQ e por consultas SPARQL. Conforme apontado no passo 2, o mapeamento foi realizado sobre *views* SQL, armazenadas em um banco de dados MySQL separado dos dados relacionais originais da conferência (Oracle Corporation, [201-]).

(7) Realizar *mashup* automatizado com outras fontes de dados: com o objetivo de alcançar as cinco estrelas propostas pelo modelo de Berners-Lee (2009), faz-se necessária a ligação de recursos da base a ser publicada com outros recursos na *Web*, utilizando-se fontes de dados e predicados pertinentes (rdfs:seeAlso, owl:sameAs, *etc.*). Para a pesquisa de fontes de dados, podem ser utilizadas iniciativas como *The Linked Open Data Cloud* (Abele *et al*., 2017). Sendo assim, a pesquisa seguiu com o processamento dos dados por meio do SPARQL *endpoint* fornecido pela plataforma D2RQ.

Uma vez que já existia uma considerável quantidade de dados na base relacional em questão, decidiu-se pela realização de um *mashup* automatizado. Ele foi feito com o emprego da ferramenta *Silk Framework* (Isele *et al*., [2009]; Petrovski *et al*., 2014) para estabelecer o *mashup* automático entre recursos RDF extraídos pelo mapeamento D2RQ e recursos da DBpedia. Esta última é uma das fontes de dados mais expressivas no grafo global da *Web* Semântica, conforme é possível perceber em Abele *et al.* (2017), e foi selecionada para este projeto como destino para dados ligados.

A escolha da ferramenta *Silk* para este projeto se justifica por fornecer, entre outros, uma gama de filtros de transformação, comparação e agregação de dados. Estes podem ser combinados em *workflows* a serem aplicados a diferentes fontes de dados, visando à geração de dados ligados, de forma automatizada, utilizando predicados pré-definidos pelo usuário. Além disso, ela já foi utilizada em diversos experimentos com resultados promissores (Isele *et al*., 2010; Halaç *et al*., 2013; Isele; Bizer, 2013; Petrovski *et al*., 2014).

Sendo assim, foi realizado o *download* das triplas RDF contendo títulos (*labels*) e resumos longos (*long abstracts*) da versão 10/2016 da DBpedia, disponível para *download* na língua portuguesa, e importadas em um banco de dados RDF (triple store) específico para *mashup*. Para tal, foram experimentados dois bancos de dados RDF (*triple stores*), a saber: GraphDB (Ontotext, 2018) e Stardog (Stardog Union, 2018). Dentre os bancos de dados RDF avaliados para este projeto, esses dois foram selecionados pela facilidade de instalação, configuração e uso, além de ambos oferecerem suporte à inferência ontológica.

Em seguida, foi criado um *workflow* na ferramenta *Silk* para encontrar similaridades entre os nomes das palavras-chave das publicações e das áreas de conhecimento da conferência e os títulos dos recursos da DBpedia. Os resultados encontrados pelo *Silk* vieram a compor triplas RDF com o predicado owl:sameAs, que significa que dois URI diferentes representam o mesmo recurso do mundo real, ligando os recursos da conferência com os da DBpedia. As triplas de *mashup* geradas foram importadas nos dois bancos de dados RDF de *mashup*, os quais eram completamente desvinculados do banco de dados relacional original da conferência.

(8) *Publicar grafo RDF Linked Data na Web*: foram avaliadas duas abordagens para a publicação dos recursos resultantes dos mapeamentos. A primeira foi por meio de consultas federadas envolvendo o SPARQL *endpoint* disponibilizado pela plataforma D2RQ e o SPARQL *endpoint* do banco de dados RDF de *mashup*. E, na segunda, tanto as triplas RDF geradas pelo mapeamento D2RQ quanto as triplas de *mashup* com a DBpedia geradas pelo *Silk* foram importadas, em lote, em um terceiro banco de dados RDF centralizado.

### *Aplicação do método em um estudo de caso real de publicações de uma conferência acadêmica*

Esta seção descreve, em detalhes, a aplicação do método proposto ao estudo de caso em questão, ou seja, o mapeamento e a publicação do modelo relacional do banco de dados de publicações da conferência como dados RDF *Linked Data*. Trata-se de uma explanação em alto nível de abstração para facilitar a compreensão. Não obstante, todos os arquivos de implementação, como *views* SQL, especificação do mapeamento D2RQ das relações (classes) e seus campos (propriedades) em RDF, *workflows* de *mashup* do *Silk*, *etc*., podem ser obtidos no repositório do projeto no GitHub[5].

[5] Disponível em: https://github.com/adrianogoncalves/siac-ontologic-model.

*Mapeamento do modelo de classes conceitual da conferência em ontologias*

Os parágrafos seguintes apresentam uma breve descrição de cada uma das ontologias selecionadas para o estudo de caso. Em seguida, detalha-se o mapeamento do modelo de classes conceitual da conferência para as ontologias, representado por meio de um diagrama de classes UML (Figura 2).

Schema.org (Schema.org Community Group, [2011]) é uma ontologia resultante do esforço conjunto das empresas Google, Microsoft, Yahoo e Yandex (Yandex, c1997). Trata-se de um vocabulário para vários domínios de conhecimento (*crossdomain*), com o objetivo de inserir metadados estruturados em páginas *Web*, viabilizando retornos mais precisos pelas máquinas de busca. Essa ontologia foi utilizada como ponto de partida para estruturar o mapeamento proposto por este trabalho. Para abreviar URI, o prefixo "schema" foi utilizado para referenciar o *namespace* da ontologia (http://schema.org/).

*Dublin Core Metadata Initiative Metadata Terms* (DCTerms) (DCMI Usage Board, 2012) é uma das ontologias mais antigas e conhecidas. Extensão da ontologia Dublin Core, consiste em um vocabulário para descrever metadados genéricos. Para abreviar URI, o prefixo "dcterms" foi utilizado para referenciar o namespace da ontologia (http://purl.org/dc/terms/).

DBpedia *Ontology* (DBpedia, 2010) é uma ontologia crossdomain definida pela equipe da DBpedia para a descrição dos seus dados provenientes da Wikipédia. Para abreviar URI, o prefixo "dbo" foi utilizado para referenciar o namespace da ontologia (http://dbpedia.org/ontology/).

*Semantic Web Conference* (SWC) (Nuzzolese *et al.*, 2016) é uma ontologia para descrever conferências científicas baseada na ontologia SWRC, adicionando *links* para classes de ontologias conhecidas. Para abreviar URI, o prefixo "swc" foi utilizado para referenciar o namespace da ontologia (http://data.semanticweb.org/ns/swc/ontology/).

*Semantic Web for Research Communities* (SWRC) é uma ontologia que visa modelar conceitos relacionados a comunidades científicas (publicações, estudantes, universidades, entre outros) (Sure *et al.*, 2005). Ela disponibiliza uma série de classes e propriedades, mas não possui *links* para outras ontologias. Para abreviar URI, o prefixo "swrc" foi utilizado para referenciar o *namespace* da ontologia (http://swrc.ontoware.org/ontology#).

*The Bibliographic Ontology* (BIBO) é uma ontologia que provê conceitos e propriedades para descrever citações e referências bibliográficas (livros, artigos, etc.) (D'Arcus; Giasson, 2009). Para abreviar URI, o prefixo "bibo" foi utilizado para referenciar o *namespace* da ontologia (http://purl.org/ontology/bibo/).

*Friend of a Friend* (FOAF) (Brickley; Miller, 2014) é uma ontologia para descrever pessoas, suas atividades e relações com outras pessoas, grupos, entidades e documentos (Azevedo; Jacyntho, 2014). Para abreviar URI, o prefixo "foaf" foi utilizado para referenciar o *namespace* da ontologia (http://xmlns.com/foaf/0.1/).

*Simple Knowledge Organization System* (SKOS) (Miles; Bechhofer, 2009) é uma ontologia para modelar estruturas de organização de conhecimento (taxonomias, tesauros, folksonomias, *etc*.) (Azevedo; Jacyntho, 2014). Para abreviar URI, o prefixo "skos" foi utilizado para referenciar o *namespace* da ontologia (http://www.w3.org/2004/02/skos/core#).

*Academic Institution Internal Structure Ontology* (AIISO) (Styles *et al.*, 2008) é uma ontologia para descrever a estrutura interna de uma instituição acadêmica. Para abreviar URI, o prefixo "aiiso" foi utilizado para referenciar o namespace da ontologia (http://purl.org/vocab/aiiso/schema#).

A Figura 2 demonstra, por meio de um diagrama de classes UML, o mapeamento do modelo de classes conceitual, gerado a partir do modelo relacional do banco de dados da conferência, nas classes e propriedades das ontologias selecionadas.

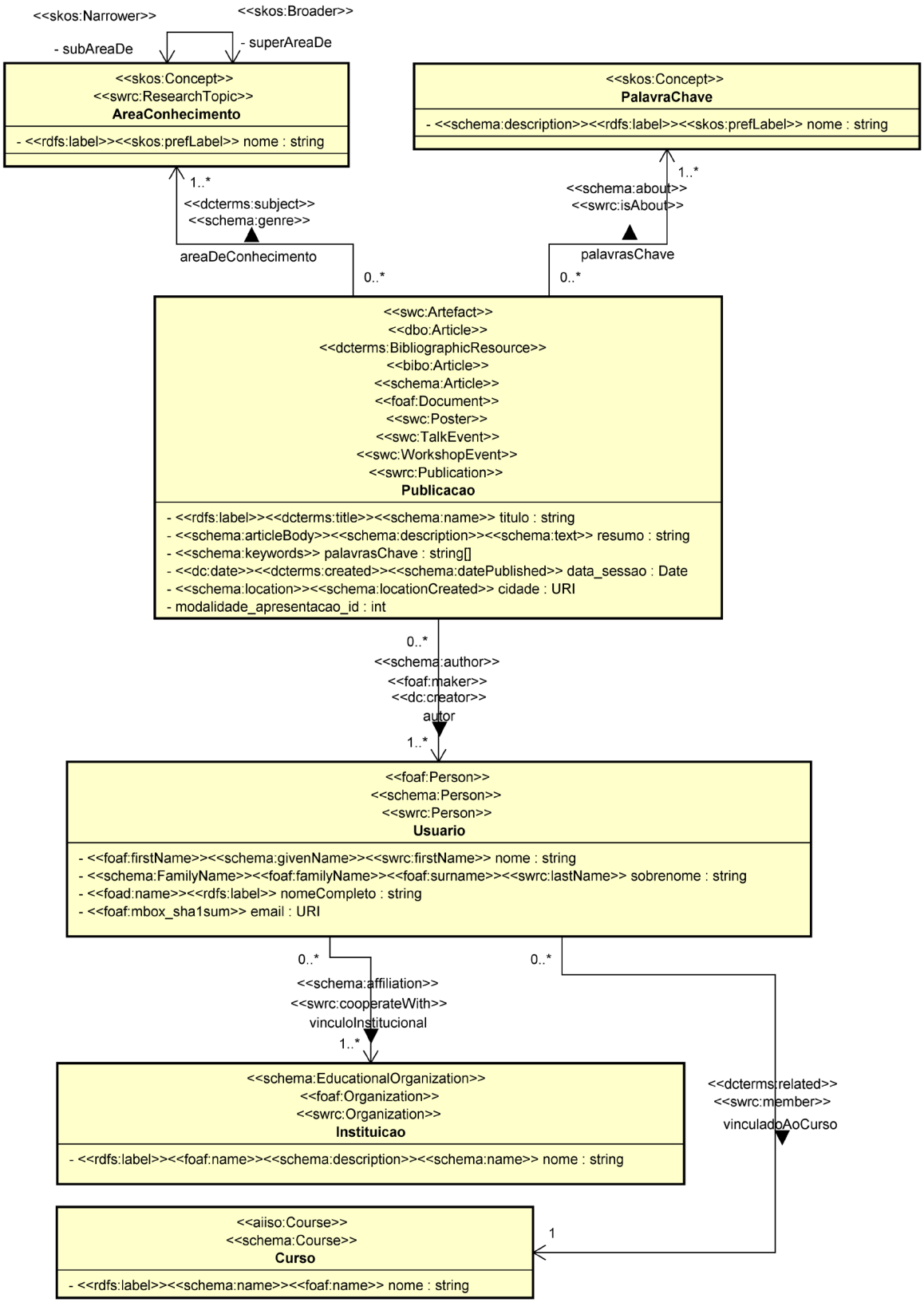


**Figura 2**. Mapeamento do modelo conceitual da conferência para as ontologias selecionadas.

Fonte: Elaborada pelos autores (2017).

O mapeamento de cada classe e suas propriedades (atributos e associações) é expresso por estereótipos UML. Por exemplo, a classe "AreaConhecimento" foi mapeada nas classes skos:Concept e swrc:ResearchTopic; o atributo "nome" da "AreaConhecimento" foi mapeado nas propriedades rdfs:label e skos:prefLabel; e a propriedade "AreaConhecimento" da classe "Publicacao", representada no diagrama pela associação entre "Publicacao" e "AreaConhecimento", foi mapeada nas propriedades dcterms:subject e schema:genre. Os outros mapeamentos seguem o mesmo raciocínio.

*Gerar visão RDF somente leitura sobre os dados relacionais*

Todo esse mapeamento representado em alto nível de abstração no diagrama da Figura 2 foi especificado em um arquivo texto, por meio da linguagem declarativa de mapeamento relacional-RDF da plataforma D2RQ (D2RQ *Mapping Language*). Como apontado no passo 6 do método proposto, o que, de fato, é mapeado são as *views* SQL criadas com base nas classes conceituais da Figura 2. Conforme dito anteriormente, a linguagem de mapeamento D2RQ é uma ontologia RDF schema simples, e o arquivo de mapeamento é um documento RDF escrito na sintaxe Turtle. É baseada nesse arquivo de mapeamento que a plataforma D2RQ possibilita acessar, *on the fly*, o banco de dados relacional como um grafo RDF virtual, apenas de leitura (*read-only*).

Para ilustrar, considerando-se novamente a classe "AreaConhecimento" (view v_area_conhecimento) do modelo apresentado na Figura 2, esta foi mapeada para as classes ontológicas skos:Concept e swrc:ResearchTopic, bem como seu atributo "nome" (campo "nome" da view v_area_conhecimento) para a propriedade rdfs:label, por meio do seguinte trecho do arquivo RDF de mapeamento[6]:

*map:area_conhecimento d2rq:ClassMap ;*

*d2rq:class skos:Concept, swrc:ResearchTopic*

*d2rq:classDefinitionLabel "area_conhecimento" ;*

*d2rq:dataStorage map:database_views ;*

*d2rq:uriPattern"area_conhecimento/@@v_area_conhecimento.id@@"*

*map:area_conhecimento__label d2rq:PropertyBridge ;*

*d2rq:belongsToClassMap map:area_conhecimento ;*

*d2rq:column "v_area_conhecimento.nome" ;*

*d2rq:property rdfs:label .*

A partir dessa especificação, o D2RQ criou um mapeamento de classe (*ClassMap*), nomeado como map:area_conhecimento, e uma ponte de propriedade (*PropertyBridge*), vinculada a esse mapeamento de classe. Dessa forma, todas as tuplas retornadas pela view v_area_conhecimento são mapeadas, na visão RDF, em recursos com a propriedade rdf:type ligada às classes ontológicas mencionadas, bem como os valores do campo v_area_conhecimento.descricao são mapeados como valores da propriedade rdfs:label desses recursos. A propriedade d2rq:uriPattern serve para definir um padrão para a formação dos URIs dos recursos gerados, assim como a propriedade d2rq:classDefinitionLabel permite definir uma etiqueta (*label*) usada no menu de acesso na representação HTML, voltada para humanos. Esse mapeamento é capaz de gerar triplas como:

*<http://example.com/area_conhecimento/1234>*

*<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>*

*<http://www.w3.org/2004/02/skos/core#Concept>.*

[6] Considerando-se os prefixos de *namespace* das ontologias selecionadas e o prefixo d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> da ontologia de mapeamento D2RQ.

*<http://example.com/area_conhecimento/1234>*
*<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>*
*<http://swrc.ontoware.org/ontology#ResearchTopic>.*

*<http://example.com/area_conhecimento/1234>*
*<http://www.w3.org/2000/01/rdf-schema#label>*
*"Ciência da Computação".*

De forma análoga, são mapeadas todas as classes e propriedades do modelo conceitual para as classes e propriedades das ontologias selecionadas, utilizando-se as views criadas no banco de dados relacional, visando-se obter a visão em grafo RDF do modelo proposto.

### *Linked Data mashup*

Para a geração do *mashup*, buscando-se encontrar ligações entre os recursos da conferência e os da DBpedia, foi utilizado o *software Silk Framework*. Esse *software* provê mecanismos para a criação de *workflows*, encadeando diversos tipos de processamento predefinidos para tratamentos e comparações entre os dados, aplicados a variados tipos de origem de dados semânticos (arquivos RDF, SPARQL endpoints, entre outros). A partir do caminho desenvolvido nesses *workflows*, os dados obtidos das duas bases de dados são comparados, e são geradas, automaticamente, novas triplas RDF, associando os recursos cujo resultado do processamento obedeça aos critérios definidos. O predicado a ser utilizado nessas novas triplas é estabelecido junto ao *workflow* do *Silk*. Neste trabalho, realizou-se o processamento usando o *workflow* descrito na Figura 3 e o predicado owl:sameAs, que significa que dois recursos interligados por ele são, na verdade, o mesmo recurso identificado por URIs distintos, ou seja, representam a mesma entidade do mundo real.

Como é possível ver na Figura 3, esse processo compara os nomes das palavras-chave e as áreas de conhecimento da conferência (mapeados na propriedade skos:prefLabel) com os nomes ou rótulos dos recursos da DBpedia (propriedade rdfs:label), aplicando-se alguns filtros (retângulos verdes) e utilizando, para comparação de cadeias de caracteres, o algoritmo da distância de Levenshtein (Bilenko *et al.*, 2003) (retângulo laranja), cuja implementação já se encontra disponível no *Silk*.

As triplas resultantes do *mashup* foram armazenadas em um banco de dados RDF separado dos dados relacionais originais.

### *Publicação dos dados*

Para publicar na *Web* os dados mapeados e ligados, foram usadas duas abordagens: (1) Disponibilização dos dois SPARQL *endpoints*: do D2RQ sobre os dados relacionais e do banco de dados RDF com o *mashup* com a DBpedia, permitindo a realização de consultas SPARQL federadas; (2) Exportação de todas as triplas geradas pelo D2RQ a partir dos dados relacionais originais, por meio do *software dump-rdf*, distribuído junto à plataforma D2RQ, e importação dessas triplas e das triplas de *mashup* geradas pelo *Silk* em um terceiro banco de dados RDF centralizado independente, a ser atualizado em lote, de tempos em tempos, divulgando-se, assim, um único SPARQL *endpoint*. Mesmo nesse caso, a interface *Web* do D2RQ precisou ser mantida para que os URI dos recursos pudessem ser dereferenciados (acessados).

Conforme apontado na metodologia, para este trabalho, foram avaliados dois bancos de dados RDF, a saber: Stardog (Stardog Union, 2018) e GraphDB (Ontotext, 2018), ambos em suas versões gratuitas. Os dois apresentaram resultados satisfatórios.

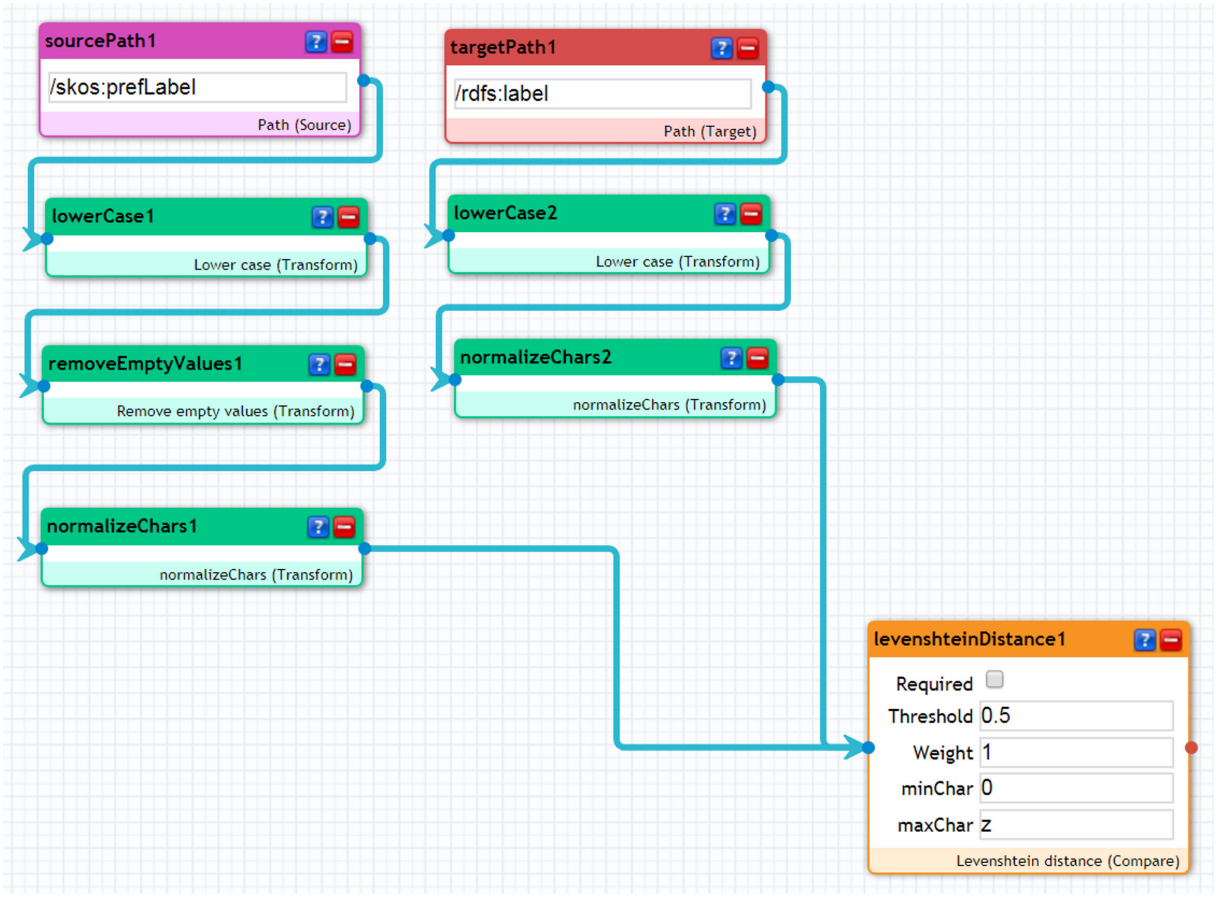


**Figura 3**. *Workflow do Silk* para procurar ligações do tipo owl:sameAs entre os recursos da conferência e os recursos da DBpedia, com base nos respectivos nomes (labels).

Fonte: Elaborada pelos autores (2017).

*Exemplo de mapeamento de uma publicação em RDF*

Para ilustrar o mapeamento em RDF dos trabalhos acadêmicos da conferência, a Figura 4 apresenta uma publicação ("Extração de regras em uma base de conhecimento por meio da navegação em um grafo") em RDF (Figura 4 a), representação para máquinas, e em HTML (Figura 4 b), representação para humanos, ambas geradas pela plataforma D2RQ. As representações RDF e HTML de um autor, não apresentadas por questões de espaço, são similares às representações da publicação, mudando apenas as propriedades e classes envolvidas.

Para efeito de testes, foi utilizado o endereço local como URI base para o banco de dados semântico (http://localhost:2020/)[7]. De novo, por razões de espaço, alguns valores textuais mais extensos foram abreviados com reticências na representação em RDF.

## Resultados e discussão

Como se pode observar nos exemplos apresentados, diante dos ensaios realizados com o modelo ontológico proposto, constatou-se ser possível navegar com sucesso pelas informações da conferência, por meio da interface

[7] Quando a aplicação for, de fato, posta em produção e publicada na Web, o endereço local http://localhost:2020 será substituído por um domínio real (http://example.com) ainda a ser definido.

**(a)**

```
<http://localhost:2020/resource/trabalhoPublicado/4333> a <http://data.semanticweb.org/ns/swc/ontology/Artefact>,
        <http://dbpedia.org/ontology/Article>,
        dcterms:BibliographicResource,
        bibo:Article,
        swrc:Publication,
        ns2:Article,
        foaf:Document,
        <swc:TalkEvent> ;
    rdfs:label "EXTRAÇÃO DE REGRAS EM UMA BASE DE CONHECIMENTO POR MEIO DA NAVEGAÇÃO EM UM GRAFO" ;
    dcterms:created "2017-10-26" ;
    dcterms:creator <http://localhost:2020/resource/usuario/1075>,
        <http://localhost:2020/resource/usuario/5625>,
        <http://localhost:2020/resource/usuario/6850>,
        <http://localhost:2020/resource/usuario/6854> ;
    dcterms:date "2017-10-26" ;
    dcterms:subject <http://localhost:2020/resource/area_conhecimento/10300007> ;
    dcterms:title "EXTRAÇÃO DE REGRAS EM UMA BASE DE CONHECIMENTO POR MEIO DA NAVEGAÇÃO EM UM GRAFO" ;
    swrc:isAbout <http://localhost:2020/resource/palavra_chave/base_de_conhecimento>,
        <http://localhost:2020/resource/palavra_chave/constru%C3%A7%C3%A3o_de_regras>,
        <http://localhost:2020/resource/palavra_chave/grafo_de_conceitos> ;
    ns2:about <http://localhost:2020/resource/palavra_chave/base_de_conhecimento>,
        <http://localhost:2020/resource/palavra_chave/constru%C3%A7%C3%A3o_de_regras>,
        <http://localhost:2020/resource/palavra_chave/grafo_de_conceitos> ;
    ns2:articleBody "<p style=\"text-align: justify;\"><span style=\"font-weight: 400;\">Bases de conhecimento comum..." ;
    ns2:author <http://localhost:2020/resource/usuario/1075>,
        <http://localhost:2020/resource/usuario/5625>,
        <http://localhost:2020/resource/usuario/6850>,
        <http://localhost:2020/resource/usuario/6854> ;
    ns2:datePublished "2017-10-26" ;
    ns2:description "<p style=\"text-align: justify;\"><span style=\"font-weight: 400;\">Bases de conhecimento comum..." ;
    ns2:genre <http://localhost:2020/resource/area_conhecimento/10300007> ;
    ns2:keywords "base de conhecimento",
        "construção de regras",
        "grafo de conceitos" ;
    ns2:location <http://dbpedia.org/resource/Rio_de_Janeiro> ;
    ns2:locationCreated <http://dbpedia.org/resource/Rio_de_Janeiro> ;
    ns2:name "EXTRAÇÃO DE REGRAS EM UMA BASE DE CONHECIMENTO POR MEIO DA NAVEGAÇÃO EM UM GRAFO" ;
    ns2:text "<p style=\"text-align: justify;\"><span style=\"font-weight: 400;\">Bases de conhecimento comum..." ;
    rdfs:isDefinedBy <http://localhost:2020/data/trabalhoPublicado/4333> ;
    foaf:maker <http://localhost:2020/resource/usuario/1075>,
        <http://localhost:2020/resource/usuario/5625>,
        <http://localhost:2020/resource/usuario/6850>,
        <http://localhost:2020/resource/usuario/6854> ;
    foaf:page <http://localhost:2020/page/trabalhoPublicado/4333> .
```

**(b)**

EXTRAÇÃO DE REGRAS EM UMA BASE DE CONHECIMENTO POR MEIO DA NAVEGAÇÃO EM UM GRAFO

Resource URI: http://localhost:2020/resource/trabalhoPublicado/4333

Home | All trabalho_publicado

| Property | Value |
|---|---|
| schema:about | <http://localhost:2020/resource/palavra_chave/base_de_conhecimento> |
| schema:about | <http://localhost:2020/resource/palavra_chave/constru%C3%A7%C3%A3o_de_regras> |
| schema:about | <http://localhost:2020/resource/palavra_chave/grafo_de_conceitos> |
| schema:articleBody | Bases de conhecimento comum são conjuntos de fatos multipropósito sobre o mundo, alguns de seus usos são no processamento de linguagem natural ou na construção de programas com comportamento inteligente que fazem inferências sobre o mundo. Apesar de geralmente apresentarem um grande conjunto de fatos, para a utilização dessas bases é necessário extrair conhecimento não explícito, sendo empregado técnicas de extração de conhecimento como Aleph [1], que gera um conjunto de hipóteses sobre fatos presentes na base, testando a veracidade dessas hipóteses para os dados armazenados ou  PRA [2], que faz uso da construção de um grafo de relações que são posteriormente transformados em regras.  No presente trabalho propõe-se um método de construção de hipóteses por meio da navegação em um grafo de conceitos interligados (formato comum utilizado por bases de conhecimento). A navegação, ao contrário do PRA, utiliza uma heurística de seleção de caminhos baseada na similaridade semântica dos termos. Para os testes foi utilizado uma base com grafos representando cenas em imagens. Essa base foi inicialmente pré-processada, visando uniformizar o conhecimento contido nela além de corrigir erros que foram encontrados durante o experimento. Como critério de avaliação medimos o precision [3] e o recall [3], comparando os valores obtidos com outros métodos de construção de regras. [1] Srinivasan, Ashwin. "The aleph manual." (2001). NBR 6023      [2] Lao, Ni, Tom Mitchell, and William W. Cohen. "Random walk inference and learning in a large scale knowledge base." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011. NBR 6023 [3] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Vol. 1. No. 1. Cambridge: Cambridge university press, 2008. NBR 6023 |
| schema:author | <http://localhost:2020/resource/usuario/1075> |
| schema:author | <http://localhost:2020/resource/usuario/5625> |
| schema:author | <http://localhost:2020/resource/usuario/6850> |
| schema:author | <http://localhost:2020/resource/usuario/6854> |
| dcterms:created | 2017-10-26 |
| dcterms:creator | <http://localhost:2020/resource/usuario/1075> |

**Figura 4**. Trecho da representação RDF (a) e da representação HTML (b) de uma publicação, geradas pela plataforma D2RQ.

Fonte: Elaborada pelos autores (2018).

fornecida pelo D2RQ, tanto para humanos, em HTML, quanto para máquinas, em RDF e SPARQL, embora a interface tenha manifestado erros (exceções) durante testes com acessos simultâneos. Até esse ponto, alcançou-se o nível quatro estrelas do modelo proposto por Berners-Lee (2009). Com o objetivo de atingir o nível cinco estrelas, o processamento das triplas RDF com a ferramenta de *mashup Silk*, baseado no *workflow* da Figura 3, conseguiu estabelecer *links* owl:sameAs entre as palavras-chave e áreas de conhecimento da conferência e os recursos da DBpedia. De 10.784 palavras-chave e 1.335 áreas de conhecimento cadastradas no sistema, todas mapeadas como instâncias da classe skos:Concept, a ferramenta *Silk* foi capaz de encontrar, automaticamente, 6.923 *links*, realizando, assim, o *Linked Data mashup* de mais de 50% dos recursos processados.

Como exemplo de resultado de *mashup* obtido com o *Silk*, considerando ainda a publicação demonstrada na Figura 4, dentre as suas três palavras-chave ("base de conhecimento", "construção de regras" e "grafos de conceitos") mapeadas como instância de skos:Concept, foi encontrada uma ligação automática para o termo "base de conhecimento" (<http://localhost:2020/resource/palavra_chave/base_de_conhecimento>) com recurso <http://pt.dbpedia.org/resource/Base_de_conhecimento> da DBpedia em português, o qual, por sua vez, está ligado, dentro da própria base da DBpedia, através do predicado owl:sameAs, ao recurso <http://dbpedia.org/resource/Knowledge_base> da DBpedia em inglês. Este último recurso possui mais informações úteis sobre o respectivo conceito. Igualmente, a área de conhecimento da mesma publicação, "Ciência da Computação" (<http://localhost:2020/resource/area_conhecimento/10300007>), foi automaticamente associada ao recurso <http://pt.dbpedia.org/resource/Ciência_da_computação>. Este, por sua vez, está ligado internamente, na DBpedia, ao recurso <http://dbpedia.org/resource/Computer_science>, que descreve as propriedades da área de conhecimento em questão.

Sobre os ensaios com as duas abordagens de publicação de dados explicadas anteriormente, foram executadas consultas SPARQL nos dois cenários, experimentando os dois bancos de dados RDF (Stardog e GraphDB), com seus respectivos raciocinadores de inferência ontológica habilitados. As consultas foram realizadas tanto usando explicitamente a propriedade owl:sameAs (Figura 5 a) quanto usando inferência com base na semântica da propriedade owl:sameAs (Figura 5 b). As Figuras 5 a e 5 b, apresentam os códigos das consultas SPARQL executadas, recuperando todos os artigos da conferência que estão relacionados, seja por palavra-chave ou por área de conhecimento, ao recurso <http://pt.dbpedia.org/resource/Aprendizado> da DBpedia. A Figura 5 c, mostra o resultado obtido com ambas as consultas no GraphDB, demonstrando, assim, a aplicação bem-sucedida do conceito de *Linked Data* à base de dados da conferência. Esse resultado promove a reutilização dos dados da DBpedia, disponibilizando dados inteligíveis por máquinas e os incorporando ao grafo mundial da *Web* de Dados Ligados. O alcance dessa ligação permite, por exemplo, a associação dos artigos da conferência a outros de outras instituições de pesquisa pelo mundo que também possuam recursos de suas bases de dados ligados à DBpedia, fomentando, portanto, o crescimento e a distribuição do conhecimento científico e o apoio à descoberta desse conhecimento.

As duas abordagens de publicação dos dados obtiveram os resultados esperados, sendo que a segunda abordagem, centralizada, sem a necessidade da realização de consulta federada, apresentou melhor desempenho e menor chance de erros de comunicação e de protocolos. É importante mencionar que o banco de dados RDF Stardog, diante da consulta demonstrada na Figura 5 b, não foi capaz de inferir automaticamente a ligação direta, pelas propriedades schema:about e schema:genre, entre as publicações e os recursos da DBpedia, funcionando de forma correta apenas com a consulta exposta na Figura 5 a, com a propriedade owl:sameAs explícita. Não foram encontrados problemas nos demais cenários avaliados.

Foram realizados ainda testes de processamento buscando localizar *links* semânticos entre a base de dados da conferência e a DBpedia, por meio de outro *workflow* do *Silk* que buscasse similaridades entre os conteúdos dos títulos e resumos dos recursos de ambas as bases de dados, estabelecendo ligações do tipo rdfs:seeAlso ("veja também", em português). Todavia, o processamento se mostrou extremamente custoso e demorado, e não

**(a)**

```
SELECT ?publicacao ?titulo_publicacao
{
    ?k owl:sameAs <http://pt.dbpedia.org/resource/Aprendizado> .
    SERVICE <http://localhost:2020/sparql> {
      ?publicacao schema:name ?titulo_publicacao .
      {
        ?publicacao schema:about ?k
      }
      UNION {
        ?publicacao schema:genre ?k
      }
    }
}
```

**(b)**

```
SELECT DISTINCT ?publicacao ?titulo_publicacao
{
      ?publicacao a foaf:Document .
      ?publicacao schema:name ?titulo_publicacao .
      {
          ?publicacao schema:about <http://pt.dbpedia.org/resource/Aprendizado>
      }
      UNION {
          ?publicacao schema:genre <http://pt.dbpedia.org/resource/Aprendizado>
      }
}
```

**(c)**

| | publicacao | titulo_publicacao |
|---|---|---|
| 1 | http://localhost:2020/resource/trabalhoPublicado/2506 | AVALIAÇÃO DO APRENDIZADO DE CEFALOMETRIA COM ALUNOS DO CURSO DE GRADUAÇÃO EM ODONTOLOGIA |
| 2 | http://localhost:2020/resource/trabalhoPublicado/4955 | RELATO DE EXPERIÊNCIA NA OFICINA ABATJOUR |
| 3 | http://localhost:2020/resource/trabalhoPublicado/4545 | INTEGRAÇÃO UNIVERSIDADE-ESCOLA: UM ENCONTRO COM A CIÊNCIA |

**Figura 5**. (a) Consulta SPARQL federada (banco de dados RDF de *mashup* e *endpoint* D2RQ) e sem inferência. (b) Mesma consulta SPARQL, porém não federada (banco de dados RDF centralizado com todas as triplas) e com inferência. (c) Resultados de ambas as consultas.

Fonte: Elaborada pelos autores (2018).

foram encontrados *links* nessa abordagem. Pretende-se realizar uma análise mais aprofundada, visando a uma reestruturação desse *workflow*, em trabalhos futuros.

Halaç *et al.* (2013) utilizam a filosofia *Linked Data* para integrar dados distribuídos entre diferentes sistemas de uma Universidade, propondo uma arquitetura que viabiliza a ligação entre as informações dos bancos de dados desses sistemas, além da criação de uma aplicação para auxiliar nessa integração de dados. O projeto utiliza a plataforma D2RQ para mapear bancos de dados relacionais de alguns dos sistemas, e a ferramenta *Silk Framework* para a descoberta de *links* entre os recursos de diferentes fontes de dados. Com a evolução da pesquisa, optou-se pela alimentação periódica em lote de uma base RDF centralizada, a partir dos bancos de dados dos sistemas em questão, por meio de uma aplicação que realiza a extração de dados, a conversão e a carga de forma automática, com o auxílio de ferramentas como D2RQ, *Silk* e *Triplister* (Rogers, 2011 *apud* Halaç *et al.*, 2013), mantendo os dados ligados atualizados de forma dinâmica.

Segarra *et al.* (2016) propõem uma arquitetura de integração entre repositórios digitais distribuídos, em um modelo *Linked Data* baseado em ontologias consagradas, como *DCTerms, Bibo, Schema* e FOAF, utilizando um enforque virtual através de consultas federadas SPARQL. O modelo proposto foi aplicado às bases de publicações acadêmicas de universidades equatorianas, angariando, desse modo, escalabilidade e aplicabilidade na integração, assim como sua fácil expansão a outros sistemas de informação.

Em Santarém Segundo *et al.* (2017), é feita uma análise de como os padrões e tecnologias da *Web* Semântica contribuem no processo de construção de redes semânticas e na organização de informações, visando prover informações de mais relevância aos usuários. Para tal, é realizado um estudo na colaboração digital de publicações acadêmicas, com foco na plataforma VIVO (*Duraspace, Phoenix, Arizona*, Estados Unidos da América), a qual utiliza tecnologias da *Web* Semântica, como RDF e OWL, para descrever e relacionar recursos. Conclui-se, então, que a *Web* Semântica, com suas tecnologias, aumenta o campo de visão e de relacionamentos de conceitos, oferecendo aos usuários resultados mais ricos do ponto de vista de relações semânticas, sendo a plataforma estudada (VIVO) um bom exemplo do uso dessas tecnologias.

Freitas Junior e Jacyntho (2016) propõem a construção de um protótipo para a catalogação semântica de publicações científicas, seguindo os princípios *Linked Data* da *Web* Semântica. Foi desenvolvida e publicada uma aplicação, na qual foram realizados experimentos com publicações reais. A aplicação desenvolvida disponibiliza uma interface amigável ao usuário, que deve cadastrar os dados manualmente no sistema.

Em comparação com esses outros trabalhos, esta proposta se diferencia por fornecer, além de um método para publicar, de forma sistemática, dados relacionais em grafo RDF, uma seleção de ontologias *Linked Data* bastante completa para o domínio de conhecimento de trabalhos acadêmicos, bem como por oferecer duas abordagens automatizadas de como publicar dados relacionais como *Linked Data*, uma federada e outra centralizada, usando ferramentas bem conhecidas pela comunidade da *Web* Semântica. As ontologias e diretrizes/ferramentas usadas nas duas abordagens podem, perfeitamente, ser empregadas em outros projetos similares.

## Conclusão

Neste trabalho, foi apresentado um método para a publicação de bases de dados relacionais segundo os princípios *Linked Data*, validado por meio de um estudo de caso real de trabalhos acadêmicos. Foi usado, como estudo de caso, o mapeamento de um banco de dados relacional de uma conferência interna de uma universidade federal brasileira para o modelo de dados RDF, utilizando diversas ontologias conhecidas. Além disso, foi feita a geração automatizada de *Linked Data mashup*, visando interligar a base de dados da conferência com o grafo global da chamada *Web* de Dados Ligados, por meio da fonte de dados DBpedia. Diante dos desafios propostos, os resultados obtidos foram satisfatórios, atingindo-se o objetivo de se publicar uma visão *Linked Data* cinco estrelas completa sobre dados relacionais originais, sem precisar alterá-los em absolutamente nada. A publicação dos dados da conferência foi realizada em formato compreensível por máquina, em URIs dereferenciáveis, e foram feitas consultas SPARQL ad hoc envolvendo recursos oriundos da base de dados da conferência e da DBpedia, considerando-se as ligações estabelecidas no *mashup*.

Procurando tornar a proposta ainda mais robusta, outros workflows devem ser desenvolvidos futuramente na ferramenta *Silk*, viabilizando a descoberta de novas ligações automáticas de *mashup*, inclusive com outras fontes de dados *Linked Data*, além da DBpedia. No escopo desse trabalho futuro, poder-se-ia considerar também a possibilidade da aplicação de tradução automática de termos para a língua inglesa, com o objetivo de facilitar o *mashup* com fontes de dados em inglês.

O estudo de caso voltou-se à publicação das informações dos artigos e autores. Uma proposta interessante seria incluir, no mapeamento, informações sobre a conferência em si e a universidade, utilizando as mesmas ontologias já apresentadas nesta pesquisa, e, se necessário, outras complementares. Por fim, é importante aprofundar ainda mais a análise quali-quantitativa dos resultados encontrados, definindo mecanismos para medir a relevância e fidedignidade dos *links* RDF obtidos automaticamente com a ferramenta *Silk*.

## Referências

Abele, A. *et al. Linking open data cloud diagram 2017*. [*S.l.*]: The Linked Open Data Cloud, 2017. Available from: http://lod-cloud.net/. Cited: Dec. 13 2017.

Antoniou, G. *et al. A semantic Web primer*. 3rd. ed. [*S.l.*]: The MIT Press, 2012.

Azevedo, R. S. N.; Jacyntho, M. D. A. Um modelo baseado em ontologias linked data para catalogação de projetos de software. *In*: Conferências Ibero-Americanas WWW/Internet e Computação Aplicada. *Anais* [...]. Porto: IADIS, 2014.

Beckett, D. *et al. RDF 1.1 Turtle*. Cambridge: W3C, 2014. Available from: https://www.w3.org/TR/turtle/. Cited: July 27 2018.

Berners-Lee, T. *et al.* The semantic Web. *Scientific American*, v. 284, n. 5, p. 34-43, 2001.

Berners-Lee, T. *Linked data*. Cambridge: W3C, 2009. Available from: https://www.w3.org/DesignIssues/LinkedData.html. Cited: Nov. 29 2017.

Bilenko, M. *et al.* Adaptive name matching in information integration. *IEEE Intelligent Systems*, v. 18, n. 5, p. 16-23, 2003.

Bizer, C. Cyganiak, R. Gauß T. The RDF book mashup: from Web APIs to a Web of data. *In*: ESWC'07: Workshop on Scripting for the Semantic Web, Innsbruck. *Proceedings* [...]. Innsbruck: CEUR Workshop Proceedings, 2007. Available from: http://ceur-ws.org/Vol-248/paper4.pdf. Cited: Oct. 27 2018.

Bizer, C. *et al.* DBpedia: a crystallization point for the Web of data. *Journal of Web Semantics*, v. 7, n. 3, p. 154-165, 2009.

Bizer, C. *et al. D2RQ*: accessing relational databases as virtual RDF graphs. Berlin: D2RQ. Available from: http://d2rq.org/. Cited: Mar. 30 2018.

Brickley, D. *et al. RDF Schema 1.1*: W3C Recommendation. Cambridge: W3C, 2014. Available from: https://www.w3.org/TR/rdf-schema/. Cited: Nov.12 2018.

Brickley, D.; Miller, L. *FOAF Vocabulary Specification 0.99*. [*S.l., s.n.*], 2014. Available from: http://xmlns.com/foaf/spec/. Cited: Mar. 30 2018.

Camilo, C. O.; Silva, J. C. *Um estudo sobre a interação entre mineração de dados e ontologias*. Goiânia: Universidade Federal de Goiás, 2009. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_002-09.pdf. Acesso em: 1 nov. 2018.

Cyganiak, R. *et al. The D2RQ Mapping language*. Berlin: D2RQ, 2012. Available from: http://d2rq.org/d2rq-language. Cited: Nov. 29 2017.

D'Arcus, B.; Giasson, F. *The Bibliographic ontology*. [*S.l.*]: The Bibliographic Ontology, 2009. Available: http://bibliontology.com/. Cited: Mar 30. 2018.

DBpedia. *DBpedia*: towards a public data infrastructure for a large, multilingual, semantic knowledge graph. Berlin: DBpedia, 2017. Available from: http://dbpedia.org. Cited: June 16, 2018.

DBpedia. *DBpedia mappings*. Berlin: DBpedia, 2010. Available from: http://mappings.dbpedia.org/. Cited: Mar. 30, 2018.

Dublin Core Metadata Initiative. *DCMI Usage Board*. Ohio: DCMI, 2012. Available from: http://dublincore.org/documents/dcmi-terms/. Cited: Mar. 30, 2018.

Freitas Junior, N.; Jacynto, M. D. A. Um protótipo linked data para catalogação semântica de publicações. Perspectivas em *Ciência da Informação*, v. 21, n. 4, p. 48-65, 2016.

Gandon, F.; Schreiber, G. *RDF 1.1 XML Syntax*. Cambridge: W3C. Available from: https://www.w3.org/TR/rdf-syntax-grammar/. Cited: Mar. 27 2018.

Gruber, T. R. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, v. 43, p. 907-928, 1995.

Halaç, T. G. *et al.* Publishing and linking university data considering the dynamism of datasources. *In*: International Conference on Semantic Systems, 9th, 2013, Graz. Proceedings [...]. Graz: ACM, 2013. Available from: http://dl.acm.org/citation.cfm?id=2506182.2506202. Cited: Mar. 27 2018.

Heath, T.; Bizer, C. *Linked Data*: evolving the *Web* into a global data space. San Raphael, CA: Morgan & Claypool, 2011. v. 1.

Horrocks, I. *et al. Web ontology language*. Cambridge: W3C, 2012. Available from: https://www.w3.org/OWL/. Cited: Nov. 10 2018.

Isele, R. *et al. Silk*: the linked data integration framework. Mannheim: Silk, 2009. Available from: http://silkframework.org/. Cited: Mar. 30 2018.

Isele, R. *et al. Silk server*: adding missing links while consuming linked data. *In*: First International Workshop on Consuming Linked Data, 2010, Shanghai. *Proceedings* [...]. Shanghai: CEUR Workshop Proceedings, 2010. Available from: http://ceur-ws.org/Vol-665/IseleEtAl_COLD2010.pdf. Cited: Mar. 30 2018.

Isele, R.; Bizer, C. Active learning of expressive linkage rules using genetic programming. *Journal of Web Semantics*, v. 23, p. 2-15, 2013.

Jacyntho, M. D. A. *Um modelo de bloqueio multigranular para RDF. 277*. Tese (Doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, 2012.

Jacyntho, M. D. A.; Azevedo, R. S. N. Uma Arquitetura *Linked Data* para criação de repositórios semânticos auto-atualizáveis de

projetos de software. *In*: Encontro Mineiro de Engenharia de Produção, 6, 2015. *Anais* [...]. São João da Barra: EMEPRO, 2015.

Jacyntho, M. D. A.; Schwabe, D. A multigranularity locking model for RDF. *Journal of Web Semantics*, v. 39, p. 25-46, 2016.

Jaenicke, N. *et al.* Triplify: documentation. *Leipzig*: Triplify, 2010. Available from: https://web.archive.org/web/20150208025440/http://triplify.org:80/Documentation. Cited: Nov. 11 2018.

Miles, A.; Bechhofer, S. *SKOS Simple Knowledge Organization System Namespace Document*: HTML variant. Cambridge: W3C, 2009. Available from: https://www.w3.org/2009/08/skos-reference/skos.html. Cited: Nov. 30 2017.

Nuzzolese, A. G. *et al.* Semantic Web conference ontology: a refactoring solution. In: European Semantic Web Conference, 2016, Heraklion, Greece. *Proceedings* [...]. Heraklion: Springer, 2016. Available from: http://link.springer.com/10.1007/978-3-319-47602-5_18. Cited: Nov. 30 2017.

Object Management Group. *Unified modeling language*. Needham: OMG, c1997. Available from: http://www.uml.org/. Cited: Nov. 12 2018.

Ontotext. *GraphDB*. Sofia: Ontotext, 2018. Available from: http://graphdb.ontotext.com/. Cited: Mar. 30 2018.

Oracle Corporation. *MySQL community edition*. Redwood City: Oracle, [201-]. Available from: https://www.mysql.com/products/community/. Cited: Mar. 30 2018.

Petrovski, P. *et al.* Integrating product data from Websites offering microdata markup. International Conference on World Wide Web: WWW'14 Companion, 23. *Proceedings* [...]. Seoul: ACM, 2014. Available from: http://dl.acm.org/citation.cfm?doid=2567948.2579704. Cited: Nov. 30 2017.

RDF Working Group. *Resource Description Framework*. Cambridge: W3C, 2014. Available from: https://www.w3.org/RDF/. Cited: Nov. 12 2018.

Rogers, D. Tripliser. [*S.l.*]: Dave Rog, 2011. Available from: http://daverog.github.io/tripliser/. Cited: Mar. 30 2018. – Autor, Referência não encontrada no texto.

Santarém Segundo, J. E. *et al.* Conceitos e tecnologias da *Web* semântica no contexto da colaboração acadêmico-científica: um estudo da plataforma Vivo. *Transinformação*, v. 29, n. 3, p. 297-309, 2017.

Schema.org Community Group. *Schema.org*. [*S.l.*]: Schema, [2011]. Avaialable from: http://schema.org/. Cited: Mar. 30 2018.

Segarra, J. *et al.* Integration of digital repositories through federated queries using semantic technologies. *In*: Latin American Computing Conference, 43, 2016. *Proceedings* [...]. Valparaíso: IEEE Xplore, 2016. Available from: http://ieeexplore.ieee.org/document/7833406/. Cited: June 10 2018.

Souza, A. N. *A Web semântica na definição de um modelo de dados estruturado para apoiar pesquisas no Lago Batata (Oriximiná/PA)*. Dissertação (Mestrado em Engenharia de Produção e Sistemas Computacionais) – Universidade Federal Fluminense, Rio das Ostras, 2016.

SPARQL Working Group. *SPARQL 1.1 overview*. Cambridge: W3C. Available from: https://www.w3.org/TR/sparql11-overview/. Cited: Dec. 13 2017.

Sporny, M. *et al. JSON-LD 1.0*: A JSON-based serialization for linked data. Cambridge: W3C, 2014. Available from: https://www.w3.org/TR/json-ld/. Cited: Mar. 27 2018.

Stardog Union. *Stardog*. Arlington: Stardog, 2018. Available from: https://www.stardog.com/. Cited: Mar. 30, 2018.

Styles, R. *et al. Academic Institution Internal Structure Ontology*. [*S.l.*]: AIISO, 2008. Available from: http://vocab.org/aiiso/. Cited: Mar. 30, 2018.

Sure, Y. *et al.* The SWRC ontology: semantic web for research communities. *In*: Portuguese Conference on Artificial Intelligence EPIA, 12, 2005, Covilhã, Portugal. *Proceedings* [...] Covilha: Springer, 2005. Available from: https://link.springer.com/chapter/10.1007/11595014_22. Cited: Nov. 30 2017.

Vandenbussche, P. Y. *et al.* Linked open vocabularies: a gateway to reusable semantic vocabularies on the web. *Semantic Web*, v. 8, n. 3, p. 437-452, 2016.

W3C. *About W3C*. Cambridge: W3C, [201-]. Available from: https://www.w3.org/Consortium/. Cited: Mar. 30 2018.

W3C OWL Working Group. *OWL 2 web ontology language document overview*. 2nd. ed. Cambridge: W3C, 2012. Available from: https://www.w3.org/TR/owl2-overview/>. Cited: Nov. 23 2017.

Yandex. *About Yandex*. Moscow: Yandex, c1997. Available from: https://yandex.com/company/. Cited: Mar. 30 2018.

Yu, L. *A Developer's guide to the semantic web*. Berlin: Springer Science & Business Media, 2011.