*Open*SBBD: Usando *Linked Data* para Publicação de Dados Abertos sobre o SBBD

Mateus Gondim Romão Batista, Bernadette Farias Lóscio

Centro de Informática, Universidade Federal de Pernambuco, Brasil
{mgrb, bfl}@cin.ufpe.br

Abstract. The Semantic Web was designed to expand the current Web we know, enabling the large amount of data available on the Web to be processed not only by humans, but also by machines. Along this process, the Semantic Web gathered new technologies which allow the modeling and description of data embebbed with semantic definitions. In addition to the adoption of these technologies, the concept of Linked Data emerged. Linked Data is a set of principles and techniques created with the purpose of interlinking data from distinct data sources. On the other hand, the Open Data Movement has encouraged data publishing on the Web, without restrictions from copyright, patents or other mechanisms of control. The combination of Semantic Web technologies, the principles of Linked Data and the Open Data movement has played a major role in data publishing. In this context, this work proposes: I) the creation of an interlinked open dataset, available in RDF and consistent with the principles of Linked Data, about the 27 editions of the Brazilian Symposium on Databases; II) the creation of a SPARQL Endpoint available to be queried using SPARQL; III) The development of a Web application with the purpose of providing more user-friendly views about this dataset.

Resumo. A Web Semântica foi projetada para expandir a Web que conhecemos atualmente, possibilitando que a imensa quantidade de dados disponíveis na Web possa ser compreendida não só por pessoas, mas também por máquinas. Neste processo, a Web Semântica agregou novas tecnologias que permitem a modelagem e descrição de dados com definições semânticas associadas. Em conjunto com a adoção dessas tecnologias, surgiu o conceito de *Linked Data*, um conjunto de princípios e técnicas para publicação de dados estruturados na Web, criado com o objetivo de possibilitar a interligação de dados de fontes de dados distintas. Por outro lado, o movimento de Dados Abertos (Open *Data*) tem incentivado a disponibilização de dados na Web sem restrições de copyright, patentes e outros mecanismos de controle. A combinação das tecnologias da Web Semântica, os princípios de *Linked Data* e o movimento de Dados Abertos tem desempenhado um papel importante na publicação de dados na Web. Nesse contexto, este artigo propõe: i) a criação de um conjunto de dados abertos, disponível em RDF e de acordo com os princípios de *Linked Data*, sobre as 27 edições do Simpósio Brasileiro de Banco de Dados; ii) a disponibilização de um *SPARQL Endpoint* para a execução de consultas SPARQL sobre os dados e metadados do SBBD; iii) o desenvolvimento de uma aplicação Web com o intuito de oferecer visões mais amigáveis destes dados.

Keywords: Semantic Web, Linked Data, Open Data

1. INTRODUÇÃO

A Web Semântica é a extensão da *World Wide Web* que permite às pessoas compartilharem conteúdo além dos limites de aplicações e *websites* [10]. Encorajando a inclusão de conteúdo semântico em páginas Web, a

Web Semântica visa converter a Web atual dominada por documentos estruturados e semi-estruturados em uma "Web de dados". Para tornar a Web Semântica uma realidade, uma série de novas tecnologias foram adotadas, como RDF, OWL e XML. Conectado com estes novos atores da Web, surgiu um conjunto de princípios e tecnologias chamado *Linked Data*, que de uma maneira geral, refere-se a empregar o RDF e o *Hypertext Transfer Protocol* (HTTP) para publicar dados estruturados na Web, interligando dados de diferentes fontes de dados [4]. O movimento de Dados Abertos (Open Data), que incentiva a disponibilização de dados a todos, também contribui para a disponibilização de conjuntos de dados que podem ser interligados no formato *Linked Data*.

Seguindo estes conceitos, este trabalho aborda o problema de criação de uma base de dados e metadados sobre o Simpósio Brasileiro de Banco de Dados (SBBD), o maior evento na América Latina para a apresentação e discussão de resultados de pesquisa relacionados à área de Banco de Dados [11]. A cada edição do SBBD são gerados diversos dados sobre a comunidade brasileira de Banco de Dados, os quais podem ser de grande utilidade para a obtenção de informações relacionadas ao perfil dos participantes, às áreas de pesquisa, às instituições participantes, dentre outras. Entretanto, na maioria das vezes, estes dados são disponibilizados em documentos textuais ou até mesmo apenas de forma impressa, não permitindo o seu processamento direto por algum aplicativo. Dessa forma, este trabalho tem como objetivo a criação de um conjunto de dados abertos, denominado *Open*SBBD, com os dados mais relevantes sobre todas as 27 edições já realizadas do SBBD. Os dados do *Open*SBBD são estruturados de acordo com o modelo RDF seguindo os princípios de *Linked Data*.

Como principais contribuições deste trabalho, destacam-se: i) A criação de um conjunto de dados seguindo os princípios de *Linked Data* com alto potencial para ser reutilizado por futuros trabalhos, em conjunto com a criação de uma ontologia, denominada SBC*Event*, com novos termos sobre o domínio de conferências e com reuso de vocabulários existentes; ii) Disponibilização de um SPARQL *Endpoint* para a realização de consultas SPARQL sobre o conjunto de dados criado e iii) Criação de visualizações dos dados do SBBD em formatos amigáveis, como gráficos e tabelas. O restante do artigo está estruturado da seguinte forma. A Seção 2 aborda a contextualização deste trabalho. A Seção 3 descreve a criação da ontologia SBC*Event*. A Seção 4 aborda a criação do conjunto de dados *Open*SBBD. A Seção 5 descreve a criação de um SPARQL *Endpoint* e o desenvolvimento de uma aplicação para a visualização dos dados do *Open*SBBD. Por fim, a Seção 6 expõe a conclusão deste trabalho, além de sugestões para projetos futuros.

2. CONTEXTUALIZAÇÃO

Nesta seção, apresentamos alguns conceitos importantes para o entendimento do restante do trabalho.

2.1 Dados Abertos

O termo Dados Abertos refere-se a dados que podem ser livremente usados, reusados e distribuídos por qualquer pessoa – sujeito apenas, e no máximo, ao requisito de atribuição e compartilhamento pela mesma licença [7]. Os princípios básicos da abertura de dados são [8]: i) *Disponibilidade e acesso*: os dados devem estar disponíveis como um todo, preferencialmente via download na internet; ii) *Reuso e redistribuição:* os dados devem ser fornecidos sob termos que permitam reuso e redistribuição, incluindo a interligação com outros conjuntos de dados e suporte a processamento por máquinas e iii) *Participação universal:* qualquer pessoa deve ser capaz de usar, reusar e redistribuir os dados.

Os principais formatos utilizados para publicar dados abertos são JSON, XML, CSV e RDF. Este último é o formato recomendado pelo W3C[1], e tem atraído grande interesse daqueles que publicam dados de forma

[1] http://www.w3.org/

aberta. Um exemplo disso é o grande número de conjuntos de dados publicados em *Linked Data* atualmente disponíveis no portal de dados governamentais abertos do Reino Unido[2]. A principal motivação é a facilidade de combinar dados neste formato com múltiplas fontes de dados, interligando-se a outras iniciativas *Open Data* na Web [6].

2.2 Considerações de projeto de conjuntos de dados *Linked Data*

Na publicação de dados em *Linked Data*, existem aspectos que devem ser levados em conta para que o projeto possa tornar-se eficiente e alcançar seus objetivos. Dentre esses aspectos, destacaremos dois: a escolha do formato das URIs (*Uniform Resource Identifier*) e o reuso de vocabulários. A escolha do formato das URIs, identificadores únicos de recursos na Web, é um ponto que merece destaque em um projeto de um conjunto de dados em *Linked Data*. As principais recomendações nesta escolha são: i) Utilizar URIs HTTP, pois o esquema "http://" é o único esquema de URIs amplamente suportado pelas infraestrutura dos dias atuais; ii) Definir URIs em um *namespace* HTTP sob controle, onde se pode implementar o que for necessário para torná-las dereferenciáveis, ou seja, prover conteúdo quando as URIs forem acessadas e iii) Tentar manter as URIs estáveis e persistentes, considerando que trocar as URIs em um momento posterior irá quebrar todos os *links* existentes.

Outro elemento importante no que diz respeito à criação de um conjunto de dados em *Linked Data* são os vocabulários. Vocabulários são descrições de conceitos de um domínio (geral ou específico), materializadas em um conjunto de termos. Os termos de um vocabulário, juntamente com os seus relacionamentos, podem ser descritos por meio de ontologias e seus componentes (classes e propriedades), os quais, assim como os outros recursos, possuem uma URI, o que possibilita o seu reuso. De uma maneira geral, os vocabulários são usados para a descrição dos metadados de um conjunto de dados. Com o objetivo de tornar a tarefa de processamento de dados o mais simples possível, o reuso de termos de vocabulários conhecidos e estabelecidos como um padrão é fortemente recomendado. É importante notar que novos termos devem ser criados apenas no caso dos vocabulários já existentes não fornecerem os termos desejados [3]. É uma prática comum incluir termos de diferentes vocabulários na criação de uma nova ontologia para descrever todos os conceitos relacionados a um conjunto específico de dados.

3. ONTOLOGIA SBC*EVENT*

O primeiro passo para a criação do conjunto de dados *Open*SBBD foi a criação da ontologia que representasse os termos usados para a descrição dos dados do SBBD. Para a criação desta ontologia, inicialmente, foram avaliados quais conceitos referentes às conferências promovidas pela SBC (Sociedade Brasileira de Computação) seriam importantes para uma modelagem coerente do domínio. Desta forma, a ontologia poderia ser utilizada para representar não só o SBBD, mas também outros simpósios promovidos pela SBC. Durante a criação da ontologia, duas etapas principais foram realizadas, como descrito a seguir.

- *Análise e reuso de vocabulários:* após a definição dos principais conceitos da ontologia, buscou-se analisar vocabulários conhecidos e estáveis com o objetivo de reaproveitar o maior número de termos possível. Consultando tanto vocabulários de caráter geral como vocabulários com um foco específico no mundo acadêmico (eventos, publicações, etc.), foi possível reutilizar 13 classes e 17

[2] http://www.data.gov.uk

propriedades. Os vocabulários reusados foram os seguintes: FOAF[3], DUBLIN CORE[4], EVENT[5], GEONAMES[6], BIBO[7], AKT[8], SWC[9], SWRC[10].

- *Criação de novos termos:* Apesar dos vocabulários existentes cobrirem uma ampla parte dos termos necessários para a criação da ontologia SBC*Event*, ainda foi necessário definir alguns novos termos, já que em alguns casos não foi encontrado nenhum termo existente refletindo a semântica do conceito que se desejava representar. Desta forma, um conjunto de termos foi criado com o objetivo de complementar o esquema da ontologia.

Utilizando o *Protégé*[11], a ontologia foi criada em OWL. Todos os novos termos foram ligados a outros vocabulários, utilizando *tags* de RDF *Schema*. Uma classe criada que merece destaque é a *sbc:Participation*. Esta classe foi criada para representar um relacionamento entre três conceitos: *Pessoa*, *Organização* e *Conferência* (equivalente a um relacionamento ternário em modelagens Entidade Relacionamento). Alguns vocabulários possuem propriedades que ligam uma organização a uma pessoa, e uma pessoa a uma conferência. No entanto, a afiliação de participantes em simpósios não é definitiva, ou seja, ao longo de diferentes edições do SBBD, autores, por exemplo, podem estar afiliados a diferentes universidades. Desta forma, a modelagem da ontologia deveria oferecer suporte à representação deste tipo de situação de forma consistente. Seguindo uma recomendação do W3C [5], uma classe (*sbc:Participation*) foi criada para representar essa relação ternária. Com isso, uma instância desta classe representa uma instância da relação entre indivíduos das três classes relacionadas.

4. CRIAÇÃO DO CONJUNTO DE DADOS *OPEN*SBBD

A criação do conjunto de dados *Open*SBBD baseou-se em duas linhas de ação: a conversão de dados em bases relacionais para o modelo RDF e a extração manual de dados dos *websites* e anais das edições passadas do SBBD. Para o mapeamento entre o modelo relacional e o modelo RDF, foi utilizada a plataforma D2RQ[12], que fornece um ambiente integrado com múltiplas opções para acessar dados relacionais como grafos RDF virtuais de leitura. Entre os métodos de acesso possíveis, destacam-se os Dumps RDF(*RDF/XML*), as APIs RDF (para aplicações Java) e um SPARQL *Endpoint* [1]. A base relacional utilizada é proveniente de um trabalho sobre uma rede de coautoria baseado em artigos publicados no SBBD, publicado no próprio SBBD em 2010[9]. Para adequá-la a este trabalho, foram realizadas algumas modificações. Um exemplo foi a padronização de valores literais referentes ao sexo de uma pessoa. No banco original os valores eram "M" e "F", mas depois da observação de outros conjuntos de dados, concluiu-se que o padrão mais largamente utilizado é "*male"* e "*female"*.

[3] http://xmlns.com/foaf/spec/
[4] http://purl.org/dc/terms/
[5] http://purl.org/NET/c4dm/event.owl#
[6] http://www.geonames.org/ontology/ontology_v3.1.rdf
[7] http://purl.org/ontology/bibo/
[8] http://www.aktors.org/publications/ontology/portal#
[9] http://data.semanticweb.org/ns/swc/ontology#
[10] http://ontoware.org/swrc/swrc/SWRCOWL/swrc_updated_v0.7.1.owl#
[11] http://protege.stanford.edu/
[12] http://d2rq.org/

Um dos componentes da plataforma D2RQ é o *generate-mapping*. A partir da análise do esquema do banco de dados, esta ferramenta cria um arquivo de mapeamentos no formato *Turtle*[13], no qual é possível configurar os mapeamentos de tabelas e colunas do banco para classes e propriedades de ontologias. Assim, os mapeamentos foram definidos utilizando tanto os termos criados como os termos reusados, previamente definidos, além do mapeamento para dados de outros conjuntos de dados, como o GeoNames, interligando os respectivos conjuntos de dados. A partir deste arquivo, um dump RDF do banco de dados relacional foi gerado, utilizando a serialização *RDF/XML*, através de outro componente da plataforma D2RQ, a ferramenta *dump-rdf*. Após executar o *dump-rdf*, um *dump* RDF foi criado de acordo com o esquema definido na Seção 3. Além destes dados, outros dados foram adicionados ao conjunto de dados manualmente. A partir de uma análise dos *websites* e anais dos eventos passados, foram criados *scripts* SQL para a inserção desses dados no banco que, por sua vez, foram convertidos para RDF através do D2RQ.

5. BUSCA E VISUALIZAÇÃO DOS DADOS DO *OPEN*SBBD

Para facilitar a busca dos dados disponíveis no *Open*SBBD, foi utilizada uma ferramenta chamada *D2R-Server*, da plataforma D2RQ, que permite a publicação do conteúdo de bases relacionais na Web Semântica. Além disso, disponibiliza um SPARQL *Endpoint* para a submissão de consultas SPARQL. A partir desta interface, foram realizadas consultas explorando as classes e propriedades definidas na ontologia. Como exemplo, temos a consulta *Q1*: *Quais instituições de ensino já publicaram no SBBD e quantos autores publicaram por cada uma delas (ordem descendente)?* A consulta SPARQL correspondente é mostrada na Figura 1 e os dez primeiros resultados são apresentados na Figura 2.

Para a visualização de dados foi desenvolvida uma aplicação com o objetivo de fornecer informações relevantes sobre o SBBD, por meio de gráficos e tabelas. As principais tecnologias utilizadas foram a linguagem de programação *Python*[14], e o framework para desenvolvimento Web *Django*[15], e a API *SPARQL Endpoint interface to Python*[16], utilizada para submeter consultas SPARQL ao *Endpoint* do *D2R-Server*. As figuras 3 e 4 ilustram exemplos de gráficos gerados na aplicação a partir de consultas realizadas. O esquema final da ontologia, o conjunto de dados e o SPARQL *Endpoint* podem ser acessados através do link http://www.cin.ufpe.br/~mgrb/opensbbd/.

```
SELECT DISTINCT ?name (COUNT(?person) as ?authors) WHERE {
  ?institution a akt:Learning-Centred-Organization.
  ?institution foaf:name ?name.
  ?person a foaf:Person.
  ?article dcterms:creator ?person.
  ?participation sbbd:participationPerson ?person.
  ?participation sbbd:participationOrganization ?institution.
  ?participation sbbd:participationConference ?conference.
  ?article bibo:presentedAt ?conference.
}
GROUP BY ?institution ?name
ORDER BY DESC(?authors)
```

Fig. 1. - Consulta SPARQL Q1

?name	?authors
Universidade Federal do Rio Grande do Sul	100
Universidade Federal do Rio de Janeiro	76
Pontifícia Universidade Católica (RJ)	71
Universidade Federal de Pernambuco	64
Universidade Estadual de Campinas	47
Universidade Federal da Paraíba	45
Universidade Federal de Minas Gerais	43
Universidade de São Paulo	28
Instituto Militar de Engenharia	26
Universidade Federal de São Carlos	13

Fig. 2. - Resultado da Consulta Q1

[13] http://www.w3.org/TR/turtle/
[14] http://www.python.org/
[15] https://www.djangoproject.com/
[16] http://sparql-wrapper.sourceforge.net/

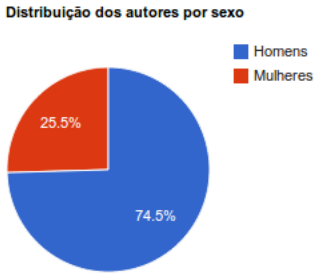

Fig. 3. Gráfico de distribuição de autores por sexo

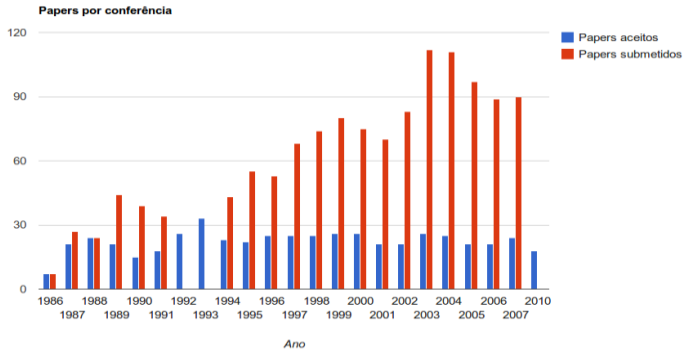

Fig. 4. Gráfico de artigos por conferência ao longo das edições do SBBD

6. CONCLUSÃO

Este trabalho dividiu-se essencialmente em três partes: a criação de um conjunto de dados em RDF, interligado com grandes conjuntos de dados conhecidos; a publicação deste conjunto de dados em um *SPARQL Endpoint*, disponível para a execução de consultas; o desenvolvimento de uma aplicação Web para a visualização de gráficos e tabelas gerados a partir do conjunto de dados criado. Com este conjunto de dados sobre o SBBD, as informações das edições passadas deste grande evento passam a fazer parte da Web de dados. Além disso, aplicações relacionadas ao SBBD agora têm a possibilidade de utilizar este conjunto como uma de suas fontes de dados. Este material também pode incentivar futuros trabalhos na área de Web Semântica e *Linked Data* sobre eventos da SBC, ou de forma mais genérica, sobre eventos acadêmicos em geral. Entre linhas de ações futuras, podemos destacar as seguintes: i) Ampliar o conjunto de dados com mais dados que não foram coletados durante a realização deste trabalho; ii) Mudança de ferramenta para disponibilização do *SPARQL Endpoint*, uma vez que um dos pré-requisitos para utilizar o *SPARQL Endpoint* oferecido pelo *D2R-Server* é a existência de uma base relacional, para que os mapeamentos sejam feitos em tempo real; iii) Enriquecimento da aplicação de visualização de dados.

REFERÊNCIAS

[1] BIZER, C.; CYGANIAK, R. "D2RQ-lessons learned." *W3C Workshop on RDF Access to Relational Databases*. 2007.

[2] BIZER, C.; CYGANIAK, R.; HEATH, T. "How to publish linked data on the web." (2008).

[3] BIZER, C.; HEATH, T.; BERNERS-LEE, T. "Linked data-the story so far." *International Journal on Semantic Web and Information Systems (IJSWIS)* 5.3 (2009): 1-22.

[4] BIZER, C.; HEATH, T.; IDEHEN, K.; BERNERS-LEE, T. Abril, 2008. Linked data on the web (LDOW2008). In *Proceeding of the 17th international conference on World Wide Web* (pp. 1265-1266). ACM.

[5] NOY, N.; et al. "Defining n-ary relations on the semantic web." *W3C Working Group Note* 12 (2006): 4.

[6] OPEN DATA HANDBOOK. File Formats, 2013. Disponível em <http://opendatahandbook.org/en/appendices/file-formats>. Acesso em: 10 jul. 2013.

[7] OPEN DEFINITION, 2013. Disponível em: <http://opendefinition.org/>. Acesso em: 4 jul. 2013.

[8] OPEN KNOWLEDGE FOUNDATION, 2013. Disponível em: <http://okfn.org/opendata/>. Acesso em: 4 jul. 2013.

[9] PROCÓPIO, P. S.; LAENDER, A. H. F.; MORO, M. M. "Análise da rede de coautoria do simpósio brasileiro de bancos de dados." SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, Florianópolis, 2011. Proceedings... Florianópolis (2011).

[10] SEMANTIC WEB WIKI, 2013. Disponível em: <http://semanticweb.org/wiki/Main_Page>. Acesso em: 9 jan. 2013.

[11] SIMPÓSIO BRASILEIRO DE BANCO DE DADOS. SBBD, 2012. Disponível em: <http://sws2012.ime.usp.br/sbbd/>. Acesso em: 25 jan. 2013.