Título:

Uma Ferramenta para suportar a publicação de visões RDF de dados relacionais

Nome do Aluno:

Luís Eufrasio Teixeira Neto

Nome da Orientadora:

Vânia Maria Ponte Vidal

Nível:

Mestrado

Nome do Programa de pós-graduação:

Mestrado e Doutorado em Ciência da Computação

Nome da Universidade:

Universidade Federal do Ceará

E-mail de contato do aluno:

luiseufrasio@lia.ufc.br

E-mail de contato da orientadora:

vvidal@lia.ufc.br

Ano de ingresso no programa:

2010

Data da defesa da proposta:

05/05/2011

Época esperada de conclusão:

29/02/2012

Etapas já concluídas:

Proposta Defendida

Etapas futuras:

Defesa de Dissertação

Resumo em português:

Neste artigo propomos uma ferramenta web para publicação de bases relacionais na forma de ontologias RDF. Para tanto, utilizamos a linguagem de mapeamento adotada pela w3c: r2rml (RDB to RDF Mapping Language), que será suportada pela nossa solução. Assim, o usuário terá uma interface amigável para criar suas ontologias e mapeamentos. Também exploramos as funcionalidades já implementadas pela ferramenta D2R e identificamos em quais cenários é mais vantajoso virtualizar ou materializar as visões geradas.

Palavras-chave:

Linked Data, Web Semântica, RDF, R2RML, D2R Server, Publicação de Dados

Uma Ferramenta para suportar a publicação de visões RDF de dados relacionais

Vânia Maria P. Vidal[1], Luís Eufrasio T. Neto[1]

[1]Departamento de Computação – Universidade Federal do Ceará (UFC)
Qd Campus Pici, 1 – Pici – 60455-900 – Fortaleza – CE – Brazil
{vvidal,luiseufrasio}@lia.ufc.br

Abstract. *In This paper we propose a web tool for publishing relational databases in the form of RDF ontologies. For this purpose, we use the mapping language adopted by w3c: R2RML (RDB to RDF Mapping Language), which is supported by our solution. Thus, the user will have a friendly interface to create ontologies and their mappings. We also explore the features already implemented by the tool D2R and identify scenarios in which it is more advantageous to virtualize or materialize the visions generated.*

Resumo. *Neste artigo propomos uma ferramenta web para publicação de bases relacionais na forma de ontologias RDF. Para tanto, utilizamos a linguagem de mapeamento adotada pela w3c: R2RML (RDB to RDF Mapping Language), que será suportada pela nossa solução. Assim, o usuário terá uma interface amigável para criar suas ontologias e mapeamentos. Também exploramos as funcionalidades já implementadas pela ferramenta D2R e identificamos em quais cenários é mais vantajoso virtualizar ou materializar as visões geradas.*

1. Introdução

O padrão Linked Data foi estabelecido como a melhor prática para expor, compartilhar e conectar partes de dados, informações e conhecimentos na Web Semântica usando URIs e RDF. No entanto, a grande maioria dos dados corporativos, inclusive dados da Web, permanecem armazenados em SGBDs relacionais. Para tornar as informações armazenadas em bancos relacionais disponíveis para a Web de dados, é preciso publicá-las no modelo RDF. Essa publicação pode ser materializada (triplificação dos dados) ou Virtual o que é feito por ferramentas como o D2R Server através da criação de endpoints SPARQL/RDF.

Para atender ao requisito da web semântica de publicação dos dados usando um vocabulário comum que permite melhor reuso e a integração da base com outras fontes RDF, criam-se visões RDF utilizando vocabulários de uma ontologia de domínio e um mapeamento da ontologia local para a ontologia de domínio. Em seguida, é preciso traduzir esse mapeamento para um formato que possa ser interpretado por uma ferramenta de publicação de dados RDF. Dessa forma, adotamos a linguagem padrão R2RML para criação desses mapeamentos.

Neste trabalho, tratamos do problema de gerar automaticamente tais arquivos de mapeamento. Para tanto, propomos uma ferramenta web que disponibiliza para o usuário a criação da configuração de acesso às bases de dados, a geração da ontologia local, a criação da ontologia de aplicação usando um vocabulário que é subconjunto do vocabulário de ontologia de domínio, a criação do mapeamento de alto nível entre as

ontologias local e de aplicação e a geração automática do arquivo de mapeamento r2rml a partir de regras de mapeamento de alto nível.

O restante deste artigo está organizado da seguinte forma: a seção 2 apresenta trabalhos relacionados e o diferencial da nossa solução. A seção 3 contextualiza o problema definindo nosso foco dentro de um problema maior no qual esse trabalho está inserido. A seção 4 apresenta o D2R Server e introduz a linguagem de geração de mapeamentos padrão da W3C: R2RML. A seção 5 apresenta a arquitetura da ferramenta, seus componentes principais e o estado atual do trabalho. A seção 6 conclui expondo oportunidades de trabalhos futuros que foram gerados com essa pesquisa.

2. Trabalhos Relacionados

Existem algumas ferramentas disponíveis já na web com a finalidade semelhante à nossa. Dentre elas podemos citar: Fuseki e Virtuoso. Podemos citar como diferencial da nossa solução: a nossa ferramenta é a única Web, suporta o padrão W3C R2RML e busca resolver problemas de performance nas consultas SPARQL geradas.

3. Contextualização do Problema

Para publicarmos fontes de dados na Web de acordo com os princípios do Linked Data, é necessário seguir uma sequência de passos como definido abaixo:

1. Primeiro criamos uma *source ontology* S = (V_S , C_S), que modela os dados a serem publicados.
2. Então selecionamos uma *domain ontology* D = (V_D , C_D), que modela o domínio da aplicação. De fato, D pode ser uma combinação de ontologias cobrindo domínios distintos.
3. Dando procedimento, criamos um mapeamento *source-to-domain* γ de S para D (note que γ pode não conter definições para todos os símbolos em V_D) e geramos uma *exported (application) ontology*.
4. Publicação e Registro da exported ontology. Os dados podem ser exportados de duas formas: Enfoque virtual onde é criado um SPARQL end-point ou materializado, que exporta e mantém os dados em RDF.

Nesse trabalho são abordados os problemas relativos ao último passo utilizando o enfoque virtual adotado pelo D2R para publicar os dados. No entanto, é necessário encontrar formas de traduzir os diversos mapeamentos γ para a linguagem de mapeamento padrão r2rml. A definição dessa tradução e sua implementação são as principais contribuições desse trabalho, juntamente com a definição de estratégia para tomada de decisão sobre qual tipo de visão deve ser criada: virtual ou materializada.

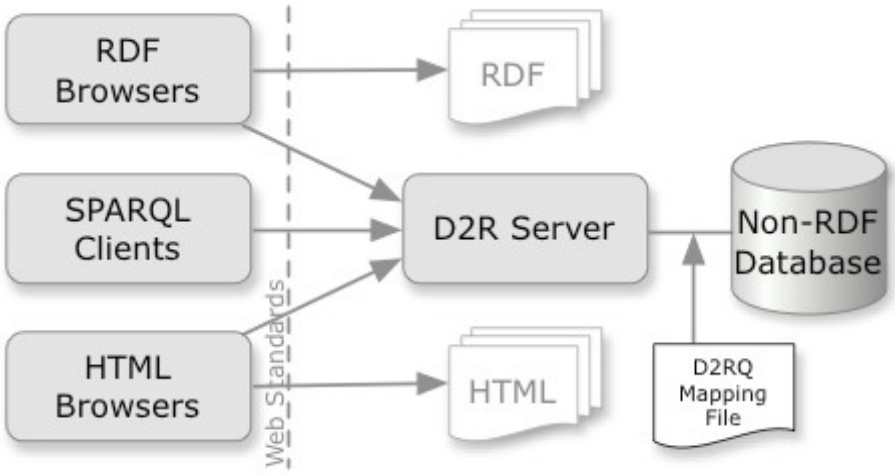
4. Fundamentação Teórica

4.1. Introdução ao D2R Server

O D2R Server é uma ferramenta para publicar o conteúdo de bases de dados relacionais na Web Semântica, um espaço de informação global consistindo de dados ligados.

Ele pode ser acessado de 3 formas diferentes conforme mostra a figura abaixo:

Figura 1. Formas de acesso ao D2R Server



Sua interface Linked Data torna as descrições RDF de recursos individuais disponíveis no protocolo HTTP, a interface SPARQL torna possível que aplicações pesquisem e consultem o banco de dados usando a linguagem de consulta SPARQL sobre o protocolo SPARQL e uma interface HTML que oferece acesso a navegadores Web conhecidos.

Requisições oriundas da Web são reescritas em consultas SQL usando o mapeamento. Essa tradução “on-the-fly” permite a publicação de grandes bases de dados no formato RDF e elimina a necessidade de replicação desses dados.

Nesse trabalho tentamos reutilizar as funcionalidades já implementadas pelo D2R para criação do Endpoint SPARQL.

4.2. Introdução ao R2RML (RDB to RDF Mapping Language)

Uma linguagem para expressar mapeamentos customizados de bancos de dados relacionais para datasets RDF. Tais mapeamentos possibilitam visualizar os dados relacionais no modelo de dados RDF, expressos na estrutura e vocabulário escolhidos pelo autor do mapeamento.

Cada mapeamento R2RML é adaptado para um esquema de banco de dados específico e para um vocabulário alvo. A entrada para um mapeamento R2RML é um banco de dados relacional que está de acordo com esse esquema. A saída é um dataset RDF [SPARQL], conforme definido no SPARQL, que usa predicados e tipos do vocabulário alvo. O mapeamento é conceitual; processadores R2RML são livres para materializar os dados de saída, ou para disponibilizá-los de forma virtual através de uma interface que consulta o banco de dados de origem, ou para oferecer qualquer outra forma de acesso ao dataset RDF de saída.

Um R2RML mapping define um mapeamento a partir de um banco de dados relacional para um dataset RDF. Esse mapeamento é representado como um grafo RDF. Em outras palavras, RDF é usado não apenas como o modelo de dados alvo do mapeamento, mas também como um formalismo para representar o mapeamento R2RML. Um documento de mapeamento R2RML é qualquer documento escrito na sintaxe Turtle [tartaruga] RDF.

Um mapeamento R2RML consiste em uma ou mais estruturas chamadas TriplesMaps. Cada TriplesMap contém uma referência a uma tabela lógica no banco de

dados de entrada. A tabela lógica pode ser um dos itens seguintes:

1. A tabela base que existe no esquema SQL de entrada.
2. Uma visão que existe no esquema SQL de entrada.
3. Uma consulta SQL válida no esquema de entrada.

Como o R2RML tornou-se recentemente a linguagem padrão adotada pela W3C, resolvemos gerar nossos mapeamentos nessa notação.

5. A Ferramenta

Nessa seção introduzimos as principais funcionalidades da ferramenta web e sua integração com o D2R para criação de endpoits SPARQL a partir dos mapeamentos criados. Atualmente o D2R não suporta R2RML, porém possui sua própria linguagem de mapeamento conhecida como D2RM. Portanto haverá um esforço adicional de traduzir mapeamentos R2RML para D2RM. A seguir definimos a arquitetura da solução e seus principais componentes.

5.1. Arquitetura Proposta

Com o intuito de aproveitar as várias soluções já existentes, resolvemos criar a camada de apresentação utilizando a ferramenta GWT do Google juntamente com o framework JENA para criação de ontologias de domínio e dos arquivos de mapeamento R2RML. Além do D2R para acesso às bases relacionais e criação dos endpoints SPARQL.

A figura abaixo descreve de forma gráfica a arquitetura escolhida e as dependências entre seus componentes:

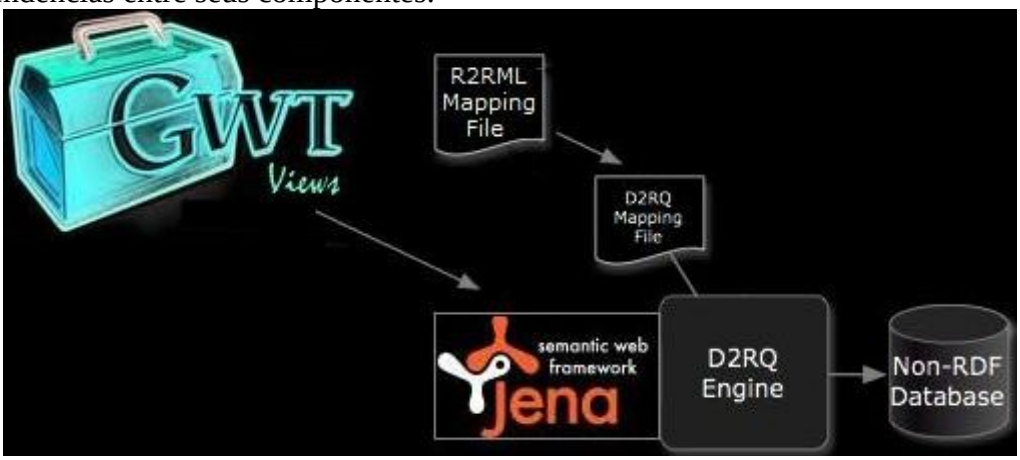


Figura 2. A arquitetura escolhida

5.2. Componentes Principais

Uma das principais contribuições desse trabalho é tornar fácil para o usuário exportar seus dados relacionais em RDF através de uma interface amigável que esconde a complexidade de geração e manutenção dos mapeamentos.

Para tornar isso possível, é preciso criar mecanismos para:

1. Criação de Ontologias de Domínio;
2. Configuração de acesso ao banco de dados via D2R;

3. Geração de mapeamentos R2RML;
4. Tradução de R2RML para D2R Mapping Language;
5. Criação do Endpoint SPARQL via D2R Server;

Cada um desses componentes possui seus algoritmos específicos para implementar o mecanismo ao qual ele está proposto a atender. Destacamos aqui o componente de geração dos mapeamentos R2RML que soluciona como representar as possíveis regras de mapeamento nesse formato, e o componente de tradução de R2RML para D2RM, que define um DE-PARA complexo para tratar tal transformação.

5.3. Estado Atual

Com a arquitetura definida, o trabalho encontra-se na fase de especificação dos algoritmos a serem implementados.

6. Trabalhos Futuros

Um vez que a solução abordada até aqui preocupa-se com a simples publicação dos dados relacionais em RDF de forma virtual. Como propostas futuras, pensamos em realizar testes para medir a performance dessas publicações e tentar definir uma estratégia de como publicar os dados ou até mesmo parte deles. Ou seja, queremos descobrir quais subconjuntos dos dados devem ser virtualizados e quais devem ser materializados. Dessa forma definiríamos uma solução híbrida que atenda a requisitos não funcionais de performance nas consultas aos Endpoints SPARQL.

7. Referências

Percy E. Salas, Karin K. Breitman, Marco A. Casanova, José Viterbo: StdTrip: An a priori design approach and process for publishing Open Government Data.

Sacramento, E., Vidal, V. M., Macêdo, J. A., Lóscio, B. F., Lopes, F. L. R., Casanova, M. A., and Lemos, F.: Towards automatic generation of application ontologies. In: Journal of Information and Data Management, Vol. V, No. N, Month 20YY, pp.1-16, 2010.

Bizer, C., Cyganiak, R.: D2R server – publishing relational databases on the Semantic Web. Disponível em: <http://www4.wiwiss.fu-berlin.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf>.

Jena - A Semantic Web Framework for Java. Disponível em: <http://jena.sourceforge.net>

Das Souripriya, Sundara Seema, Cyganiak Richard (2011), R2RML: RDB to RDF Mapping Language. Disponível em <http://www.w3.org/TR/r2rml/>