Semantic Data Services

Uma abordagem para leitura e atualização de dados semânticos

Hermano Albuquerque Lira
Serviço Federal de Processamento de Dados
Av. Pontes Vieira, 832
Fortaleza, CE, Brasil
hermano.lira@serpro.gov.br

Pedro Porfírio Muniz Farias
Universidade de Fortaleza
Av. Washington Soares, 1321, J-30
Fortaleza, CE, Brasil
porfirio@unifor.br

ABSTRACT

Nos últimos anos a quantidade de dados disponibilizados na web e aderentes aos princípios dos dados interligados (*Linked data*) cresceu bastante. Esse crescimento foi impulsionado pelo projeto de dados abertos interligados (*Linking Open Data Project*[1]), além de iniciativas de vários países para publicar dados do setor público. No entanto, apesar desse crescimento, os serviços provedores de dados interligados ainda carecem de padronização dos métodos de acesso e manipulação de dados. Este trabalho visa definir uma abordagem, intitulada *Semantic Data Services*, para a construção de serviços de dados estruturados de acordo com o modelo RDF, cuja finalidade é apresentar uma padronização para serviços que possua mecanismos de leitura e atualização de dados.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services

General Terms

Padronização, Serviço de Dados Semânticos

Keywords

serviço de dados, RDF

1. CONTEXTO TEÓRICO

Segundo Tim Berners-Lee, a web semântica (também chamada de web de dados) é uma evolução da web atual [6]. Ela visa fornecer estruturas e atribuir semântica aos dados disponíveis na web, criando um ambiente onde máquinas e humanos possam trabalhar de forma cooperativa. De acordo com o W3C (*World Wide Web Consortium*)[2], para tornar a web de dados uma realidade, é importante que uma grande quantidade de dados esteja disponível na web em um modelo padrão, processável por máquina e gerenciável mediante as ferramentas da web semântica[3].

O modelo padrão proposto pela W3C é o RDF (*Resource Description Framework*) [11], um modelo aderente aos princípios dos dados interligados (*Linked data* [7]), cuja semântica é codificada em conjuntos de triplas sujeito-predicado-objeto.

Apesar do aumento na quantidade de dados disponíveis na web alinhados aos princípios dos dados interligados, uma parte expressiva ainda está armazenada em bases de dados com outros modelos e formatos (i.e. relacionais, XML, HTML, etc). A disponibilização destes dados na web de dados depende não apenas da existência de um modelo comum de representação dos dados, mas também de serviços capazes de fazer a conversão entre modelos e prover acesso de leitura e escrita sobre estas bases. Os serviços de dados são serviços web que funcionam como uma camada de acesso às bases de dados, tornando os dados acessíveis e fornecendo os meios para a sua publicação e atualização.

Neste trabalho, é proposta a abordagem *Semantic Data Services* (SDS), uma padronização para a criação de serviços de dados descritos por interfaces SERIN (*Semantic RESTful Interface*) [12] e capazes de manipular dados no modelo RDF por meio do protocolo HTTP.

A interface SERIN é uma ontologia escrita em *Web Ontology Language* (OWL), cujas classes recebem anotações. Todos os dados providos por um SDS são instâncias das classes definidas na ontologia, como ilustrado na figura 1, e as anotações têm o papel de descrever quais operações do protocolo HTTP são permitidas sobre estes dados. Portanto, as classes anotadas e suas propriedades caracterizam semanticamente os dados disponíveis na base de dados e, assim, qualquer provedor que deseje aderir a essa interface, deve disponibilizar recursos de acordo com o descrito na ontologia.

A abordagem SDS facilita a integração de dados, pois utiliza o conceito de interface para disponibilizar uma visão unificada dos dados residentes em diferentes bases e define um método de acesso e manipulação para estes dados. Esta abordagem torna-se significativa em uma variedade de situações, que incluem tanto aplicações comerciais (quando uma empresa precisa cotar preços em seus fornecedores) quanto científicas (combinação de resultados de pesquisa de difer-

[1] `http://linkeddata.org/`
[2] `www.w3.org`
[3] `http://www.w3.org/standards/semanticweb/data`

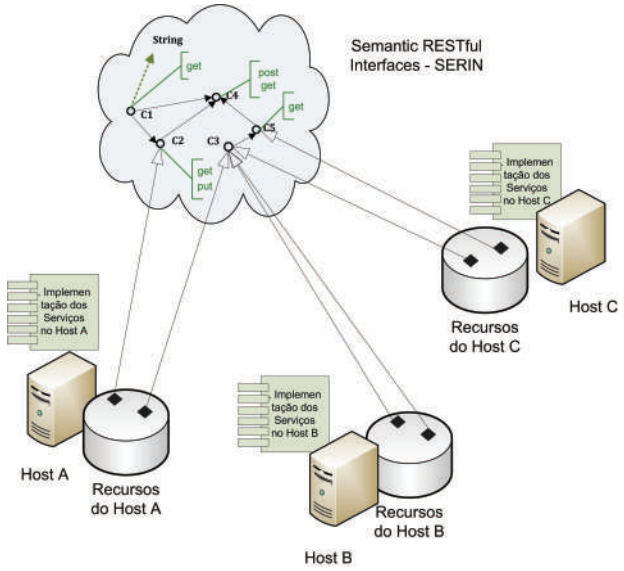

Figure 1: Os recursos apresentados em cada *host* são instâncias de classes e podem ser acessados segundo serviços de dados que adotam a interface semântica especificada

entes repositórios de bioinformática, por exemplo).

2. IDENTIFICAÇÃO DO PROBLEMA

A web é um vasto repositório de dados e uma quantidade crescente desses dados é gerada e armazenada em bases de dados todos os anos [8]. A publicação dessas bases na web é de suma importância em várias situações, por exemplo:

- quando a academia publica suas pesquisas para que outros pesquisadores possam colaborar com a construção do conhecimento;
- quando o governo publica seus dados para dar publicidade e transparência de seus atos aos cidadãos;
- quando as empresas publicam seus dados para que os acionistas possam acompanhar seus investimentos.

Os serviços de dados proveem visões de bases de dados que estão, provavelmente, em constante mudança. No entanto, tais serviços carecem de uma padronização quanto à sintaxe e quanto à semântica dos dados publicados. No aspecto sintático, é possível encontrar serviços cuja representação do formato dos dados está codificada em XML[4], JSON[5], ATOM[6], entre outros. No aspecto semântico a heterogeneidade é ainda maior, já que cada provedor de dados define seus próprios modelos de representação.

Com uma incontável quantidade de fontes de dados, a web claramente impõe importantes desafios para a gestão de dados e descoberta de conhecimento. Neste cenário, verifica-se a dificuldade de integração dos dados, em virtude das heterogeneidades sintáticas e semânticas. A construção de serviços de dados que estejam alinhados aos princípios dos dados interligados pode contribuir para a integração dos dados, por prover aos usuários uma visão única, uniforme e homogênea.

[4] *Extensible Markup Language* (XML)
disponível em: `http://www.w3.org/XML`
[5] *JavaScript Object Notation* (JSON)
disponível em: `http://www.json.org`
[6] *The Atom Syndication Format*
disponível em: `http://www.ietf.org/rfc/rfc4287.txt`

3. OBJETIVOS

Este trabalho tem como objetivo geral definir uma abordagem para a construção de serviços de dados semânticos. Tais serviços são alinhados aos princípios dos dados interligados e ao estilo arquitetural REST [10].

A partir do objetivo geral, pretende-se alcançar os seguintes objetivos específicos:

- Estender a especificação SERIN, adicionando novas anotações para prover a verificação de restrições de integridade de dados;
- Estender a especificação SERIN, adicionando suporte a *queries* URI [5] com capacidade para ordenar e filtrar dados baseado em qualquer critério, navegar entre relacionamentos e paginar dados.
- Implementar suporte a controle de concorrência e a transações baseado no padrão de projeto *try-confirm-cancel* [14];

4. METODOLOGIA ADOTADA

Inicialmente foi feito o levantamento do referencial teórico sobre os assuntos a serem tratados neste trabalho. Mediante pesquisas na literatura disponível foram determinados os trabalhos relacionados, a fim de possibilitar a identificação do estado da arte e do problema de pesquisa. Em seguida foi determinado o objetivo e as contribuições esperadas do trabalho.

Após a definição do objetivo de pesquisa foi feita a verificação da viabilidade técnica do trabalho por meio da implementação de um protótipo. A linguagem Java foi escolhida para a implementação por ser de uso geral, madura, possuir uma plataforma robusta para o desenvolvimento de serviços web e dispor de uma ampla variedade de *frameworks* que implementam as especificações dos padrões da web semântica.

Implementado o protótipo inicial do serviço de dados semânticos, a arquitetura da solução será refinada para incorporar os objetivos específicos deste trabalho, como as novas anotações para a interface SERIN, o controle de concorrência e o suporte às transações.

Como cenário de uso, pretende-se publicar dados públicos governamentais em formato RDF. Será utilizado o catálogo de repositórios de dados públicos disponível no portal brasileiro de dados abertos[7].

Por fim, será utilizada a solução implementada para validar os objetivos deste trabalho. A partir do cenário de uso citado acima, será avaliado o alinhamento do serviço de dados ao

[7] `www.dados.gov.br`

estilo arquitetural REST, o conformidade da estrutura de dados aos princípios dos dados interligados e a adesão da solução aos padrões da web semântica. Também será feita uma comparação entre a solução proposta neste trabalho com trabalhos relacionados, a fim de destacar as melhorias desta solução em relação aos trabalhos anteriores. Assim, os resultados obtidos nesse estudo de caso possibilitarão a conclusão do trabalho.

5. ESTÁGIO ATUAL DO TRABALHO

A revisão da literatura e um estudo detalhado da especificação SERIN e seus conceitos relacionados foram concluídos, possibilitando um alinhamento de ideias em torno da criação dos serviços de dados semânticos. Uma versão inicial do *framework* de serviços de dados semânticos foi implementada e está disponível *online*[8]. Nessa versão inicial só estão implementadas as anotações relativas às operações HTTP (GET, PUT, POST e DELETE). Uma segunda versão está em andamento a cerca de mecanismos de verificação de integridade de dados, bem como mecanismos para suporte a consultas via *queries* URI, ambos a serem incorporados na especificação SERIN.

6. TRABALHOS RELACIONADOS

As pesquisas sobre serviços de dados permeiam várias áreas de estudos, como gestão do conhecimento, a integração de dados, a web semântica e os serviços web. Nesta seção serão analisados importantes trabalhos feitos nessas áreas sobre os serviços de dados [3][13][9][15]. A tabela 1 apresenta um estudo comparativo entre os aspectos analisados neste trabalho.

De acordo com Auer [4], a forma mais comum de publicar recursos na web de dados segue o modelo RDF e usa URIs para identificação de recursos. O acesso a esses recursos pode se dar de três formas diferentes, como ilustrado na figura 2:

- Acesso por consulta, o que significa que o agente envia uma consulta SPARQL para um *endpoint* e processa o resultado da consulta devolvido pelo serviço;
- Acesso em nível de entidade, o que significa que o agente executa um HTTP GET em uma URI identificadora de um recurso RDF e processa o resultado (normalmente o resultado é um grafo RDF que representa uma instância e suas propriedades);
- Acesso por *Dump*, o que significa que o agente executa um HTTP GET e obtêm todo o grafo RDF como resultado (e.g. processos de Extração, Transformação e Carga - ETL[9]).

O *Open Data Protocol* (OData) [3] é um protocolo de serviço de dados construído sobre as tecnologias HTTP, AtomPub [1] e JSON. Dentre as principais características do OData, também presentes no SDS, estão a independência de fontes de dados, o alinhamento ao REST, a existência de uma interface descritora de metadados (*Conceptual Schema Definition Language* - CSDL [2]) e a definição de uma linguagem

[8] `www.activeontology.com.br`

[9] Sigla em inglês para *Extract, Transform and Load.*

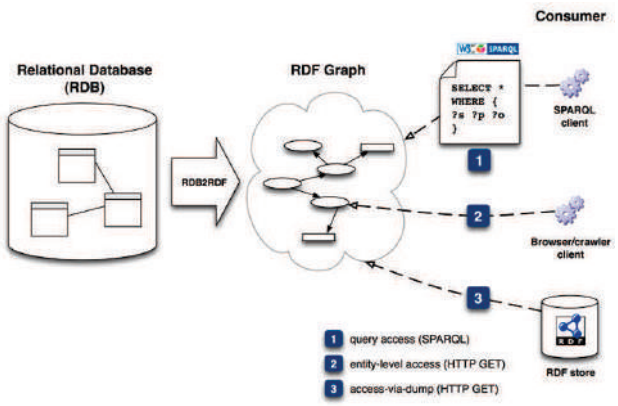

Figure 2: Tipos de acesso a serviços de dados voltados para a web de dados. (Esta figura foi originalmente publicada pelo *RDB2RDF Working Group* em [4].)

de consulta via *queries* URI. Entretanto, o modelo de dados do OData, chamado de EDM (*Entity Data Model*) [2], e seu protocolo de comunicação (AtomPub) não possuem suporte aos padrões da web semântica. Atualmente, os dois formatos de serialização de dados suportados pelo OData são o JSON e o Atom, ambos não aderentes aos princípios dos dados interligados.

O *SPARQL 1.1 Graph Store HTTP Protocol* [13] é uma especificação proposta pelo W3C que descreve um protocolo de nível de aplicação para atualização e recuperação de grafos RDF em um *Graph Store* via HTTP. Este protocolo especifica, portanto, a semântica das operações HTTP para o gerenciamento de um *Graph Store*. Em particular, ele fornece operações para a remoção e criação de grafos RDF, além de substituição e inserção de triplas ao conteúdo desses grafos. Porém, apesar de adotar o protocolo HTTP, o princípios do estilo arquitetural REST não são totalmente implementados. Uma das razões para isso, é que as URIs que identificam os grafos frequentemente não são desreferenciáveis[10]. Esta abordagem não apresenta uma linguagem de consulta que permita filtrar, ordenar ou paginar seus dados, nem possui uma interface de metadados associada ao serviço.

Enquanto o *SPARQL 1.1 Graph Store HTTP Protocol* foi concebido com uma forte ênfase na gestão de repositórios *Graph Store*, uma outra especificação do W3C, *SPARQL 1.1 Protocol* [9], foi elaborada para o foco na manipulação dos dados propriamente ditos. Contudo, o *SPARQL 1.1 Protocol* utiliza uma abordagem diferente tanto do SDS como do OData, pois enquanto o primeiro é um serviço de manipulação de dados via consultas SPARQL, os últimos são serviços de dados orientados a entidades, isto é, o OData representa seu dados como entidades EDM e o SDS como entidades (ou indivíduos RDF). O *SPARQL 1.1 Protocol*, portanto, descreve os meios para transmitir consultas SPARQL de recuperação e atualização de dados para um serviço processar e retornar os resultados da consulta via HTTP. No

[10] Os agentes podem usar uma URI para acessar o recurso referenciado, o que é chamado de desreferenciação de URI.

Table 1: Comparação entre especificações de serviços de dados

Possui suporte a:	REST	Linked Data	Escrita	Tipo de Acesso
Semantic Data Services	Sim	Sim	Sim	Em nível de entidade
Open Data Protocol (OData)	Sim	Não	Sim	Em nível de entidade
Graph Store HTTP Protocol	Parcial	Sim	Sim	Por *Dump*
SPARQL *Protocol*	Parcial	Sim	Sim	Por consulta
Linked Data Services (LIDS)	Parcial	Sim	Não	Em nível de entidade

entanto, apesar esta abordagem possuir uma linguagem de consulta bastante expressiva para filtros e ordenação de dados, ela carece de interfaces descritoras de metadados, o que a torna menos adequada para a automação por máquinas.

A abordagem *Linked Data Services* (LIDS) [15] é uma padronização para serviços web compatível com os princípios dos dados interligados e que retorna dados RDF via HTTP. LIDS é baseado em padrões estabelecidos da web, incluindo HTTP, RDF e SPARQL. Diferentemente das abordagens anteriores, cujos serviços atuam como uma camada de abstração sobre fontes de dados, a abordagem LIDS oferece uma abstração sobre serviços web subjacentes. Assim, a proposta de LIDS é a criação de envoltórios (*Wrappers*) sobre serviços web pré-existentes (e.g. twitter, facebook, etc). Em razão disto, o acesso aos dados é indireto e consequentemente LIDS não oferece suporte a escrita de dados.

7. CONTRIBUIÇÕES ESPERADAS

As contribuições esperadas deste trabalho são: a construção de serviços, voltados para a manipulação de dados em RDF mediante o alinhamento com os princípios do estilo arquitetural REST e alinhamento com os princípios dos dados interligados; oferecer uma interface de acesso a dados na web que seja aderente aos padrões recomendados pelo W3C para web semântica (i.e. RDF, OWL); contribuir para a realização da visão da web de dados por meio de uma solução que trabalhe com dados em formato processável por máquinas para, assim, melhorar a interoperabilidade e a integração de dados; fornecer uma solução, segundo uma arquitetura bem definida, para o gerenciamento dos dados e que forneça uma camada de abstração que esconda a complexidade do acesso as fontes de dados subjacentes; promover o uso de ontologias OWL como interfaces de serviço de dados; incentivar as empresas, organizações e indivíduos a publicar seus dados livremente, em um formato padrão e aberto.

8. AGRADECIMENTOS

Os autores agradecem ao SERPRO, Serviço Federal de Processamento de Dados, pelo apoio financeiro, em forma de bolsa, para a pesquisa.

9. REFERENCES

[1] RFC 5023, Atom Publishing Protocol (AtomPub). Request For Comments (RFC), Oct. 2007.

[2] Conceptual Schema Definition File Format V3. `http://download.microsoft.com/download/B/0/B/B0B199DB-41E6-400F-90CD-C350D0C14A53/[MC-CSDL].pdf`, July 2012. [Disponivel online; Accessado em: 26 jul. 2013].

[3] Open Data Protocol (OData) Specification V3. `http://download.microsoft.com/download/9/5/E/95EF66AF-9026-4BB0-A41D-A4F81802D92C/[MS-ODATA].pdf`, Jan. 2013. [Disponivel online; Accessado em: 26 jul. 2013].

[4] S. Auer, L. Feigenbaum, D. Miranker, A. Fogarolli, and J. Sequeda. Use Cases and Requirements for Mapping Relational Databases to RDF. `http://www.w3.org/TR/rdb2rdf-ucr/`, June 2010. [Disponivel online; Accessado em: 26 jul. 2013].

[5] T. Berners-Lee, R. Fielding, and L. Masinter. RFC 3986, Uniform Resource Identifier (URI): Generic Syntax. Request For Comments (RFC), 2005.

[6] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34 – 43, May 2001.

[7] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, Mar 2009.

[8] K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. Structured databases on the web: Observations and implications. *ACM SIGMOD Record*, 33(3):61–70, 2004.

[9] L. Feigenbaum, G. T. Williams, K. G. Clark, and E. Torres, editors. *SPARQL 1.1 Protocol*. W3C Recommendation. World Wide Web Consortium, Mar. 2013.

[10] R. T. Fielding. *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, Irvine, 2000.

[11] F. Manola and E. Miller, editors. *RDF Primer*. W3C Recommendation. World Wide Web Consortium, Feb. 2004.

[12] B. D. A. Muniz, L. M. Chaves, J. C. C. Neto, J. R. Villela, and P. P. M. Farias. SERIN - SEMANTIC RESTFUL INTERFACES. In *Proceedings of the IADIS International Conference on WWW/Internet*, pages 463–467, Rio de Janeiro, Brazil, 2011. IADIS Press.

[13] C. Ogbuji, editor. *SPARQL 1.1 Graph Store HTTP Protocol*. W3C Recommendation. World Wide Web Consortium, Mar. 2013.

[14] G. Pardon and C. Pautasso. Towards distributed atomic transactions over RESTful services. In E. Wilde and C. Pautasso, editors, *REST: From Research to Practice*, pages 507–524. Springer New York, Jan. 2011.

[15] S. Speiser and A. Harth. Taking the LIDS off data silos. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 44:1—-44:4, New York, NY, USA, 2010. ACM.