Um Framework para a Publicação de Dados Abertos Governamentais a partir de Bases de Dados Relacionais

Clayton Martins Pereira
Mestrando em Engenharia Eletrônica e Computação
Instituto Tecnológico de Aeronáutica, Brasil
clayton@ita.br

José Maria Parente de Oliveira
Professor da Divisão de Ciência da Computação
Instituto Tecnológico de Aeronáutica, Brasil
parente@ita.br

ABSTRACT
This work aims to define a framework, based on Semantic Data Layer of DIGO architecture for open government data, for the purpose of automating the generation and publication of open government data (semantic data) obtained from structured data maintained in Relational Databases.

RESUMO
Este trabalho visa definir um *framework*, baseado na camada de dados semânticos (*Semantic Data Layer*) da arquitetura de dados abertos governamentais DIGO, com a finalidade de automatizar os processos de geração e publicação de dados abertos governamentais (dados semânticos) obtidos a partir de dados estruturados mantidos em Bases de Dados Relacionais.

Categories and Subject Descriptors
D.3.3 [**Programming Languages**]: Language Constructs and Features – frameworks, modules, packages.

General Terms
Management, Design, Experimentation, Languages.

PALAVRAS-CHAVE
Dados Abertos Governamentais, *Resource Description Framework*, Ontologias, Bases de Dados Relacionais.

INFORMAÇÕES
Categoria: Mestrado
Universidade: Instituto Tecnológico de Aeronáutica
Programa: Pós-Graduação em Engenharia Eletrônica e Computação – Área Informática
Início: 2011.2
Previsão de Defesa: 2012.2

1. CONTEXTO TEÓRICO
A *Web* Semântica representa um novo campo de pesquisa e desenvolvimento na área da Tecnologia da Informação, que tem por objetivo trazer estrutura para o conteúdo significativo de páginas da *Web*, dado que o conteúdo disponibilizado na *Web* atualmente, de uma forma geral, é destinado apenas à leitura humana, não possibilitando que este seja reutilizado ou manipulado por computadores e seus *softwares*. Com a *Web* Semântica, pretende-se que as informações a serem disponibilizadas na *Web* sejam estruturadas, com significados bem definidos, de forma a habilitar que computadores, tais como *desktops* e dispositivos portáteis, sejam capazes de processar e entender automaticamente estes dados, os quais hoje são apenas mostrados em tela ou impressos [1].

Dados estruturados são aqueles organizados segundo um padrão rígido e predefinido, respeitando diversos critérios como campos (ou atributos) de dados, extensão do campo, domínio (valores possíveis) do dado, tipo de dado, etc. Este é o caso, por exemplo, dos dados mantidos em tabelas de Bases de Dados Relacionais (BDR ou *RDB – Relational DataBases*), usados pelos sistemas de informação na maioria das instituições [2].

O *RDF* (*Resource Description Framework*) é um modelo que permite esta representação estruturada de dados e informações, onde significados são codificados em conjuntos de triplas objeto-atributo-valor, chamadas de declarações, que podem ser comparadas a uma oração onde temos: sujeito, predicado e objeto. O próprio usuário define sua terminologia através de uma linguagem chamada *RDF Schema* (que não tem qualquer relação com o *XML Schema*). Nela é possível definir um vocabulário, as propriedades de espécies de objetos e os valores que podem assumir, bem como descrever relacionamento entre estes [3].

As Ontologias, como artefato de engenharia, consistem em estruturas formais de conceitos e relações entre conceitos, além de um conjunto de axiomas que restringe a interpretação dessa estrutura e proporciona a derivação de conhecimento do conhecimento factual representado na estrutura. Através de Linguagens Ontológicas os usuários podem gravar conceituações explícitas e formais de modelos de domínio. Os principais requisitos de uma linguagem ontológica são: a sintaxe bem definida, o suporte a raciocínio eficiente, uma semântica formal, poder expressivo suficiente e conveniência de expressão. As linguagens ontológicas mais utilizadas atualmente são a *OWL* (*Ontology Web Language*) e a *RDFS* (*Resource Description Framework Schema*) [4].

O termo "dados abertos" é definido pelo *World Wide Web Consortium - W3C* como a publicação de dados em seu formato bruto, com semântica embutida, de forma que possam ser interpretados por máquina dentro do contexto semântico para os quais foram definidos, permitindo assim seu reuso por outras aplicações. Isso é possível através da disponibilização na *Web* de *datasets* na forma de triplas *RDF*, acessíveis por meio de

Identificadores Únicos de Recursos (*URI*, similar a uma *URL HTTP*), e consultáveis por uma *Query Language* (*SparQL*, considerada a linguagem *SQL* da *Web* semântica) [5].

Dados Abertos Governamentais referem-se à disponibilização de dados em formato aberto por órgãos e entidades governamentais, de maneira que possam ser prontamente publicados e acessados por todos os interessados, além de permitir seu reuso por outras aplicações [2]. Atualmente, as instituições governamentais publicam parte dos seus dados em Portais *Web*, utilizando as linguagens e tecnologias da *Web* atual, o que não oferece facilidades para reutilização desses dados na geração de novas informações. Sendo assim, o acesso à informação relevante, precisa e passível de reutilização por outros aplicativos, torna-se cada vez mais complexo.

Um modelo proposto para permitir a publicação de dados abertos governamentais e a correspondente forma de acesso a esses dados é apresentado em [2]. A Figura 1 apresenta a arquitetura geral de apoio ao modelo, denominada DIGO - Disponibilizando Informações de Governo. Ela é um modelo comum e tem por objetivo estipular um acordo semântico entre fontes heterogêneas de dados e permitir a fusão de dados de Portais *Web*. Ela está dividida em cinco camadas. A Figura 1 ilustra as suas camadas. A Camada 1 é denominada Geração de dados e Conhecimento. A Camada 2 é denominada Dados Sintáticos. A Camada 3 é denominada Dados Semânticos. A Camada 4 é denominada Fusão de dados. A Camada 5 é a camada de Informação. As linhas tracejadas em preto delimitam as camadas.

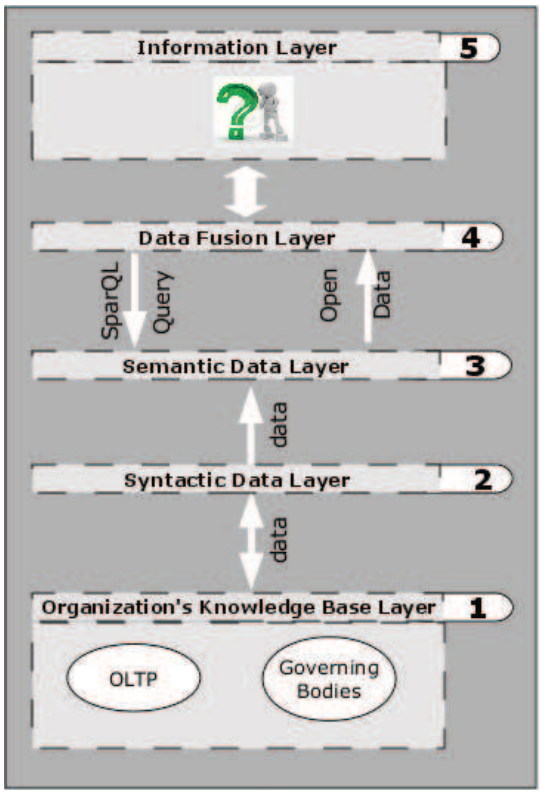

Figura 1. Camadas da arquitetura DIGO. Fonte: [2].

O escopo deste trabalho está em alguns elementos da Camada 3 da arquitetura (dados estruturados). A Camada 3 é a camada de disponibilização dos dados abertos e é responsável por disponibilizar os dados contendo a semântica embutida neles, mantendo o contexto semântico para o qual foram definidos em sua fonte originadora. Ela é composta por quatro subcamadas, apresentadas na Figura 2: Camada de Extração de dados, Camada de Transformação e Carga, Camada de Persistência dos Dados Semânticos e Camada de Manipulação de dados.

2. IDENTIFICAÇÃO DO PROBLEMA

Os processos de geração e publicação de dados abertos semânticos (em formato de triplas *RDF*) a partir de bases de dados relacionais (dados estruturados) contam atualmente com uma série de ferramentas de *software*, grande parte delas produtos de projetos de pesquisa acadêmica (vinculadas a trabalhos de mestrado ou de doutorado), as quais, no entanto, ainda exigem um elevado nível de conhecimento técnico e de interação manual do usuário para sua instalação, configuração e operação. Além disso, algumas dessas ferramentas possuem limitações de compatibilidade com alguns dos diversos sistemas gerenciadores de banco de dados (SGBDs) e sistemas operacionais atualmente encontrados no mercado, ou ainda implementam linguagens específicas para o processo de mapeamento entre ontologias ou vocabulários e as tabelas e colunas das bases de dados relacionais, o que dificulta o suporte e as eventuais necessidades de alterações ou correções manuais nesse mapeamento.

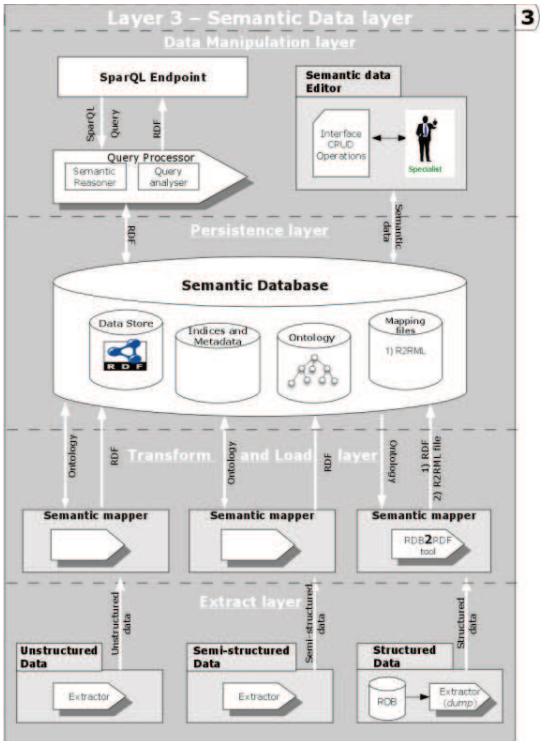

Figura 2. Visão detalhada da camada 3 da arquitetura DIGO e suas subcamadas. Fonte: [2].

Neste cenário, verifica-se a carência de *frameworks*/*toolkits* para automatizar as tarefas de extração, mapeamento e conversão de dados estruturados (em bases de dados relacionais) para dados abertos semânticos (geração de triplas *RDF*), de armazenamento (em *open datasets*) e de manipulação (publicação e consulta) das triplas geradas, de forma a oferecer uma solução única que integre métodos e ferramentas de *software* já desenvolvidas para cada

uma dessas tarefas, através de uma interface amigável para o usuário, que seja multiplataforma e compatível com os principais SGBDs disponíveis no mercado.

Com a entrada em vigor, em maio de 2012, da Lei de Acesso à Informação[1], que regula o acesso às informações sob a guarda de órgãos e entidades públicas de todos os poderes e entes federativos, todos estes deverão, ao divulgarem suas informações, "utilizar todos os meios e instrumentos legítimos de que dispuserem, sendo obrigatória a divulgação em sítios oficiais da rede mundial de computadores (*internet*)"[2], sendo que tais sítios deverão "possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina"[3] [6]. Considerando que grande parte dos dados a serem divulgados por tais órgãos e entidades são processados por Sistemas de Informações e persistidos em bases de dados relacionais, justifica-se a relevância do problema ora identificado.

3. OBJETIVO

O trabalho de mestrado ora apresentado visa definir um *framework*, baseado na camada de dados semânticos (*Semantic Data Layer*) da arquitetura de dados abertos governamentais DIGO [2], com a finalidade de automatizar os processos de geração, armazenamento e publicação de dados abertos governamentais (dados semânticos) obtidos a partir de dados estruturados mantidos em bases de dados relacionais (BDR).

Tal trabalho tem por objetivos específicos: 1) integrar, através de uma interface gráfica a ser desenvolvida em linguagem *Java*, as ferramentas de *software* selecionadas, após uma etapa prévia de pesquisa e avaliação, para executarem as funções de extração, mapeamento e conversão de dados estruturados (persistidos em BDR) para dados semânticos (geração de triplas *RDF*), de armazenamento das triplas *RDF* geradas (*RDF Store*), e de publicação (visualização e consulta) de dados semânticos, sendo que estas ferramentas deverão ser aderentes aos padrões do *W3C* (*RDF*, *OWL* e *R2RML*); 2) obter um método e desenvolver uma ferramenta de *software*, a ser implementada na referida interface gráfica, para a integração de uma ontologia *OWL* (em formato *turtle*) ao arquivo de mapeamento semântico da BDR, a fim de conferir maior expressividade e poder de inferência na geração e publicação dos dados semânticos (triplas *RDF*); 3) Validar o *framework* através de sua implementação em dois estudos de caso, um deles utilizando uma base de dados (SGBD *MySQL*) de teses defendidas em um dos programas de pós-graduação do Instituto Tecnológico de Aeronáutica (ITA), e outro utilizando uma base de dados (SGBD *MS-SQL Server*) do Sistema Único de Saúde (SUS) sobre tratamento e mortalidade de câncer de cérebro.

4. CONTRIBUIÇÕES ESPERADAS

Como contribuições esperadas deste trabalho estão: a obtenção de um modelo e de uma interface gráfica para integrar métodos e ferramentas de *software* voltados à geração e publicação de dados abertos governamentais, com base na camada de dados semânticos da arquitetura DIGO [2]; possibilitar ao usuário, através da referida interface gráfica, configurar as ferramentas do *framework* e fazer a integração de uma ontologia *OWL* ao mapeamento semântico da BDR, de forma a não exigir deste usuário elevados conhecimentos técnicos; facilitar e estimular a publicação de dados abertos governamentais, bem como a composição de aplicações para o consumo destes nas camadas de fusão e de informação da referida arquitetura; oferecer uma solução automatizada para que os órgãos e entidades da Administração Pública disponibilizem seus dados em formato aberto, de forma que os usuários possam obtê-los através de consultas customizadas de acordo com suas necessidades, bem como reaproveitá-los para a combinação com outras diferentes fontes de dados a fim de gerar novas informações (*mashup*), cumprindo assim o disposto na Lei de Acesso à Informação.

5. METODOLOGIA ADOTADA

Inicialmente foi conduzido um estudo detalhado da arquitetura DIGO [2] e dos conceitos relacionados, seguido de pesquisa na literatura e na *internet* em busca, respectivamente, dos trabalhos relacionados e das ferramentas existentes (determinação do estado da arte), a fim de possibilitar a identificação do problema de pesquisa e a análise de sua viabilidade técnica para, em seguida, determinar o objetivo e as contribuições esperadas do trabalho.

Assim, verificada a viabilidade técnica do trabalho, foram realizados alguns experimentos em laboratório a fim de avaliar as ferramentas encontradas, possibilitando a escolha daquelas que irão compor a versão inicial do *framework*. Para a realização dos experimentos foram testadas amostras das duas bases de dados a serem utilizadas nos referidos estudos de caso. As ferramentas foram escolhidas com base na aderência destas aos padrões do *W3C*, na compatibilidade com os SGBD utilizados (*MySQL* e *MS-SQL Server*), e na possibilidade de edição do arquivo de mapeamento semântico gerado, para a integração da ontologia.

Escolhidas as ferramentas que irão compor a versão inicial do *framework*, será então desenvolvida a interface gráfica, em linguagem *Java*, para integração destas ferramentas. Em paralelo, será elaborado o método e desenvolvida a ferramenta, a ser implementada na referida interface gráfica, para integração da ontologia *OWL* ao arquivo de mapeamento semântico da BDR.

Por fim, o *framework* será validado através de sua implementação nos dois estudos de caso citados anteriormente, onde, além de comprovar o funcionamento da interface gráfica desenvolvida, será possível comparar as visualizações e as consultas (*queries SparQL*) de dados semânticos obtidos a partir de duas situações: primeiramente através do mapeamento padrão de cada uma das BDR utilizadas (*default* da ferramenta) e, em seguida, através da integração da ontologia *OWL* do respectivo domínio a estes mapeamentos. Os resultados obtidos nestes estudos de caso possibilitarão a conclusão do trabalho.

6. ESTÁGIO ATUAL DO TRABALHO

A revisão da literatura e os experimentos de avaliação das ferramentas encontradas foram concluídos, possibilitando a seleção daquelas que irão compor a versão inicial do *framework*, as quais são relacionadas na tabela 1.

As ontologias de domínio para cada um dos estudos de caso foram obtidas, assim como o método para integração destas ao mapeamento semântico da BDR foi elaborado, possibilitando o projeto, especificação e início do desenvolvimento da respectiva

[1] Lei 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal.

[2] §2º do art. 8º da Lei 12.527/11.

[3] Inciso III do §3º do art. 8º da Lei 12.527/11.

ferramenta. Neste momento, também encontra-se em desenvolvimento a interface gráfica do *framework*.

Tabela 1. Ferramentas e padrões selecionados para comporem a versão inicial do *framework*.

Ferramenta/Padrão	Objetivo
D2RQ Engine	Mapeamento, extração e conversão de dados estruturados (geração de dados semânticos / triplas *RDF*).
D2R-Server	Visualização e consulta, via console *Web*, de dados semânticos gerados a partir de dados estruturados.
Jena API	Interface com aplicações locais *Java* e Armazenamento de triplas *RDF* (*RDF store*).
OWLtoD2RQ-Mapping tool[4]	Ferramenta para integração de ontologia *OWL* ao arquivo de mapeamento semântico do BDR.
Turtle	Formato do arquivo de mapeamento semântico do BDR e da ontologia *OWL* a ser integrada.

7. COMPARAÇÃO COM TRABALHOS RELACIONADOS

O *LOD2 Stack*[5] [7] é um *framework*, desenvolvido por um consórcio de empresas, centros de pesquisas e universidades da Europa e da Ásia, que agrupa uma série de ferramentas para a publicação de Dados Abertos Ligados[6] (ou *Linked Open Data - LOD*), abrangendo as tarefas de extração, consulta e exploração, criação, descoberta semi-automática de *links* entre as fontes de dados e, enriquecimento e reparação de bases de conhecimento. Dentre as características principais do *LOD2 Stack*, também encontradas no *framework* proposto neste trabalho, estão a adoção da ferramenta *D2RQ* para a extração de dados semânticos a partir de bases de dados relacionais, bem como para a visualização e consulta (através de console *SparQL*) destes, e sua interação com o usuário através de interface gráfica. Entretanto, a aplicação não permite a integração de ontologias ao mapeamento semântico da base de dados relacional, e está disponível para instalação somente na plataforma *Linux*, devendo ainda ser acessada através de um *browser Web*, pois apesar de algumas das ferramentas serem instaladas localmente no usuário, as demais ferramentas oferecidas são aplicações *on-line* na *internet*, diferentemente do *framework* proposto neste trabalho, que será multiplataforma, com todas suas ferramentas instaladas localmente no usuário.

O *Neon Toolkit*[7] [8] é um *framework* para a construção e o gerenciamento de ontologias, desenvolvido por um consórcio de instituições europeias, que conta com uma grande quantidade de ferramentas e *plugins* para estas finalidades. Apesar de seu propósito ser diferente ao deste trabalho, o *Neon* conta com um *plugin* (*ODE Mapster*) [9] para o mapeamento (através de interface gráfica) entre uma ontologia e uma base de dados relacional, porém compatível somente com os SGBDs *MySQL* e *ORACLE*, além de gerar o arquivo de mapeamento semântico em linguagem não aderente ao padrão *W3C* (*R2O*), mas que servirá de referência para o desenvolvimento da aplicação que visa integrar ontologia ao arquivo de mapeamento semântico da base de dados relacional, proposta neste trabalho.

O *Virtuoso*[8] [10] é uma plataforma, desenvolvida pela *OpenLink Software*, que tem por objetivo integrar dados, serviços e processos de negócio através de uma arquitetura que possibilita, em um único sistema, oferecer serviços de gerenciamento e integração de dados (com suporte a dados abertos semânticos), de integração de aplicações (*Web* services e *SOA*) e de integração e gerenciamento de processos. Possui uma versão comercial e uma *open-source* (com várias limitações). Possibilita a publicação de dados abertos gerados a partir do SGBD interno do produto (em ambas as versões) ou de um SGBD externo (somente na versão comercial).

8. CONCLUSÕES

Este trabalho vem preencher uma lacuna encontrada nas ferramentas disponíveis para a geração e publicação de dados semânticos a partir de dados estruturados persistidos em bases de dados relacionais, que é a falta de uma interface gráfica que automatize todo o processo e integre as diferentes ferramentas necessárias, bem como gerar as triplas *RDF* com base em uma ontologia, de forma a conferir maior expressividade e poder de inferência às visualizações e consultas desses dados semânticos.

O método para integração de ontologia *OWL* ao mapeamento semântico da BDR, nos testes efetuados manualmente, mostrou-se eficaz no sentido de tornar mais amigável para o usuário a visualização das triplas *RDF*, ao rotular sujeito, predicado e objeto de cada tripla com o respectivo termo da ontologia, o que facilita a formulação das *queries SparQL*.

A conclusão do desenvolvimento da interface gráfica e da ferramenta de integração de ontologia ao mapeamento semântico do BDR, e a implementação dos estudos de caso, permitirão aferir se o *framework* atingiu os objetivos e contribuições esperadas.

9. REFERÊNCIAS

[1] Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American vol. 284, Mai. 2001, pg. 28-37.

[2] Machado, A. L; Parente de Oliveira, J. M. DIGO: An Open Data Architecture for e-Government. In: 15th IEEE International Enterprise Distributed Object Computing Conference, 2011, Helsinki. 15th IEEE International Enterprise Distributed Object Computing Conference - Workshop Proceedings. IEEE Computer Society Press, 2011.

[3] Antoniou, G.; Harmelen, F. A Semantic Web Primer. 2ed. Cambridge: MIT Press, 2008.

[4] Guizzardi, G. Ontological Foundations for Structural Conceptual Models. Telematica Instituut Fundamental Research Series No. 015 (TI/FRS/015). Enschede: Universal Press, 2005.

[4] Nome provisório da ferramenta, em desenvolvimento pelo autor.

[5] http://demo.lod2.eu/lod2demo

[6] Dados abertos ligados refere-se à criação de *links* entre uma fonte de dados abertos e outras fontes disponíveis na *internet*, de forma a permitir que sejam combinadas e produzam novas informações e aplicações.

[7] http://neon-toolkit.org/

[8] http://virtuoso.openlinksw.com/

[5] W3C (e-Gov). eGovernment at W3C: improving access to government through better use of the Web, Online 2009. Disponível em: <http://www.w3.org/2007/eGov>.

[6] Cartilha Acesso à Informação Pública: Uma introdução à Lei nº 12.527, de 18 de novembro de 2011. Brasília: Controladoria Geral da União, 2011. Disponível em: <http://www.acessoainformacao.gov.br/acessoainformacao gov/publicacoes/CartilhaAcessoaInformacao.pdf.>

[7] Auer, S. et al. Managing the Life-Cycle of Linked Data with the LOD2 Stack. Disponível em: <http://lod2.eu/Blog Post/1214-paper-about-lod2-stack-accepted-for-iswc.html>.

[8] Haase, P. et al. Ontology engineering and plugin development with the neon toolkit. In: The 6th International Semantic Web Conference. ISWC 2007 Tutorial. Busan, Korea: Semantic Web Science Association, nov. 2007.

[9] Rodriguez, J. B.; Gómez-Pérez, A. Upgrading relational legacy data to the semantic web. Proceedings of the 15th international conference on World Wide Web. Anais...: WWW '06. New York, NY, USA: ACM, 2006. Disponível em: <http://doi.acm.org/10.1145/1135777. 1136019>.

[10] Erling, O.; Mikhailov, I. Virtuoso: RDF Support in a Native RDBMS. In: De Virgilio, R.; Giunchiglia, F.; Tanca, L. (Eds.). Semantic Web Information Management. Berlin, Heidelberg: Springer, Berlin Heidelberg, 2010. p. 501-519.