SemanticSUS: Um Portal Semântico baseado em Ontologias e Dados Interligados para Acesso, Integração e Visualização de Dados do SUS

Matheus Mayron Lima da Cruz, Caio Viktor Silva Avila, Vânia Maria Ponte Vidal, Narciso Moura Arruda Junior

[1] Universidade Federal do Ceará (UFC)
Fortaleza, CE - Brasil

{matheusmayron,arlaass,vaniap.vidal,narcisoarruda}@gmail.com

Abstract. *This demo presents SemanticSUS, a semantic portal for access, analysis and visualization of large amounts of data from the SUS (Sistema Único de Saúde). The SemanticSUS Portal is based on an approach that combines Ontologies and Linked Data to address the challenges in developing applications where there is a need to integrate heterogeneous data sources. SemanticSUS aims to provide an ontological layer, semantically connected to the data, and to allow integrated access to the data. The platform also offers the semantic integration service based on the pay-as-you-go approach, which guarantees sufficient flexibility and extensibility so that new data sources can be added to the portal. Another feature of the portal is the tool "Mashup Buider", which allows the construction of Data Mashup in a simple and automatic way.*

Resumo. *Esta demo apresenta SemanticSUS, um portal semântico para acesso, análise e visualização de grande quantidade de dados do Sistema Único de Saúde (SUS). O Portal SemanticSUS é baseado em um enfoque que combina ontologias e Dados Interligados para enfrentar os desafios no desenvolvimento de aplicações onde existe a necessidade de integrar fontes de dados heterogêneas. O SemanticSUS tem como principal objetivo oferecer uma camada ontológica, conectada semanticamente aos dados, e que permita o acesso integrado aos dados. A plataforma disponibiliza também o serviço de integração semântica baseado na abordagem "pay-as-you-go", o que garante flexibilidade e extensibilidade suficientes para que novas fontes de dados possam ser adicionadas ao portal. Outra facilidade do portal é a ferramenta "Mashup Buider", a qual permite a construção de Mashup de Dados de forma simples e automática.*

1. Introdução

A exploração e análise dos dados do SUS (Sistema Único de Saúde) são fundamentais para a descoberta de conhecimentos que podem ser utilizados no desenvolvimento de políticas públicas com impacto direto sobre a saúde da população. No entanto, a eficácia da descoberta de conhecimento depende da preparação adequada dos dados e interpretação dos resultados, o que apresenta alguns desafios, dado que as fontes de dados são heterogêneas e distribuídas.

Um dos principais desafios na integração dos dados do SUS está relacionado ao problema da interoperabilidade de dados presentes em fontes heterogêneas [Bishr 1998].

Para que se possa alcançar a interoperabilidade de dados é necessário integrar semanticamente as fontes de dados. Chamamos de *integração semântica* o processo que faz uso de uma representação conceitual dos dados e seus relacionamentos para eliminar possíveis heterogeneidades. A representação conceitual pode ser feita por meio do uso de ontologias que são, por definição, uma representação formal e explicita de uma conceitualização compartilhada [Studer et al. 1998].

Esta demo apresenta ***SemanticSUS***, um portal semântico para acesso, análise e visualização de grande quantidade de dados do SUS. O Portal *SemanticSUS* é baseado em um enfoque que combina ontologias e Dados Interligados para enfrentar os desafios no desenvolvimento de aplicações onde existe a necessidade de integrar semanticamente fontes de dados heterogêneas. O principal objetivo do portal é oferecer uma camada ontológica que se conecta semanticamente aos dados e permite o acesso integrado as fontes de dados. O acesso a essa camada através do portal pode ocorrer por meio de diferentes tipos de interfaces de consulta, de forma que o portal possa atender a diferentes demandas de acesso e tipos de usuário. A plataforma também disponibiliza o serviço de integração semântica baseada na abordagem *pay-as-you-go* [Madhavan et al. 2007], que garante flexibilidade e extensibilidade suficientes para que novas fontes de dados possam ser adicionadas ao portal. No estado atual, o *SemanticSUS* publica duas bases de dados do SUS disponíveis na plataforma GISSA[1] [Freitas et al. 2017].

O restante do artigo está organizado como se segue. Seção 2 descreve a arquitetura do portal semântico. Seção 3 discute o processo de integração semântica no *SemanticSUS*. Seção 4 mostra como consultar as fontes de dados através da camada ontológica do Portal. Seção 5 ilustra a construção de um *mashup* com a ferramenta Mashup Builder. Por fim, seção 6 apresenta as conclusões.

2. Arquitetura do Portal Semântico

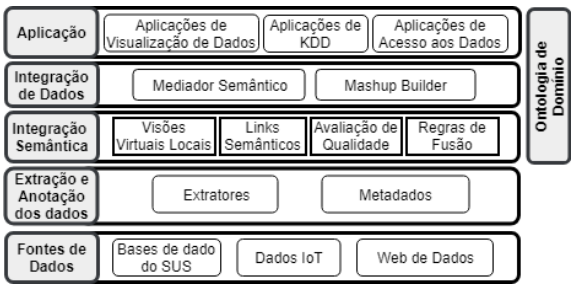

Figura 1. Arquitetura do Portal Semântico

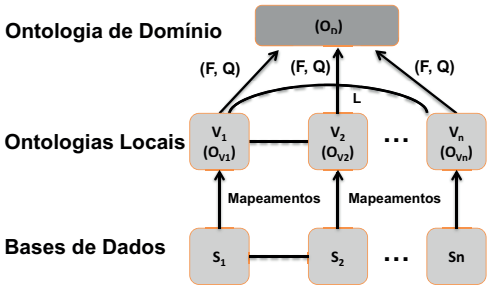

Figura 2. Framework para Integração Semântica

Figura 1 mostra a arquitetura do SemanticSUS. Cada uma das camadas tem sua descrição apresentada a seguir:

- **Fontes de dados:** Nesta camada estão as fontes de dados disponibilizadas pelo portal. As fontes de dados podem ser de diferentes tipos e proveniência;
- **Extração e anotação de dados:** Nesta camada é realizada a extração e anotação das fontes de dados disponíveis no Portal. As anotações são metadados que descrevem

[1]GISSA é um sistema que disponibiliza, para gestores de saúde pública municipal, informações contextualizadas e em tempo responsivo.

desde o formato da fonte até comentários sobre os significados de elementos do esquema utilizado pela fonte;

- **Ontologia de Domínio:** Essa ontologia é usada como um meio para estabelecer um vocabulário formal e explícito a ser compartilhado para a anotação semântica das fontes de dados. A linguagem utilizada para modelagem da ontologia é a linguagem OWL[2];
- **Integração Semântica (IS):** O resultado da IS define uma visão virtual integrada das múltiplas fontes de dados. Na IS usamos um *framework* baseado em 3 camadas [Vidal et al. 2015], como sumarizado na Figura 2. De acordo com esse *framework*, o resultado da IS é uma n-tupla $\lambda = (O_D, V_1, ..., V_n, L, F, Q)$, onde:
  - O_D é a ontologia de Domínio;
  - $V_1, ..., V_n$ são especificações de visões virtuais locais descritas pelas ontologias locais $O_{V_1}, ..., O_{V_n}$;
  - L é um conjunto de regras de *linkage* (links semânticos) definidas entre as classes semanticamente semelhantes das ontologias locais;
  - F é um conjunto de regras de fusão as quais especificam o processo de fusão de objetos relacionados através de um link *owl:sameAs* em uma única representação;
  - Q é um conjunto de métricas de avaliação de qualidade, que são usadas para quantificar a qualidade das fontes de dados.
- **Integração de Dados:** O Portal disponibiliza dois serviços de acesso às fontes de dados locais através da visão ontológica:
  - **Mediador Semântico:** Os usuários podem definir consultas *ad-hoc* sobre a ontologia de Domínio, e o *mediador semântico*, com base na ontologia e nos mapeamentos, irá reformular a consulta em termos de consultas sobre as fontes de dados. O resultado das sub-consultas são integrados e consolidados, e o resultado final é retornado para o usuário;
  - **Mashup Builder:** Os usuários podem especificar uma visão de *Mashup* especializada sobre a ontologia de domínio através de uma interface visual, e o *Mashup Builder* irá materializar a visão de *mashup* de forma automática, baseado no resultado da integração semântica;
- **Aplicação:** Na última camada estão os usuários e aplicações que fazem uso dos serviços de integração dos dados.

3. Integração Semântica no SemanticSUS

No SemanticSUS, o processo de Integração Semântica (IS) é baseado na abordagem *pay-as-you-go* [Madhavan et al. 2007]. Essa abordagem permite que a IS possa ser realizada de forma incremental, e a medida que mais esforço é investido na integração semântica de novas fontes de dados, o portal poderá atender as necessidades de um maior número de aplicações. O processo de IS de uma fonte S segue os seguintes passos:

1. Especificação da visão virtual local $V_S = < O_S, M_S >$ onde O_S é a ontologia da visão, e M_S é um conjunto de regras que mapeiam o vocabulário de O_S no vocabulário da fonte S. Note que O_S descreve a fonte de dados usando o vocabulário da Ontologia de domínio, portanto O_S é um subconjunto de O_D. Na construção da ontologia de domínio também adotamos o enfoque *pay-as-you-go*;

[2]https://www.w3.org/OWL/

2. Especificação das Regras de *Linkage* as quais serão usadas para gerar Links semânticos, que mapeiam instâncias de S com instâncias de outras fontes de dados que representam o mesmo objeto do mundo real (*resolução de entidades*). Em OWL, esses links são estabelecidos através da propriedade *owl:sameAs*. O processo de identificação desses *links* é conhecido como *linkage* de dados;
3. Especificação das regras de fusão para as propriedades de S. Estas regras são definidas com base na qualidade da fonte S, e são necessárias na resolução de conflitos gerados por inconsistências nos dados [Mendes et al. 2012].

No estado atual, o portal *SemanticSUS* publica duas bases de Dados do SUS: o SIM [3] e o SINASC[4]. Essa fontes estão integradas semanticamente e já podem ser acessadas através da visão ontológica do portal. O resultado da Integração semântica das fontes de dados SIM e SINASC pode ser visualizados através do link *http://tiny.cc/semanticsus_is*. Abaixo apresentamos uma descrição do resultado:

- A ontologia de Domínio, e ontologias locais das fontes de dados do SIM e SINASC são representadas em OWL;
- Os mapeamentos dos esquemas da fontes para as ontologias locais estão especificados em R2RML;
- As regras de ligação estão especificadas usando a ferramenta SILK [Volz et al. 2009]; A regra de *linkage* é usada para gerar links *owl:sameAs* entre instâncias das classes *sim:PessoaMorta* e *sinasc:RN* (recém nascido). Essa regra é determinística e usa a propriedade número da declaração de nascido vivo para gerar os links *owl:sameAs*;
- As regras de fusão estão especificadas usando a ferramenta SIEVE [Mendes et al. 2012].

4. Consultando as fontes de Dados através da Visão Ontológica

O Mediador Semântico (MS) é um sistema que permite que sejam realizadas consultas às fontes de dados através da Visão Ontológica definida pela integração semântica. A visão ontológica provê um único ponto de acesso aos dados, e permite que consultas sejam formuladas em termos da ontologia de domínio, de forma que o usuário não precisa entender das fontes de dados, nem das relações entre elas, e a resposta é recebida de forma inteligível.

Outra vantagem de consultas mediadas por ontologias, é que o conhecimento semântico provido pela ontologia pode ser explorado durante a consulta, e portanto, pode prover respostas mais completas para as consultas [Calvanese et al. 2017]. Para demostrar o fluxo seguido pelo MS no processamento de uma consulta, suponha que um usuário deseja obter informações sobre o parto das crianças que vieram a óbito com menos de 28 dias e tiveram asfixia ao nascer. Para responder essa requisição é necessário consultar a fonte de dados do SIM, para identificar as crianças que vieram a óbito com menos de 28 dias, e a fonte de dados do SINASC, para obter as informações sobre o parto dessas crianças. Além disso, é preciso explorar os axiomas da ontologia para inferir sobre "asfixia no parto", uma vez que essa informação não está disponível de forma explícita nas fontes de dados. Os passos envolvidos no processamento de uma consulta são sumarizados a seguir (mais detalhes em *http://tiny.cc/semanticsus_ms*):

[3] Sistema de Informações sobre Mortalidade
[4] Sistema de Informações sobre NASCidos vivos

1. O usuário define a consulta SPARQL Q sobre a ontologia de domínio da visão ontológica;
2. A consulta Q é traduzida em sub-consultas SQL definidas sobre o esquema das fontes de dados locais. A tradução é feita baseada nos mapeamentos entre as fontes de dados e a ontologia de domínio. Note que é necessário o uso de "mapeamentos completos", o que significa que os axiomas, restrições de integridade, e as regras de ligação são todos compilados nos mapeamentos;
3. As sub-consultas SQL são executadas pelo SGBD local;
4. O resultado das sub-consultas são integrados usando os links *owl:sameAs*, e o resultado final é traduzido em RDF, de acordo com a consulta SPARQL;
5. O resultado é então retornado para o usuário.

5. Construindo Mashup de Dados Especializado

Um ***mashup* de dados** é uma visão materializada construída através da transformação e integração de dados presentes em diferentes bases. O processo de criação de *mashup* é uma tarefa complexa se as fontes de dados não foram integradas semanticamente[Vidal et al. 2015].

A seguir, descrevemos a utilização do *Mashup Builder* na criação do *mashup* RMN(Risco Morte Neonatal) que foi utilizado no desenvolvimento de um modelo preditivo para estabelecer o risco de morte neonatal. O mashup RMN deverá conter as seguintes informações sobre as crianças que morreram antes de completarem 28 dias de vida (período neonatal): peso, idade gestacional, Apgar, e prévios relatórios de natimorto da mãe. O *mashup* foi construído utilizando as fontes de dados SIM e SINASC disponíveis na base de dados do sistema GISSA [Freitas et al. 2017] como discutimos a seguir. A construção do *mashup* RMN usando a ferramenta Mashup Builder é realizada em 3 passos (Figura 3):

1. Primeiro, o usuário deve especificar a necessidade de informação da aplicação através de uma interface de especificação da visão de mashup definida sobre a ontologia de domínio;
2. Em seguida, baseado na especificação da visão definida no passo 1 e no resultado da integração semântica do SIM e SINASC, a ferramenta gera automaticamente a especificação da visão do *mashup* sobre as fontes de dados. Essa especificação é dada por:
   $\lambda = (V_{RMN}, E_{SIM}, E_{SINASC}, L_{RMN}, F_{RMN}, Q_{RMN})$, onde:
   - V_{RMN} é o vocabulário da visão do Mashup. V_{RMN} = {*o:RN, o:PessoaMorta, o:sexo, o:temMae, osus:Mae, o:qtdNascMortos, o:infoNasc, o:parto, o:gestação, o:diasVividos, o:tempoDeGestacao, o:apgar1min, o:apgar5min, o:pesoAoNascer*};
   - E_{SIM} e E_{SINASC} são especificações das visões exportadas pelas fontes de dados;
   - L_{RMN} contém as regras de *linkage* da visão semântica em e que são relevantes para as visões exportadas;
   - F_{RMN} contém as regras de fusão da visão semântica que são relevantes para a visão do *mashup*;
   - Q_{RMN} contém a avaliação da qualidade das visões exportadas.
3. Por fim, a ferramenta realiza a materialização da Visão do *mashup* em 3 passos: (1) primeiro, as visões exportadas são materializadas usando mapeamentos R2RML que mapeiam os dados das fontes de dados SIM e SINASC em instâncias das visões exportadas por estas fontes . (2) Em seguida, os links *owl:sameAs* são gerados pela

ferramenta SILK, de acordo com a regra de *linkage* L_{RMN} definida na integração semântica. (3) Por último, é realizada a fusão dos objetos - relacionados através de um link *owl:sameAs* - em uma única representação. O processo de fusão é realizado pela ferramenta SIEVE.

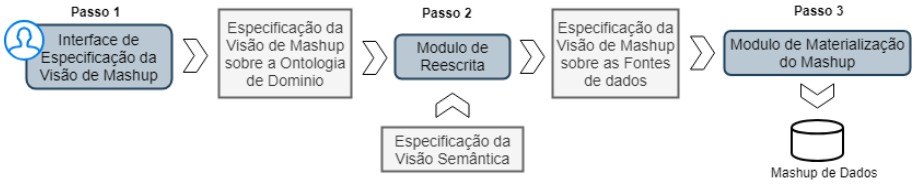

Figura 3. Construção de *Mashup* de Dados Especializado

As especificações geradas para esse exemplo, bem como mais detalhes sobre o Mashup Builder podem ser encontrados em *http://tiny.cc/semanticsus_mb*

6. Conclusão

Nesse artigo, apresentamos *SemanticSUS*, um portal semântico para acesso, integração, análise e visualização de dados do SUS. O Portal *SemanticSUS* é baseado em um enfoque que combina ontologias e Dados Interligados para enfrentar os desafios de integrar fontes de dados heterogêneas. O portal oferece uma arquitetura flexível e escalável que permite a fácil adição de novas fontes. No estado atual, o portal integra semanticamente duas fontes de dados do SUS disponíveis na plataforma GISSA. Na Seção 4 mostramos como consultar essas fontes de dados usando o Mediador Semântico. Na Seção 5 mostramos a geração de um *mashup* de dados especializado usando o *Mashup Builder*. Um vídeo ilustrando o funcionamento do portal está disponível em *http://tiny.cc/semanticsus_video*

Como trabalhos futuros, pretendemos melhorar o ambiente de Integração semântica, disponibilizar interfaces de consulta semântica e em linguagem natural, desenvolver ferramentas de visualização dos dados, assim como realizar a integração semântica de nova fontes do SUS no *SemanticSUS*.

Referências

[Bishr 1998] Bishr, Y. (1998). Overcoming the semantic and other barriers to gis interoperability. *International journal of geographical information science*, 12(4):299–314.

[Calvanese et al. 2017] Calvanese, D. et al. (2017). Ontop: Answering sparql queries over relational databases. *Semantic Web*, 8(3):471–487.

[Freitas et al. 2017] Freitas, R. et al. (2017). Using linked data in the data integration for maternal and infant death risk of the sus in the gissa project. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, pages 193–196. ACM.

[Madhavan et al. 2007] Madhavan, J. et al. (2007). Web-scale data integration: You can only afford to pay as you go.

[Mendes et al. 2012] Mendes, P. N. et al. (2012). Sieve: linked data quality assessment and fusion. In *2012 Joint EDBT/ICDT Workshops*, pages 116–123. Citeseer.

[Studer et al. 1998] Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197.

[Vidal et al. 2015] Vidal, V. M. et al. (2015). Specification and incremental maintenance of linked data mashup views. In *CAiSE*, pages 214–229. Springer.

[Volz et al. 2009] Volz, J. et al. (2009). Silk-a link discovery framework for the web of data. *LDOW*, 538.