

# Construção de Linked Data Mashup para Integração de Dados da Saúde Pública

Gabriel Lopes<sup>1</sup>, Vânia Vidal<sup>2</sup>, Mauro Oliveira<sup>1</sup>

<sup>1</sup>Instituto Federal do Ceará (IFCE)  
Fortaleza, CE

<sup>2</sup>Universidade Federal do Ceará (UFC)  
Fortaleza, CE

`gabriel.lopes@ppgcc.ifce.edu.br`, `vvidal@lia.ufc.br`,

`amauroboliveira@gmail.com`

**Abstract.** *Linked Data promotes the publication of structured data, easing the development of an homogenized-view over heterogeneous sources, called Linked Data Mashup view (LDM view). This article describes the processes of specification and materialization of a LDM view of two heterogeneous bases from Brazilian Public Health System (SUS): Information System on Live Births (SINASC) and SUS electronic (e-SUS). From this process, it was possible to obtain an integrated-view of the bases previously isolated. This integrated-view will be useful for analyzing the correlation between mother's information during pregnancy with deaths and anomalies in newborns.*

**Resumo.** *Linked Data promove a publicação de dados na Web de forma estruturada, facilitando a criação de uma visão homogênea sobre fontes heterogêneas, chamada de visão de Linked Data Mashup (visão LDM). Esse artigo descreve os processos de especificação e materialização de uma visão LDM sobre duas bases heterogêneas do Sistema Único de Saúde (SUS): Sistema de Informações sobre Nascidos Vivos (SINASC) e SUS eletrônico (e-SUS). A partir desse processo, foi possível obter uma visão integrada das bases anteriormente isoladas. Essa visão será útil para analisar a correlação entre informações da mãe durante a gravidez, as causas de óbitos e as anomalias-congênicas em recém-nascidos.*

## 1. Introdução

Diversos tipos de dados podem ser encontrados na Web, desde informações sobre a venda de determinados produtos até orçamentos governamentais. Existem iniciativas que promovem a publicação de dados abertos na Web. O *Open Government Data*, por exemplo, é um movimento que está sendo aderido por diversos países, como o Brasil, os EUA e o Reino Unido para incentivar a publicação de dados governamentais sobre diversas áreas: Saúde, Educação, Clima, Finanças, dentre outras. Tais iniciativas impulsionam o desenvolvimento de aplicações comerciais, pesquisas, apoio a tomada de decisões complexas, dentre outros.

Porém, essas bases de dados geralmente estão publicadas em formatos diferentes e proprietários, i.e., além de não utilizarem um padrão para publicação, possuem *schemas* (modelos de dados) distintos, fazendo com que cada base pareça isolada em relação com

as demais. Desenvolver uma visão integrada desses dados é um processo complexo que envolve diversas variáveis, como o tempo de desenvolvimento, a manutenibilidade e a remoção da heterogeneidade dos dados [Ziegler and Dittrich 2007].

Nesse contexto, a iniciativa *Linked Data* [Bizer et al. 2009a] promove a publicação de bases de dados anteriormente isoladas como fontes RDF [RDF 2014] interligadas. O *Linked Data* está impulsionando a evolução da Web atual, que utiliza arquivos HTML para descrever objetos do mundo real, inteligíveis apenas por humanos, para uma Web em formato de grafo, onde cada nó é uma fonte de dados no formato RDF e que pode ser acessada também por máquinas. Este novo conceito de Web é conhecido por Web de Dados [Heath and Bizer 2011]. Além disso, o *Linked Data* trouxe novas oportunidades para criação de aplicações semânticas, que utilizam visões integradas de dados no formato *Linked Data* chamadas de *visões Linked Data Mashup* (LDM). Uma visão LDM oferece novas funcionalidades para aplicações semânticas, combinando, agregando e transformando dados de fontes heterogêneas disponíveis na Web [Hoang et al. 2014].

A motivação desse trabalho surgiu com a necessidade da prefeitura da cidade de Tauá, Ceará - Brasil, de investigar as causas dos óbitos em recém-nascidos por meio da análise de informações sobre suas mães: uso de drogas, tabaco ou álcool durante a gravidez; doenças crônicas, como a diabetes e hipertensão; possíveis complicações em gestações anteriores e a quantidade de consultas pré-natal. No entanto, tais informações estão distribuídas em bases heterogêneas, e o Sistema de Apoio a Tomada de Decisão (ou *Decision Support System* - DSS) utilizado pela prefeitura não disponibiliza uma visão integrada dessas bases. Com isso, um gestor de saúde é impossibilitado de verificar, por exemplo, se determinada mãe é portadora de diabetes ou se é usuária de drogas.

Esse artigo descreve o processo de desenvolvimento de uma visão *Linked Data Mashup* que integra dados de duas bases do Sistema Único de Saúde (SUS) brasileiro: o Sistema de Informações sobre Nascidos Vivos (SINASC) e o SUS eletrônico (e-SUS). Esse *mashup* permite o desenvolvimento de aplicações sobre os dados heterogêneos, como *dashboards*, gráficos, etc. Esse artigo representa a etapa inicial de um trabalho, que tem como objetivo aumentar o poder de decisão de um DSS a partir da disponibilização de uma visão integrada utilizando *Linked Data Mashup*. A integração dos dados foi realizada utilizando o *framework* conceitual abordado em [Vidal et al. 2015]. Esse *framework* apresenta um processo formal para especificar visões LDM. A vantagem de utilizar essa abordagem, dentre as várias existentes na literatura (e.g. [Schultz et al. 2011], [Jarrar and Dikaiakos 2008]), é que a especificação gerada pode ser reutilizada para integrar outras bases semelhantes, como as bases e-SUS e SINASC de outras cidades. Com a materialização do *mashup*, foi possível disponibilizar uma visão integrada das bases, acessível a partir de uma *interface web* de consultas sobre bases RDF. Além das questões levantadas na motivação, também foi possível analisar a relação dos hábitos da mãe durante a gravidez com as anomalias congênitas do recém-nascido.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta o *Framework* utilizado no desenvolvimento da visão LDM. Na Seção 3 são apresentados os passos realizados para a materialização da visão integrada. Finalmente, a Seção 4 contém a conclusão e os trabalhos futuros.

## 2. Framework para a Especificação e Materialização da visão LDM

Nesta seção, é apresentado o *framework* [Vidal et al. 2015] que foi utilizado para a construção da visão *Linked Data Mashup* descrita no presente artigo. Este *framework* descreve o processo de construção de um *mashup* em duas etapas: (i) especificação de uma visão LDM e (ii) materialização dos dados. Primeiro, especifica-se um *mashup* seguindo o *framework* baseado em ontologias, resumido na Figura 1. Em seguida, a materialização dos dados é realizada com o auxílio de ferramentas específicas seguindo a especificação previamente definida.

Na *Camada de Mashup*, a *Ontologia de Aplicação* ( $O_{MV}$ ) representa a conceptualização da motivação para criação do *mashup*. Nessa camada também são definidas as regras de fusão dos dados ( $\mu$ ) e as regras para a avaliação de qualidade ( $Q$ ) das fontes. Durante a etapa de fusão, múltiplas representações de um mesmo objeto do mundo real são combinadas numa única representação. Na *Camada de Dados*, cada fonte  $S_i$  é representada por uma *Ontologia Fonte*  $O_{S_i}$ . Cada *Visão Exportada*  $E_i$  é composta por uma ontologia  $O_{E_i}$ , que é um subconjunto de  $O_{MV}$ , e um conjunto de regras  $M_{E_i}$  que mapeia os conceitos de  $O_{S_i}$  para  $O_{E_i}$ . Também é na camada das visões exportadas que o conjunto dos links *sameAs* ( $ML_i$ ) é especificado e materializado. Esses *links* representam a similaridade entre dois objetos do mundo real que estão em bases distintas. Essa especificação da visão LDM será utilizada para a materialização dos dados através do uso de ferramentas específicas em cada etapa. Nas subseções seguintes são abordadas as etapas de especificação e materialização.

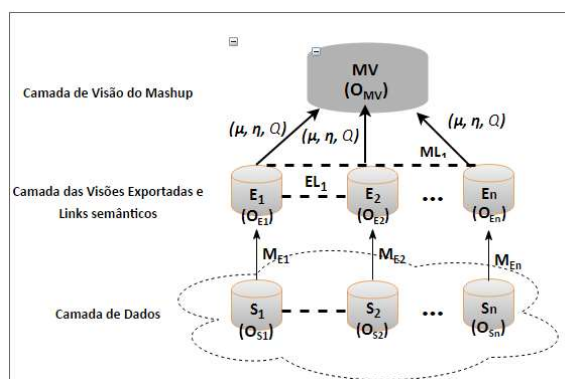


Figura 1. *Framework 3-Camadas*

O processo de geração de uma visão LDM é uma tarefa complexa que envolve 5 desafios: i) modelagem de uma ontologia de *mashup*; ii) mapeamento das fontes de dados heterogêneas para as ontologias exportadas (visões exportadas); iii) geração das especificações para identificação dos links *sameAs*; iv) definição das regras para quantificar a qualidade das fontes de dados e v) combinação e fusão de múltiplas representações de um mesmo objeto do mundo real para uma única representação.

A etapa de materialização é feita automaticamente a partir de uma especificação. Para isso, são utilizadas ferramentas específicas em cada etapa. A materialização de uma visão LDM consiste de três passos: 1. **Materialização das visões exportadas**: esse passo utiliza os mapeamentos  $M_{E_i}$  definidos na especificação da visão exportada  $E_i$ , para traduzir os dados de  $S_i$  para o vocabulário de  $O_{E_i}$ . 2. **Materialização dos links *sameAs***: a partir da especificação de um conjunto de links, gerá-los com o auxílio de alguma ferramenta



Para a criação dos mapeamentos  $M_{E_{sinasc}}$  e  $M_{E_{esus}}$ , foi utilizada a linguagem padrão para mapeamentos de dados relacionais em RDF, o R2RML [W3C 2016]. A materialização das visões exportadas foi realizada utilizando a ferramenta *D2R-Server*, que conta com uma *engine* para interpretar os mapeamentos R2RML e gerar os dados RDF no formato *n-triple*. Ao final da materialização de  $E_{sinasc}$  e  $E_{esus}$ , foram geradas 162865 e 24841 triplas RDF respectivamente. Por conta de limitação de espaço, os *schemas*, as ontologias e os mapeamentos são omitidos nesse trabalho.

### 3.2.1. Materialização dos Links *sameAs*

Os links *sameAs* identificam que dois objetos do mundo real em fontes de dados distintas representam uma mesma entidade. Nesse artigo, utilizamos a ferramenta SILK [Bizer et al. 2009b] para a materialização dos links *owl:sameAs*. Nessa ferramenta, são especificadas heurísticas que identificam a similaridade entre dois registros. Na visão LDM desenvolvida no presente trabalho, o intuito é identificar que dois *foaf:Person* em bases distintas representam um mesmo indivíduo. Para realizar o *match*, foram utilizadas as seguintes informações: o nome da pessoa (*foaf:name*); a data de nascimento (*gissa:dataNascimento*) e o Cartão Nacional da Saúde (*gissa:cns*). Após o processamento do SILK, foram identificados 326 links entre  $O_{E_{sinasc}}$  e  $O_{E_{esus}}$ . A baixa quantidade de *links* gerados se deve à má qualidade dos dados.

### 3.2.2. Qualidade das Fontes e Materialização do mashup

Para quantificar a qualidade das bases, utilizamos alguns dos critérios abordados em [Pipino et al. 2002], como: (i) quantidade de registros *essenciais* (informações necessárias para o *match* de indivíduos, subseção 3.2.1) ausentes nas bases de dados; (ii) quantidade de registros duplicados e (iii) quantidade de registros inconsistentes, como nomes de pessoas com erros de escrita, por exemplo. Atribuímos peso 1 para cada critério, somamos 0.1 sempre que um registro se encaixasse em algum dos critérios e fizemos uma média simples para obter a pontuação de cada base. Dessa forma, a fonte de dados com maior pontuação representa a menos confiável. Na fusão, utilizamos todas as propriedades originadas da fonte mais confiável. A base  $S_{sinasc}$  obteve uma pontuação de: (i) 844.1; (ii) 130.9 e (iii) 98.8. A base  $S_{esus}$  obteve: (i) 42.3; (ii) 0.0 e (iii) 0.0. A grande taxa de registros nulos e/ou com erros de escrita da fonte SINASC, representados pelo critério (i) e (iii) respectivamente, estão relacionados com a metodologia adotada na aquisição desses dados, que é por meio do preenchimento de uma ficha, onde geralmente os indivíduos não têm os documentos necessários. A duplicidade das informações é atribuída pela falta de informatização do sistema, que não verifica se aquele usuário já realizou um cadastro.

A materialização da fusão foi realizada utilizando a ferramenta SIEVE. Nessa ferramenta, são especificados os critérios de qualidade e as regras de fusão [Mendes et al. 2012]. Após a execução de SIEVE, 177609 triplas RDF foram geradas.

## 4. Conclusão e Trabalhos Futuros

Nesse artigo foi utilizado um *framework* para especificação de uma visão *Linked Data Mashup* sobre duas bases heterogêneas do Sistema Único de Saúde (SUS): SINASC e

e-SUS. Esse trabalho tem como objetivo auxiliar um Sistema de Apoio a Tomada de Decisão, utilizado pela prefeitura da cidade de Tauá (Ceará - Brasil), a correlacionar os óbitos e as anomalias-congênicas em recém-nascidos com informações sobre a mãe durante a gravidez. Esperamos ser possível a realização de trabalhos de conscientização nas gestantes, alertando sobre a relação de maus-hábitos durante a gravidez e anomalias congênicas no recém-nascido e, com isso, alcançar uma diminuição dos casos de óbitos. Como trabalhos futuros, pretendemos realizar um trabalho de anonimização nos dados do *mashup* para que possamos disponibilizá-los à comunidade, permitindo a criação de aplicações em diversas áreas, como a Mineração de Dados. Também pretendemos integrar novas fontes de dados ao nosso *mashup*. Por exemplo, agregar informações sobre mortalidades da base SIM (Sistema de Informações sobre Mortalidades) para analisar os óbitos-maternos.

## Referências

- Bizer, C., Heath, T., and Berners-Lee, T. (2009a). Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.
- Bizer, C., Volz, J., Kobilarov, G., and Gaedke, M. (2009b). Silk - a link discovery framework for the web of data. In *18th International World Wide Web Conference*.
- Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition.
- Hoang, H. H., Cung, T. N., Truong, D. K., Hwang, D., and Jung, J. J. (2014). Semantic information integration with linked data mashups approaches. *IJDSN*, 2014.
- Jarrar, M. and Dikaiakos, M. D. (2008). Mashql: A query-by-diagram topping sparql. In *Proceedings of the 2Nd International Workshop on Ontologies and Information Systems for the Semantic Web, ONISW '08*, pages 89–96, New York, NY, USA. ACM.
- Mendes, P. N., Mühleisen, H., and Bizer, C. (2012). Sieve: Linked Data Quality Assessment and Fusion. In *2nd International Workshop on Linked Web Data Management (LWDM 2012) at the 15th International Conference on Extending Database Technology, EDBT 2012*, page to appear.
- Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*, 45(4):211–218.
- RDF (2014). Resource description framework.
- Schultz, A., Matteini, A., Isele, R., Bizer, C., and Becker, C. (2011). Ldif : Linked data integration framework.
- Vidal, V. M. P., Casanova, M. A., Arruda, N., Roberval, M., Leme, L. P., Lopes, G. R., and Renso, C. (2015). *Advanced Information Systems Engineering: 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8-12, 2015, Proceedings*, chapter Specification and Incremental Maintenance of Linked Data Mashup Views, pages 214–229. Springer International Publishing, Cham.
- W3C (2016). *R2RML RDB to RDF Mapping Language*. available at <https://www.w3.org/TR/r2rml/>.
- Ziegler, P. and Dittrich, K. R. (2007). Data integration – problems, approaches, and perspectives.