Using Linked Data in the Data Integration for Maternal and Infant Death Risk of the SUS in the GISSA Project

Renato Freitas
Instituto Federal do Ceará (IFCE)
Aracati, Ceará, Brasil
jrenatosfreitas@gmail.com

Cleilton Lima
Instituto Atlântico
Fortaleza, Ceará, Brasil
cleilton_rocha@atlantico.com.br

Oton Braga
Instituto Federal do Ceará (IFCE)
Aracati, Ceará, Brasil
otoncbraga@gmail.com

Gabriel Lopes
Instituto Federal do Ceará (IFCE)
Fortaleza, Ceará, Brasil
gabriellopes9102@gmail.com

Odorico Andrade
Congresso Nacional
Brasília, Distrito Federal, Brasil
odorico0811@gmail.com

Mauro Oliveira
Instituto Federal do Ceará (IFCE)
Aracati, Ceará, Brasil
amauroboliveira@gmail.com

ABSTRACT

Making good governance decisions is a constant challenge for Public Health administration. Health managers need to make data analysis in order to identify several health problems. In Brazil, these data are made available by DATASUS. Generally, they are stored in distinct and heterogeneous databases. The *Linked Data* approach allow a homogenized view of the data as a unique basis. This article proposes a ontology-based model and *Linked Data* to integrate datasets and calculate the probability of maternal and infant death risk in order to give support in decision-making in the GISSA project.

KEYWORDS

Ontology; Linked Data; Public Health System; SUS Database

1 INTRODUÇÃO

Tomar boas decisões de governança é um desafio constante para administração de qualquer atividade profissional, não sendo diferente na Saúde Pública. Devido a interdependência entre os diversos domínios envolvidos em sistemas de saúde (clínico epidemiológico, administrativo, normativo, etc.) [7], gestores precisam analisar a relação entre os dados destes domínios a fim definir as melhores estratégias, seja para a prevenção ou para a solução de problemas. SINASC[1] e e-SUS[2] são exemplos de bases de dados de saúde pública, heterogêneas e distintas, disponibilizadas pelo Departamento de Informática do SUS - DATASUS.

A análise da relação entre os dados das diversas bases do DATASUS é uma atividade dispendiosa e massiva, mesmo fazendo-se uso clássico de computadores. Para que um gestor tenha uma visão completa de um problema em saúde pública, os dados das diversas bases disponíveis devem passar por uma integração, i.e., tornarem-se um conjunto homogêneo. Contudo, integrar dados não é um processo trivial. Faz-se, portanto, necessário se dispor de mecanismos computacionais mais elaborados, capazes de integrar dados e extrair informações relevantes que auxiliem gestores de saúde a tomarem boas decisões [8]. Sistemas baseados em ontologias e *Linked Data* [1][4][5] e suas tecnologias associadas, tais como RDF e SPARQL, são capazes de integrar fontes de dados e inferir novas informações a partir de bases heterogêneas de conhecimento [2][9].

Neste contexto, tem-se o GISSA[3], um sistema inteligente de governança para o apoio à tomada de decisão em ambientes de saúde, desenvolvido a partir do *framework* LARIISA [8]. Trata-se de um projeto financiado pela FINEP[4] que atende o Programa Rede Cegonha do Ministério da Saúde, cujo o objetivo é preservar a saúde da mãe e da criança, em especial nos primeiros anos de vida [5] [10].

Este artigo propõe e implementa um modelo baseado em ontologias e *Linked Data* que, usando dados clínicos e sociais do DATASUS, promove a integração de dados e calcula a probabilidade do risco de óbito materno e infantil para o GISSA. Com isso, o novo modelo fortalece a gestão de conhecimento, apoiando profissionais e tomadores de decisão no SUS.

A organização deste artigo é apresentada a seguir. Na seção dois, são apresentados os trabalhos relacionados que usam *Linked Data* para saúde. Na seção três é descrito o modelo proposto, sua arquitetura, integração de dados e cálculos dos riscos de óbito materno e infantil. Por fim, na secão quatro, a conclusão e as aspirações futuras deste artigo.

2 TRABALHOS RELACIONADOS

Nesta seção apresentamos alguns trabalhos que propõem soluções computacionais de suporte à tomada de decisão em sistemas de saúde.

Em [4] é apresentado um modelo de suporte à tomada de decisão na gestão de resíduos, baseado em raciocínio sobre regras e ontologias. As ontologias foram criadas a partir dos dados abertos de 30 empresas e representam a taxonomia de resíduos, classificando-os pelo grau de efeitos nocivos sobre o meio ambiente. Resultados mostram as melhores estratégias de gerenciamento de resíduos

[1]Sistema de Informações sobre Nascidos Vivos
[2]SUS eletrônico

[3]https://www.gissa.com.br/
[4]Financiadora de Estudos e Projetos

WebMedia '17, October 17–20, 2017, Gramado, Brazil

ACM ISBN 978-1-4503-5096-9/17/10…$15.00
https://doi.org/10.1145/3126858.3131606

com custo mínimo, aumentando a eficiência do sistema de gerenciamento de resíduos em Volgograd, Rússia.

O sistema proposto por [2], baseado em *Linked Data*, destina-se à seleção de métodos de tratamento de câncer. Ele faz uma integração dos dados dos hospitais e dados abertos no campo de ciência da vida, *Linked Life Data* (LLD), e os dispõem num espaço de dados global. Depois, ele usa um algorítmo de seleção para encontrar casos de tratamento de câncer com base na similaridade na classificação do paciente. Já o trabalho de [1] disponibiliza informações sobre infestação do mosquito Aedes Aegypti no município de Cuiabá, Brasil, através de *mashup*. Para isso o autor utiliza *Linked Open Data* e SPARQL. Os resultados desse trabalho, além de trazerem benefícios à comunidade, através de informações de saúde pública, proporcionam uma ferramenta de auxílio aos gestores na tomada de decisão nos casos de surtos epidemiológicos.

Em [3], foi desenvolvida uma aplicação que usa dados integrados de informações sobre medicamentos. As fontes foram selecionadas de acordo com as necessidades de conhecimento dos médicos. Para integrá-las, foram usados os princípios *Linked Data* e Processamento de Linguagem Natural (PLN). Os resultados dessa aplicação implicaram a otimização do tempo do médico e uma ferramenta de suporte à tomada de decisão que ajuda a reduzir erros nas prescrições de medicamentos. Embora esse trabalho use os princípios *Linked Data*, ele não faz uso de ontologia de domínio ou de aplicação para representar o conhecimento que o médico necessita.

Diferente dos trabalhos relacionados aqui apresentados, este artigo apresenta uma ontologia de risco desenvolvida a partir das heurísticas de especialistas em saúde onde estão os fatores de riscos relevantes para o cálculo da probabilidade de óbito materno-infantil.

3 MODELO BASEADO EM ONTOLOGIA E *LINKED DATA*

Para a construção do nosso modelo, seguimos as especificações de materialização apresentadas em [5]. Essa materialização resulta num *mashup*, i.e., uma visão homogeneizada dos dados, do qual é utilizada para realizar inferências. A criação desse modelo envolve cinco etapas:

1. Selecionar as fontes de dados que alimentarão a aplicação.
2. Extrair e transformar os dados das fontes selecionadas, possivelmente heterogêneos, em grafos RDF.
3. Identificar *links* semânticos entre as fontes de dados.
4. Combinar e fundir representações do mesmo objeto em fontes distintas numa visão homogeneizada.
5. Realizar consultas parametrizadas a fonte de dados integrada usando o vocabulário da O_D e obter o cálculo da probabilidade do risco de óbito-infantil.

3.1 Arquitetura

O modelo proposto neste artigo está estruturado numa arquitetura de 5 camadas, exibida na Figura 1.

A camada Bases de Dados é formada pelas bases de dados SIM, e-SUS, SINASC E SINAN, todas disponibilizadas pelo do DATASUS. Cada base de dados db_i é descrita por uma ontologia fonte O_{dbi}, Figura 2. Na camada de Acesso e Transformação de Dados é realizado o mapeamento das bases de dados relacionais para RDF,

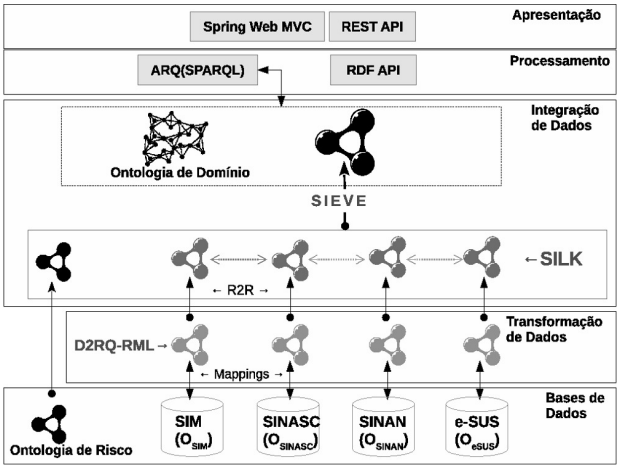

Figure 1: Arquitetura do modelo baseado em ontologia e *Linked Data.*

através dos *frameworks* D2RQ[5] e R2RML. Na camada de Integração dos Dados, a interligação das fontes RDF é realizada por *links* semânticos através do SILK - *Link Discovery Framework*. Na camada de Processamento de Dados é utilizada a linguagem SPARQL para realizar consultas parametrizadas à fonte de dados integrada. Na camada de Apresentação, uma aplicação web usa *dashboards* para exibir as informações inferidas pela camada de processamento de dados.

3.2 *Knowledge Base*

Modelos baseados em ontologias possuem, geralmente, base de conhecimento composta por ontologia de domínio, ontologias de aplicação e regras de inferência. Para a base de conhecimento do GISSA não foram especificadas regras de inferência. Foram criadas, como pode ser visto na Figura 2, uma ontologia de domínio O_D para representar uma demanda de governança e uma ontologia de risco O_{Risk} para representar as heurísticas dos especialistas.

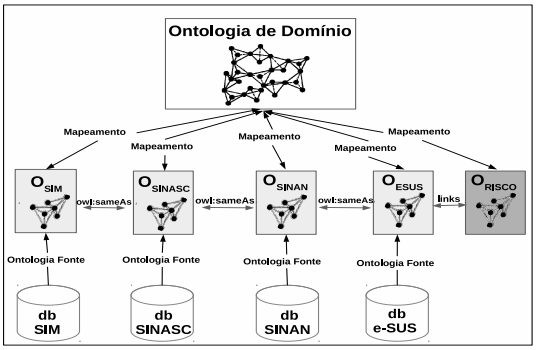

Figure 2: Mapeamento das ontologias.

[5] http://d2rq.org/

3.2.1 Ontologia de Domínio. A ontologia de domínio O_D trata-se de uma ontologia de referência. Ela especifica todos os conceitos necessários ao modelo. Essa ontologia contém o vocabulário geral para integrar os dados exportados em RDF através dos mapeamentos e *links* semânticos.

3.2.2 Ontologia de Risco (O_{Risk}). Essa ontologia, O_{Risk}, foi desenvolvida a partir das heurísticas dos especialistas em saúde materno-infantil (Figura 3). Ela é dividida em dois domínios: clínico e social. Nesses domínios, usando-se *Linked Data* na integração de dados, estão os fatores de risco relevantes para o cálculo da probabilidade para risco de óbito materno-infantil. Além disso, também são descritos os eventos relacionados a gestação e ao parto. Assim, a Ontologia de Risco O_{Risk} representa uma coleção de riscos, tais como, "uma mãe que tenha baixa escolaridade, que não recebe bolsa-família (riscos sociais), que teve rubéola (risco clínico)", etc. Levando em consideração que alguns tipos de riscos materno têm influência direta no bebê, essa correlação está representada na O_{Risk}. Por exemplo, se uma mãe teve rubéola ou tétano neonatal, se o parto foi induzido, se a gestação foi múltipla, então o risco de óbito do bebê aumenta consideravelmente. Essa ontologia tem 51 tipos de riscos e cada risco tem um peso. Esse peso foi definido por especialistas, mediante o relato de suas experiências e pesquisas, em conformidade com a gravidade do risco. Na figura 3 é apresentada parte da O_{Risk}, os riscos clínicos do bebê.

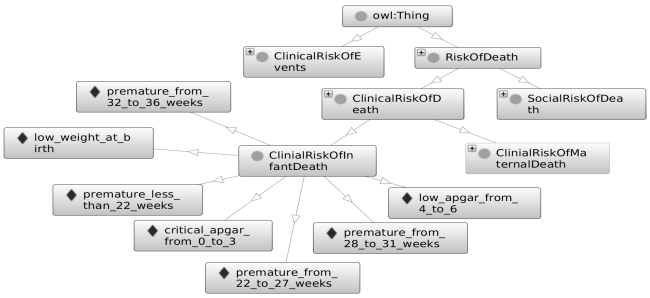

Figure 3: Parte da Ontologia de Risco.

3.3 Transformação dos Dados

Para transformar uma base de dados em grafo RDF foram utilizadas no GISSA duas ferramentas: i) R2RML, uma linguagem de mapeamento de dados relacionais para RDF. Definiu-se, de acordo com o padrão das triplas RDF, uma coluna-chave que identifica um registro para ser o URI do sujeito. Da base de dados SIM, por exemplo, usou-se a coluna "numerodo", e da base de dados SINASC, o atributo "numerodn". As demais colunas foram mapeadas para serem as propriedades do referido URI e o valor literal das colunas foi mapeado para ser o objeto na tripla; ii) D2RQ-server acessa as bases de dados através da ferramenta *generate-mapping*. Ele interpreta os mapeamentos do R2RML e gera os RDF populados.

3.4 Integração do Dados

Foi usada a SILK[6], uma linguagem de especificação de links no padrão XML, para se identificar os relacionamentos entre entidades dentro das fontes RDF. Utilizando-se heurísticas, foi verificada se existe uma relação semântica entre entidades para o processo de integração dos dados [11]. A Listing 1 representa a estrutura básica da especificação de *links* semânticos entre as fontes usadas na aplicação GISSA. Na linha 3 são denominados todos os prefixos que referenciam as URI's das fontes de dados. Na linha 5 são configurados um `<DataSource>` para cada fonte de dados RDF_i e, dentro dele, os parâmetros `name="endpointURI"` e `name="graph"`.

Listing 1: Estrutura da especificação SILK

```
1  <?xml version="1.0" encoding="utf-8" ?>
2  <Silk>
3      <Prefixes ... />
4        ...
5      <DataSources ... />
6        ...
7      [<Blocking ... />]
8        ...
9      <Interlinks ... />
10       ...
11     [<Outputs ... />]
12       ...
13 </Silk>
```

Foram configuradas (linha 9) a *tag* `<LinkType>` com a propriedade **owl:sameAs**; a *tag* `<SourceDataset>` como a fonte origem; a *tag* `<TargetDataset>` como a fonte alvo. Para a *tag* `<LinkageRule>` foram passadas as regras de comparação através da propriedade `<Comparemetric= "levenshteinDistance" threshold= "1"`, comparando as entradas `<Input path= "?a/sim:dtnasc"/>` e `<Input path= "?b/sinasc:dtnasc"/>` por suas *labels*. Por fim, na linha 11 foram configurados os parâmetros `<Param name= "format" value= "ntriples"/>` para definir o tipo de saída do arquivo. A fusão dos dados fora realizada pelo *framework* SIEVE[6].

3.5 Cálculo dos Riscos

Como já comentado, e mostrado na Figura 3, o cálculo do risco de óbito infantil e materno foram divididos em dois domínios: clínico e social. Essa divisão permite identificar as principais causas de óbito em cada domínio, além de possibilitar que decisões e ações específicas sejam realizadas. O risco clínico de óbito infantil é o mais complexo, pelo fato de que alguns riscos clínicos presentes nas mães influenciam direta e imediatamente no risco do bebê. Além dos riscos clínicos da mãe, também foram analisados os riscos identificados nos eventos gestação e parto que impactam diretamente na vida do bebê, tais como, se o parto foi induzido ou não, se a gestação foi única, dupla ou múltipla, dentre outros. Uma mãe ou um bebê podem ser classificados em baixo, médio ou alto risco, considerando o cálculo do percentual de risco para cada indivíduo. As faixas de valores para classificar um indivíduo são: entre 0% e 10% corresponde a baixo risco, entre 10% e 20% o indivíduo é considerado em risco intermediário, e acima de 20% é considerado em alto risco. Elas levam em consideração a quantidade de riscos existentes em um indivíduo, visto que para se atingir o critério de alto risco são necessários vários riscos presentes em um indivíduo.

Em seguida é descrito como é realizado o cálculo do fator de risco de óbito materno e infantil nos domínios clínico e social. O cálculo do percentual de óbito materno considerando os fatores sociais é dado por:

[6]http://silkframework.org/

Definição 1. *∀ mãe* $\boldsymbol{M}$, *∃ um cojunto de riscos* $R_M = \{r_i, r_{i+1}, \cdots, r_k\} \in R_{TM} = \{r_i, r_{i+1}, \cdots, r_n\}$, *tal que,* $0 < i \leq k \leq n$, *onde* R_{TM} *são todos os riscos possíveis para M.*

Definição 2. *Cada fator de risco* $r_i \in R_{TM}$ *tem um* ***peso*** $0 \leq w \leq 20$.

Desta forma, o risco social total de $m \leftarrow M$, i.e. R_m, pode ser encontrado pela Equação 1.

$$RiscoSocialDaMae(m) = \sum_{i=1}^{k} f(r_i),$$
$$f(r_i) = \begin{cases} PesoDoRiscoSocial(r_i), & \text{se a mãe apresentar o risco social } r_i \\ 0, & \text{caso contrário} \end{cases} \quad (1)$$

Para obter o máximo dos riscos que poderiam está presentes em uma mãe, R_{TM}, calcula-se a Equação 2.

$$MaxRiscoSocialDaMae(m) = \sum_{i=1}^{n} PesoDoRiscoSocial(r_i),\ r_i \in R_{TM} \quad (2)$$

A probabilidade de óbito materno considerando os riscos sociais é dada pela Equação 3.

$$ProbabObitoRiscoSocial(m) = \frac{RiscoSocialDaMae(m)}{MaxRiscoSocialDaMae(m)} \quad (3)$$

Os mesmos cálculos foram feitos para encontrar a probabilidade de óbito materno observando os riscos dos fatores clínicos. Para se calcular o risco de óbito infantil foram considerados a influência de fatores de risco do ascendente do indivíduo e de eventos que o envolvem diretamente. A probabilidade de óbito infantil é encontrada pela Equação 4. O resultado de *RiscoDaMae(m)* envolve, a fusão entre os totais do risco social e risco clínico da mãe. E o resultado de *RiscoDosEventos(m)* compreende os riscos existentes no parto e gestação.

$$P^7 = \frac{RiscosDaCrianca(c) + RiscosDaMae(m) + RiscosDosEventos(m)}{MaxRiscoBebe(c) + MaxRiscoMae(m) + MaxRiscoEventos(m)} \quad (4)$$

Todos os cálculos deste modelo do GISSA foram feitos na camada de negócio. Para realizar consultas sobre os dados integrados, foi usada a API Jena, um *framework* Java para construir aplicações para Web Semântica e *Linked Data*. Ela usa protocolos SPARQL e o vocabulário da O_D nos *scripts* de consultas. Essas consultas parametrizadas com dados da mãe e do bebê à base integrada, retornam a probabilidade de óbito infantil por meio de métodos definidos para tal procedimento.

4 CONCLUSÃO

Este trabalho apresentou um modelo baseado em ontologias e *Linked Data Mashup* que integra bases de dados distintas e heterogêneas do SUS, baseado na metodologia desenvolvida em [5]. Este modelo fornece a probabilidade de risco de óbito materno e infantil, fornecendo indicadores a gestores de saúde pública. Paralelamente, um trabalho com objetivo similar feito usando mineração de dados [10] foi implementado no projeto GISSA. Atualmente, o GISSA busca a definição de seu modelo de inteligência onde estes dois trabalhos têm papel relevante. Apesar da probabilidade ter sido validada pelos especialistas em saúde, não existe ainda uma ferramenta matemática, como Matriz de Confusão, na validação de algoritmos para validar as ontologias neste trabalho, ou vice-versa. Assim, a expectativa é de que o modelo final de inteligência do *framework* GISSA seja espelhado em um modelo híbrido onde seja determinante o trabalho aqui apresentado.

[7] Probabilidade de óbito infantil

ACKNOWLEDGMENTS

Os autores agradecem à Lucelia Ribeiro[8] e Charlys Pinheiro[9], do Instituto Atlântico[10], e à Dra Ivana Barreto, da Fundação Oswaldo Cruz (Fiocruz), que muito contribuíram com as heurísticas dos riscos de óbito materno e infantil, bem como aos demais participantes e dirigentes do projeto GISSA. Agradecimentos especiais à Profa. Vânia Vidal que conduziu as pesquisa sobre ontologia. Este artigo foi apoiado pela FINEP e pela Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico, no âmbito do Programa de Incentivo à Interiorização e Inovação Tecnológica - BPI, FUNCAP, edital *n*○ 09/2015.

REFERENCES

[1] Patricia Graziely Antunes de Mendonça, Cristiano Maciel, and José Viterbo Filho. 2014. Visualizing Aedes Aegypti Infestation in Urban Areas: A Case Study on Open Government Data Mashups. In *Proceedings of the 15th Annual International Conference on Digital Government Research (dg.o '14)*. ACM, New York, NY, USA, 186–191. https://doi.org/10.1145/2612733.2612751

[2] J. Hu, H. Cai, B. Xu, and C. Xie. 2014. A Linked Data Based Decision Support System for Cancer Treatment. In *2014 Enterprise Systems Conference*. 39–44. https://doi.org/10.1109/ES.2014.15

[3] Jakub Kozák, Martin Nečaský, Jan Dědek, Jakub Klímek, and Jaroslav Pokorný. 2013. Linked Open Data for Healthcare Professionals. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services (IIWAS '13)*. ACM, New York, NY, USA, Article 400, 10 pages. https://doi.org/10.1145/2539150.2539195

[4] M. Kultsova, R. Rudnev, A. Anikin, and I. Zhukova. 2016. An ontology-based approach to intelligent support of decision making in waste management. In *2016 7th International Conference on Information, Intelligence, Systems Applications (IISA)*. 1–6. https://doi.org/10.1109/IISA.2016.7785401

[5] Gabriel Lopes, Vânia Vidal, and Mauro Oliveira. 2016. A Framework for Creation of Linked Data Mashups: A Case Study on Healthcare. In *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web (Webmedia '16)*. ACM, New York, NY, USA, 327–330. https://doi.org/10.1145/2976796.2988213

[6] Pablo N Mendes, Hannes Mühleisen, and Christian Bizer. 2012. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. ACM, 116–123.

[7] Luiz Odorico Monteiro de Andrade. 2012. Inteligência de Governança para apoio à Tomada de Decisão. *Ciência & Saúde Coletiva* 17, 4 (2012).

[8] Mauro Oliveira, Carlos Hairon, Odorico Andrade, Regis Moura, Claude Sicotte, JL Denis, Stenio Fernandes, Jerome Gensel, Jose Bringel, and Herve Martin. 2010. A context-aware framework for health care governance decision-making systems: A model based on the brazilian digital tv. In *World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a*. IEEE, 1–6.

[9] Solange Oliveira Rezende. 2003. *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

[10] Cristiano Silva, Joyce Quintino, Ronaldo Ramos, Odorico Monteiro, and Mauro Oliveira. 2017. LAÍS, um Analisador Baseado em Classificadores para a Geração de Alertas Inteligentes em Saúde, Victoria E. Herscovitz, Cesar A. Z. Vasconcellos, and Erasmo Ferreira (Eds.). XXXV Simpósio Brasileiro de Redes de Computadores (SBRC) - I Workshop de Computação Urbana (CoUrb), Belém, Pará, Brasil, 1–13.

[11] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. 2009. Silk-A Link Discovery Framework for the Web of Data. *LDOW* 538 (2009).

[8] Enfermeira e Mestre em Saúde Pública pela Universidade Federal do Ceará

[9] Bacharel em Telemática pelo IFCE e Analista de Sistemas no IA

[10] http://www.atlantico.com.br/