Integração Semântica das Bases de Dados do Município de São Paulo: Um Estudo de Caso com Anomalias Congênitas*

Débora Lina N. Ciriaco Pereira[1], **Renata Wassermann**[1]**(Orientadora),**
Lais Salvador [2]**(Orientadora)**

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (IME-USP)
São Paulo – SP – Brasil

[2]Universidade Federal da Bahia – Salvador – BA – Brasil

{dciriaco, renata}@ime.usp.br, laisns@dcc.ufba.br

Abstract. *The lack of semantic information is a big challenge, even in context-driven areas like Healthcare, characterized by established terminologies. Here semantic data integration is the solution to provide precise information and answers questions like: how many individuals diagnosed with a congenital anomaly live in the city of São Paulo? This case study will evaluate two ontology based data access frameworks (Linked Data Mashup and Ontop) regarding integration quality (precision, completeness, and processing time to answer competence questions), time dispended and the possibility of scalability and replicability. We have a partnership with the city of São Paulo's Health Department to access identified databases related to congenital anomalies.*

Resumo. *A ausência do uso de contexto nos dados tem sido um desafio mesmo para áreas com terminologias estabelecidas como a da saúde. A integração semântica dos dados é uma solução para obter informações mais precisas e responder perguntas como: Quantos indivíduos com anomalia congênita vivem no município de São Paulo? Este estudo de caso avaliará dois frameworks de acesso a dados baseado em ontologias (Linked Data Mashup e Ontop), quanto à qualidade das junções (precisão, cobertura e processamento para responder às questões de competência), ao tempo demandado e à possibilidade de escalabilidade e replicabilidade. A Secretaria Municipal de São Paulo disponibilizará o acesso às bases identificadas referentes a anomalias congênitas.*

1. Introdução

A área da saúde é conhecida por manter grandes repositórios de dados, glossários e padrões terminológicos. No entanto, atualmente os registros eletrônicos em saúde pertencentes ao setor público, referentes às bases de dados de estatísticas vitais e assistência à saúde, não se utilizam de tecnologias semânticas para interpretar e processar os dados por meio de contexto. A ideia da utilização e unificação do conteúdo semântico, principalmente através da *Web*, é difundida por [Bizer et al. 2009], onde o conteúdo está

*This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, CAPES proc. 88887.136422/2017-00, and FAPESP proc. 14/50937-1 and FAPESP proc. 15/24485-9.

integrado por meio de seus significados. Para acessar esse conteúdo surguiu o paradigma *Ontology Based Data Access* (OBDA). No OBDA uma ontologia define o esquema global de alto nível de um conjunto de dados (já existente, normalmente em bancos de dados relacionais) e promove um vocabulário para realização de buscas pelo usuário [Kontchakov et al. 2013].

Segundo a iniciativa *Open Data Barometer*[1], em 2016 o Brasil era o 18^o colocado no ranque de 144 países avaliados quanto à disponibilidade e qualidade dos dados públicos de seus municípios em áreas como: saúde, educação, finanças, legislação. Embora bem colocado e com diversas iniciativas de publicação dos dados[2], é perceptível a necessidade de melhora da qualidade das fontes, no que diz respeito à presença de licenças para uso aberto, identificadores-chave para os registros e formatos legíveis por máquinas. Este cenário demonstra a fragilidade semântica dos dados disponibilizados.

O Brasil possui cerca de 140 sistemas de informação em saúde, desenvolvidos pelo Ministério da Saúde para a notificação de eventos relacionados à atenção à saúde [Brasil et al. 2017]. Entretanto, por terem surgido a partir de demandas diferentes, os sistemas são independentes, não havendo um método fácil e automático para busca de registros interbases, não existindo em todas as bases o mesmo identificador único. Outra característica é o fato dos registros serem centrados em eventos, como procedimentos realizados durante uma internação ou medicamentos dispensados por um estabelecimento de saúde, e não no paciente. Somado a isso, a compreensão dos dados se dá pelo acesso a diversos dicionários de dados e codificações onde cada variável corresponde a uma codificação diferente, como a CID-10[3] (Código Internacional de Doenças) e o SIGTAP[4] (que possui o Catálogo de Procedimentos do Sistema Único de Saúde). Este panorama dificulta encontrar respostas a perguntas como: Quantos pacientes são atendidos pela rede de saúde? E, quais foram os procedimentos realizados pelo paciente na última internação?

A redundância e fragmentação das informações é evidenciada em temas como o do reporte das anomalias congênitas. Segundo [SMS-SP 2012], anomalias congênitas são malformações de órgãos ou partes do corpo durante o desenvolvimento intra-uterino, detectáveis ao nascer. Estas exibem manifestações clínicas diversificadas, desde pequenas alterações morfológicas até defeitos complexos de órgãos ou segmentos corporais, correspondendo ao conjunto de CIDs-10 Q00-Q99 e D18. O principal instrumento de registro desses casos é a Declaração de Nascido Vivo (DN), instrumento base do Sistema de Informações de Nascidos Vivos (SINASC)[5]. No entanto, seus dados também estão distribuídos por outras sete bases independentes: SIM[6] (de mortalidade), SIH[7] (de

[1]Open Data Barometer: `https://opendatabarometer.org/?_year=2016&indicator=ODB`

[2]Algumas iniciativas nacionais para a disponibilização de dados:
Portal Brasileiro de dados abertos: `http://dados.gov.br/`
Governo Aberto SP: `http://www.governoaberto.sp.gov.br/`
Dados Abertos, Prefeitura de São Paulo: `http://dados.prefeitura.sp.gov.br/`

[3]CID-10: `http://www.cid10.com.br/`

[4]SIGTAP: `http://sigtap.datasus.gov.br/`

[5]SINASC: `http://datasus.saude.gov.br/sistemas-e-aplicativos/eventos-v/sinasc-sistema-de-informacoes-de-nascidos-vivos`

[6]SIM: `http://datasus.saude.gov.br/sistemas-e-aplicativos/eventos-v/sim-sistema-de-informacoes-de-mortalidade`

[7]SIH: `http://datasus.saude.gov.br/sistemas-e-aplicativos/hospitalares/`

informações hospitalares), TANU (de triagem auditiva neonatal), bases de gestão do sistema de saúde (GSS e SIGA) e de produção ambulatorial (BPA-I e APAC). Isso resulta em pacientes presentes na base do TANU, mas não na do SINASC, por exemplo. Ter as informações dispersas entre as bases significa não saber ao certo quantos pacientes foram diagnosticados e quais estão em tratamento. A falta de informações leva à sub-notificação, o que impede a elaboração de indicadores demográficos e de saúde acurados, bem como o melhor desenvolvimento de sistemas de vigilância e estabelecimento de políticas de saúde [SMS-SP 2012].

Ao inquirir soluções, é observado que a compreensão das terminologias utilizadas para o preenchimento dos registros e a necessidade de análises extremamente contextualizadas dificultam iniciativas de profissionais de fora da área da saúde. Por outro lado, lidar com o tamanho das bases e a complexidade computacional que as permeiam é um fator limitador para os profissionais da saúde. Isto é agravado quando a população alvo é a do município de São Paulo, que em 2010 era de mais de 11 milhões de habitantes[8], equiparando a complexidade de seu sistema de saúde a uma unidade da federação. Deste modo não há relatos do desenvolvimento de uma solução de integração de bases de dados escalável para todo o município.

Para o desenvolvimento deste projeto de mestrado foi reconhecida a necessidade de uma equipe multidisciplinar. Assim, foi realizada uma parceria com a Secretaria Municipal de Saúde de São Paulo (SMS-SP), que: disponibilizará o acesso às bases de dados identificadas (SINASC, SIM e SIH) com informações referentes às anomalias congênitas; auxiliará na compreensão dos conceitos e na eleição das questões de competências utilizadas para validar as ontologias. As questões de competência elencadas até agora foram:

- Qual a prevalência de anomalias congênitas na cidade de São Paulo?
- Quem são os pacientes com anomalia congênita? Quais são suas características?
- Qual o caminho de cuidado a que cada paciente tem se submetido?

2. Proposta e Objetivos

Este estudo de caso visa, a longo prazo, integrar as bases de dados identificadas, relacionadas às anomalias congênitas, SINASC, SIM e SIM, pertencentes à Secretaria Municipal de Saúde de São Paulo. No entanto, para isso é necessário eleger uma metodologia escalável para o problema. Assim, inicialmente haverá a comparação de dois *frameworks* de OBDA: Ontop e *Linked Data Mashup* (LDM). Estes serão medidos quanto à precisão, cobertura e tempo de processamento ao responder às questões de competência e às capacidades de escalabilidade e replicabilidade. O *framework* Ontop[9], desenvolvido pela Free University of Bozen-Bolzano, foi escolhido por fazer parte do estado da arte. A seleção do LDM [Vidal et al. 2015], por sua vez, ocorreu por ter sido empregado em uma prova de conceito para integração das bases de dados e-SUS (que contém dados clínicos dos indivíduos) e SINASC no município de Tauá-CE [Lopes et al. 2016], cuja população em 2010 era de 55.716 habitantes[10], possuindo assim características similares ao presente trabalho, embora em menor escala.

---

sihsus

[8]IBGE, São Paulo, SP https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama

[9]Framework Ontop: http://ontop.inf.unibz.it/

[10]IBGE, cidade de Tauá-CE: https://cidades.ibge.gov.br/brasil/ce/taua/panorama

Os objetivos específicos são:

- Mapear as terminologias utilizadas nas bases de dados SINASC, SIM e SIH;
- Mapear as similaridades entre os dois *frameworks*: Ontop e LDM;
- Criar as ontologias, contendo as terminologias e a estrutura das bases de dados;
- Integrar o conteúdo das bases de dados SINASC e SIM através do número da declaração de óbito;
- Integrar o conteúdo das bases de dados SINASC e SIH através das variáveis pessoais, tais como: nome, sexo e data de nascimento;
- Realizar consultas através das ontologias;
- Avaliar a qualidade da integração;
- Extrair um método de trabalho;
- Disponibilizar os dados não identificados, já públicos, em formato RDF;

3. Metodologia

O desenvolvimento do projeto se dará seguindo duas etapas. Até que ocorra a aprovação para uso dos dados identificados por meio dos comitês de ética do IME e da SMS-SP, serão utilizados os dados públicos sem identificação. Nesta fase se dará o processo de criação das ontologias seguindo os dois *frameworks* a serem testados. Com isso, é esperado que todas as terminologias relacionadas às bases de dados estejam em formato ontológico para o recebimento dos dados identificados. Também é prevista a publicação das bases de dados em formato RDF. Na segunda etapa, iniciada após as autorizações, serão finalizados os mapeamentos semânticos, realizadas as consultas de acordo com as questões de competência e analisados os dois *frameworks*, como pode ser observado na Tabela 1.

Tabela 1. Apresentação da metodologia

<table>
<tr><td rowspan="2">Aprovações</td><td rowspan="2">Etapas</td><td rowspan="2">Framework Ontop</td><td colspan="3">Framework LDM</td></tr>
<tr><td>Especificação</td><td>Materialização</td><td>Visão</td></tr>
<tr><td rowspan="5">Fase pré-aprovação</td><td>Edição do documento para o comitê de ética</td><td colspan="4">–</td></tr>
<tr><td>Utilização de dados públicos</td><td colspan="4">Questões de competência + Disponibilização dos dados em LOD</td></tr>
<tr><td>Construção das ontologias</td><td>X</td><td>X</td><td>X</td><td>–</td></tr>
<tr><td rowspan="2">Realização dos mapeamentos</td><td rowspan="2">X</td><td rowspan="2">X</td><td rowspan="2">X</td><td rowspan="2">–</td></tr>
<tr><td rowspan="3">Fase pós-aprovação</td></tr>
<tr><td>Realização das consultas</td><td>X</td><td>–</td><td>–</td><td>X</td></tr>
<tr><td>Análise das respostas</td><td colspan="4">Fase de avaliação</td></tr>
</table>

3.1. Fases de modelagem pelo *framework Linked Data Mashup*

O *framework* LDM integra as bases de dados percorrendo três etapas:

- **Especificação de um LDM:** É a etapa do processo ligada ao mapeamento e projeção da integração semântica. Aqui, serão selecionados os conjuntos de dados, modeladas as ontologias de domínio, escolhidos os conceitos essenciais que unirão as ontologias, especificados os links semânticos e as regras de fusão do conjunto de dados.
- **Materialização de um LDM:** É na materialização onde de fato ocorrerá a criação das ontologias e a integração semântica. Nesta etapa, as bases de dados serão traduzidas em ontologias através do *plugin* R2RML *processor*[11] e serão acessadas a partir do banco de dados pelo *plugin* D2RQ[12]. Os links semânticos serão criados e importados para a ontologia por meio do SILK[13]. Por fim, as regras de fusão materializarão a integração e avaliarão sua qualidade através do *plugin* SIEVE[14].
- **Visão de aplicação de um LDM:** Uma vez criadas as ontologias e realizadas as ligações semânticas será possível realizar buscas específicas nas bases. Nesta etapa serão selecionados os conceitos e filtrados os dados para fazer consultas de acordo com as questões de competência.

3.2. Fases de modelagem pelo *framework* Ontop

O *framework* Ontop engloba as fases de *linkage* das bases de dados utilizando apenas o software de modelagem de ontologias Protégé[15] e seu *plugin* neste software [Calvanese et al. 2017]. Assim, para obter os dados integrados será necessário seguir os estágios de:

- **Construção da ontologia:** Os conceitos já presentes nas bases de dados serão modelados em ontologias correspondentes a cada base. Em seguida, será construída uma ontologia de ligação que será utilizada pelo *plugin* do Ontop no Protégé para realizar a integração dos dados.
- **Integração dos dados:** Para realizar a integração dos dados será necessário realizar uma conexão com a base de dados e executar o mapeamento dos links semânticos por meio do assistente de mapeamentos do Ontop.
- **Consultas SPARQL[16]:** As consultas feitas a partir das questões de competência serão realizadas na própria interface do Protégé para consultas em SPARQL.

3.3. Avaliação dos *frameworks*

Uma vez triplificadas as bases, os *frameworks* serão avaliados. A análise considerará a qualidade das junções quanto à precisão, cobertura e tempo de processamento ao responder às questões de competência e as capacidades de escalabilidade e replicabilidade dos métodos.

Em uma comparação não metodológica, é sabido que o *framework* LDM possui a facilidade da descrição do método e de ter sido testado em duas bases de dados do Sistema Único de Saúde. No entanto, isso ocorreu com poucos registros, dado o tamanho do município em que o caso de uso foi realizado. Por outro lado, o *framework* Ontop é

[11]R2RML: `https://www.w3.org/TR/r2rml/`
[12]D2RQ: `http://d2rq.org/`
[13]SILK: `http://silkframework.org/`
[14]SIEVE: `http://sieve.wbsg.de/`
[15]Protégé: `https://protege.stanford.edu/`
[16]SPARQL `https://www.w3.org/TR/rdf-sparql-query/`

completamente integrado, não sendo necessário utilizar *plugins* dispersos. Entretanto não houve o teste para bases de dados como as que serão integradas neste trabalho, embora em sua documentação haja um apelo à escalabilidade.

4. Contribuições

As contribuições propostas para este projeto estão relacionadas à comparação de dois *frameworks* que utilizam conceitos de OBDA para e a integração semântica dos conteúdos das bases de dados SINASC, SIM e SIH. Assim, é esperado realizar:

- Comparação dos *frameworks* de OBDA: LDM e Ontop, mapeando as similaridades e extraindo um método de trabalho;
- Integração das bases de dados identificadas SINASC e SIM, SINASC e SIH da cidade de São Paulo pertencentes à SMS-SP, para indivíduos cujos CIDs-10 sejam iguais a Q00-Q99 ou D18, referentes às anomalias congênitas;
- Disponibilização das bases de dados SINASC, SIM e SIH, não identificadas do município de São Paulo em formato RDF.

Com a integração semântica das três bases de dados (SINASC, SIM e SIH) é esperado poder seguir os pacientes por todo o fluxo de cuidado, localizá-los, aumentar a quantidade de informações sobre um indivíduo, melhorar o acompanhamento clínico e aumentar a confiabilidade dos dados, acarretando num auxilio na tomada de decisão dos gestores de saúde. Por outro lado as ontologias desenvolvidas poderão ser reutilizadas em outras cidades que utilizam as mesmas bases.

Referências

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

Brasil, da Saúde, M., and da Estratégia e Saúde, C. G. (2017). Estratégia de e-saúde para o brasil. Online, `http://portalarquivos.saude.gov.br/images/pdf/2017/julho/12/Estrategia-e-saude-para-o-Brasil.pdf`.

Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., and Xiao, G. (2017). Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, 8(3):471–487.

Kontchakov, R., Rodriguez-Muro, M., and Zakharyaschev, M. (2013). Ontology-based data access with databases: A short course. In *Reasoning web: semantic technologies for intelligent data access*, pages 194–229. Springer.

Lopes, G., Vidal, V., and Oliveira, M. (2016). Construção de linked data mashup para integração de dados da saúde pública. In *SBBD*, pages 145–150.

SMS-SP, S. M. d. S. d. S. P. (2012). *Declaração de Nascido Vivo - Manual de Anomalias Congênitas*. São Paulo: Secretaria Municipal da Saúde.

Vidal, V. M., Casanova, M. A., Arruda, N., Roberval, M., Leme, L. P., Lopes, G. R., and Renso, C. (2015). Specification and incremental maintenance of linked data mashup views. In *International Conference on Advanced Information Systems Engineering*, pages 214–229. Springer.