Integração Semântica das Bases de Dados do Sistema Único de Saúde: Um Estudo de Caso com o Município de São Paulo*

Débora Lina Ciriaco[1], Alexandre Pessoa[1], Laís Salvador[2], Renata Wassermann[1]

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

[2]Departamento de Ciência da Computação – Universidade Federal da Bahia
Salvador – BA – Brasil

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

Abstract. *Database integration plays a crucial role in understanding the healthcare domain. In this work, a semantic integration methodology was adapted to the birth and mortality database systems of the Brazilian Unified Health System. A case study was carried out to develop a health indicator in the context of maternal and child health in the municipality of São Paulo. Three layers of ontologies were created for the solution, containing specifications and mappings. The ontologies were populated and evaluated for their ability to answer competence questions elected by domain experts. The solution proved to be useful in the integration process, presenting a global view of the data and their relationships.*

Resumo. *Integração de bases de dados tem um papel crucial para compreender o domínio da saúde. Neste trabalho, uma metodologia de integração semântica foi adaptada para as bases do sistemas de nascimento e mortalidade do Sistema Único de Saúde. Foi realizado um estudo de caso voltado ao desenvolvimento de um indicador de saúde no contexto da saúde materno-infantil do município de São Paulo. Foram criadas três camadas de ontologias para a solução, contendo especificações e mapeamentos. As ontologias foram povoadas e avaliadas quanto à capacidade de responder às questões de competência eleitas pelos especialistas do domínio. A solução mostrou-se útil no processo de integração apresentando uma visão global dos dados e seus relacionamentos.*

1. Introdução

No Brasil, a informatização dos sistemas de informação de saúde começou antes do surgimento do próprio Sistema Único de Saúde - SUS, que é de 1990. O Departamento de Informática do SUS - DATASUS, ligado ao Ministério da Saúde criou e mantém, ao longo dos 27 anos de existência, mais de 140 sistemas ligados a gestão de saúde no país. A maior parte dos sistemas está relacionada à notificação de eventos do cuidado, tais como nascimento - SINASC, óbito - SIM, número de vacinas ministradas e notificação de

*Esta pesquisa é parte do INCT da Internet do Futuro para Cidades Inteligentes financiado pelo CNPq proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, FAPESP proc. 14/50937-1, e FAPESP proc. 15/24485-9.

agravos, como AIDS, sífilis, tuberculose, dengue, entre outros. Há também sistemas voltados para a gestão do gasto público, portanto restritos aos eventos financiados pelo SUS, tais como os sistemas para notificação de internações hospitalares - SIH e atendimentos ambulatoriais - SIA [SIS 2015].

Por terem surgido a partir de demandas diferentes, os sistemas são independentes, não havendo um método fácil e automático para busca de registros inter-bases. Apesar das iniciativas recentes, também não há o mesmo identificador único em todas as bases. Outro fator dificultante numa análise global é o fato dos registros não estarem centrados nos pacientes mas em eventos, como procedimentos realizados durante uma internação ou medicamentos dispensados [SIS 2015]. A esse quadro é somado o vasto vocabulário utilizado, exigindo conhecimento do domínio para compreensão dos dados.

Muitos desses sistemas têm seus dados disponibilizados no site do DATASUS[1] e podem ser adquiridos em formato CSV a partir de uma interface online de filtragem, o TABNET[2]. Caso seja necessária a aquisição da base completa o acesso se dá por outra página, a de arquivos de dados[3]. Nela, os arquivos são disponibilizados em formato DBF ou DBC. Os dados disseminados publicamente são devidamente anonimizados a fim de não comprometer a identidade dos pacientes e profissionais que atuaram no atendimento. Como consequência, a tarefa de ligação dos registros das bases, que depende do acesso aos dados identificados, deve estar relacionada a um órgão mantenedor dos registros originais, geralmente as Secretarias Municipais da Saúde.

Dado o contexto de desafios computacionais e do domínio da saúde, foi perceptível a necessidade de desenvolver a solução dentro de uma equipe multiprofissional. Assim, para a realização desta pesquisa, foi formada uma parceria com a Secretaria Municipal da Saúde de São Paulo - SMS-SP e com a Faculdade de Saúde Pública da Universidade de São Paulo - FSP-USP. Com o auxílio da equipe de especialistas (composta por aproximadamente 20 profissionais, entre especialistas nas bases de dados e técnicos da SMS-SP, professores e pós-graduandos da FSP-USP) foi possível compreender o domínio e levantar questões. Os especialistas guiaram a escolha das bases de dados (SINASC e SIM), do caso de uso (saúde materno-infantil voltada para o desenvolvimento do indicador de saúde DPGP, utilizando os dados do município de São Paulo) e auxiliaram no processo de validação da solução.

O desenvolvimento do indicador de saúde se dá no contexto do projeto: "Dias Potenciais de Gravidez Perdidos (DPGP): uma medida inovadora da idade gestacional, para avaliar intervenções e resultados de saúde materno-infantil". O projeto busca desenvolver um indicador de saúde que medirá a perda no tempo de gestação, comparando-a com o tempo de gestação esperado [Diniz et al. 2019]. No entanto, a elaboração do indicador depende da integração de algumas bases de dados do SUS, dentre elas as de natalidade - SINASC e de mortalidade - SIM, da compreensão das variáveis e de como se dá o seu preenchimento. A necessidade de técnicas de integração que levam em conta o uso de contexto foi determinante na escolha deste estudo de caso.

---

[1]DATASUS: `http://www2.datasus.gov.br/DATASUS/index.php?area=02`

[2]TABNET: `http://www2.datasus.gov.br/DATASUS/index.php?area=060804`

[3]Arquivos de dados DATASUS: `http://www2.datasus.gov.br/DATASUS/index.php?area=0901`

A abordagem técnica utilizada para a realização da integração semântica de bases de dados foi influenciada pelos trabalhos de [Ekaputra et al. 2017, Ferronato et al. 2016, Ristoski and Paulheim 2016, Stoilos et al. 2018b, Stoilos et al. 2018a, Zhang et al. 2007, Bauer et al. 2016], mas foram os trabalhos de [Vidal et al. 2015, Lopes et al. 2016] que mais trouxeram contribuições. Neles, os autores desenvolveram um *framework* baseado em ontologias para a especificação formal de visões de dados integrados, seguindo uma metodologia de integração de bases de dados por meio de ontologias - OBDI (*Ontology-Based Data Integration*).

O presente estudo tem como objetivo realizar a integração semântica das bases de dados dos sistemas de natalidade - SINASC e mortalidade - SIM do SUS, ambas relacionadas à saúde materno-infantil e ao desenvolvimento do indicador de saúde DPGP - Dias Potenciais de Gravidez Perdidos. A integração foi testada com os dados do município de São Paulo, advindos tanto do site do DATASUS (dados não integrados do SINASC e do SIM) quanto da parceria com a SMS-SP e com a FSP-USP (dados do SINASC e do SIM *linkados* e anonimizados, nomeados de DNDO), filtrando as variáveis que são úteis para o problema da pesquisa. Para isso, foi eleita e adaptada uma metodologia de OBDI. O presente trabalho tem a aprovação do comitê de ética da Faculdade de Saúde Pública da USP, com o parecer de número 2.958.248.

A seção 2 descreve a metodologia de desenvolvimento do sistema de integração semântica, incluindo o protocolo utilizado para a escolha da metodologia e arquitetura do sistema. A seção 3 contém a metodologia e a discussão do desenvolvimento das ontologias e mapeamentos que são independentes do domínio central da aplicação. Nesse caso, são apresentadas as ontologias das bases do SINASC e do SIM. Já a seção 4 apresenta a discussão e o desenvolvimento das ontologias dependentes de contexto, onde, as ontologias do SIM e do SINASC no contexto da saúde materno-infantil relacionada ao indicador de saúde DPGP. Finalmente, a seção 5 traz as considerações finais e os trabalhos futuros.

2. Metodologia de desenvolvimento da solução

Para o desenvolvimento da solução de integração, as fontes de dados (SINASC e SIM) e variáveis foram escolhidas, bem como compreendido o contexto em que o sistema seria utilizado. Em seguida as metodologias de ODBI foram analisadas a fim de encontrar a mais adequada para o projeto.

Em [Wache et al. 2001] e [Ekaputra et al. 2017] são apresentadas as principais metodologias de OBDI. Quanto à arquitetura, elas são: i) ontologia única, onde um vocabulário global é construído e, a partir dele, são realizados mapeamentos para as fontes de dados; ii) múltiplas ontologias, onde uma ontologia local é desenvolvida para cada fonte de dado e são criados mapeamentos semânticos entre as ontologias para a integração dos vocabulários; iii) ontologia híbrida, onde os mecanismos anteriores são utilizados, havendo a construção do vocabulário compartilhado numa ontologia global e a partir dele ocorre a criação das ontologias locais; e iv) GAV, baseada na abordagem de integração de dados relacionais *Global-as View*, as ontologias locais e a ontologia global são desenvolvidas, mas de maneira independente em termos de vocabulário.

Segundo os autores, todas as abordagens possuem pontos negativos e positivos, desse modo, sua adoção dependerá da complexidade e organização das fontes de dados, bem como do tempo disponível para o desenvolvimento da solução e de como

ela será utilizada. Desse modo, seguindo a árvore de recomendação apresentada em [Ekaputra et al. 2017], Figura 1, foi constatado que a metodologia de OBDI híbrida era a mais adequada para o problema de pesquisa, uma vez que comporta bem a adição de novas fontes de dados, algo desejado a médio prazo, como discutido em [Pereira 2019]. Ao longo do desenvolvimento, foi notado que a preservação do vocabulário das fontes de dados era importante, principalmente para a reutilização das ontologias em diferentes contextos, levando assim a revisão da metodologia mais adequada, sendo adotada a GAV.

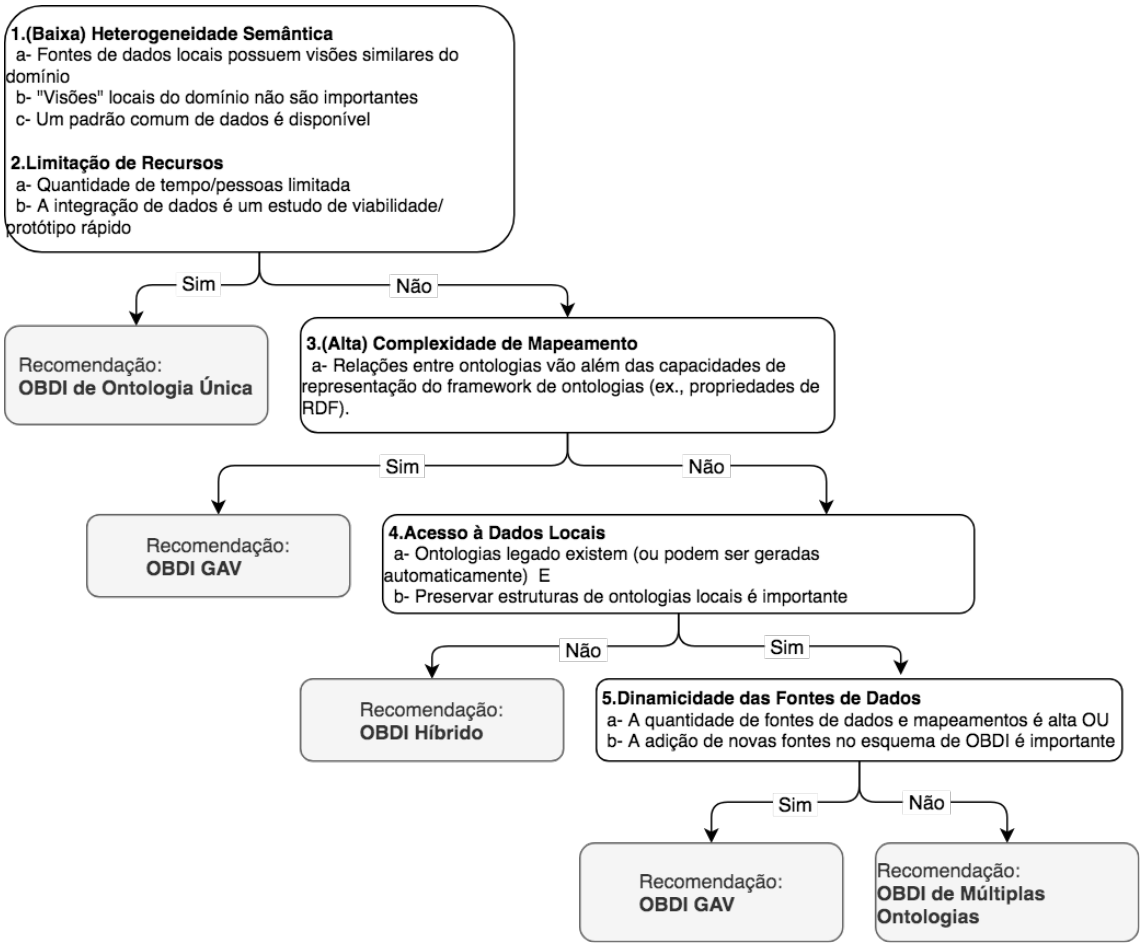

Figura 1. Árvore de Recomendação de abordagens de OBDI criada por [Ekaputra et al. 2017], adaptada pelos autores

Seguindo a árvore, temos os cinco conjuntos de decisão e seus argumentos, apresentados de acordo com o contexto do projeto e suas fontes de dados:

1. Baixa Heterogeneidade Semântica:
   (a) As fontes de dados não possuem visões similares do domínio, embora possuam conceitos em comum. Cada base de dados possui a sua visão para o evento, a do SINASC apresenta o evento do nascimento e a do SIM, o evento do óbito;
   (b) As "visões" locais do domínio são importantes pois cada profissional tem sua abordagem sobre o assunto;
   (c) As bases compartilham de um padrão comum de dados pois possuem dicionários de metadados similares;

O último item foi o único ponto que apontou para a escolha da abordagem OBDI de Ontologia Única, enquanto que os dois itens anteriores levavam a necessidade de continuar as análises.

2. Limitação de Recursos:
    (a) O tempo e a quantidade de pessoas, principalmente a equipe de especialistas do domínio, foram apresentados como adequados às soluções mais complexas;
    (b) A criação e evolução da integração dos dados não seria um protótipo rápido, devido a quantidade de bases de dados e variáveis, bem como a complexidade terminológica delas;

   As respostas das decisões 1 e 2 levam para a decisão 3.
3. Alta complexidade de Mapeamento:
    (a) Apesar da complexidade terminológica, as relações entre ontologias não vão além das capacidades de representação do *framework*;

   Esta resposta leva para a decisão 4.
4. Acesso a Dados Locais:
    (a) Não existem ontologias legadas, mas as ontologias das fontes de dados poderiam ser geradas automaticamente se existisse uma estrutura relacional dos dados. Como não houve acesso a essa estrutura foi necessário recriá-la e adaptá-la manualmente;
    (b) Foi entendido que era importante preservar a estrutura local das ontologias que representam diretamente o esquema das bases de dados. Afinal, muitos profissionais utilizam as fontes de dados com a estrutura original, embora, historicamente essa estrutura seja restrita a eventos, não modelando bem o domínio;

   Com essas respostas houve um impasse, sendo um ponto positivo para a escolha da solução ODBI Híbrido e outro para seguir na árvore de recomendação. Foi decidido então por seguir na árvore.
5. Dinamicidade das Fontes de Dados:
    (a) A quantidade de mapeamentos entre as fontes de dados é alta;
    (b) A adição de novas fontes de dados é importante para os próximos passos do projeto.

   Ao responder este passo a recomendação foi de seguir a abordagem OBDI GAV.

No entanto, apesar das indicações, foi necessário realizar ajustes na arquitetura, adicionando mais uma camada de ontologias, a fim de auxiliar no processo de harmonização de vocabulários e de integração dos dados, como apresentado em [Pinheiro 2011]. Isso também ocorreu devido ao tamanho e complexidade das fontes de dados. Desse modo, levando em consideração o impasse na árvore de recomendação, a metodologia desenvolvida é uma combinação entre as metodologias OBDI Híbrido e OBDI GAV, como apresentado na Figura 2. Ela apresenta o diagrama das suas abordagens originais e os compara com o diagrama da abordagem desenvolvida. As setas vazadas apontam as camadas similares entre as abordagens. Ao considerar a abordagem desenvolvida como híbrida, nota-se a similaridade entre o Vocabulário Compartilhado (V) e a Ontologia de Domínio (OD) e em como seus vocabulários são herdados pela camada inferior de ontologias. No entanto, na abordagem desenvolvida há a adição da camada de ontologias sobre as fontes de dados, cujo vocabulário é independente da ontologia de

domínio, característica da abordagem GAV. Por outro lado, ao considerá-la como parte da GAV, a adição da camada de ontologias de integração (I_n) que tem como vocabulário base a ontologia de domínio, a difere da arquitetura usual.

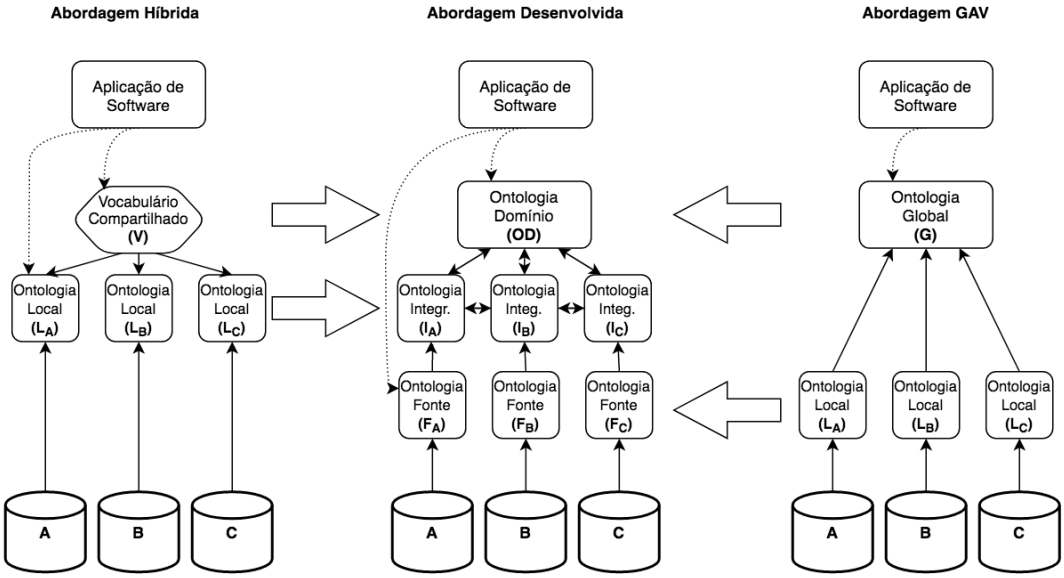

Figura 2. Comparação entre a abordagem de OBDI desenvolvida e as abordagens híbrida e GAV. As setas vazadas indicam as camadas equivalentes nas abordagens.

Assim, a arquitetura que compõe a solução proposta possui as seguintes camadas, apresentadas na Figura 2 como "Abordagem Desenvolvida"e detalhada na Figura 3:

- Camada de Fontes de Dados: que recebe diversos formatos de dados materializados ou virtualizados, integrados ou não. No estudo de caso, as fontes utilizadas foram os dados públicos não integrados do SINASC e do SIM e os dados já integrados do SINASC e do SIM num arquivo nomeado DNDO, fornecido pela SMS-SP;
- Camada de ontologias de Fonte (OF)[4]: construída a partir da replicação do vocabulário das fontes de dados. Essa camada auxilia na harmonização dos formatos de dados de entrada do sistema de integração, gerando ontologias que podem ser reutilizadas em outros contextos. Foram criadas as ontologias OF_{SINASC}, OF_{SIM} e OF_{DNDO}, cada uma originada da fonte de dados correspondente;
- Camada de Ontologias de Integração (OI)[5]: criada para realizar a integração do vocabulário contido nas camadas OF e OD, apresentando relações entre os vocabulários. Ela contém o vocabulário da camada OD distribuído em suas ontologias e mapeia os conceitos correspondentes da camada OF para a OD. Nela também são

[4]Sobre a camada de Ontologias de Fonte: Considerando a similaridade com a abordagem GAV, no trabalho de [Ekaputra et al. 2017] essa camada é chamada de camada de ontologias locais. Já em [Vidal et al. 2015] essa camada é chamada de camada de dados

[5]Sobre a camada de Ontologias de Integração: Em [Vidal et al. 2015] essa camada é chamada de camada de visões Exportadas e Links Semânticos. Em [da Cruz et al. 2019] os autores nomeiam de camada de ontologias locais e em [Barisevičius et al. 2018] de camada de base de conhecimento. Ao considerar essa abordagem como híbrida, parte do que ocorre na camada OI é similar ao que ocorre na camada de ontologias locais, de [Ekaputra et al. 2017]

estabelecidas as relações de similaridade entre conceitos (sinônimos como os conceitos gravidez e gestação) e entre entidades (estabelecimento da mesma entidade paciente em fontes de dados distintas). As ontologias criadas foram: OI_{SINASC}, OI_{SIM}, OI_{DNDO};

- Camada de Ontologia de Domínio (OD)[6]: é a camada ontológica que estabelece o vocabulário para uma determinada aplicação, onde as consultas são realizadas, sendo composta por uma única ontologia. Foi criada a ontologia OD_{SMI}, relacionada ao domínio da saúde materno-infantil;
- Camada de Acesso às Aplicações, porta para o acesso aos resultados das consultas realizadas nas camadas OD e na OF.

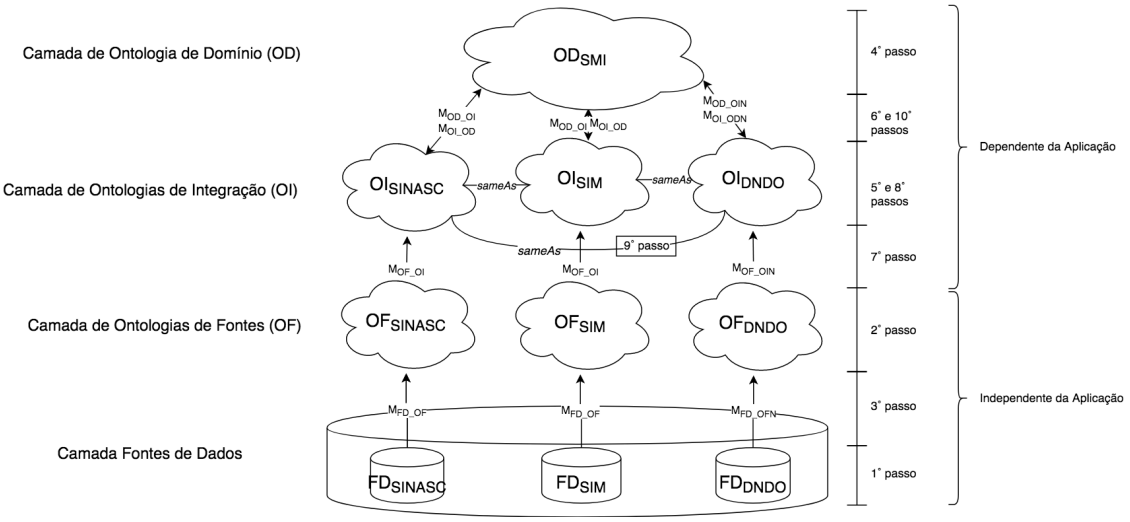

Figura 3. Arquitetura da solução OBDI desenvolvida.

A Figura 3 também contém a arquitetura da solução. A imagem de pilha corresponde às fontes de dados; o ícone de nuvem, às ontologias, nomeadas de acordo com suas camadas correspondentes OX_1 até OX_N; os arcos, indicam a presença do mesmo elemento entre ontologias; e as setas, nomeadas de acordo com a origem e o destino do mapeamento $M_{Origem\_Destino}$, representando a direção do mapeamento dos conceitos nas ontologias. A mesma gravura ainda apresenta a ordem de criação de cada camada, ilustrada pelos passos ordenados, e se a camada é dependente ou independente de uma aplicação específica. A parte independente da aplicação pode ser reutilizada em outros contextos enquanto que a dependente possui vocabulário próximo ao domínio (no estudo de caso ao de saúde materno-infantil no contexto do indicador DPGP) tornando-se mais especializada.

3. Desenvolvimento das ontologias e mapeamentos independentes da aplicação

Essa seção apresenta o desenvolvimento das ontologias da camada de Ontologias de Fonte (OF) e a caracterização e mapeamentos das fontes de dados para as ontologias dessa

[6]Sobre camada de Ontologia de Domínio: Em [Vidal et al. 2015] essa camada é chamada de camada de Integração Semântica. Em [da Cruz et al. 2019], os autores nomeiam de camada de ontologia de domínio e em [Ekaputra et al. 2017] de camada ontológica global. Já em [Barisevičius et al. 2018] a camada é referida como camada da ontologia de alto nível - *upper level ontology*

camada. As fontes de dados disponíveis para o desenvolvimento da solução estavam em formato de planilha eletrônica. No entanto, foi constatado que estas já possuíram uma estrutura de dados relacional e que os dados eram capturados a partir de fichas de preenchimento (Declaração de Nascido Vivo - DN, para o SINASC e Declaração de Óbito - DO para o SIM) com campos agrupados em blocos temáticos.

Diante desse histórico, a construção das ontologias de fonte, OF_{SINASC} e OF_{SIM}, seguiu o algoritimo de conversão de bases de dado relacionais em ontologias, de [Haw et al. 2017]. Para isso, o diagrama entidade relacionamento - DER das fontes de dados foi reconstruído, tendo como base os blocos das fichas de preenchimento. Cada bloco foi traduzido como uma entidade no DER e posteriormente, uma classe na ontologia. Cada variável presente nos blocos foi convertida em um atributo no DER e uma propriedade de dados, na ontologia.

A Figura 4 apresenta o bloco da DN referente à gestação e parto, a classe correspondente na ontologia OF_{SINASC}, e as propriedades de dados relacionadas a essa classe. A partir dos DERs foram criados os primeiros diagramas que representavam os conceitos e as relações que estariam presentes nas ontologias da camada OF, sendo considerado a primeira versão das classes e propriedades das ontologias. Esses diagramas também serviram para apresentar o fluxo de conceitos presente nos dados. O recorte do diagrama contendo a classe `GestaçãoEParto` é apresentado na Figura 5.

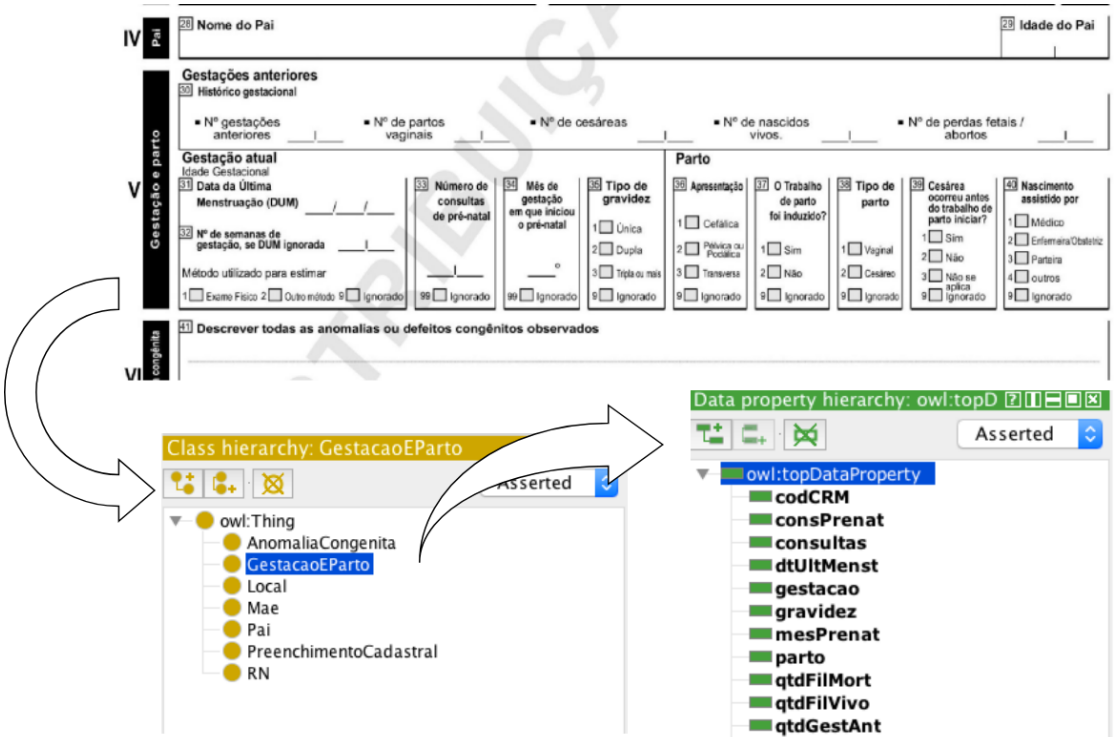

Figura 4. Conversão de um bloco da ficha de preenchimento do SINASC para uma classe na ontologia OF_{SINASC}.

O desenvolvimento da ontologia OF_{DNDO}, no entanto, baseou-se na importação das ontologias OF_{SINASC} e OF_{SIM} e na adição e modificação de alguns conceitos. A DNDO foi utilizada para materializar os dados *linkados*, dado que não foi possível efetuar o *linkage* com os dados anonimizados. Os dados da DNDO vieram da parceria

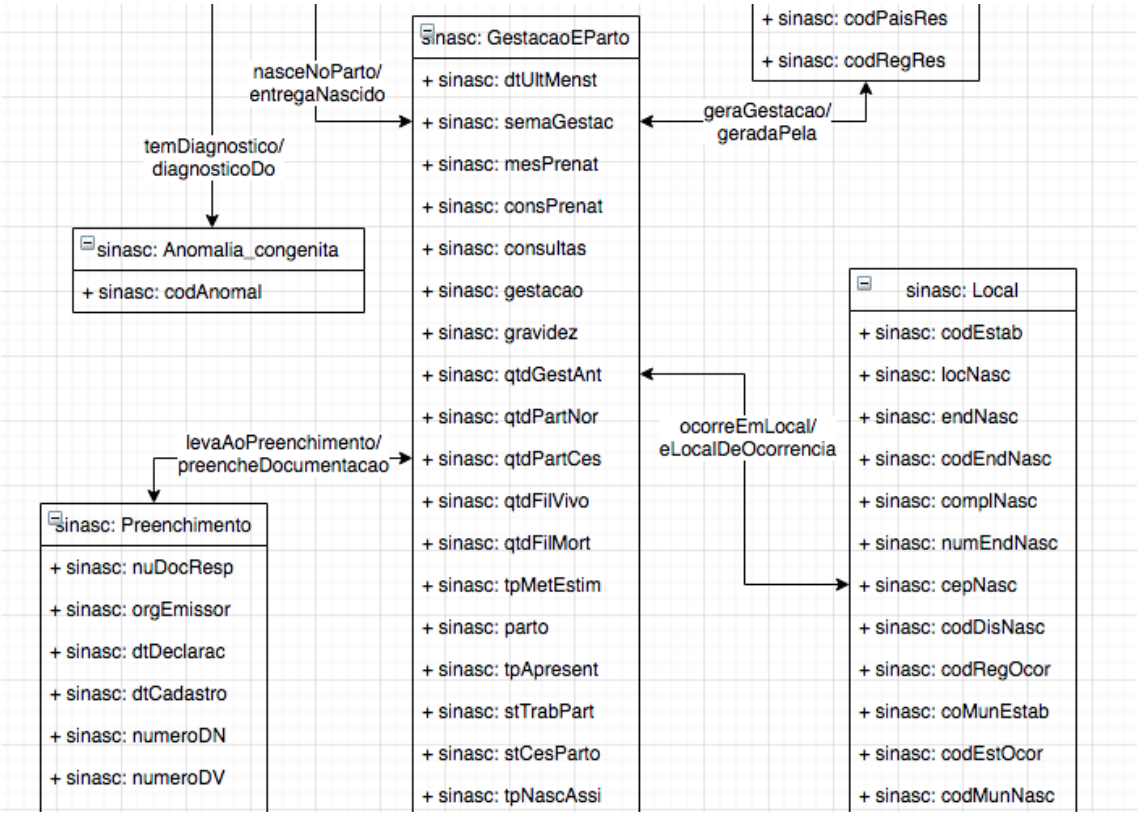

Figura 5. Recorte do diagrama com conceitos e relações correspondentes a ontologia OF_{SINASC}, apresentação da classe `GestacaoEParto`.

com a SMS-SP, que os forneceu por meio de uma planilha com a integração das bases de dados do SINASC e do SIM e a adição de algumas variáveis com a conversão de valores que eram mais interessantes para o projeto. Haviam duas opções para a utilização dos dados da DNDO, modificar as ontologias do OF_{SINASC} e OF_{SIM} com as conversões realizadas ou então, o que foi escolhido, criar uma nova ontologia e assim testar o uso de uma ontologia já com dados integrados.

A abordagem apresentada supõe que o processo de *linkage* pode ocorrer previamente (como a DNDO) ou durante o processo de integração, desde que haja um identificador único correspondente em ambas as fontes de dados. Isso dependerá de como as fontes de dados estão estruturadas. Caso as fontes tenham passado pelo processo de *linkage*, como a DNDO, resta executar a integração semântica. Caso não, mas exista um identificador comum, é possível efetuar a integração combinada, realizada na camada OI. Para atestar isso, foi decidido por apresentar as duas possibilidades. Então, na camada OF foram criadas três ontologias (OF_{SINASC}, OF_{SIM}, com expressividade $\mathcal{ALCHI}(\mathcal{D})$ e OF_{DNDO}, com $\mathcal{ALUI}(\mathcal{D})$), respectivamente correspondentes às bases de dados SINASC, SIM e DNDO.

Os mapeamentos das bases de dados para as ontologias foram realizados utilizando o plugin Ontop[7], para Protégé[8]. Em seguida as ontologias foram povoadas e foram realizadas consultas para testar sua consistência. Estas ontologias foram avaliadas com

[7]Plugin Ontop: `https://github.com/ontop`

[8]Aplicativo utilizado para a construção das ontologias: `https://protege.stanford.edu/`

um raciocinador e pelo algoritmo do OOPS![9] Os arquivos relativos às ontologias foram disponibilizados num repositório online[10]. Essas ontologias podem ser reutilizadas, uma vez que possuem o vocabulário original das fontes de dados SINASC e SIM e são independentes do contexto de saúde materno-infantil.

4. Desenvolvimento das ontologias e mapeamentos dependentes da aplicação

Esse bloco da solução apresenta as ontologias e os mapeamentos que são dependentes da aplicação e que assim foram desenvolvidos a partir das questões de competências elencadas pelos especialistas do domínio pensando na saúde materno-infantil e na produção do indicador de saúde DPGP. Fazem parte desta etapa o desenvolvimento das ontologias das camadas OD e OI, assim como seus devidos mapeamentos.

O desenvolvimento da ontologia da camada OD seguiu a metodologia apresentada em [Grüninger and Fox 1995] e resultou na ontologia OD_{SMI} - ontologia de domínio da saúde materno-infantil, com expressividade $\mathcal{SRI}(\mathcal{D})$. Foram realizadas reuniões com os 20 especialistas que definiram 58 questões de competências[11] que a solução de integração deveria responder por meio da ontologia da camada OD. Esta ontologia recebeu todos os conceitos dessas questões de competência, no entanto apenas 20 delas são respondidas com os dados do SINASC e do SIM. As questões foram então agrupadas por nível de dificuldade para respondê-las. Há aquelas que:

1. Podem ser respondidas de um único modo e que necessitam de:
   (a) Uma variável de apenas uma base, como: Qual o peso ao nascer do recém nascido X?
   (b) Mais de uma variável da mesma base, geralmente numa apresentação que remodela os dados, como: Quais foram as causas do óbito materno na usuária Y?
   (c) Mais de uma variável de mais de uma base, como: Quais foram os registros de óbitos de recém nascidos no SIM que têm registro de anomalia congênita na base de dados do SINASC?
2. Podem ser respondidas de diversos modos pois a mesma informação está localizada em mais de uma base de dados, assim, pode conter até as três subdivisões descritas no modo 1. Como em: Houve aborto ou perdas fetais antes desta gestação? Onde o resultado pode ser encontrado tanto na base de dados do SINASC quanto do SIM.

Essas questões foram estudadas e tiveram seus conceitos mapeados e transformados em classes e propriedades na ontologia da camada OD. Para isso foi desenvolvido o mesmo tipo de diagrama com conceitos e relações apresentado nas ontologias da camada OF e exemplificado pela Figura 5. Uma vez desenvolvida a ontologia, esta, assim como as ontologias da OF, foi avaliada com um raciocinador e pelo algoritmo do OOPS! e depois foram realizadas as consultas derivadas das questões de competência, tais como a questão apresentada a seguir. As respostas foram revisadas pelos especialistas que concordaram com o raciocínio apresentado.

[9]OOPS! OntOlogy Pitfall Scanner! `http://oops.linkeddata.es/`

[10]Repositório que contém as ontologias e os diagramas citados: `https://gitlab.com/beliefchangetools/diasus`

[11]A lista completa das questões de competência também está disponível no repositório do projeto.

A questão de competência sobre Nascido Vivo, *"Houve óbito de algum recém nascido que tem registro de anomalia congênita no SINASC? Apresentar o recém nascido, o número da declaração de nascido vivo, o número da declaração de óbito e o código da anomalia congênita"* tem seu resultado apresentado na Figura 6, tendo sido realizada a partir da ontologia OD_{SMI}.

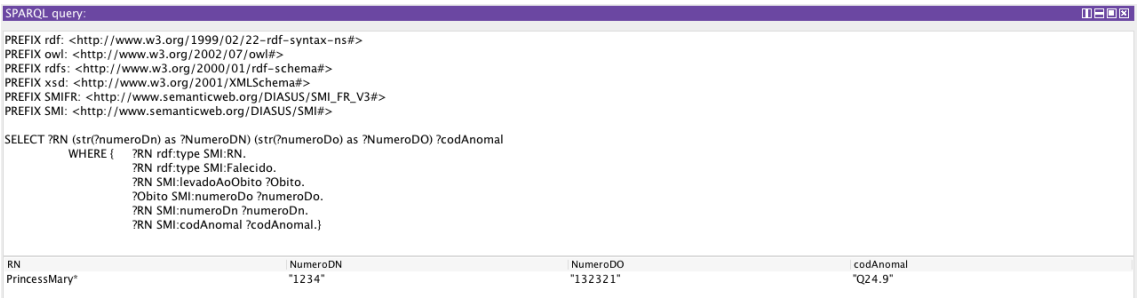

Figura 6. Resposta da questão de competência sobre óbito de recém nascido com diagnóstico de anomalia congênita no SINASC, ontologia OD_{SMI}

A Figura 7 apresenta o caminho percorrido através das ontologias para responder à questão de competência Nascido Vivo. Nela, Ob. significa Óbito; RN, recém nascido vivo; e ACon, Anomalia Congênita. A solicitação da consulta é recebida pela OD_{SMI} que passa a consulta para a camada de integração que a particionará de acordo com os conceitos presentes em cada uma das ontologias. Também é nessa camada que os conceitos similares são mapeados e integrados. Então a solicitação da consulta segue para a camada de ontologias de fonte que mapeará os conceitos nas fontes de dados. Essas retornarão a consulta com o resultado. A resposta da consulta seguirá o caminho inverso, sendo apresentado em cada uma das camadas e verificado na transição da camada de integração para a de domínio.

As ontologias da camada OI integram as fontes de dados (ontologias da camada OF) e os conceitos presentes nas questões de competência (ontologia da camada OD). Assim, a construção dessas ontologias depende da compreensão da presença e estrutura dos dados nas ontologias de fonte e de como o conhecimento está representado na ontologia de domínio. Foram desenvolvidas três ontologias (OI_{SINASC}, OI_{SIM}, OI_{DNDO}, com expressividade $\mathcal{SRI}(\mathcal{D})$), ambas possuem o vocabulário da ontologia OD_{SMI}, assim essa ontologia foi importada em cada uma delas. Em seguida os conceitos presentes na ontologia OD_{SMI} foram agrupados de acordo com os conceitos presentes em cada uma das ontologias da camada OF. Nessas etapas foram identificadas as diferenças e similaridades de vocabulário das fontes de dados e este foi harmonizado com o vocabulário adotado na ontologia da camada OD. Os mapeamentos relacionados a camada OI foram realizados no nível conceitual. Assim, a implementação do acesso aos conceitos e dados entre as camadas de ontologia serão realizados num segundo momento do projeto.

5. Considerações finais

Durante o processo de desenvolvimento da solução, houve a compreensão do esforço semântico necessário para a interpretação dos dados e das relações entre os conceitos por parte dos profissionais da SMS-SP e da FSP-USP. Foi apresentada também uma visão geral dos conceitos relacionados às duas fontes de dados, mostrando o fluxograma dos

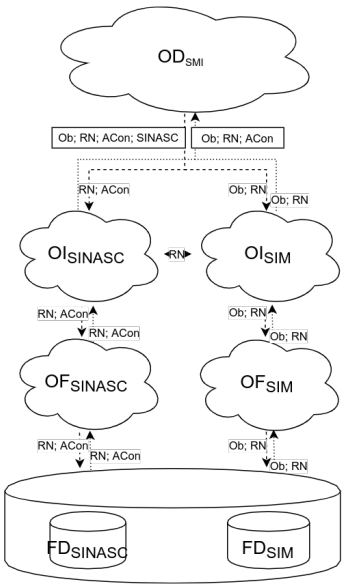

Figura 7. Diagrama representando o caminho percorrido para responder à questão de competência Houve óbito de algum recém nascido que tem registro de anomalia congênita no SINASC?

dados e como eles são interpretados. Por fim, as bases que já eram públicas, referentes ao município de São Paulo no ano de 2015, foram remodeladas e também disponibilizadas em formato de dados abertos, junto com as ontologias do modelo e os mapeamentos entre as camadas de ontologia da solução.

Este trabalho elegeu uma metodologia e arquitetura de integração semântica compatível com os dados do SUS, favorecendo futuras integrações de outras bases de dados do Sistema. Assim, é esperado que as ontologias, do mesmo modo, possam ser reutilizadas por outras iniciativas e seu uso estendido para além dos dados do município de São Paulo, uma vez que todos os municípios brasileiros possuem o mesmo formato de dados. A maior expectativa é de que Secretarias Municipais e Estaduais da Saúde do país tenham acesso a esses mapeamentos, uma vez que possuem os dados identificados, tornando o processo de integração possível. No entanto, pesquisadores fora dessa esfera também podem fazer uso desse conhecimento para compreenderem os conteúdos das bases de dados e como eles se relacionam.

Como trabalho futuro, planeja-se modelar e adicionar, nas ontologias, os metadados presentes nas bases de dados do SINASC e SIM, tais como Código Brasileiro de Ocupações - CBO e o Código Internacional de Doenças - CID. Será também feita a implementação das interfaces entre as camadas das ontologias. Pretende-se ainda, utilizar os dados integrados e o conhecimento inferido para gerar dados para o projeto do indicador de Dias Potenciais de Gravidez Perdidos. A médio prazo, será feita a adição de novas bases de dados, como a de internações hospitalares - SIH. Também ocorrerá a integração da solução desenvolvida com visualizadores de dados.

Referências

(2015). *Sistemas de Informação da Atenção à Saúde: Contextos Históricos, Avanços e Perspectivas no SUS*. Brasil, Ministério da Saúde, Secretaria de Atenção a Saúde,

Departamento de Regulação, Avaliação e Controle.

Barisevičius, G., Coste, M., Geleta, D., Juric, D., Khodadadi, M., Stoilos, G., and Zaihrayeu, I. (2018). Supporting digital healthcare services using semantic web technologies. In *International Semantic Web Conference*, pages 291–306. Springer.

Bauer, C., Ganslandt, T., Baum, B., Christoph, J., Engel, I., Löbe, M., Mate, S., Stäubert, S., Drepper, J., Prokosch, H.-U., Winter, A., and U, S. (2016). Integrated data repository toolkit (IDRT). A suite of programs to facilitate health analytics on heterogeneous medical data. *Methods of information in medicine*, 55(02):125–135.

da Cruz, M. M. L., Avila, C. V. S., Vidal, V. M. P., and Junior, N. M. A. (2019). Semanticsus: Um portal semântico baseado em ontologias e dados interligados para acesso, integração e visualização de dados do sus. In *Anais Estendidos do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 13–18. SBC.

Diniz, C., Reis-Queiroz, J., Kawai, C., Queiroz, M., Bonilha, E., Niy, D., Sena, B., and Lansky, S. (2019). *"Dias potenciais de gravidez perdidos" (DPGP): uma medida inovadora da idade gestacional*. Revista de Saúde Pública. No prelo.

Ekaputra, F. J., Sabou, M., Serral, E., Kiesling, E., and Biffl, S. (2017). Ontology-based data integration in multi-disciplinary engineering environments: A review. *Open Journal of Information Systems (OJIS)*, 4(1):1–26.

Ferronato, A. C. C., Pires, F. R., and Bernardini, F. C. (2016). Um modelo para integração e disponibilização de dados na área de saúde governamental. In *Anais do XII Simpósio Brasileiro de Sistemas de Informação*, pages 124–127.

Grüninger, M. and Fox, M. S. (1995). Methodology for the design and evaluation of ontologies. *Proceedings of IJCAI'95*.

Haw, S.-C., May, J. W., and Subramaniam, S. (2017). Mapping relational databases to ontology representation: A review. In *Proceedings of the International Conference on Digital Technology in Education*, pages 54–58. ACM.

Lopes, G., Vidal, V., and Oliveira, M. (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*, pages 327–330. ACM.

Pereira, D. L. N. C. (2019). Integração semântica das bases de dados do Sistema Único de Saúde: um estudo de caso com o município de São Paulo. Master's thesis, Universidade de São Paulo.

Pinheiro, J. C. (2011). *Processamento de consulta em um framework baseado em mediador para integração de dados no padrão de Linked Data*. PhD thesis, Universidade Federal do Ceará, Fortaleza - CE.

Ristoski, P. and Paulheim, H. (2016). Semantic web in data mining and knowledge discovery. *Journal of Web Semantics*, 36(C):1–22.

Stoilos, G., Geleta, D., Shamdasani, J., and Khodadadi, M. (2018a). A novel approach and practical algorithms for ontology integration. In *International Semantic Web Conference*, pages 458–476. Springer.

Stoilos, G., Geleta, D., Wartak, S., Hall, S., Khodadadi, M., Zhao, Y., Alghamdi, G., and Schmidt, R. A. (2018b). Methods and metrics for knowledge base engineering and integration. In *WOP@ ISWC*, pages 72–86.

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.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In *IJCAI-01 workshop: ontologies and information sharing*, volume 2001, pages 108–117. Citeseer.

Zhang, X., Hu, C., Zhao, Q., and Zhao, C. (2007). Semantic data integration in materials science based on semantic model. In *Third IEEE International Conference on e-Science and Grid Computing (e-Science 2007)*, pages 320–327.