Integração de Dados Abertos em Saúde com o modelo OBDA: Um Estudo de Caso na Área de Cirurgia Bariátrica

Samuel L. B. Bispo[1], **Vinícius K. Fukace**[1], **Ana Heloísa B. Mazur**[1], **Raqueline R. M. Penteado**[1], **Heloise M. P. Teixeira**[1]

[1]Departamento de Informática – Universidade Estadual de Maringá (UEM)
87.020-900 – Maringá – PR – Brasil

{ra103643,ra115672,ra118003,rrmpenteado, hmpteixeira}@uem.br

Abstract. *With the increasing availability of open health data on the Datasus portal, several opportunities for studies arise in order to generate information to support decision-making and the elaboration of health actions. A challenge for computer systems is to run queries in different data sources, requiring the implementation of a solution capable of handling the semantic differences of the data. This research describes a case study of open data integration in health through OBDA (Ontology Based Data Access) tools, using an ontology in the field of bariatric surgery as a conceptual layer over the data sources. The main motivation of the research is to present and discuss the use of ODBA tools through a practical experiment in DATASUS open data integration.*

Resumo. *Com a crescente disponibilização de dados abertos em saúde no portal do DATASUS, surgem várias oportunidades de estudos no sentido de gerar informações para subsidiar a tomada de decisão e elaboração de ações em saúde. Um desafio para sistemas computacionais é executar consultas em diferentes fontes de dados, sendo necessária a implementação de uma solução capaz de tratar as diferenças semânticas dos dados. Esta pesquisa descreve um estudo de caso de integração dados abertos em saúde por meio de ferramentas OBDA (Ontology Based Data Access), utilizando uma ontologia no domínio da cirurgia bariátrica como camada conceitual sobre os dados. A principal contribuição da pesquisa é apresentar e discutir o uso de ferramentas OBDA por meio de um experimento prático de integração de dados do DATASUS.*

1. Introdução

O Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS) disponibiliza informações que podem servir para subsidiar análises e tomadas de decisão baseadas em evidências e elaboração de programas de ações em saúde. No portal do DATASUS podem ser encontrados vários conjuntos de dados em diferentes formatos, por exemplo, sobre coronavírus, dados de morbidade, diagnósticos, entre outros. Também são encontradas informações sobre assistência à saúde da população, cadastros (rede assistencial), redes hospitalares e ambulatoriais, cadastro dos estabelecimentos de saúde, além de informações sobre recursos financeiros e informações demográficas e socioeconômicas [Brasil 2022].

Iniciativas da Rede Nacional de Dados em Saúde (RNDS) buscam oferecer serviços essenciais de Saúde Digital e incentivam o desenvolvimento da interoperabilidade entre sistemas de informação de saúde de todos os setores [Brasil 2020]. Segundo [Brasil 2020]: "*..os serviços de terminologia viabilizam a interoperabilidade entre sistemas, promovem a qualidade da informação – sobretudo a clínica – e resultam em melhor atendimento ao paciente, maior capacidade de gestão da qualidade da saúde, em benefício de cidadãos, profissionais e gestores*". O mesmo documento ressalta a importância do desenvolvimento de repositórios de terminologias de modo a oferecer serviços tais como o mapeamento entre terminologias, acesso a ontologias, mapeamento semântico e semelhança entre termos.

Dessa grande quantidade de dados, quando integrados e processados por uma ferramenta computacional, pode-se extrair informações valiosas para a quantificação e a avaliação das informações em saúde. Normalmente isso não é uma tarefa simples, havendo a necessidade de consultas mais complexas, executadas em múltiplas bases de dados. Para facilitar o acesso aos dados, uma solução é adotar o paradigma OBDA (*Ontology Based Data Access*) [Xiao et al. 2018], onde as ontologias são utilizadas como camada conceitual sobre os dados, oferecendo maior expressividade na descrição do domínio e facilitando o acesso à informação por sistemas computacionais, sem a necessidade de um especialista da área.

O presente trabalho apresenta um estudo prático sobre integração de dados abertos do DATASUS, aplicando conceitos do paradigma OBDA, onde o domínio de interesse é representado em uma ontologia sobre cirurgia bariátrica, denominada *OBaS - Ontology for Bariatric Surgery*, proposta em [Nunes 2021]. O estudo buscou responder as seguintes perguntas com base nos dados abertos disponibilizados pelo DATASUS: *1) As comorbidades dos pacientes implicam no custo da cirurgia bariátrica?* e *2) A prática de atividades físicas e alimentação saudável pelos pacientes implica no custo da cirurgia bariátrica?* Para implementar essas consultas, foi necessária a utilização de ferramentas como o *Protégé*[1], *Ontop*[2], *TabWin*[3] e o H2[4]. Foram consultadas duas fontes de dados abertas na área de cirurgia bariátrica, uma do Sistema de Informações Ambulatoriais do SUS (SIASUS) e outra do Sistema de Informações Hospitalares do SUS (SIHSUS).

A principal motivação da pesquisa é apresentar conceitos e discutir o uso de ferramentas OBDA, por meio de um experimento prático de integração de dados do DATASUS. Discutir a aplicação de ferramentas e métodos sobre integração de dados em saúde é um tema relevante para se atingir os objetivos de acesso à informação e conhecimento para tomada de decisão clínica e administrativa. O estabelecimento de padrões para a troca de informações em saúde é fundamental para a expansão dos serviços e funcionalidades que venham ao encontro dos interesses de todos os setores da saúde [Brasil 2020].

As seções seguintes estão organizadas como segue. A Seção 2 apresenta os principais conceitos relacionados a integração de dados e ferramentas utilizadas no estudo. A Seção 3 descreve o estudo de caso e, por fim, a Seção 4 apresenta as considerações finais.

[1]https://protege.stanford.edu/
[2]https://ontop-vkg.org/
[3]http://siab.datasus.gov.br/DATASUS/index.php?area=060805
[4]https://www.h2database.com/html/main.html

2. Fundamentação teórica

Esta seção apresenta os principais conceitos teóricos bem como as ferramentas utilizadas no estudo de caso.

2.1. Ontologia e Abordagem OBDA

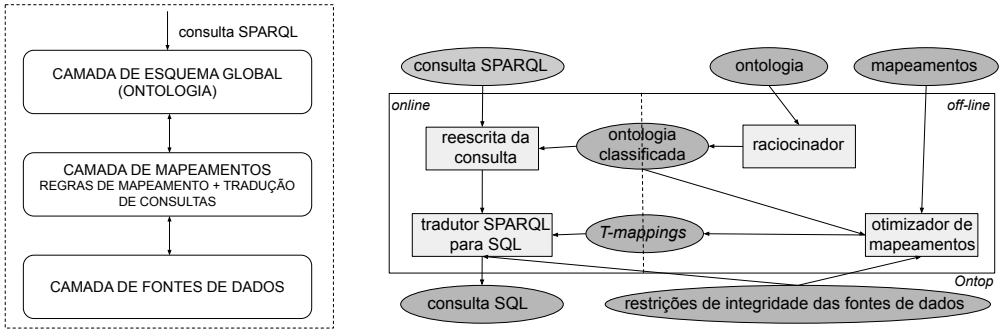
O termo *ontologia* surge na filosofia, nomeando o ramo da metafísica que estuda a natureza e estrutura da realidade. Na computação, ontologia é definida como uma especificação explícita e formal de uma conceitualização compartilhada [Studer et al. 1998]. Entende-se por conceitualização os conceitos, objetos, entidades e relacionamentos presentes em um determinado domínio. Para sistemas de conhecimento, o que "existe" é exatamente o que pode ser representado [Gruber 1993].

Ontologias tem sido uma solução interessante para integração de dados, pois provê um vocabulário rico e predefinido que serve como uma interface conceitual para os bancos de dados, que é independente do esquema da base. Conforme estudo descrito em [Alkhamisi and Saleh 2020], existem muitas propostas na literatura que utilizam ontologias para integrar dados em diferentes domínios, como a área da saúde [Zhang et al. 2018]. Podem ser utilizadas para tratar problemas de semântica, usadas como modelos de referência para mapear conceitos envolvidos em diferentes aplicações e organizações, explicitado de forma clara o significado dos componentes envolvidos. Servem como esquemas de metadados, fornecendo um vocabulário controlado de conceitos, cada um com uma semântica explicitamente definida de modo que possa ser processada por algoritmos computacionais. Desta forma, as ontologias modelam formalmente a estrutura de um sistema, isto é, as classes e relações relevantes que emergem de sua observação e que são úteis para determinados propósitos [Guarino et al. 2009].

A abordagem OBDA (*Ontology Based Data Access*) viabiliza a integração semântica de dados por meio de ontologias e mapeamentos. A ontologia é usada como uma visão conceitual de alto-nível dos repositórios de dados, possibilitando que um usuário tenha acesso aos dados sem ter o conhecimento específico de como eles estão organizados em suas fontes. Conforme ilustra a Figura 1(a), um sistema que utiliza esta abordagem é constituído por três componentes [Calvanese et al. 2017b]: 1) esquema global, 2) fonte de dados e 3) mapeamento entre esquema global e fonte de dados.

A camada de esquema global (denominada *ontologia* em sistemas OBDA), fornece uma representação formal de alto nível do domínio de interesse, e constitui o componente no qual os clientes do sistema de informação (humanos e programas de software) interagem. A camada da fonte de dados, que representa as bases de dados existentes no sistema de informação, é gerenciada por processos e serviços que operam sobre os seus dados. O mapeamento entre as duas camadas refere-se a uma representação explícita do relacionamento entre as fontes de dados e a ontologia. Mapeamentos são usados para traduzir as operações na ontologia (isto é, as consultas) em termos de ações concretas nas fontes de dados. Além disso, Calvanese et al. [Calvanese et al. 2017a] destacam que mapeamentos também viabilizam a integração de dados.

Na área da saúde, [da Cruz et al. 2019] apresentam um portal denominado SemanticSUS para contribuir no desenvolvimento de aplicações onde existe a necessidade de integrar fontes de dados heterogêneas. Segundo os autores, o principal objetivo do



(a) (b)

Figura 1. *Framework OBDA*. Fonte: [Fathy et al. 2019] (a); Fluxo de trabalho do *Ontop* baseado em [Calvanese et al. 2017a] (b)

portal é oferecer uma camada ontológica que se conecta semanticamente aos dados e permite o acesso integrado às 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. No entanto, segundo informado no Portal SemanticSUS, por questões de segurança e dados sensíveis, o acesso à ferramenta denominada *Mashup Builder* é restrito para pessoas autorizadas.

2.2. Ferramentas *Protégé* e *Ontop*

O *Ontop* é uma ferramenta OBDA de código livre, que da suporte à ontologias $OWL2QL$[5] e mapeamentos $R2RML$[6] para a integração semântica de fontes de dados que adotam o modelo relacional. A ferramenta está disponível como um *plugin* para a ferramenta *Protégé*, que fornece uma interface gráfica para várias funcionalidades do *Ontop* como, por exemplo, edição de mapeamentos e execução de consultas SPARQL. De uma forma geral, o *Ontop* expõe bases relacionais como grafos RDF virtuais relacionando os termos da ontologia global (classes e propriedades) com as fontes de dados através de mapeamentos. A linguagem de consulta SPARQL (*SPARQL Protocol and RDF Query Language*) é utilizada para formular consultas sobre essa representação. Uma vez que as bases de dados são relacionais, consultas SPARQL são traduzidas para SQL (por meio de mapeamentos) para que ocorra a execução da consulta nas bases relacionais. O processo de tradução é transparente ao usuário do sistema [Calvanese et al. 2017a].

A Figura 1(b) mostra o fluxo de trabalho do *Ontop* dividido em dois estágios, *off-line* e *online*. O estágio *off-line* processa a ontologia, os mapeamentos e as restrições de integridade das fontes de dados. A ontologia é classificada pelo *raciocinador*. O *otimizador de mapeamentos* otimiza mapeamentos existentes, além de gerar novos mapeamentos a partir da ontologia, caso necessário. No estágio *online*, consultas SPARQL submetidas à ontologia são *reescritas*, se necessário, e *traduzidas* para consultas SQL explorando mapeamentos e restrições de integridade. Por fim, consultas SQL são submetidas às *engines* das fontes de dados e os resultados retornados das consultas são traduzidos para termos RDF, o modelo de dados retornado por consultas SPARQL.

A definição de um mapeamento envolve uma *origem*, da qual uma consulta SQL

[5]https://www.w3.org/ns/owl-profile/data/QL
[6]https://www.w3.org/TR/r2rml/

recupera valores da base de dados, e um *alvo*, que constrói triplas RDF com valores da *origem*. Por exemplo, considere a classe *Internacao* de uma ontologia e a fonte de dados `fonte2` da Figura 3. O mapeamento *RDPR*$\{N_AIH\}$ *a Internacao; :custo* $\{val_tot\}$. $\leftarrow$ *SELECT FROM RDPR*, relaciona a *origem* com o *alvo*, a fonte *RDPR* e classe *Internacao*, respectivamente. A integração semântica de dados pode envolver entidades identificadas de maneira equivalente ou não. Quando entidades são identificadas de maneira equivalente, a integração no *Ontop* pode ser feita por meio de mapeamentos tradicionais das fontes de dados para a ontologia em questão.

A pesquisa de [Ciriaco et al. 2020] apresenta um estudo de caso de integração de dados na área da saúde utilizando o *Ontop*, porém, o artigo não descreve detalhes sobre o processo de integração entre as fontes de dados autônomas utilizadas. Observa-se que os trabalhos na literatura não descrevem ou discutem os desafios no uso das ferramentas OBDA, bem como os passos necessários para integração no *Ontop*. Portanto, neste artigo, buscou-se demonstrar com mais detalhes o uso da ferramenta, visto que o processo não é trivial, conforme é descrito na seção seguinte.

3. Estudo de Caso

Esta seção descreve o experimento prático de integração de dados abertos em saúde na especialidade da cirurgia bariátrica. O tema foi escolhido por ter disponível no portal do DATASUS dados abertos no domínio e por sua relevância, visto que, conforme a Organização Mundial da Saúde (OMS), existem meio bilhões de pessoas obesas ao redor do mundo [ABESO 2016]. Além consequências físicas, a obesidade também está associada a afastamentos do trabalho, absenteísmo e aposentadorias mais precoces dos indivíduos obesos. Na prática, as informações que podem contribuir com a tomada de decisão sobre a cirurgia bariátrica pelos profissionais e gestores de saúde normalmente estão distribuídas entre os diversos sistemas utilizados por instituições públicas e privadas, como laboratórios, hospitais, planos de saúde entre outros. Nestes sistemas, desenvolvidos utilizando tecnologias diferentes, os dados podem ser armazenados e estruturados utilizando vocabulários e padrões distintos. Sendo assim, a motivação deste estudo é aplicar ferramentas computacionais que possibilitam o acesso à informação e conhecimento que podem contribuir para tomada de decisão clínica e administrativa no domínio.

As subseções seguintes descrevem as etapas para o processo de integração utilizado neste estudo, a saber: definição da ontologia; definição das fontes de dados abertas; o processo de integração e, por fim, a execução de consultas.

3.1. Definição dos dados abertos e da Ontologia em Cirurgia Bariátrica

Conforme descrito na Seção 2, para implementar a integração de fontes de dados utilizando-se o modelo OBDA, é necessário que o conhecimento do domínio seja representado em uma ontologia. Para o presente estudo, utilizou-se a ontologia desenvolvida nos trabalhos de [Nunes and Berardi 2020], [Nunes 2021] e [Silva et al. 2021].

A Figura 2 ilustra a hierarquia das principais classes modeladas em [Nunes 2021] na ferramenta *Protégé*. A ontologia utilizada, denominada *Ontology for Bariatric Surgery* (OBaS), usa como base a *Ontology for General Medical Science* (OGMS), que por sua vez tem como base a *Basic Formal Ontology* (BFO), ambas desenvolvidas pelo *Open Biological and Biomedical Ontology (OBO) Foundry*, um consórcio para desenvolver ontologias

abertas e interoperáveis para as ciências biológicas. A ontologia pode ser acessada em `https://github.com/glaubernunes/ontology-for-bariatric-surgery` [Nunes 2021].

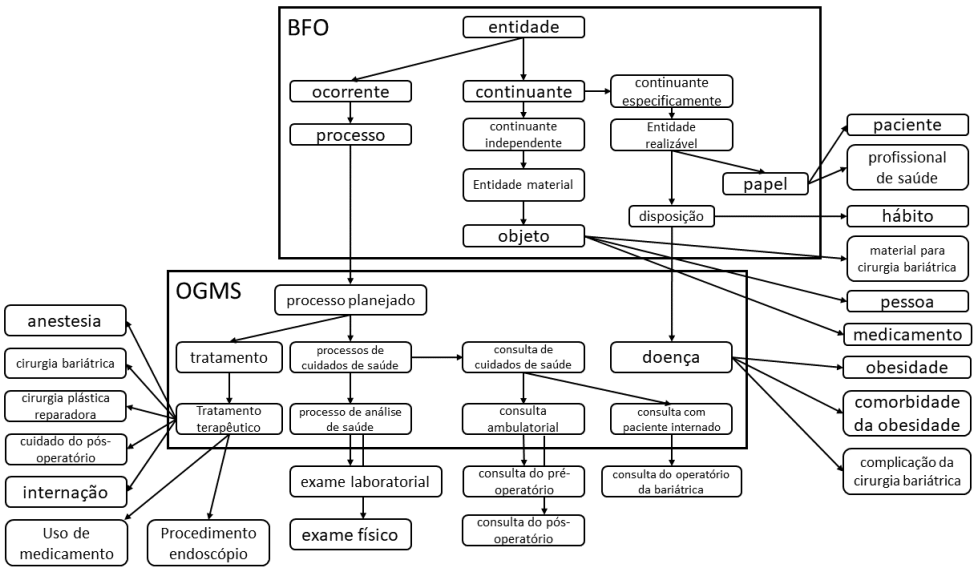


Figura 2. Visão Geral da Hierarquia de Classes. Fonte: [Nunes 2021]

As bases de dados utilizadas nesta pesquisa foram duas, uma do SIASUS (Sistema de Informações Ambulatoriais do SUS) e outra do SIHSUS (Sistema de Informações Hospitalares do SUS). Estes sistemas contêm dados de caráter predominantemente administrativo, utilizados para gerenciar recursos e registrar atendimentos realizados, sendo que dados sigilosos não são registrados ou se encontram criptografados, a fim de preservar a privacidade e segurança dos cidadãos. A base do SIASUS possui informações sobre a produção ambulatorial através de laudos de APAC (Autorização de Procedimento de Alta Complexidade).O estudo considerou exclusivamente APACs de acompanhamento à cirurgia bariátrica. Já a base do SIHSUS contém os registros de AIH (Autorização de Internação Hospitalar) de todos os atendimentos provenientes de internações hospitalares financiados pelo SUS. Da mesma forma, considerou-se somente as internações relacionadas à cirurgias bariátricas.

A Figura 3 ilustra as duas fontes de dados provenientes do SIASUS e SIHSUS, denominadas `fonte1` e `fonte2`, respectivamente. A `fonte1` possui dados sobre APACs de cirurgias bariátricas e a `fonte2` sobre as AIHs resultantes das cirurgias. Ambas possuem o atributo *equivalente* número de Autorização de Internação Hospitalar, atributo utilizado para a integração das bases no estudo de caso. Na `fonte1` esse número é o atributo AB_NUMAIH e na `fonte2` é o atributo N_AIH. O atributo AP_CNSPCN representa o número do Cartão Nacional de Saúde (CNS) do paciente e encontra-se criptografado a fim de proteger a identidade dos pacientes, uma vez que a fonte está disponibilizada publicamente. Neste estudo, considerou-se que cada criptografia é única.

A Tabela 1 mostra os atributos das fontes utilizadas e suas respectivas descrições. Além dos atributos referentes ao número de autorização de internação em ambas as fontes, os atributos da `fonte1` referem-se aos pacientes e os da `fonte2` ao custo da internação.

ABOPR

AB_NUMAIH	AP_CNSPCN	AP_CID_C1
4120105402380	❖❖❖❖❖❖❖	I10
4120105402457	❖❖❖❖❖❖❖	I10
4120105402468	❖❖❖❖❖❖❖	I10

fonte1

RDPR

N_AIH	VAL_TOT
4120105402380	6661.82
4120105402457	6623.72
4120105402468	6623.72

fonte2

Figura 3. Fontes de dados que identificam autorização para internação pelos atributos AB_NUMAIH na `fonte1` e N_AIH na `fonte2`.

`fonte1`	AB_NUMAIH	Número de autorização da internação
	AP_CNSPCN	Número do CNS do paciente
	AP_CID_C1	Se o paciente possuir Hipertensão Arterial, estará preenchido com I10
	AP_CID_C2	Se o paciente possuir Diabetes Mellitus, estará preenchido com O243.
	AP_CID_C3	Se o paciente possuir Dispidemia, estará preenchido com E78
	AP_CID_C4	Se o paciente possuir Artrose, estará preenchido com M199
	AP_CID_C5	Se o paciente possuir Apnéia, estará preenchido com G473
	AP_ATV_FIS	Se o paciente pratica (1) ou não (0) atividade física
	AP_ADESAO	Se o paciente aderiu (1) ou não (0) uma alimentação saudável
`fonte2`	N_AIH	Número de autorização da internação
	VAL_TOT	Valor total da internação

Tabela 1. Descrições dos atributos das fontes de dados.

3.2. Processo de Integração

Para o processo de integração, foi necessária uma pequena extensão da ontologia utilizada. Para isso, a ontologia foi importada para um novo projeto através da ferramenta de edição de ontologias *Protégé*, de modo que pudesse ser alterada. A importação da ontologia exigiu a geração de prefixos para as suas classes. A coluna *Classe* da Tabela 2 mostra as classes da ontologia (Figura 2) utilizadas neste estudo. A coluna *Prefixo* mostra o prefixo gerado para cada classe utilizada. No *Protégé*, a opção para a geração de prefixos encontra-se na página inicial, em *"Ontology prefixes:"*. A Tabela 3 exibe as propriedades de objeto utilizadas no estudo, que relacionam classes da ontologia.

A ontologia foi então estendida com a propriedade de objeto *temPessoa* que relaciona *Internação* e *Pessoa* e a propriedade de dado *custo* para a classe *Internação*. A adição das novas propriedades foi realizada de forma que possam ser referenciadas sem a necessidade de uso de prefixos não usuais, assim pode-se utilizar no mapeamento, por exemplo, a propriedade *temPessoa*, como simplesmente *:temPessoa*.

Classe	Prefixo
Internação	:OBAS_0000156
Pessoa	:OBAS_0000169

Tabela 2. Classes e seus prefixos

Propriedade de objeto	Prefixo
Possui comorbidade	:OBAS_0000172
Possui hábito	:OBAS_0000173

Tabela 3. Propriedades de objeto e seus prefixos

3.2.1. Armazenamento das fontes de dados

Para o armazenamento dos dados, foi necessária a conversão das fontes de dados abertas para relações (tabelas) que foram armazenadas no sistema de banco de dados relacional

H2. A bases de dados foram obtidas no site do Datasus[7] e os dados utilizados são referentes ao estado do Paraná, do mês de Abril de 2020. Este período foi escolhido para evitar inconsistência nos dados devido ao retardo de notificação. As fontes de dados *.dbc* foram convertidas para *.csv* por meio da ferramenta *TabWin* (desenvolvida e distribuída gratuitamente pelo Datasus). Por fim, os arquivos *.csv* foram convertidos para tabelas no H2, sendo elas: a ABOPR para a `fonte1` (com 4141 registros) e a RDPR para a `fonte2` (com 179 registros). Ambas não foram indexadas.

O armazenamento dos dados no H2 foi feito de maneira centralizada, ou seja, em uma única máquina. O suporte à execução de consultas em bases alocadas em máquinas distintas por meio do *Ontop* requer sistemas tais como Teiid[8] ou Denodo[9] para integrar múltiplas base de dados. Estes sistemas expõem para o *Ontop* um conjunto de tabelas oriundas de diferentes bases de dados. A utilização das fontes alocadas no H2 pelo *Ontop* exige o estabelecimento de uma conexão entre as ferramentas. Essa conexão é feita por meio do *driver* "$org.h2.Driver$" disponível no *Protégé* e acessado em *File − Preferences − JDBC Drivers*, a partir do menu superior do *Protégé*. Para que a conexão seja estabelecida com sucesso, o serviço de banco de dados do H2 deve estar ativo.

3.2.2. Definição dos mapeamentos

Para o estudo foram definidas duas consultas SPARQL: *Consulta 1: As comorbidades dos pacientes implicam no custo da cirurgia bariátrica?"* e *Consulta 2: A prática de atividades físicas e alimentação saudável pelos pacientes implica no custo da cirurgia bariátrica?* Conforme mostra a Tabela 4, foi necessário definir dez mapeamentos das fontes de dados para ontologia, sendo oito para a consulta 1 e dois para a consulta 2. O primeiro mapeamento envolveu a `fonte1` e a classe *Pessoa* da ontologia. A origem do $Map01$ foi definida pela consulta SQL *SELECT * FROM ABOPR*, recuperando todos os dados da `fonte1`. Para o alvo foi utilizada a classe *Pessoa*, representada por *:abo/{AP_CNSPCN} a :OBAS_0000169* no mapeamento. O mesmo tipo de mapeamento foi feito em $Map02$, onde a origem é a tabela RDPR da `fonte2` e o alvo a classe *Internação*. O número de autorização da internação é utilizado como identificador da internação.

A integração das fontes foi estabelecida com $Map03$, conforme a Tabela 4. Observa-se que o *script* SQL utilizado na origem envolve ambas as fontes. Este mapeamento define a relação entre uma internação da `fonte2` com uma pessoa da `fonte1`. As comorbidades das pessoas foram mapeadas para a ontologia com os mapeamentos $Map04$ a $Map08$. Definiu-se que pessoa (*AP_CNSPCN*) possui comorbidade (*:OBAS_0000172*) envolvendo as comorbidades (*AP_CID_C1*, *AP_CID_C2*, *AP_CID_C3*, *AP_CID_C4* e *AP_CID_C5*). É importante destacar que a comorbidade pode ou não estar preenchida na fonte de dados. Logo, no mapeamento, deve-se retornar apenas as pessoas que tenham esse campo diferente de vazio. O $Map04$ definiu a relação entre pessoa e a comorbidade Hipertensão Arterial. Seguindo o mesmo raciocínio, definiu-se os mapeamentos para Diabetes Mellitus ($Map05$), Dislipidemia ($Map06$), Artrose ($Map07$) e Apneia ($Map08$).

Para a consulta 2, foram necessários mais dois mapeamentos. O $Map09$ que re-

[7]https://datasus.saude.gov.br/transferencia-de-arquivos/
[8]https://teiid.io/
[9]https://www.denodo.com/en

$Map01$	:abo/{AP_CNSPCN} a :OBAS_0000169 .
	SELECT * FROM ABOPR
$Map02$	:rdpr/{N_AIH} a :OBAS_0000156 ; :custo {VAL_TOT} .
	SELECT * FROM RDPR
$Map03$	**:rdpr/{N_AIH} :temPessoa :abo/{AP_CNSPCN} .**
	SELECT * FROM ABOPR, RDPRWHERE ABOPR.AB_NUMAIH = RDPR.N_AIH
$Map04$	:abo/{AP_CNSPCN} :OBAS_0000172 :abo/{AP_CID_C1} .
	SELECT * FROM ABOPR WHERE AP_CID_C1 != ”;
$Map05$	:abo/{AP_CNSPCN} :OBAS_0000172 :abo/{AP_CID_C2} .
	SELECT * FROM ABOPR WHERE AP_CID_C2 != ”;
$Map06$	:abo/{AP_CNSPCN} :OBAS_0000172 :abo/{AP_CID_C3} .
	SELECT * FROM ABOPR WHERE AP_CID_C3 != ”;
$Map07$	:abo/{AP_CNSPCN} :OBAS_0000172 :abo/{AP_CID_C4} .
	SELECT * FROM ABOPR WHERE AP_CID_C4 != ”;
$Map08$	:abo/{AP_CNSPCN} :OBAS_0000172 :abo/{AP_CID_C5} .
	SELECT * FROM ABOPR WHERE AP_CID_C5 != ”;
$Map09$	:abo/{AP_CNSPCN} :OBAS_0000173 :abo/{AP_ATV_FIS} .
	SELECT * FROM ABOPR WHERE AP_ATV_FIS = 1;
$Map10$	:abo/{AP_CNSPCN} :OBAS_0000173 :abo/{AP_ADESAO} .
	SELECT * FROM ABOPR WHERE AP_ADESAO = 1;

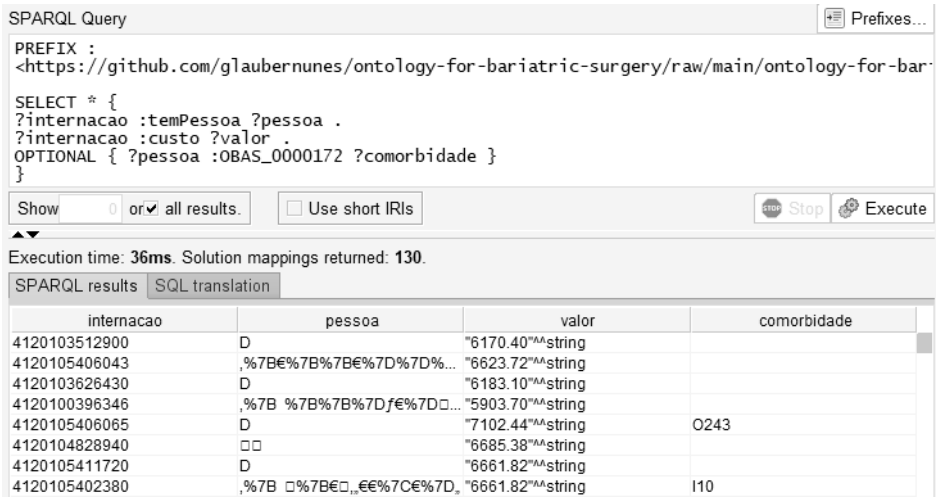
Tabela 4. Mapeamentos das fontes de dados para a ontologia.

laciona pessoa com o hábito (*:OBAS_0000173*) de fazer atividade física (*AP_ATV_FIS*) e o $Map10$ que relaciona pessoa com a adesão de um hábito de uma alimentação saudável (*AP_ADESAO*). Na fonte de dados, pessoas com as colunas referentes à hábitos preenchida com 1 representam aquelas que têm o hábito saudável. Logo, essas pessoas foram as selecionadas na fonte do mapeamento.

3.3. Processamento das consultas

A consulta 1 foi expressa em SPARQL com o seguinte *script*: “*SELECT * {?internacao :temPessoa ?pessoa . ?internacao :custo ?valor . OPTIONAL {?pessoa :OBAS_0000172 ?comorbidade}}*”. Esta consulta recupera o código da internação (AIH), o código da pessoa internada (CNS), o valor da internação (custo) e a comorbidade da pessoa, caso a pessoa tenha comorbidade. Do mesmo modo, a consulta 2 foi expressa conforme o *script*: “*SELECT * {?internacao :temPessoa ?pessoa . ?internacao :custo ?valor . OPTIONAL {?pessoa :OBAS_0000173 ?habito}} ORDER BY ?habito DESC(?valor)*”. Esta consulta recupera o código da internação (AIH), o código da pessoa internada (CNS), o valor da internação (custo) e os hábitos da pessoa, caso existam na fonte de dados. Os resultados são ordenados de acordo com o hábito do paciente.

Para execução, cada consulta SPARQL é traduzida pelo *Ontop* para SQL, uma vez que os dados estão armazenados em um banco de dados relacional. O *script* em SQL é enviado para o H2, que retorna o resultado para o *Ontop*, finalizando o ciclo de processamento da consulta. A Figura 4 ilustra a interface do *Ontop* para execução e visualização do resultado gerado para a consulta 1, que retorna o custo da internação e a presença/ausência de comorbidade na pessoa internada. Esta consulta tem como objetivo buscar a relação entre custo de internação e comorbidades de pacientes que fazem a cirurgia. Devido à criptografia no atributo número de CNS, o resultado retornou valores não usuais. Observa-se na Figura 4 que existem múltiplas entradas na coluna *Pessoa* com representação visual parecida, porém, os dados são de pessoas diferentes. Detalhes do *script* SQL gerado para a consulta pode ser consultado em `https://bit.ly/3D4Mrr4`.



SPARQL Query

```
PREFIX :
<https://github.com/glaubernunes/ontology-for-bariatric-surgery/raw/main/ontology-for-bar

SELECT * {
?internacao :temPessoa ?pessoa .
?internacao :custo ?valor .
OPTIONAL { ?pessoa :OBAS_0000172 ?comorbidade }
}
```

Execution time: **36ms**. Solution mappings returned: **130**.

internacao	pessoa	valor	comorbidade
4120103512900	D	"6170.40"^^string	
4120105406043	,%7B€%7B%7B€%7D%7D%...	"6623.72"^^string	
4120103626430	D	"6183.10"^^string	
4120100396346	,%7B %7B%7B%7Dƒ€%7D□...	"5903.70"^^string	
4120105406065	D	"7102.44"^^string	O243
4120104828940	□□	"6685.38"^^string	
4120105411720	D	"6661.82"^^string	
4120105402380	,%7B □%7B€□„€€%7C€%7D„	"6661.82"^^string	I10

Figura 4. Interface no Ontop do *script* e resultado da consulta 1

3.4. Discussão

Esta pesquisa combinou duas fontes de dados abertas do DATASUS para obter informações sobre cirurgia bariátrica. Com o estudo prático, observou-se que ontologias são complementares a um esquema de banco de dados. A principal vantagem do modelo OBDA com relação ao modelo tradicional é que se os requisitos mudam, o uso da ontologia torna a definição das regras mais flexível e de senso comum, pois é definida de acordo com especialistas no domínio.

Para implementação de consultas simples, que combinaram apenas duas diferentes fontes de dados, foram necessárias várias etapas e estudos de como as ferramentas funcionam. Na literatura, não foram encontrados trabalhos que apresentam e discutem esse processo em detalhes usando ontologias e o padrão OBDA. Para utilizar os dados abertos de acordo com o objetivo proposto nesta pesquisa, foi necessário filtrar e converter as bases num formato adequado às ferramentas OBDA utilizadas.

Para definir os mapeamentos, o *plugin* do *Ontop* para o *Protégé* fornece as funções necessárias, porém, não foi encontrada documentação sobre suas funcionalidades e a prevenção de erros do *plugin* fica a desejar. Ao definir os mapeamentos, quando ocorreu um erro, este não foi informado, dificultando identificar o problema ocorrido. Por exemplo, se a entidade necessária para o mapeamento não existir ou for definida com um nome incorreto, não é permitido salvar o mapeamento e nenhum alerta do erro é informado. Caso a classe da ontologia seja referenciada errada no mapeamento, o *plugin* do *Ontop* reconhece inconsistência, porém não auxilia na resolução. Caso erros de sintaxe sejam encontrados nos mapeamentos, a ferramenta apresenta o erro, mas em um formato confuso. Neste cenário, desenvolvedores sem experiência podem ter dificuldades em tratar esses erros simples, que podem ser facilmente solucionados quando informados.

Com relação ao tempo de execução das consultas, observou-se que o desempenho pode ser comprometido no momento da conversão de SPARQL para SQL. Essa constatação foi observada em outras pesquisas, como a de [Calvanese et al. 2017a], por exemplo, que afirma que a complexidade da combinação entre ontologia e mapeamentos implica diretamente no desempenho do *Ontop*. Na presente pesquisa observou-se o

tempo de execução das consultas utilizando um Dell G3 3590, Intel Core i5-9300H (CPU 2.40GHz) com 8 GB de RAM. As consultas 1 e 2 levaram, respectivamente, 39 milissegundos e 44 milisegundos para o término de suas execuções. Constatou-se que, o tempo de resposta foi alto. Portanto, sugere-se um estudo mais aprofundado sobre o desempenho das consultas em sistemas OBDA. Estudos práticos na literatura que usam ontologia relatam que um hardware com boa capacidade de processamento e memória é necessário para viabilizar o uso da camada ontológica [Freitas et al. 2019]. Da mesma forma, este estudo prático mostrou que para se usar OBDA na prática se faz necessário explorar também meios de otimizar o tempo de resposta das consultas.

A grande vantagem observada é que o uso da ontologia permite que a semântica dos dados seja expressa de modo que possibilite consulta a dados distribuídos em diferentes sistemas e unidades de saúde pública. É importante destacar que o objetivo deste trabalho não foi a análise dos dados obtidos com o processo de integração, mas sim estudar o processo em si. Para interpretar os resultados das consultas para gerar conhecimento, se faz necessário o envolvimento de especialistas na área da saúde.

4. Considerações Finais

Este artigo descreveu um estudo de caso sobre integração de dados por meio de uma ontologia com a ferramenta *Ontop*. O estudo contribui a compreender o processo da integração de dados abertos, visto que envolveu a integração de duas fontes de dados abertas na área de cirurgia bariátrica, obtendo-se uma fonte global. A principal contribuição da pesquisa é a descrição e discussão do processo de integração de dados bem como a execução de consultas usando ferramentas OBDA e dados abertos do DATASUS.

Com os resultados obtidos foi possível verificar que o processo de integração não é trivial e também que a utilização de dados abertos impõem desafios e exige conhecimento técnico tanto para o uso das ferramentas como também na área de ontologias. Com a crescente publicação de dados abertos pelo DATASUS em diferentes setores da saúde, discutir a aplicação de ferramentas tecnológicas e métodos sobre integração de dados é fundamental para explorar e obter conhecimento útil para tomada de decisão em saúde.

Como trabalho futuro, pretende-se explorar também assuntos pertinentes ao armazenamento de dados e otimização das consultas OBDA. Pretende-se aprofundar estudos sobre otimização de consultas em sistemas OBDA e a execução de consultas em bases de dados fisicamente distribuídas.

Agradecimentos
Ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC/CNPq-FA-UEM) pela bolsa concedida ao segundo autor e ao PIC-UEM que possibilitou a participação do primeiro e terceiro autores na pesquisa.

Referências

ABESO (2016). Diretrizes brasileiras de obesidade. `https://abeso.org.br/wp-content/uploads/2019/12/Diretrizes-Download-Diretrizes-Brasileiras-de-Obesidade-2016.pdf`. Acessado em: 15-12-2021.

Alkhamisi, A. and Saleh, M. E. (2020). Ontology opportunities and challenges: Discussions from semantic data integration perspectives. *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, pages 134–140.

Brasil (2020). Ministério da Saúde. Secretaria-Executiva. Departamento de Informática do SUS. Estratégia de Saúde Digital para o Brasil 2020-2028. http://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf.

Brasil (2022). Ministério da Saúde. Banco de dados do Sistema Único de Saúde-DATASUS. http://www.datasus.gov.br. Acessado em: 24-03-2022.

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

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., and Rosati, R. (2017b). *Ontology-Based Data Access and Integration*, pages 1–7. Springer New York, New York, NY.

Ciriaco, D. L., Pessoa, A., Salvador, L., and Wassermann, R. (2020). Integração semântica das bases de dados do sistema Único de saúde: Um estudo de caso com o município de são paulo. In da Silva Lemos, D. L., Sales, T. P., Campos, M. L. M., and Fiorini, S. R., editors, *Proceedings of the XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS 2020)*, volume 2728 of *CEUR Workshop Proceedings*, pages 61–74. CEUR-WS.org.

da Cruz, M. M., Avila, C., Vidal, V. M., and Junior, N. 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, Porto Alegre, RS, Brasil. SBC.

Fathy, N., Gad, W., and Badr, N. (2019). A unified access to heterogeneous big data through ontology-based semantic integration. pages 387–392.

Freitas, A. L. S., Freitas, A. L. S., and Teixeira, H. M. P. (2019). Sistema de alerta baseado em ontologia para lousa eletrônica em um hospital público. *Revista Brasileira de Computação Aplicada*, 11(3):99–109.

Gruber, T. R. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In Guarino, N. and Poli, R., editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands. Kluwer Academic Publishers.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on Ontologies*.

Nunes, G. M. O. (2021). Multionto: Método de construção de ontologia considerando heterogeneidade de fontes e tipos de conhecimentos - um estudo de caso sobre cirurgia bariátrica. Dissertação de mestrado, Universidade Tecnológica Federal do Paraná.

Nunes, G. M. O. and Berardi, R. C. G. (2020). Ontological model for decision support about bariatric surgery. In da Silva Lemos, D. L., Sales, T. P., Campos, M. L. M., and Fiorini, S. R., editors, *Proceedings of the XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS 2020)*, volume 2728 of *CEUR Workshop Proceedings*, pages 280–285. CEUR-WS.org.

Silva, J. E. C., Tonon, M. M., and Teixeira, H. M. P. (2021). Aquisição e representação do conhecimento científico sobre cirurgia bariátrica - fase 1. In *30o Encontro Anual de Iniciação Científica (EAIC)*, Maringá, Brasil.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data and knowledge engineering*, 25(1):161–198.

Xiao, G., Hovland, D., Bilidas, D., Rezk, M., G. M., and Calvanese, D. (2018). Efficient ontology-based data integration with canonical iris. In *European Semantic Web Conference*, pages 697–713. Springer.

Zhang, H., Guo, Y., Li, Q., George, T. J., Shenkman, E., Modave, F., and Bian, J. (2018). An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Medical Informatics Decis. Mak.*, 18(S-2):129–147.