

Integração de dados de saúde pública para apoio à tomada de decisão: uma abordagem baseada em grafo de conhecimento semântico

Sandro de Carvalho Franco¹, Nacles Bernardino Pirajá Gomes¹ and
Laís do Nascimento Salvador¹

¹Programa de Pós-graduação em Ciências da Computação (PGCOMP)
Universidade Federal da Bahia - Salvador/BA

Resumo

Este artigo aborda a complexidade e os desafios dos sistemas de informação em saúde pública no Brasil, caracterizados por bases de dados diversas e desconectadas. Para solucionar o problema do acesso unificado a dados distribuídos em múltiplas plataformas, propomos uma abordagem inovadora baseada no uso de ontologias. Ao definir um vocabulário comum que estabelece correspondências semânticas entre diferentes bases de dados, as consultas tornam-se mais simples. O texto descreve a implementação de uma integração semântica apoiada por Grafos de Conhecimento Semântico (Semantic Knowledge Graphs — SKGs). O estudo concentra-se em uma ontologia desenvolvida para responder às principais questões levantadas por especialistas em saúde pública. O trabalho define indicadores baseados em dados críticos dos principais sistemas do Sistema Único de Saúde (SUS), incluindo informações sobre nascimentos, óbitos e notificações de COVID-19, utilizando o município de Camaçari, Bahia, como estudo de caso. O objetivo é permitir que gestores de saúde analisem indicadores essenciais, como mortalidade, imunização, hospitalizações e notificações, destacando a eficácia dos SKGs na melhoria da gestão em saúde pública. Dessa forma, integramos fontes de dados relacionais em uma base unificada e virtualizada, onde consultas podem ser executadas para responder às diferentes Questões de Competência (QCs) propostas pela equipe técnica de Camaçari-BA.

Keywords

Integração de dados, grafos de conhecimento semântico, ontologias, bases de dados heterogêneas.

Abstract

This paper addresses the complexity and challenges of public health information systems in Brazil, which are characterized by diverse and disconnected databases. To solve the problem of unified access to data spread across multiple platforms, we propose an innovative approach using ontologies. By defining a common vocabulary that establishes semantic correspondences among different databases, queries become simpler. The text describes the implementation of semantic integration supported by Semantic Knowledge Graphs (SKGs). The study focuses on an ontology developed to respond to key questions raised by public health specialists. The project defines indicators based on critical data from the main systems of the Unified Health System (Sistema Único de Saúde - SUS), including information on births, deaths, and COVID-19 notifications, using data from the municipality of Camaçari, Bahia, as a case study. The goal is to allow health administrators to analyze essential indicators such as mortality, immunization, hospitalizations, and notifications, highlighting the effectiveness of SKGs in improving public health management. In this way, we integrate relational data sources into a unified and virtualized database, where queries can be executed to answer different Competency Questions (CQs) proposed by the technical team of Camaçari-BA.

Keywords

Data integration, semantic knowledge graphs, ontologies, heterogeneous databases.

1. Introdução

Nas últimas décadas, o conceito de ontologia vem adquirindo destaque em diversas áreas do conhecimento, transpondo suas raízes filosóficas para se tornar uma ferramenta essencial na ciência da computação. No contexto da organização e interpretação de dados, as ontologias fornecem uma estrutura formal e compartilhada para representar o conhecimento de um domínio, definindo um conjunto

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

✉ sandrofranco@ufba.br (S. d. C. Franco); nacles.gomes@ufba.br (N. B. P. Gomes); laisns@ufba.br (L. d. N. Salvador)

🆔 0009-0003-3497-1171 (S. d. C. Franco); 0000-0002-4326-2188 (N. B. P. Gomes); 0000-0001-7441-057X (L. d. N. Salvador)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

estruturado de termos e suas relações (Guarino et al. (2009), Grenon and Smith (2011)). Essa capacidade de integrar semanticamente informações de fontes, domínios e contextos diversos tornou-se criticamente relevante em um cenário de crescimento exponencial de dados, permitindo não apenas a organização conceptual, mas também a inferência e a extração de conhecimento novo, facilitando a interoperabilidade entre sistemas de informação heterogêneos (Grenon and Smith, 2011, de Cerqueira, 2016).

Este estudo insere-se em um contexto de inovação, propondo uma metodologia avançada para a integração de dados heterogêneos, com foco especial no Sistema Único de Saúde (SUS) dos municípios brasileiros. Por meio da construção de um Grafo de Conhecimento Semântico (GCS), este trabalho visa superar os desafios impostos pela diversidade das fontes de dados. Tal abordagem promove uma compreensão mais aprofundada e acessível das informações relativas a registros de óbitos, nascimentos e notificações de casos de COVID-19. Expandindo a pesquisa anterior realizada por (Gomes et al., 2022), este estudo aplica uma metodologia híbrida onde um vocabulário compartilhado acessa a camada ontológica, que por sua vez, acessa as fontes de dados, conforme caracterizado por (Ekaputra et al., 2017), para a elaboração do GCS. Esta técnica destaca o valor dessas ferramentas em descobrir conexões e conhecimentos implícitos nos dados.

Portanto, este estudo objetiva desenvolver um Grafo de Conhecimento Semântico (GCS), ONTOVID II, para integrar bases de dados heterogêneas do Sistema Único de Saúde (SUS) — especificamente do SIM (Sistema de Informações de Mortalidade), SINASC (Sistema de Informações de Nascidos Vivos) e e-SUS Notifica (Sistema de Notificação do Ministério da Saúde). A proposta visa fornecer aos gestores municipais uma visão unificada, servindo como uma ferramenta crucial para a tomada de decisão em saúde pública. Consequentemente, a questão de pesquisa que norteia esta investigação é: Como a integração semântica via GCS pode transformar dados dispersos das bases como SIM, SINASC e e-SUS Notifica em conhecimento acionável para a gestão municipal de saúde?

A próxima seção apresenta abordagens para integração de dados utilizando ontologias, conceitos sobre OBDA (*Ontology-Based Data Access*), OBDI (*Ontology-Based Data Integration*) e Grafos de Conhecimento Semântico (GCS). A seção 3 traz a revisão literária de trabalhos relacionados ao tema proposto. Já a seção 4 apresenta as fontes de dados utilizadas e a construção das ontologias. A seção 5 aborda a definição e utilização do GCS. A seção 6 apresenta os resultados obtidos e a seção 7 traz a conclusão do texto e proposta de trabalhos futuros.

2. Integração de Dados

Considerado um dos problemas mais antigos relacionado a abordagens com dados, o estudo sobre integração teve seu início na década de 90, desde então, uma grande variedade de aspectos referentes a estes problemas foram pesquisados tanto na área acadêmica quanto na indústria, dentre eles a possibilidade de utilizar a integração semântica de dados. Usar semântica significa desenvolver sistemas de integração de dados onde a semântica dos dados é explicitamente especificada e é levada em consideração para o planejamento de todas as funcionalidades do sistema (Flesca et al., 2018). Nas últimas duas décadas, essa ideia se tornou cada vez mais crucial para uma ampla variedade de aplicativos de processamento de informações e tem recebido muita atenção nas comunidades de IA, banco de dados, web e mineração de dados (Natalya F. Noy, 2005). Nesse aspecto, o uso de ontologias, como artefato computacional que modela estruturas de sistemas de informação, vem sendo aplicado na web semântica, promovendo a interoperabilidade entre sistemas. Ontologias computacionais, formalizadas e compartilhadas, são essenciais para organizar e gerenciar conhecimento, definindo entidades e suas relações de maneira que possam ser processadas por máquinas. Essas estruturas não apenas facilitam a comunicação e compreensão entre sistemas e usuários, mas também auxiliam em todas as etapas do ciclo de vida do software, desde a análise de requisitos até a manutenção, por meio de uma especificação clara e consensual dos conceitos e suas inter-relações (Calhau and Falbo, 2010). A consulta a dados utilizando ontologias podem ser realizadas em bases que lidam com o mesmo domínio, neste caso a consulta pode ser mais fácil de ser realizada pois o contexto seria o mesmo, ou sobre bases que lidam com

domínios diferentes, o que poderia tornar as consultas mais difíceis de se realizar. Para os casos mais complicados, pode-se utilizar integração semântica sobre os dados a serem consultados com o objetivo de assegurar que informações distintas, armazenadas em bases distintas, possam ser integradas por possuírem equivalência semântica, ou seja, são relacionados ao mesmo assunto.

Segundo (Gardner, 2005), a chave para ser capaz de integrar informações de forma reutilizável é o uso da semântica, que descreve o significado de uma palavra ou conceito. Neste projeto foi possível responder questões referentes às bases SIM, SINASC e e-SUS Notifica, tais como, por exemplo: Quantos foram os indivíduos imunizados que vieram a óbito por COVID-19 (SIM x COVID-19), Quantos foram os indivíduos nascidos e contaminados com COVID-19 (SINASC x COVID-19), Quantos foram os indivíduos nascidos e contaminados com COVID-19 que vieram a óbito (SIM x SINASC x COVID-19), entre outras possibilidades. Embora a palavra indivíduos seja utilizada em diversas questões, é possível perceber que, mesmo entre bases distintas, há um conceito comum compartilhado entre elas. A partir desse conceito, torna-se viável inferir informações relevantes a respeito desses dados heterogêneos. (Gardner, 2005) diz que a capacidade de distinguir entre estes sinônimos, homônimos e termos relacionados é essencial ao integrar dados de diferentes repositórios.

2.1. OBDA (Ontology-Based Data Access)

Organizações complexas geralmente possuem grande quantidade de dados armazenados em diferentes bases, o que demanda a implementação de consultas unificadas. É o que acontece com as secretarias municipais e estaduais de saúde no acesso às bases do SUS. Uma forma de mitigar este problema é usar a abordagem OBDA que conecta uma ontologia (TBox) a dados de um banco de dados relacional (ABox) por meio de uma camada de mapeamento, que posteriormente pode ser utilizada por um raciocinador automatizado, como ilustrado na Figura 1. Segundo (Bagosi et al., 2014) ODBA é um paradigma de acesso a dados por meio de uma camada conceitual. Essa camada é expressa na forma de uma ontologia, no formato *Web Ontology Language* (OWL), e os dados armazenados em bancos de dados relacionais.

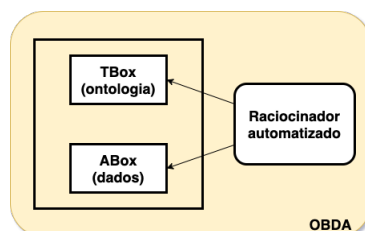


Figura 1: Esquema OBDA (Keet, 2020)

2.2. OBDI (Ontology-Based Data Integration)

Considerando um exemplo em que mesma pessoa é representada em um banco de dados por meio de nome e data de nascimento, num segundo banco de dados por meio do CPF (Cadastro de Pessoa Física) e num terceiro ela seja representada pela CNH (Carteira Nacional de Habilitação). Com a intenção de combinar informações sobre esta mesma pessoa que estão armazenados em bancos de dados diferentes, é preciso encontrar uma maneira em que se possa representá-la e combinar suas informações. Para este tipo de integração de dados armazenados em bancos de dados distintos pode-se utilizar a abordagem *Ontology-Based Data Integration* (OBDI). Segundo (Calvanese et al., 2015), OBDI é uma extensão do OBDA em que os dados não são armazenados em um único banco de dados, mas em uma infinidade de bancos de dados que precisam ser consultados de forma integrada, mas mantendo a mesma arquitetura conceitual baseada em mapeamentos. Uma das principais características dos modelos que empregam o OBDI é a independência conceitual em relação aos dados, o que permite agregar novas visualizações e fontes de dados. O modelo OBDI possui quatro abordagens que podem auxiliar na construção de uma ontologia. Para (Wache et al., 2002) as principais abordagens podem ser classificadas como: a)

abordagem de ontologia única; b) abordagem de múltiplas ontologias; e c) abordagem híbrida. Existe uma quarta abordagem chamada de GAV (Global-as-View), variante adicional proposta por (Ekaputra et al., 2017).

2.3. Grafos de Conhecimento Semântico - GCS

(Soylu et al., 2018) nos definem que os GCS são um mecanismo para consolidar e integrar semanticamente um grande número de fontes de dados heterogêneas em um espaço de dados abrangente. Os GCS são reconhecidos como uma abordagem eficaz no processo de integração semântica de dados heterogêneos, apesar de serem considerados como um paradigma relativamente novo. As ontologias desempenham um papel essencial neste cenário sendo utilizadas como um vocabulário unificado que permite a combinação e enriquecimento de informações para a realização de consultas complexas e inferências sobre os dados integrados (Calvanese et al., 2018). Segundo (Junior, 2021) um GCS é destinado a acumular e transmitir conhecimento do mundo real, cujos nós representam entidades de interesse e cujas arestas representam relações entre essas entidades. Dentre as múltiplas tarefas envolvidas no processo de integração de dados, uma das mais importantes é a compreensão dos conceitos e relações que existem por trás dos registros (Cheatham and Pesquita, 2017).

3. Trabalhos Relacionados

Para iniciar o desenvolvimento deste trabalho foi feita uma revisão na literatura buscando trabalhos que tivessem relação direta ou indireta com o tema da pesquisa. Inicialmente, examinamos pesquisas que demonstravam a aplicação da integração semântica em domínios variados. Por exemplo, (Auceli et al., 2019) exploraram a correlação entre rendimento escolar e ocorrências policiais em Curitiba, integrando dados de educação e segurança pública. No domínio jurídico, (de Oliveira, 2017) aplicou ontologias para recuperar acórdãos do Supremo Tribunal Federal (STF), oferecendo uma alternativa aos formulários tradicionais. Em um escopo mais técnico e generalista, trabalhos como os de (Hajmoosaei and Abdul-Kareem, 2007), (Zhao and Ichise, 2014) e (Xiao et al., 2018) propuseram frameworks e soluções para integração de múltiplos bancos de dados relacionais, abordando desafios de heterogeneidade semântica e de esquema.

No domínio público não específico de saúde, por exemplo, o estudo de (Vidal et al., 2021) se destaca por propor a construção de um Grafo de Conhecimento Semântico (GCS) a partir de fontes heterogêneas como Receita Federal, IBGE e Correios, demonstrando a viabilidade técnica para criar uma representação semântica unificada de dados governamentais. Dentre os trabalhos focados em saúde pública, identificamos importantes esforços que servem como base para esta pesquisa. (Pereira, 2019) e a plataforma SemanticSUS¹ (Lima da Cruz et al., 2019) se concentram na integração semântica entre os sistemas SIM e SINASC, estabelecendo uma base metodológica crucial para a unificação de dados vitais. Aproximando-se do contexto pandêmico, (Maddalena and Baião, 2021) desenvolveram uma ontologia para a COVID-19 baseada na UFO (Unified Foundation Ontology), oferecendo um modelo conceptual valioso. Complementarmente, (Fernantes, 2012) visou à integração de dados epidemiológicos para minimizar a intervenção humana na análise, objetivo que dialoga com a motivação de automatização deste projeto.

Embora essa base literária seja robusta, observa-se uma lacuna crítica: a integração simultânea e tríplice entre dados vitais (SIM, SINASC) e dados epidemiológicos de notificação (e-SUS Notifica). Os trabalhos relacionados concentram-se primordialmente na integração de pares de bases ou na modelagem de um domínio único. Portanto, este projeto diferencia-se ao propor um GCS que consolida e inter-relaciona essas três fontes heterogêneas, com o propósito explícito de gerar insights acionáveis para a tomada de decisão municipal em tempo crítico, bem como facilitar a inclusão de novas fontes de dados do SUS, ampliando o GCS.

¹<https://semanticsus.github.io/semanticSUS/index.html>

4. Fontes de dados

Para dar início à implementação das ontologias inicialmente firmou-se uma parceria entre a UFBA e a Secretaria Municipal de Saúde de Camaçari. Esta parceria foi necessária para que os dados pudessem ser acessados e com isso a ontologia pudesse ser construída com o objetivo de ser utilizada na Secretaria. Seguindo a Lei Geral de Proteção de Dados (LGPD), Lei nº 13.709/2018, este trabalho não apresenta dados pessoais, apenas informações quantitativas. As bases utilizadas neste trabalho foram as do SIM, SINASC e e-SUS Notifica, elas representam a Camada de Fonte de Dados da Figura 2. Na figura pode-se observar

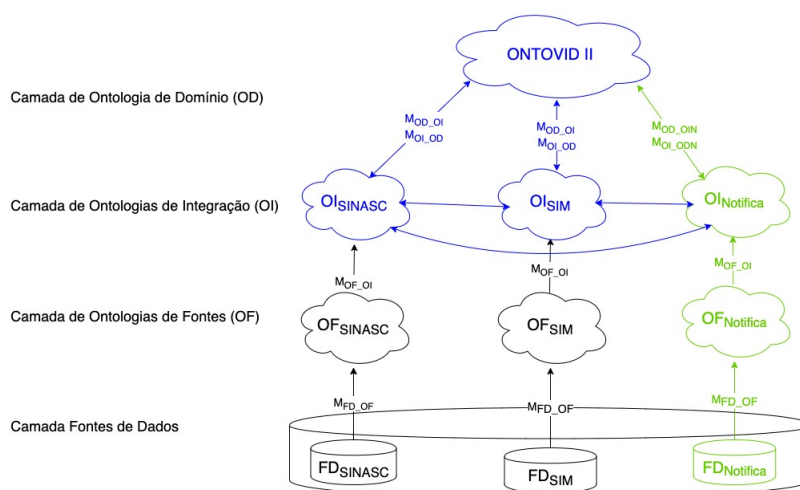


Figura 2: Fluxo proposto para este projeto. (Fonte: autor)

o modelo proposto para este projeto onde, em verde, tem-se a inclusão da ontologia de Notificações de COVID-19 e em azul a realização da integração semântica entre as ontologias SINASC x SIM x COVID-19. Com estas integrações, na Camada de Ontologia de Domínio (OD) encontra-se a ontologia capaz de responder questões de competência referentes às fontes de dados apresentadas. As bases do SIM e SINASC foram configuradas em um banco de dados Firebird², rodando no sistema operacional Windows³. Por se tratar de um banco de dados antigo, foram disponibilizados dois arquivos de extensão **.fdb**, um para cada base de dados, não foi disponibilizado o acesso direto a base da secretaria. Para acesso a base de dados do e-SUS Notifica, foi disponibilizada uma *view* em PostgreSQL⁴ que roda num servidor Linux OpenSUSE⁵.

4.1. SINASC

A base de dados do SINASC contém informações sobre nascimentos e, assim como os dados relacionados ao SIM, apenas a tabela principal **TB_DN** (tabela sobre dados de nascimento) foi utilizada. Desta tabela consegue-se obter informações necessárias para construir a ontologia de fonte OF_{SINASC}, são elas: APGAR⁶ primeiro minuto, APGAR quinto minuto, CNS, data de nascimento, ID, idade, nome completo, número de declaração de nascido vivo, peso ao nascer, quantidade de nascidos mortos, sexo, tempo de gestação e tipo de parto.

²<https://firebirdsql.org/>

³<https://www.microsoft.com/pt-br/windows>

⁴<https://www.postgresql.org/>

⁵<https://www.opensuse.org/>

⁶APGAR (Aparência, Pulso, Gesticulação, Atividade, Respiração) é uma escala que avalia a saúde do recém-nascido nos primeiros minutos de vida

4.2. SIM

A base de dados do SIM contém informações sobre mortalidade, para este projeto não foi utilizada toda a estrutura de dados do SIM, apenas a tabela principal **TB_DO**. Desta tabela consegue-se obter informações necessárias para construir a ontologia de fonte OF_SIM, são elas: CID, CNS, data do atestado de óbito, data do óbito, data de nascimento, dias vividos, ID, idade, nome completo, número de declaração de nascido vivo, número da declaração do óbito, quantidade de nascidos mortos e sexo.

4.3. e-SUS Notifica

A base de dados do e-SUS Notifica contém as informações de notificação de COVID19. Esta base trata-se de uma *view* que contém informações necessárias para construção da OF_NOTIFICA, são elas: bairro, CEP, cidade, classificação final, CNS, ID, CPF, data do teste, data do teste Lamp, data do teste PCR, data de nascimento, data da primeira dose, data da segunda dose, CBO, e-mail, estado, estado do teste, evolução do caso, laboratório da primeira dose, laboratório da segunda dose, nome completo, número da notificação, primeira dose, segunda dose, resultado do teste, sexo, sintoma e tipo do teste.

4.4. Ontologias de Fonte

Neste trabalho, as ontologias de fonte são as ontologias que representam as fontes de dados a elas associadas, sendo assim tem-se OF_SIM, OF_SINASC e OF_NOTIFICACOES, estando elas relacionadas as bases dados do SIM, SINASC e e-SUS Notifica respectivamente. Todas as ontologias foram criadas utilizando a linguagem OWL 2 com sintaxe RDF/XML e estas representam a Camada de Ontologias de Fonte(OF). A da Figura 3 mostra um exemplo de ontologia de fonte, a OF_SIM. Nesta ontologia temos as

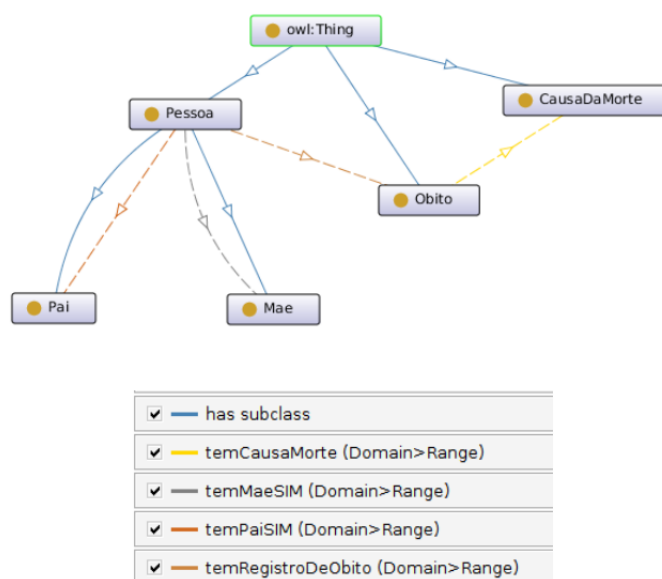


Figura 3: Representação da ontologia de fonte OF_SIM construída no Protégé. (Fonte: autor)

classes: i) **CausaDaMorte**, que representa a causa do óbito; ii) **Óbito**, que possui as informações sobre o óbito de uma pessoa; iii) **Pessoa**, que representa uma pessoa na base do SIM; iv) **Mãe**, que possui dados sobre a mãe de uma pessoa que veio a óbito e; v) **Pai**, que possui dados sobre o pai de uma pessoa que veio a óbito. A classe **Pessoa** tem as subclasses (**has subclass**) **Mae** e **Pai**, as linhas pontilhadas representam as relações entre as classes, onde i) Pessoa **temRegistroDeObito** da classe Obito; ii) Pessoa pode ter informações do pai (**temPaiSIM**); iii) Pessoa tem informações da mãe (**temMaeSIM**) e; iv) Obito **temCausaMorte** da classe CausaMorte.

4.5. Ontologias de Integração

As ontologias de integração criadas para este projeto dão continuidade ao trabalho iniciado com as ontologias de fonte. As ontologias de integração também foram criadas no Protégé onde, cada ontologia foi criada a partir da importação da ontologia de fonte correspondente. Assim temos: i) OI_SIM importada de OF_SIM; ii) OI_SINASC importada de OF_SINASC e iii) OI_NOTIFICACOES importada de OF_NOTIFICACOES. As ontologias de integração construídas representam a Camada de Ontologias de Integração (OI) da Figura 2 da Proposta. Utilizando recurso de importação, qualquer alteração que seja realizada na ontologia de fonte automaticamente se reflete na ontologia de integração, isso reduz a necessidade de alteração em ambas ontologias. Diferentemente das ontologias de fonte, as ontologias de integração possuem ligação com bases de dados onde o Protégé se conecta às bases utilizando *drivers JDBC*. A estratégia de vinculação de registros utilizada na Ontovid II baseia-se no conceito de “indivíduos” como elementos de referência compartilhada entre as bases. Para cada entidade, são priorizados identificadores disponíveis como CPF, data de nascimento, sexo e município de residência. Nos casos em que algum identificador está ausente, a instância é criada apenas com os atributos disponíveis, preservando a consistência semântica, ainda que com menor capacidade de seleção. Não foram aplicados, nesta fase, algoritmos avançados de resolução de conflitos, aspecto reconhecido como limitação e direcionado para evolução em trabalhos futuros.

5. ONTOVID II

A necessidade de produzir informações integradas e de fácil acesso tem impulsionado o desenvolvimento de novas soluções que facilitem o processo de integração entre estes dados. Este tipo de integração pode ser complexa e onerosa devido à heterogeneidade semântica ou distribuição de dados, e isso também acontece com dados da área de saúde. Soluções para este tipo de integração tem se mostrado eficiente com abordagens que utilizam ontologias. Este trabalho tem como objetivo realizar a integração semântica entre as bases SIM, SINASC e e-SUS Notifica por meio de grafos de conhecimento semântico, de forma a minimizar as dificuldades recorrentes enfrentadas por gestores do SUS no acesso a dados de saúde integrados. A proposta visa disponibilizar indicadores consolidados que apoiem análises relacionadas a nascimentos, óbitos e notificações de COVID-19, fortalecendo a capacidade de planejamento e gestão em saúde pública. Para isso foram utilizadas soluções semelhantes para criação da ontologia com domínio relacionado a dados sobre Nascimentos, Óbitos e Notificações de COVID-19. Para além, utilizou-se também o cenário 4 da metodologia NeOn ⁷, neste processo foram desenvolvidas as ontologias determinando seu domínio e escopo. Uma vez construídas as ontologias de fonte e de integração, o passo subsequente foi a implementação da ontologia de domínio a ser utilizada pelo Grafo de Conhecimento Semântico (GCS). No modelo proposto, representado na Camada Semântica da Figura 6, esta ontologia de domínio corresponde à camada ilustrada na Figura 2. A ontologia de domínio resultante, denominada ONTOVID II e representada na Figura 4, consolida os conceitos das três fontes de dados (e-SUS Notifica, SINASC e SIM).

Sua estrutura pode ser analisada da seguinte forma: a) *Bloco 1*: Agrupa as classes e subclasses necessárias para representar o domínio do e-SUS Notifica, como *TesteCovid19*, *Notificacao*, *ProfissionalDeSaude* e *Vacina*; b) *Bloco 2*: Contém as classes e subclasses referentes aos dados do SINASC, como *RegistroNascidoVivo*, *Gestacao*, *Nascimento* e *Mae*; e c) *Bloco 3*: Apresenta as classes e subclasses do SIM, centradas na classe *Obito* e *CausaDaMorte*. Todas essas classes herdam propriedades e restrições de suas respectivas ontologias locais, garantindo a consistência semântica. Finalmente, as classes *Pessoa*, *Pai* e *Mae* (à direita) desempenham um papel crucial como classes de unificação, agregando e harmonizando as propriedades comuns provenientes das três ontologias locais, o que permite uma visão integrada e coerente do cidadão nos diferentes sistemas.

⁷Cenário 4 da metodologia NeOn é um dos nove cenários propostos para o desenvolvimento de ontologias, focado especificamente na reutilização e reengenharia de recursos ontológicos não ontológicos.

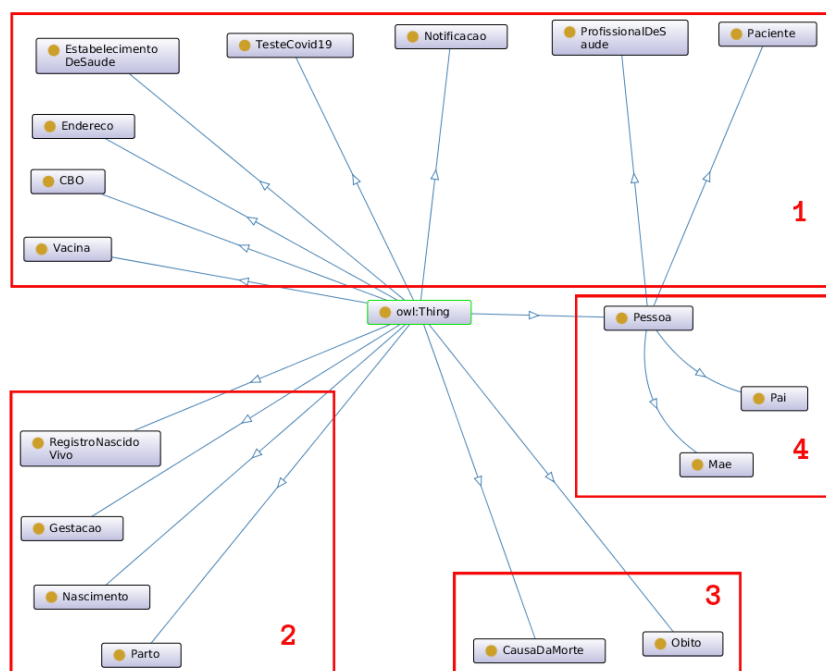


Figura 4: Ontologia de domínio Ontovid II. (Fonte: autor)

A Figura 5 mostra o modelo conceitual do GCS que representa a ontologia de domínio presente na Figura 2 e, a seguir serão apresentadas as classes das ontologias locais, que permearam o acesso as fontes de dados do SIM, SINASC e e-SUS Notifica:

- SIM: i) Óbito e; ii) Causa da Morte
- SINASC: i) Gestação; ii) Parto; iii) Nascimento e; iv) Registro de Nascido Vivo
- e-SUS Notifica: i) Paciente; ii) Endereço; iii) Profissional de Saúde; iv) Vacina; v) Notificação; vi) Classificação Brasileira de Ocupação (CBO) e vii) Estabelecimento de Saúde

Após levantamento dos modelos conceituais das bases, observou-se que as classes **Pessoa**, **Pai** e **Mãe** são conceitos que se relacionam a todas ontologias das fontes de integração. Com relação ao SIM observa-se que *Pessoa* que tem *Mãe*, tem *Pai*, tem registro de *Óbito* e este tem a *Causa da Morte*. Com relação ao SINASC, uma *Pessoa* nascida possui *Registro de Nascido Vivo* e neste registros encontram-se informações do *Parto*, do *Nascimento*, da *Gestação*, da *Mãe* e do *Pai*. Com relação às Notificações, *Profissional de Saúde* é uma *Pessoa* que aplica *Vacina* e o mesmo pode se vacinar, tem *CBO*, realiza as *Notificações* e tem vínculo com *Estabelecimento de Saúde*. *Paciente* é uma *Pessoa* que tem dados da *Notificação* e a notificação tem o registro de *Vacina* e do *TesteCovid19*.

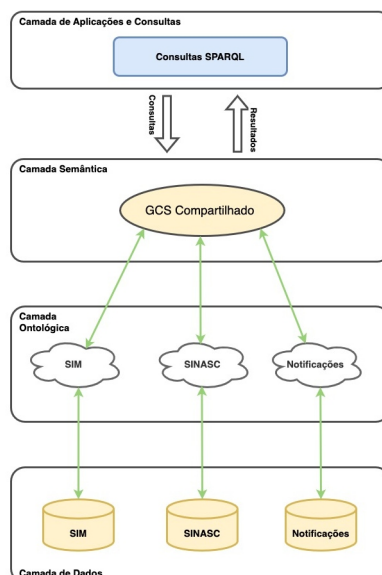


Figura 6: Modelo ONTOVID II em camadas.(Fonte: autor).

Para este caso foram formuladas **seis** Questões de Competência (QCs) envolvendo somente a base de Notificações e **oito** envolvendo as bases do SIM, SINASC e Notificações. Seguem as Questões de Competência levantadas para este projeto e seus resultados em 19 de julho de 2023. Observando os resultados relacionados na Tabela 1, vale ressaltar que as consultas realizadas para compor o Painel de Mortalidade (SIM), o Painel de Natalidade (SINASC) e o Painel de Notificações de COVID-19 podem ser consideradas consultas simples pois suas ontologias acessam diretamente uma única base de dados. Já as consultas realizadas para compor o Painel de Saúde Materno-Infantil (SIM - SINASC) e o Painel de Notificações x SIM x SINASC podem ser consideradas consultas complexas pois estas acessam duas ou mais bases de dados diferentes. Com relação a estas consultas mais complexas, o uso de ontologias se torna um facilitador pois seu dicionário de dados abstrai as relações entre estas bases de dados distintas, assim, conhecendo este dicionário, construir consultas que relacionem estas bases se torna uma tarefa mais simplificada de se realizar. As Questões de Competência (QCs) descritas na Tabela 1 foram construídas em linguagem SPARQL e executadas no Protégé. Todos os resultados foram apresentados à Secretaria Municipal de Saúde de Camaçari para validação dos gestores. As consultas foram executadas no ambiente de desenvolvimento e os responsáveis pela validação analisaram e aceitaram⁸ como válidos os dados informados nas consultas, com isso, há um forte indício para utilização da ONTOVID-II em ambiente de produção. A avaliação realizada concentrou-se na validação qualitativa dos resultados junto à equipe técnica da Secretaria Municipal de Saúde de Camaçari. Essa abordagem foi suficiente para demonstrar a correção das consultas formuladas, mas não contemplou métricas quantitativas como tempo médio de execução, escalabilidade ou indicadores de precisão no processo de vinculação de registros. O reconhecimento dessas limitações é importante para orientar pesquisas futuras, nas quais se pretende ampliar os experimentos e adotar medidas comparativas de desempenho.

⁸https://drive.google.com/drive/folders/1Ow_Y41NJERS2TWEGGK71n9U_zgVUzEH1?usp=sharing

Tabela 1

Questões de Competência para ONTOVID II

(Fonte: autor)

Painel de Mortalidade (SIM)	Resultado
QC 01: Número de óbitos nos anos 2021, 2022 e 2023	2021: 2088 2022: 1580 2023: 576
QC 02: Mortalidade por COVID-19 - CID B342	Total: 799
Painel de Natalidade (SINASC)	Resultado
QC 01: Número de nascidos vivos nos anos 2021, 2022 e 2023	2021: 3880 2022: 2220 2023: 1505
QC 02: Taxa bruta de natalidade nos anos 2021, 2022 e 2023 - Município IBGE 290570 (Camaçari)	2021: 5,52% 2022: 3,16% 2023: 2,14%
Painel de Notificações de COVID-19	Resultado
QC 01: Óbitos que ocorreram em função do COVID-19	Total: 565
QC 02: Total de pessoas se recuperaram plenamente do COVID-19	Total: 26216
QC 03: Total de pessoas vacinadas pegaram COVID-19 (confirmados em laboratório)	Total: 410
Painel de Saúde Materno-Infantil (SIM - SINASC)	Resultado
QC 01: Número de Mulheres que vieram a óbito em Idade Fértil (faixa etária de 10 a 49 anos)	1418
QC 02: Taxa de Mortalidade Infantil por 1000 nascidos vivos - Município IBGE 290570 (Camaçari)	2021: 0,81% 2022: 0,77% 2023: 0,30%
Painel de Notificações x SIM x SINASC	Resultado
QC 01: Óbitos com registro no SIM	Total: 439
QC 02: Mortalidade por COVID-19 com registro no SIM	Total: 389
QC 03: Mortalidade por COVID-19 com registro no SIM por ano de nascimento	2020: 158 2021: 277 2022: 9

7. Conclusão e Trabalhos Futuros

O objetivo principal deste trabalho foi realizar a integração semântica entre as bases heterogêneas SIM, SINASC e e-SUS Notifica, utilizando grafos de conhecimento semântico, com o propósito de apoiar gestores do SUS no acesso a informações integradas sobre nascimentos, óbitos e notificações de COVID-19.

Esse objetivo foi alcançado por meio do desenvolvimento da ontologia Ontovid II, construída a partir de uma abordagem híbrida de integração. A solução permitiu responder a questões de competência formuladas pela Secretaria Municipal de Saúde de Camaçari-BA, com resultados validados pela equipe técnica, demonstrando seu potencial de aplicação prática em ambiente produtivo. Assim, pode-se afirmar que a questão de pesquisa que orientou este estudo — “Como integrar semanticamente bases heterogêneas de saúde pública (SIM, SINASC e e-SUS Notifica), de modo a disponibilizar informações integradas que apoiem gestores do SUS na tomada de decisão?” — foi respondida com a construção e validação da ontologia Ontovid II.

Apesar dos resultados alcançados, este estudo apresenta algumas limitações. A solução foi validada apenas com dados do município de Camaçari-BA, o que restringe sua generalização imediata para outros contextos. Além disso, não foi desenvolvido um protótipo de interface para uso direto pelos gestores, o que limita a aplicabilidade imediata em ambiente produtivo. Por outro lado, os resultados evidenciam relevantes implicações práticas: a utilização de grafos de conhecimento semântico mostrou-se eficaz para integrar dados heterogêneos de saúde, reduzindo a complexidade técnica do acesso às informações e oferecendo suporte direto para gestores municipais no acompanhamento de indicadores de mortalidade, natalidade e notificações de COVID-19.

Como trabalhos futuros, pretende-se: i) incorporar outras bases de dados de saúde, como a da Campanha Nacional de Vacinação contra a COVID-19; ii) realizar o alinhamento das ontologias desenvolvidas com ontologias de alto nível, como DOLCE ou BFO; iii) desenvolver um protótipo funcional para apresentar os resultados em ambiente produtivo e; iv) implementar métricas quantitativas, avaliação de desempenho e testes de escalabilidade para que se possa avaliar a robustez da integração sob uso em larga escala. Além disso, pretende-se disponibilizar as ontologias, arquivos de mapeamento e consultas SPARQL em repositório aberto, acompanhados de conjuntos de dados sintéticos anonimizados, a fim de favorecer a reprodutibilidade e a transparência metodológica. Também se planeja expandir os experimentos para outros municípios, explorando a generalização da abordagem em cenários com diferentes perfis e padrões de qualidade de dados.

Agradecimentos

Os autores gostariam de agradecer o suporte financeiro recebido da Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), por meio do projeto "UAware: Uma Plataforma de Comunicação de Crise Ubíqua e Sensível ao Contexto de apoio ao Gerenciamento de Emergências em Cidades e Áreas Industriais"(Termo de Outorga TIC 0008/2015).

Agradecemos também à Prefeitura Municipal de Camaçari, em especial à Secretaria de Saúde e à Diretoria de Planejamento, pelo indispensável acesso aos dados, pelo compartilhamento de conhecimento técnico do domínio e pelo constante apoio durante a realização desta pesquisa.

Declaração sobre o Uso de Inteligência Artificial Generativa

Durante a preparação deste trabalho, os autores utilizaram o ChatGPT (GPT-4, OpenAI) e o GitHub Copilot para apoio no aprimoramento textual, verificação gramatical e ortográfica, bem como para sugestões de código. Após a utilização dessas ferramentas, o conteúdo foi integralmente revisado e editado pelos autores, que assumem total responsabilidade pelo conteúdo da publicação.

Referências

- N. Guarino, D. Oberle, S. Staab, What is an ontology?, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 1–17. URL: https://doi.org/10.1007/978-3-540-92673-3_0. doi:10.1007/978-3-540-92673-3_0.
- P. Grenon, B. Smith, Foundations of an ontology of philosophy, *Synthese* 182 (2011) 185–204. doi:10.1007/s11229-009-9658-x.
- L. D. de Cerqueira, Uma Abordagem Baseada em Ontologias para Integração Semântica de Sistemas na Camada de Processos, Master's thesis, Universidade Federal do Espírito Santo, 2016.
- N. Gomes, S. Franco, L. Salvador, Ontovid - uma abordagem para construção de grafos de conhecimento semântico com enfoque em notificações e Óbitos relacionados ao novo coronavírus (covid-19), 2022, pp. 425–436. doi:10.5753/sbcas.2022.222723.
- F. Ekaputra, M. Sabou, E. Serral, E. Kiesling, S. Biffl, Ontology-based data integration in multi-disciplinary engineering environments: A review, *Open Journal of Information Systems (OJIS)* 4 (2017) 1–26.
- S. Flesca, S. Greco, E. Masciari, D. Saccà, A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, volume 31, 2018. doi:10.1007/978-3-319-61893-7.
- A. Y. H. Natalya F. Noy, AnHai Doan, Semantic integration (editorials), *ai magazine*, 2005, p. 26(1). doi:<https://doi.org/10.1609/aimag.v26i1.1794>.
- R. Calhau, R. Falbo, An ontology-based approach for semantic integration, 2010, pp. 111 – 120. doi:10.1109/EDOC.2010.32.
- S. P. Gardner, Ontologies and semantic data integration, *Drug Discovery Today* 10 (2005) 1001–1007. URL: <https://www.sciencedirect.com/science/article/pii/S135964460503504X>. doi:[https://doi.org/10.1016/S1359-6446\(05\)03504-X](https://doi.org/10.1016/S1359-6446(05)03504-X).
- T. Bagosi, D. Calvanese, J. Hardi, S. Komla-Ebri, D. Lanti, M. Rezk, M. Rodríguez-Muro, M. Slusnys, G. Xiao, The ontop framework for ontology based data access, in: D. Zhao, J. Du, H. Wang, P. Wang, D. Ji, J. Z. Pan (Eds.), *The Semantic Web and Web Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 67–77.
- C. M. Keet, An introduction to ontology engineering, volume v1.5, 2020, pp. 165–175.
- D. Calvanese, M. Giese, D. Hovland, M. Rezk, Ontology-based integration of cross-linked datasets, in: M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, K. Thirunarayan, S. Staab (Eds.), *The Semantic Web - ISWC 2015*, Springer International Publishing, Cham, 2015, pp. 199–216.
- H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, S. Ubner, Ontology-based integration of information - a survey of existing approaches, *Proceedings of the IJCAI'01 Workshop on Ontologies and Information Sharing*, Seattle, Washington, USA, Aug 4-5 (2002).
- A. Soylu, O. Corcho, E. Simperl, D. Roman, F. Martínez, C. Taggart, I. Makgill, B. Elvesaeter, B. Symonds, H. McNally, G. Konstantinidis, Y. Zhao, T. Lech, Towards integrating public procurement data into a semantic knowledge graph, 2018.
- D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, *Ontology-Based Data Access and Integration*, Springer International Publishing, 2018, p. 2590–2596.
- A. B. B. Junior, Um framework baseado em conhecimento de senso comum para sistemas de perguntas e respostas sobre grafo de conhecimento, 2021. URL: <https://repositorio.ufrn.br/handle/123456789/45676>.
- M. Cheatham, C. Pesquita, Semantic data integration. in *handbook of big data technologies*, Springer International Publishing (2017) 263–305. doi:<https://doi.org/10.1007/978-3-319-49340-4>.
- P. Auceli, R. Berardi, N. Kozievitch, Integração semântica entre dados dos domínios da educação e segurança: um caso em curitiba, in: *Anais da XV Escola Regional de Banco de Dados*, SBC, Porto Alegre, RS, Brasil, 2019, pp. 141–150. URL: <https://sol.sbc.org.br/index.php/erbd/article/view/8487>. doi:10.5753/erbd.2019.8487.
- R. B. de Oliveira, Utilização de Ontologias Para Busca em Base de Dados de Acórdãos do STF, Master's thesis, Universidade de São Paulo, São Paulo, 2017.
- A. Hajmoosaei, S. Abdul-Kareem, An ontology-based approach for resolving semantic schema conflicts

- in the extraction and integration of query-based information from heterogeneous web data sources, in: *Proceedings of the Third Australasian Workshop on Advances in Ontologies - Volume 85, AOW '07*, Australian Computer Society, Inc., AUS, 2007, p. 35–43.
- L. Zhao, R. Ichise, Ontology integration for linked data, *Journal on Data Semantics* 3 (2014) 237–254. URL: <https://doi.org/10.1007/s13740-014-0041-9>. doi:10.1007/s13740-014-0041-9.
- G. Xiao, D. Hovland, D. Bilidas, M. Rezk, M. Giese, D. Calvanese, Efficient ontology-based data integration with canonical iris, in: A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), *The Semantic Web*, Springer International Publishing, Cham, 2018, pp. 697–713.
- T. Vidal, C. Viktor, S. Avila, R. Mariano, T. Calixto, P. Ivo, J. Filho, A. Brayner, M. Vidal, *Uso das tecnologias da web semântica na construção de grafos de conhecimento semântico baseado no enfoque híbrido*, 2021.
- D. L. N. C. Pereira, *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, São Paulo, 2019.
- M. M. Lima da Cruz, C. Viktor, V. Vidal, N. Arruda Junior, *Semanticsus: Um portal semântico baseado em ontologias e dados interligados para acesso, integração e visualização de dados do sus*, 2019, pp. 13–18. doi:10.5753/sbcas.2019.6277.
- L. Maddalena, F. Baião, *Ontocovid: Aplicando sabio para a modelagem conceitual bem fundamentada no domínio da covid-19. (ontocovid: Applying sabio to conceptual modeling well grounded in the covid-19 domain)*, in: F. Farinelli, R. C. G. Berardi, J. L. Carbonera, D. Schmidt (Eds.), *Proceedings of the XIV Seminar on Ontology Research in Brazil (ONTOBRAS 2021) and V Doctoral and Masters Consortium on Ontologies (WTDO 2021)*, Online, Brazil, November 16-19, 2021, volume 3050 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 259–266. URL: <http://ceur-ws.org/Vol-3050/Short5.pdf>.
- F. L. G. D. Fernantes, *Integração de Dados Baseada em Ontologias e Raciocínio Automático: Estudo de Caso com Dados Públicos de Saúde*, 2012. URL: <https://repositorio.ufpe.br/handle/123456789/10973>.