Integração Semântica de Dados Públicos: O Caso do Observatório dos Sertões de Crateús com o Framework OBSCRAT

Yuri Cristian P. de Oliveira[1], Emanoel B. Alves[1], Amanda Drielly Pires Venceslau[2], Lisieux Marie M. dos S. Andrade[1], Marciel Barros Pereira[1], Giovanna Maria V. Xavier[1]

[1]Campus Crateús – Universidade Federal do Ceará (UFC), Crateús - CE - Brazil

[2]Instituto Universidade Virtual – Universidade Federal do Ceará (UFC), Fortaleza - CE - Brazil

yuricristiam12@gmail.com, emanoel-alves09@hotmail.com, amanda.pires@ufc.br

lisieuxandrade@ufc.br, marciel@crateus.ufc.br, giovannaverissimox@alu.ufc.br

Abstract. *According to the Brazilian Federal Constitution, access to information is a fundamental right guaranteed to all citizens. This right contributes to public transparency, combats misinformation, and encourages social participation. Based on this principle, the Open Data Observatory of the Sertões de Crateús was created. This work aims to enrich the data made available by this Observatory. To this end, a semantic integration framework called OBSCRAT was developed. As a case study, the possible correlation between COVID-19 cases and mortality from respiratory diseases was investigated using two heterogeneous databases. The queries performed indicated that there were no cases of underreporting in the region for the period analyzed. Therefore, the experiment demonstrates that the proposal successfully addressed a social problem through data integration and enrichment.*

Resumo. *De acordo com a Constituição Federal Brasileira, o acesso à informação é um direito fundamental garantido a todos os cidadãos. Esse direito contribui para a transparência pública, combate a desinformação e incentiva a participação social. Com base nesse princípio, foi criado o Observatório de Dados Abertos dos Sertões de Crateús. Este trabalho tem como objetivo enriquecer os dados disponibilizados por esse Observatório. Para isso, foi desenvolvido um framework de integração semântica chamado OBSCRAT. Como estudo de caso, investigou-se a possível correlação entre casos de COVID-19 e mortalidade por doenças respiratórias, utilizando duas bases de dados heterogêneas. As consultas realizadas indicaram que não houve casos de subnotificação na região no período analisado. Portanto, o experimento demonstra que a proposta abordou com sucesso um problema social por meio da integração e enriquecimento de dados.*

1. Introdução

Diante da importância do compartilhamento de dados de forma aberta para a população do oeste do estado do Ceará, e seguindo a Lei de Acesso à Informação (LAI), surge o Observatório de Dados Abertos da Região dos Sertões de Crateús, que corresponde a um

portal web[1], que visa contribuir com a transparência das informações úteis para à sociedade dos Sertões de Crateús. Inicialmente, foram coletados dados relacionados à saúde, que correspondem a dados históricos da COVID-19 e índices de mortalidade baseado no CID-10, sendo essas heterogêneas. Atualmente, a plataforma busca centralizar, coletar e divulgar informações públicas para a população da região, entretanto, assim como outros portais de dados, ainda não dispõe de semântica sob as informações disponibilizadas.

Fontes heterogêneas necessitam de modelos conceituais para combinar informações, provendo o enriquecer dos dados para tomada de decisão. Segundo [Thor et al. 2007], *Mashups* são aplicações web interativas que combinam conteúdo de múltiplos serviços ou fontes em um novo serviço ou fonte de dados.

Com objetivo de prover integração semântica de dados, alguns *frameworks* [Schultz et al. 2011, Lopes et al. 2016, Cavalcante 2017] que fornecem *Mashup* foram investigados. Contudo, observou-se a falta de um *pipeline* de integração e enriquecimento dos dados, código proprietário, bibliotecas obsoletas, e, para algumas etapas de integração, como a descoberta de link, nenhum suporte para o tratamentos dos dados brutos e consequentemente para os dados *Resource Description Framework* (RDF)[2].

Visando apoiar o processo de integração semântica desde a coleta dos dados, normatização, construção dos grafos e vinculação, foi elaborado o OBSCRAT. Um *framework* que provê uma infraestrutura para integração de grafos semânticos, fornecendo ferramentas para enriquecimento semântico de dados. Para o estudo de caso, foram utilizados os dados disponíveis no Observatório de Dados Abertos da Região dos Sertões de Crateús, particularmente no cenário da saúde, a fim de responder perguntas sobre mortes por doenças respiratórias e casos de COVID-19 no mesmo período, uma vez que o Observatório não possuía tal recurso.

Este artigo está organizado em seis seções, em que a segunda apresenta dois trabalhos correlatos que contribuíram para a condução da pesquisa. A terceira seção, é apresentado o *framework* OBSCRAT. A quarta seção revela a metodologia adotada no presente estudo. A quinta seção apresenta os resultados e as discussões, e a sexta e última, as considerações finais.

2. Trabalhos Relacionados

Na literatura é possível encontrar trabalhos relevantes para a problemática. Em [Cavalcante 2017], os autores apresenta um *framework* baseado em mediador semântico para construção e reutilização de *Linked Data Mashups*, o MAURA. Neste, é possível criar *Mashups* baseados em parâmetros, não sendo necessário domínio sobre Integração de Dados ou Web Semântica. O estudo apresentou resultados positivos da aplicação do MAURA no contexto da gestão de dados de saúde materno-infantil, em que foi possível identificar áreas de alto risco para intervenções prioritárias e com os dados integrados e semanticamente enriquecidos, foi possível planejar intervenções mais eficazes, alocando recursos de forma mais direcionada e eficiente.

O trabalho [da Cruz 2021] apresenta uma abordagem semi-automática para construção de um *Mashup* de dados como uma visão sobre um *Enterprise Knowledge*

[1] `https://observatorio-dados.crateus.ufc.br/`
[2] `https://www.w3.org/RDF/`

Graph (EKG). A arquitetura de EKG considerada é dividida em quatro camadas de abstração, onde a camada inferior representa um nível de abstração menor em relação a camada superior. A metodologia semiautomática baseada em tecnologias de web semântica faz uso de consultas facetadas e regras de fusão. Além disso, explora estudos de caso, como o portal SemanticSUS e o EKG-SefazMA, demonstrando a aplicação prática da abordagem.

Para atingir os objetivos desta pesquisa, o procedimento de *Mashup* de Dados se destaca como um importante mecanismo para a realização da integração de bases heterogêneas. A seguir será apresentado o framework proposto e sua aplicação.

3. OBSCRAT: Framework Semântico para o Observatório de Dados Abertos dos Sertões de Crateús

O framework OBSCRAT[3] possui uma arquitetura baseada na apresentada por [da Cruz 2021], e portanto, dividida em quatro camadas: Camada de Dados, Camada Semântica, Camada de Integração de Dados e Camada de Aplicações, onde cada camada representa um diferente nível de abstração. A Figura 1 representa a visão geral dessa arquitetura.

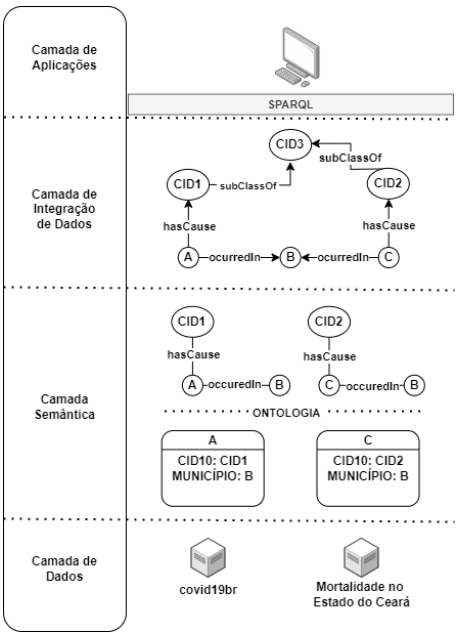

Figura 1. Arquitetura da Solução adaptada de [da Cruz 2021].

3.1. Camada de Dados

A *Camada de Dados* é composta por fontes de dados definidas para serem integradas [da Cruz 2021], nela os dados podem estar organizados em diversos formatos, tais como, banco de dados relacional, NoSQL, CSV, XML, entre outros. O repositório covid19br[4] disponibiliza um grande quantidade de dados, em formato *Comma-separated values* (CSV), relacionados à COVID-19 no Brasil, como número de casos e óbitos por

[3]https://github.com/YuriPedrosa/OBSCRAT
[4]https://github.com/wcota/covid19br

município, provenientes da junção dos dados oficias de todas as secretarias de saúde de cada estado[5].

3.2. Camada Semântica

A *Camada Semântica* possui como responsabilidade o tratamento dos problemas de integração existentes entre as fontes de dados. No trabalho [da Cruz 2021], os autores consideram os componentes: ontologia de domínio, conjunto de visões exportadas e conjunto de visões *linkset*. O OBSCRAT considerada apenas os dois primeiros para esta camada, deixando a *Camada de Integração de Dados* responsável por explicitar as ligações existentes entre os recursos. Na Figura 1, essa camada é representada pelo nível intermediário onde as informações das fontes de dados são estruturadas semanticamente. Aqui, as instâncias CID1 e CID2 são apresentadas com suas respectivas relações *hasCause* e *occurredIn* associadas ao município B. Isso demonstra como os dados brutos são transformados em uma representação semântica utilizando uma ontologia.

3.3. Camada de Integração de Dados

A *Camada de Integração de Dados* é responsável por definir links existentes entre recursos de diferentes visões exportadas provenientes da *Camada Semântica*. Tornando clara a relação de equivalência entre dois recursos diferentes, indicando quando estes se referem a um mesmo objeto ou estão correlacionados [da Cruz 2021]. Na Figura 1, essa camada é ilustrada pelo nível onde CID1 e CID2 são integrados com a ajuda da ontologia, mostrando que ambas as classes são subclasses de CID3. Além disso, as propriedades *hasCause* e *occurredIn* são usadas para conectar instâncias a causas e locais, respectivamente, estabelecendo uma rede de informações que facilitam a análise de dados de diferentes fontes.

3.4. Camada de Aplicações

A última camada é composta pelas aplicações que podem dispor de consultas SPARQL[6], definidas sobre um *Mashup* de dados. O grafo de conhecimento fornecido pela *Camada de Integração de Dados* pode ser consultado e os dados resultantes são consumidos e disponibilizados [da Cruz 2021]. Diante do que foi apresentado, a arquitetura estabelecida foi a estrutura base para a próxima fase do presente trabalho, servindo como um guia para a análise que se segue, fornecendo um caminho estruturado para investigar as relações entre a COVID-19 e a mortalidade por outras doenças respiratórias. Na Figura 1, essa camada é representada no topo, onde através de um portal consultas SPARQL podem ser executadas sobre a ontologia de domínio disponível.

4. OBSCRAT: Metodologia de Integração Semântica

A abordagem metodológica do OBSCRAT está dividida em duas etapas: a integração dos dados e a análise estatística.

[5]http://extranet.saude.ce.gov.br/tabulacao/deftohtm.exe?sim/obito.def
[6]https://www.w3.org/TR/sparql11-query/

4.1. Integração de Dados

Corresponde à etapa principal do processo metodológico adotado neste trabalho, visto que garante a coleta, organização e preparação dos dados para as análises subsequentes. Essa fase é subdividida em quatro etapas, baseadas na arquitetura proposta, as quais são: Definição e Expansão da Ontologia, Coleta e Tratamento de Dados, Mapeamento para a Ontologia Expandida, e Integração e Correlação dos Dados.

Definição e Expansão da Ontologia. Nessa etapa a ontologia *ICD-10-CM* [National Center for Biomedical Ontology 2023] foi utilizada e expandida para organizar informações de saúde de forma hierárquica através do CID-10, permitindo uma classificação detalhada e padronizada. Para isso a ferramenta Protégé[7] foi utilizada no processo de adição de novas classes que representavam eventos de mortalidade. Essa expansão foi importante para permitir a correlação entre os dados de mortalidade e outros fatores relevantes para o estudo.

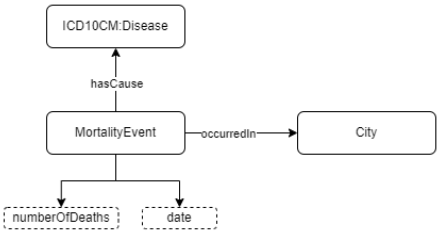

Figura 2. Ontologia Expandida.

A Figura 2 representa um extrato da ontologia, em que houve sua expansão com a adição de duas classes: *City* e *MortalityEvent*, onde *MortalityEvent* possui duas propriedades de dados que representam a data do evento (*date*) e o número de óbitos (*numberOfDeaths*), e a adição de duas propriedades de objetos: *hasCause* que estabelece uma relação entre o evento e a sua causa, representada por uma instância da classe *Disease* da ontologia *ICD-10-CM*, e *occurredIn* que estabelece uma relação entre o evento e a cidade, representada por uma instância da classe *City*.

Coleta e Tratamento de Dados. Após a expansão da ontologia, a coleta de dados, representada pela *Camada de Dados*, é realizada por meio de *scripts* automatizados em Python, que extraem as informações das fontes de dados. A escolha do Python é devido à vasta disponibilidade de bibliotecas para manipulação de dados, como Pandas, que é uma biblioteca funcional para manipulação e análise de dados. Os dados coletados foram submetidos a um processo de limpeza e normalização, que incluiu a remoção de dados inválidos, a padronização de formatos de datas e a verificação de consistência, garantindo a qualidade e a consistência dos dados para o mapeamento subsequente.

Mapeamento para a Ontologia Expandida. Com os dados tratados, foi realizado o mapeamento para a ontologia expandida, representado pela *Camada Semântica*, em que cada base de dados teve uma visão RDF exportada. A biblioteca RDFLib[8] foi utilizada para associar os dados coletados às classes e às propriedades da ontologia, garantindo a

[7] `https://protege.stanford.edu/`
[8] `https://rdflib.readthedocs.io/`

integração semântica. A biblioteca apresenta interface gráfica e um *wrapper* para *endpoints* SPARQL remotos. Esse mapeamento foi importante para a construção de um conjunto de dados coeso para o processo de análise.

Integração e Correlação dos Dados. Após a definição da ontologia e visões exportadas, iniciou-se a etapa de integração e correlação dos dados, representada pela *Camada de Integração de Dados*. Os dados foram importados para o GraphDB[9], é uma base de dados de grafos, compatível com os padrões W3C. A escolha do GraphDB foi motivada pela sua capacidade de realizar inferências, consultas complexas e pela sua compatibilidade com as tecnologias da Web Semântica, como RDF e SPARQL. Em nosso estudo de caso, a ferramenta, utilizando a ontologia definida, realizou inferências através de superclasses e subclasses projetando um grafo integrado de informações.

4.2. Análise estatística

Representada pela *Camada de Aplicações*, a análise estatística é uma etapa importante para a compreensão do impacto da COVID-19 na mortalidade por outras doenças na Região dos Sertões de Crateús. No GraphDB, são realizadas consultas SPARQL para consultar os dados de saúde pública, explorando as relações semânticas estabelecidas pela ontologia. Para esta análise, foram considerados três cenários distintos, cada um representando um período específico do ano de 2020: o ano completo de 2020, os quatro primeiros meses do ano e os meses que registraram picos de COVID-19. Essa estruturação permitiu uma investigação minuciosa dos efeitos da pandemia sobre as taxas de mortalidade, bem como a consideração de subnotificações. Dessa forma, para a análise da correlação dos dados foi aplicado o coeficiente de correlação de Pearson.

5. Resultados e Discussão

Após o processo de limpeza e normalização, que incluiu a remoção de entradas incompletas, a padronização de formatos de datas e a verificação de consistência entre os registros obteve-se um total de $294\ 586$ registros de mortalidade válidos referentes ao ano de 2020 e à região de estudo. A Figura 3 apresenta os dados obtidos.

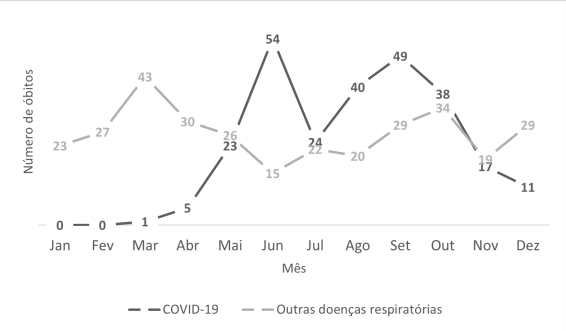

Figura 3. Óbitos causados por COVID-19 e outras doenças respiratórias.

Para o primeiro cenário de análise, referente aos primeiros quatro meses de 2020, observa-se que o coeficiente de correlação de Pearson obtido foi de $\mathbf{-0.008018496}$, apontando para uma ausência quase total de correlação entre as duas variáveis no período considerado.

[9] https://graphdb.ontotext.com/

Verificando os meses de pico da COVID-19, dos meses de junho a outubro de 2020, o coeficiente de correlação de Pearson para este cenário foi de $\mathbf{0.137513272}$, o que indica uma correlação direta, porém ainda muito fraca, entre o número de casos de outras doenças e o número de casos de COVID-19. E analisando o ano completo de 2020 o coeficiente de correlação de Pearson para este cenário foi de $\mathbf{-0.39670264}$, o que indica uma correlação inversa muito fraca entre o número de casos de outras doenças e o número de casos de COVID-19.

Portanto, os resultados obtidos indicam que, nos períodos analisados, não há evidências de uma relação significativa entre o número de casos de COVID-19 e a mortalidade por outras doenças respiratórias na região em estudo. A interpretação dos coeficientes sugere que as variações no número de casos de COVID-19 não estão fortemente associadas a mudanças na mortalidade por outras doenças respiratórias. Observa-se que, sem o processo de integração, a análise de verificação da correlação seria difícil de ser investigada, uma vez que os dados são de natureza eterogênea.

6. Considerações Finais

A presente pesquisa apresentou o *framework* OBSCRAT, que possibilitou a integração semântica de dados heterogêneos do Observatório de Dados Abertos da Região dos Sertões de Crateús. Com ele, foi demonstrado que é possível realizar a avaliação de correlação entre os dados semanticamente relacionados. Dessa forma, as principais contribuições desta pesquisa foram: (1) Destacar a importância da Web Semântica aplicada a dados abertos na pesquisa em saúde pública, contribuindo para o acesso à informação, a transparência na gestão pública e o desenvolvimento de novos estudos e pesquisas; (2) Contribuir para a Região dos Sertões de Crateús. Para pesquisas futuras, recomenda-se o contínuo aperfeiçoamento da estrutura da ontologia, com a adição de novas fontes de dados para integração semântica e a inclusão novos conceitos e propriedades.

Referências

Cavalcante, G. M. L. (2017). Maura: Um framework baseado em mediador semântico para construção eficiente de linked data mashups. Mestrado, Instituto Federal de Educação, Ciência e Tecnologia do Ceará.

da Cruz, M. M. L. (2021). Uma abordagem para construção de mashup de dados especificados como uma visão sobre um EKG. Dissertação de mestrado, Universidade Federal do Ceará.

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.

National Center for Biomedical Ontology (2023). International Classification of Diseases, 10th Edition, Clinical Modification. `https://bioportal.bioontology.org/ontologies/ICD10CM`, Acessado em 16 Jun. 2024.

Schultz, A., Matteini, A., Isele, R., Bizer, C., and Becker, C. (2011). LDIF-linked data integration framework. In *Proceedings of the Second International Conference on Consuming Linked Data-Volume 782*, pages 125–130.

Thor, A., Aumueller, D., and Rahm, E. (2007). Data integration support for mashups.