Dicionários Semânticos de Dados para Integrar Dados de Prontuários Eletrônicos de Pacientes

Marcello P. Bax[1], Evaldo de Oliveira da Silva[1]

[1]Programa de Pós-Graduação em Gestão & Organização do Conhecimento – Universidade Federal de Minas Gerais (PPGGOC-UFMG)
Caixa Postal 31.270-901 – Belo Horizonte – MG – Brasil

bax@eci.ufmg.br, evaldosilva@ufmg.br

Abstract. *In the health area, it is necessary to represent, integrate, and organize the knowledge derived from clinical data found in Electronic Medical Records (EMRs), providing advantages for developing scientific studies. Ontologies have been used in this context, can facilitate the extraction of information, disambiguate terms, and preserve the semantics of patient health variables. The usage of ontologies in the annotation of data in EMRs is a question discussed in related works. This article presents how to annotate data from EMRs, using data from the mental health field. Also, this work uses semantics data technique to generate knowledge graphs from ontologies and metadata models. Graphs can be explored later to filter base data to verify hypotheses in different scientific studies in the field of mental health.*

Resumo. *Na área da saúde é necessário representar, integrar e organizar o conhecimento oriundo de dados clínicos encontrados em Prontuários Eletrônicos de Pacientes (PEPs), proporcionando vantagens para o desenvolvimento de estudos científicos. Ontologias têm sido utilizadas neste contexto, pois podem facilitar a extração de informações, desambiguar termos e preservar a semântica das variáveis de saúde do paciente. A utilização de ontologias na anotação de dados sobre PEPs é uma questão discutida em trabalhos correlatos. Este artigo apresenta como anotar dados de PEPs, a partir de dados na área de saúde mental. Utiliza a técnica dos dicionários semânticos de dados para gerar grafos de conhecimento a partir de ontologias e de templates de metadados. Os grafos podem ser explorados posteriormente para filtrar dados necessários para verificar hipóteses em diferentes estudos científicos na área da saúde mental.*

1. Introdução

Ontologias têm sido utilizadas para representar e organizar o conhecimento a fim de anotar semanticamente dados e documentos [Gonçalves 2020]. Grafos de conhecimento podem ser gerados a partir da anotação semântica enriquecendo técnicas de recuperação de informação que auxiliam especialistas de domínio e cientistas de dados na fase de preparação dos dados para análise. Os dados a serem explorados podem ser selecionados pela navegação facetada em grafos que combinam e integram conceitualmente diferentes *datasets*, ampliando e aprofundando o escopo investigativo dos estudos científicos e levando à inferência de novos conhecimentos. O Dicionário Semântico de Dados (*Semantic Data Dictionary*, SDD), proposto por Rashid et al. (2017), é uma abordagem de anotação de dados por metadados fundamentada por ontologias que representa, normaliza e integra *datasets* de diferentes fontes. Os *datasets* anotados por SDDs geram

grafos RDF (*Resource Framework Description*). Grafos de conhecimento expressos em RDF e outros formatos têm sido propostos como base para desenvolver aplicações de IA nas mais diversas áreas: financeira, logística, educacional e na saúde [Beyer, Hare e Sallam 2020]. Especificamente, na saúde existe a necessidade de desambiguar termos, formalizar a semântica e interoperar dados relacionados à saúde dos pacientes. Prontuários Eletrônicos de Pacientes (PEPs) são fontes de extração de *datasets* que podem ser anotados por meio de SDDs, a fim de especializar a recuperação da informação neles contida. Grafos do conhecimento auxiliam na descoberta de relações entre doenças e sintomas, apoiando o diagnóstico por evidências em dados clínicos enriquecidos [Rotmensch et al. 2017]. O artigo apresenta por meio de um exemplo simplificado, o uso de SDDs para anotar semanticamente dados de PEPs na área de saúde mental. Para alcançar este objetivo, será aplicada a metodologia proposta por Gonçalves (2020). A metodologia se baseia no método DSR (*Design Science Research*) que utiliza um ciclo regulador que visa estruturar as etapas para integração dos dados semânticos [Bax 2015]. A metodologia prevê a elicitação de questões de competências que fundamentam a criação de uma ontologia de domínio caracterizada como uma versão simplificada do conhecimento para anotar dados ou que podem se basear em conceitos presentes em outras ontologias.

A seguir Seção 2 descreve os conceitos de PEP bem como os trabalhos correlatos. A Seção 3 aborda os conceitos sobre SDDs e grafos. A Seção 4 apresenta um estudo de caso simplificado de integração de dados de PEPs na área da saúde mental. A Seção 5 discute os resultados e a Seção 6 faz as considerações finais e lista os trabalhos futuros.

2. Referencial Teórico & Trabalhos Correlatos

Um PEP é um documento único que possui informações, sinais e imagens registradas, geradas a partir de fatos, acontecimentos e situações relacionadas à saúde do paciente e do atendimento realizado. Tais informações são de caráter legal, sigiloso e científico, servindo de comunicação entre membros da equipe multiprofissional e a continuidade da assistência oferecida ao paciente. PEPs são organizados com dados de anamnese, exame físico, prescrição e evolução. Hashemi et al. (2019) determinam os elementos necessários aos PEPs no campo dos transtornos mentais. Os autores realizaram uma revisão da literatura e selecionaram prontuários de pacientes com transtornos mentais para identificar uma lista preliminar de elementos de dados essenciais. Os autores identificaram sete classes tais como; dados demográficos de pacientes; dados administrativos de médicos; dados administrativos de pacientes; histórico; dados clínicos; tratamento; e dados financeiros. O Sistema e-SUS desenvolvido pelo Ministério da Saúde no Brasil, informatiza o fluxo de atendimento do cidadão utilizando o Prontuário Eletrônico do Cidadão (PEC) [SAPS 2017]. Por meio do PEC o profissional de saúde efetua o registro da consulta médica usando o método SOAP (Subjetivo, Objetivo, Avaliação e Plano). O SOAP permite registrar os dados de evolução do paciente de forma sintética e estruturada com base em questões subjetivas, além das impressões objetivas sobre o estado geral do paciente.

Diferentes motivos ensejam a anotação de dados médicos. Alguns visam melhorar a extração de informações, desambiguar termos, preservar a semântica das variáveis de saúde do paciente, além de muitos outros. Xu et. al (2016) propõem anotar dados de PEPs com metamodelos de bancos de dados. Seu método transforma os termos que não podem ser mapeados diretamente, alterando-os e mantendo sua semântica. Após a transformação

é possível obter a desambiguação dos significados a fim de garantir a precisão das informações consultadas. Chehab, Kalboussi e Kacem (2019) propõem anotar os dados com uma classificação unificada de sistemas de anotação médica. Uma ontologia preserva a semântica dos dados. Para Christen, GROß e Rahm (2016), a anotação manual é um processo complexo e a anotação automática é altamente desejável para dar suporte a anotadores humanos. Propõem uma abordagem baseada na linguagem e na reutilização, anotando documentos médicos com conceitos de ontologias. Rotmensch et al. (2017) geraram grafos de conhecimento a partir da leitura de dados sobre doenças e sintomas diretamente de PEPs. Conceitos médicos são extraídos dos PEPs utilizando cálculos de estimativas de verossimilhança com base em três modelos probabilísticos para construir automaticamente os grafos de conhecimento. Peleg, Keren, Denekamp (2008) discutem diretrizes clínicas e padrões de qualidade para o atendimento ao paciente, codificando-as em formato interpretável com PEPs permitindo recomendações ao paciente quando e onde necessário. Seu *Knowledge-Data Ontological Mapper* (KDOM) utiliza uma ontologia e um modelo de referência para mapear código de PEPs. Yamada et al. (2018) propõem a ontologia de saúde mental OSM desenvolvida para apoiar a análise de indicadores de saúde mental com dados inseridos de forma descentralizada a fim de apoiar a decisão em Sistemas de Informação em Saúde Mental (SISAM). O SISAM monitora consultas médicas, pedidos de hospitalização e a movimentação de pacientes na Rede de Atenção Psicossocial (RAPS) permitindo a integração das ações dos serviços de saúde em transtornos mentais.

A pesquisa apresentada neste artigo se diferencia dos trabalhos acima por utilizar a modelagem ontológica aplicando a abordagem SDD para anotar dados [Rashid et al. 2017]. A anotação é realizada manualmente por especialistas da área da saúde mental, tais como médicos, terapeutas ou psicólogos. Assim, grafos de conhecimento são gerados a partir de uma ontologia e de templates de metadados.

3. Anotação de Dados com Dicionário Semânticos de Dados

Rashid et. al (2017) propõem o uso de SDDs para anotação da semântica de *datasets*. O SDD é um conjunto de padrões de metadados fundamentados em ontologias que descrevem objetos (representados por dados) em classes e relacionamentos. A anotação por SDD associa os dados de um *dataset* a conceitos (ou classes) nas ontologias a fim de enriquecer semanticamente os dados presentes em um *dataset*. A anotação de dados por meio de SDD deve ser realizada por profissionais que compreendem o domínio, como por exemplo, engenheiros do conhecimento ou cientistas, familiarizados tanto com os conceitos fornecidos pela ontologia, quanto pelos *datasets* a serem anotados. A preparação da anotação é feita de forma manual e utiliza um conjunto de documentos: *InfoSheet*; *Dictionary Mapping*; *CodeBook*; *Code Mapping*; *TimeLine*; *Properties Table*. Rashid et al. (2017) recomendam a SIO (*Semanticscience Integrated Ontology*) para anotação de dados. Uma vez que a estrutura de anotação esteja pronta, a ferramenta sdd2rdf [Rashid et al 2017] interpreta o SDD e integra os dados do *dataset* descrito pelo SDD formando um grafo de conhecimento (RDF) persistido em *triplestores*. Um grafo de conhecimento representa objetos (ou recursos) de interesse e conexões entre eles [Hogan et al. 2020]. Geralmente fornece um substrato compartilhado de conhecimento, permitindo que pessoas e diferentes aplicações reutilizem as definições nele modeladas. Possibilita também a inferência de novos fatos enriquecendo o compartilhamento do conhecimento a partir do sdd2rdf, que produz consultas no formato SPARQL e regras SWRL que permitem inferências. O RDF gerado pelo sdd2rdf se fundamenta em

ontologias, e possibilita a integração semântica dos dados. A ontologia formaliza o vocabulário da anotação e abre caminho para interoperabilidade dos dados que podem ser integrados de fontes diversas.

4. Aplicação na Integração de Dados de PEPs

Datasets com dados fictícios oriundos de PEPs em saúde mental serão anotados. A Tabela 1 traz a tabela que relaciona dados assistenciais e de internação em hospitais de saúde mental. A Tabela 2 traz o extrato do Boletim Médico Hospitalar (BMH), com dados de internações de pacientes em um hospital caracterizados por variáveis como data da admissão, alta faixa etária, pelos códigos CID (Classificação Internacional de Doenças) e outros.

Tabela 1 - Dados Assistenciais.

ID	Fuma	CInternacao
36	Não	1
78	Sim	3
...	...	...

Fonte: O autor (2020)

Tabela 2 - Dados do Boletim Médico Hospitalar (BMH).

ID	DataAdm	Alta	FE	Município	Sexo	CID
36	27/04/2020	27/06/2020	ADL	Brasília	M	F30-F39
78	23/05/2020	28/06/2020	IDS	Belo Horizonte	M	F20-F29
...	...	...	...	...	...	...

Fonte: O autor (2020)

A anotação de dados é feita a partir de uma ontologia "base", *i.e.,* a primeira versão simplificada da ontologia de domínio. A ontologia utiliza conceitos da OSM e mais as colunas das tabelas acima não mapeadas para conceitos da OSM. Questões de competência ilustram problemas que poderiam ser respondidos pela análise das variáveis nas Tabelas 1 e 2. (i) Qual é a prevalência de faixa etária das internações com demência vascular? (ii) Avaliar os sintomas e monitorar o progresso dos pacientes melhora os desfechos, qualidade de vida e funcionamento social? (iii) Reinternações compulsórias estão associadas a quais transtornos mentais? Os templates de metadados em Rashid et al. (2017) incluem o *Dictionary Mapping* (DM), elaborado na Tabela 3. Os conceitos utilizados para anotação do cabeçalho das tabelas foram reutilizados da ontologia OSM, formando a ontologia base deste estudo específico. A ontologia base é enriquecida com outros conceitos relevantes para compreensão domínio discutido, tais como: "bmh" e "tipoInternação". As Tabelas 3 e 4 apresentam *Dictionary Mapping* e o *CodeBook* respectivamente.

Tabela 3 - Especificação do DM para dados explícitos e implícitos.

Column	Attribute	AtributeOf	Entity	Unit	Time	Relation	inRelationTo
ID	osm:idPaciente	??Paciente					
DataAdm	osm:dataInternacao	??Bmh		sio:Date	??admissao		
Alta	osm:dataAlta	??Bmh		sio:Date	??alta		
Município	osm:cidadePaciente	??Bmh					
Sexo	osm:Sexo	??Paciente					
CID	osm:Cid10	??Bmh					
Fuma	osm:Tabagismo	??Paciente					
CInternacao	osm:tipoInternacao	??Bmh					
FE	osm:faixaEtaria	??Bmh					
??paciente			sio:Human;osm:Paciente;			osm:hasBmh	??bmh
??bmh			osm:Bmh			osm:isBmhOf	??paciente

Fonte: O autor (2020)

Vários outros dados poderiam ter sido considerados para integração: médico, psicólogo, leito e consistência da dieta. O grafo RDF gerado pela interpretação das tabelas acima permite a navegação facetada pelos dados com a qual o pesquisador seleciona quais dados farão parte do *dataset* a ser gerado.

Tabela 4 -*Codebook*.

Column	Code	Class
FE	ADL	osm:Demograficos
FE	ADT	osm:Demograficos
FE	IDS	osm:Demograficos
CInternacao	1	osm:Internacao
CInternacao	2	osm:Internacao
CInternacao	3	osm:Internacao

Fonte: O autor (2020)

A Figura 01 apresenta o grafo RDF, em sintaxe turtle, gerado pelo script sdd2rdf, tendo como entradas os metadados das Tabelas 3 e 4 e a primeira linha dos arquivos de dados das Tabelas 1 e 2.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix sio: <http://semanticscience.org/resource/> .
@prefix example-kb: <http://example.com/kb/example#> .
@prefix osm: <http://www.w3.org/2008/05/osm#> .
example-kb:ID rdfs:subClassOf osm:idPaciente ;
  sio:isAttributeOf example-kb:Paciente .
example-kb:Sexo rdfs:subClassOf osm:sexo ;
  sio:isAttributeOf example-kb:Paciente .
example-kb:CID rdfs:subClassOf osm:Cid10 ;
  sio:isAttributeOf example-kb:Bmh .
example-kb:Fuma rdfs:subClassOf osm:Tabagismo ;
  sio:isAttributeOf example-kb:Paciente .
example-kb:CInternacao rdfs:subClassOf osm:tipoInternacao ;
  sio:isAttributeOf example-kb:Bmh .
example-kb:FE rdfs:subClassOf osm:faixaEtaria ;
  sio:isAttributeOf example-kb:Paciente .
example-kb:Paciente rdfs:label "Paciente"^^xsd:string ;
  rdfs:subClassOf sio:Human ;
  sio:inRelationTo example-kb:Bmh .
example-kb:Bmh rdfs:label "Bmh"^^xsd:string ;
  rdfs:subClassOf osm:Bmh ;
  sio:inRelationTo example-kb:Paciente .
example-kb:ID-04 a example-kb:ID , osm:idPaciente ;
  sio:isAttributeOf example-kb:Paciente-1d ;
  sio:hasValue "36"^^xsd:integer .
example-kb:Sexo-3f a example-kb:Sexo , osm:sexo ;
  sio:isAttributeOf example-kb:Paciente-1d ;
  sio:hasValue "M"^^xsd:string .
example-kb:CID-1f a example-kb:CID , osm:Cid10 ;
  sio:isAttributeOf example-kb:Bmh-18 ;
  sio:hasValue "F30-F39"^^xsd:string .
example-kb:Fuma-27 a example-kb:Fuma , osm:Tabagismo ;
  sio:isAttributeOf example-kb:Paciente-1d ;
  sio:hasValue "Nao"^^xsd:string .
example-kb:CInternacao-32 a example-kb:CInternacao , osm:Voluntario ;
  sio:isAttributeOf example-kb:Bmh-18 ;
  sio:hasValue "1"^^xsd:integer .
example-kb:FE-a5 a example-kb:FE , osm:adulto ;
  sio:isAttributeOf example-kb:Paciente-1d ;
  sio:hasValue "ADL"^^xsd:string .
example-kb:Paciente-1d a example-kb:Paciente , sio:Human ;
  rdfs:label "Paciente"^^xsd:string ;
  sio:inRelationTo example-kb:Bmh-18 .
example-kb:Bmh-18 a example-kb:Bmh , osm:Bmh ;
  rdfs:label "Bmh"^^xsd:string ;
  sio:inRelationTo example-kb:Paciente-1d .
```

Figura 01. Grafo de Conhecimento. **Fonte**: O autor, 2020.

5. Discussão de Resultados

Como contribuição o trabalho apresenta a preparação de dados[1], utilizando SDD, no âmbito de estudos científicos na área de saúde mental. Os dados anotados por metadados semânticos fundamentam a produção de grafos de conhecimento que representam conceitos e as conexões entre eles. Os elementos terminais dos grafos são valores que instanciam propriedades de objetos que, por sua vez, populam estudos científicos. Metadados semânticos utilizam a modelagem ontológica para organizar os dados. Os arquivos de dados organizados nos grafos contêm os valores em células tabulares e os templates de metadados (dicionários de dados e *codebook*) são compostos de anotações formais de propriedades (variáveis em dicionários de dados), objetos e estudos. Os templates de metadados fornecem estrutura para que especialistas no domínio identifiquem e definam explicitamente quais são os diversos objetos de um estudo científico; como tais objetos são organizados. Além disso, ontologias formalizam o entendimento acima exigindo que as definições sejam declaradas usando termos

[1] Fase do ciclo de mineração de dados

consensuados no domínio. Os conceitos e dados organizados contribuem para o reuso de dados e a reprodução dos resultados de estudos realizados

6. Considerações Finais

A relevância do problema e da abordagem apresentada advém do alto custo decorrente dos desafios técnicos enfrentados hoje para reutilizar dados de estudos científicos. O artigo descreve a abordagem de dicionário semânticos e infraestrutura técnica para anotar dados de estudos científicos realizados sobre PEPs, que permitem a normalização e harmonização sistemática das variáveis dos estudos. Os dados são organizados em grafos RDF que podem ser explorados por consultas a fim de darem origem a outros arquivos de dados. Limitações podem ser vistas a partir da aplicação apresentada neste artigo. Campos de digitação livre da orientação médica e o relato de alta não foram explorados. Os conceitos presentes nestes poderiam ter sido anotados por técnicas de processamento de linguagem natural (NLP) a fim enriquecer os conjuntos de dados das Tabelas 1 e 2. Como trabalhos futuros pretende-se utilizar técnicas de NLP para reconhecer e extrair entidades salientes que representem outras variáveis para além daquelas presentes nos arquivos de dados das Tabelas 1 e 2.

7. Referencias

Bax, M. P. Design Science: filosofia da pesquisa em ciência da informação e tecnologia. Ciência da informação, v. 42, n. 2, 2015.

Beyer, Mark; HARE, Jim; Sallam, Rita. COVID-19 Demands Urgent Use of Graph Data Management and Analytics.

Chehab, Khalil; Kalboussi, Anis; Kacem, Ahmed Hadj. An Annotation Model for Patient Record. In: HEALTHINF. 2019. p. 272-280.

Christen, Victor; Groß, Anika; Rahm, Erhard. Approaches for Annotating Medical Documents. In: LWDA. 2016. p. 227-232.

Gonçalves, José Eugênio Assis. Método de Integração Semântica de Dados Científicos Baseado em Ontologias. Tese (Doutorado) – ECI UFMG, 2020.

Hashemi, Nasim et al. Electronic Medical Records for Mental Disorders: What Data Elements Should These Systems Contain?. In: dHealth. 2019. p. 25-32.

Hogan, Aidan et al. Knowledge graphs. arXiv preprint arXiv:2003.02320, 2020.

Peleg, Mor; et. Al. Mapping computerized clinical guidelines to electronic medical records: Knowledge-data ontological mapper (KDOM). J of bio infor., v. 41, n. 1, p. 180-201, 2008.

Rashid, Sabbir M. et al. The Semantic Data Dictionary Approach to Data Annotation & Integration. In: SemSci@ ISWC. 2017. p. 47-54.

Rotmensch, Maya et al. Learning a health knowledge graph from electronic medical records. Scientific reports, v. 7, n. 1, p. 1-11, 2017.

SAPS. Secretaria de Atenção Primária à Saúde. Disponível em: https://aps.saude.gov.br/noticia/2300. 2017. Acesso em 06 de ago de 2020.

Yamada, Diego Bettiol et al. Proposal of an ontology for Mental Health Management in Brazil. Procedia computer science, v. 138, p. 137-142, 2018.

Xu, Boyi et al. Healthcare data analytics: Using a metadata annotation approach for integrating electronic hospital records. Journal of Management Analytics, v. 3, n. 2, p. 136-151, 2016.