

Especificação de um método fundamentado em modelagem ontológica para análise quali-quantitativa

Raissa Barros Dominato¹
Natalia Maria Silva Costa²
Manuela Mayworm Jens³
Eduarda Novais Falcão⁴
Rubens Amorim Souto Neto⁵
João Vitor Carvalho⁶
João Gabriel Garcia Paiva⁷
Samuel Henrique Reis do Nascimento⁸
Evaldo de Oliveira da Silva⁹
Auxiliatrice Caneschi Badaró¹⁰
Marcello Peixoto Bax¹¹

RESUMO

A análise quali-quantitativa é uma técnica de pesquisa que possui a necessidade de integrar dados qualitativos e quantitativos sob os mesmos objetos de estudo e produzidos em trilhas distintas de pesquisa. Utilizar diferentes formas de representação desses dados, permite explicitar seus relacionamentos mais complexos e assimilar melhor as informações da pesquisa. A Compreensão e a Preparação são atividades iniciais de uma pesquisa quali-quantitativa e que usam técnicas para explicitar o conhecimento sobre dados. Ontologias são comumente utilizadas para estruturar o conhecimento incorporado em *datasets* e podem facilitar extrair informações, desambiguar termos e preservar a semântica das variáveis. Neste trabalho um método sistemático baseado na anotação semântica de dados é especificado, a fim de facilitar que os dados, informações e conhecimentos gerados pelas pesquisas, sejam explorados por meio

¹ Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: raissabdominato@gmail.com

² Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: natmcosta@outlook.com

³ Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: manuelajens@hotmail.com

⁴ Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: eduardanfalcão@gmail.com

⁵ Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: rubensasn@gmail.com

⁶ Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: joaovitorcarvalhov2@outlook.com

⁷ Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: joaogarciapaiva@hotmail.com

⁸ Discente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: samuelhr.900021713@uniacademia.edu.br

⁹ Docente dos cursos de Engenharia de Software e Sistemas de Informação do Centro Universitário UniAcademia – UniAcademia. E-mail: evaldosilva@uniacademia.edu.br

¹⁰ Docente do curso de Psicologia do Centro Universitário UniAcademia – UniAcademia. E-mail: auxiliatricebadaro@uniacademia.edu.br

¹¹ Docente do curso de Pós-Graduação em Gestão e Organização do Conhecimento da Universidade Federal de Minas Gerais. E-mail: bax.ufmg@gmail.com

de grafos. Uma técnica de dicionarização semântica de dados é sugerida para gerar grafos a partir de ontologias, metadados e dados, permitindo que especialistas de domínio explorem as informações e os conhecimentos investigados explicitando-os a partir dos dados. Desta forma, o método propõe atividades que geram grafos para organizar dados anotados semanticamente para explorar e filtrar informações a fim de responder questões de interesse das pesquisas. Eles facilitam a visualização gráfica dos dados e têm potencial para aprimorar as análises quali-quantitativas. Como trabalhos futuros, propõe um estudo de caso na área de Psicologia para ilustrar e validar o método proposto em uma pesquisa quali-quantitativa que explora dados psicoterápicos reais.

Palavras-chave: Tratamento Psicoterápicos. Grafos de Conhecimento. Análise quali-quantitativa. Ontologia.

1 INTRODUÇÃO

As pesquisas que aplicam métodos mistos usam a compreensão e preparação dos dados produzidos em “trilhas” distintas (qualitativa e quantitativa). A representação conjunta dos dados produzidos é um desafio, sendo necessária a especificação de métodos que permitam compreendê-los. Diagramas, grafos e outras apresentações auxiliam na captura do conhecimento (SOMEKH; LEWIN, 2005). Ao preparar os dados, as atividades de transformação modificam a forma dos dados, tal como discutido em Driscoll *et al.* (2007).

A compreensão e preparação de dados são atividades que permitem sistematizar os diferentes conjuntos de dados, e podem abrir caminho para a inferência de novos conhecimentos (BERTHOLD *et al.*, 2010). Na área da saúde mental dados extraídos dos questionários de avaliação, usados para atribuir escalas de valores dos sinais e sintomas de ansiedade, permitem análise quantitativa para estudos científicos. Dados extraídos dos textos de tratamentos psicoterápicos podem ser codificados para análise qualitativa (BRAUN e CLARKE, 2006).

No entanto, os formatos distintos dos dados (estruturados, não-estruturados ou formatos textuais), podem dificultar as análises da eficácia dos tratamentos, se os riscos foram mitigados ou quais fatores que influenciam no tratamento, dificultando o reuso e compartilhamento do conhecimento em pesquisas nesta área. Desta forma, o objetivo do deste artigo é apresentar a especificação de um método fundamentado em modelagem ontológica para análise quali-quantitativa. O método contempla a compreensão e preparação de dados em etapas iniciais. Espera-se aplicar o método especificado para análise quali-quantitativa de dados de tratamentos psicoterápicos.

O restante do trabalho se encontra organizado em seções. A Seção 2 descreve o referencial teórico que permitiu o embasamento de conceitos abordados neste trabalho. A Seção 3 especifica o método proposto. A Seção 4 apresenta as considerações finais e trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 Ontologia

Uma ontologia é uma especificação da conceituação descrita por conceitos e relacionamentos que podem existir no mundo real. A conceituação pode ser desenvolvida contendo terminologias e vocabulários, estabelecendo propriedades e permitindo que o conhecimento seja reutilizado, evitando o retrabalho ou a redescoberta de terminologias equivalentes (GUARINO, 1998).

De acordo com Guarino (1998, p. 7), as ontologias são divididas em classificações diferentes:

- Ontologias de alto nível: incluem classes relacionadas a conceitos globais. Não pertencem a um único domínio de aplicação. Esses conceitos são genéricos e podem ser especificados por ontologias de domínio ou tarefa.
- Ontologias de domínio: modelos que representam um determinado domínio, podendo ser reutilizáveis por aplicações do domínio. Os conceitos explicitados no domínio e seus relacionamentos são representados.
- Ontologias de tarefa: representam um conjunto de conceitos usados para resolver problemas que podem ser ou não de um mesmo domínio. Dessa forma, inclui nomes de conceitos genéricos, podendo ser reaproveitados em diferentes domínios.
- Ontologias de aplicação: são ontologias criadas a partir de conceitos de um domínio particular, em certos casos, da aplicação de uma tarefa em particular em um domínio específico.

Dentre os tipos de ontologias definidos, as ontologias de domínio têm sido bastante usadas por pesquisadores da área da saúde mental (BRENAS; SHIN; SHABAN-NEJAD, 2006; CEUSTERS; SMITH, 2010; HASTINGS *et al.*, 2012; BENFARES; IDRISSE; HAMID, 2018; YAMADA *et al.*, 2018).

Ceusters e Smith (2010) apresentam uma visão da doença mental baseada no realismo ontológico e que segue os princípios incorporados na *Basic Formal Ontology* (BFO) e na aplicação da BFO na *Ontology of General Medical Science* (OGMS). Os autores analisaram afirmações sobre o que é considerado como doença mental no contexto da DSM-V (*Diagnostic and Statistical Manual of Mental Disorders*). Avaliam se a representação do conhecimento presente nas ontologias BFO e OGMS são adequadas para criar uma ontologia da saúde mental necessária para evitar interpretações erradas sobre um mesmo conceito.

Brenas, Shin e Shaban-Nejad (2019) implementaram uma ontologia formal sobre experiências adversas da infância (ACEs, *Adverse Childhood Experiences*), para integração de dados de estudos científicos na área da saúde mental. Segundo os autores, as ACEs estão relacionadas aos riscos aumentados a partir dos resultados e condições de saúde dos pacientes.

De acordo com Hastings *et al.* (2012) as ontologias podem ser usadas para aplicações sofisticadas de raciocínio automatizado. Os autores descrevem a Ontologia do Funcionamento Mental (MF) e a Ontologia da Doença Mental (MD), duas ontologias baseadas em realismo para a descrição do funcionamento mental humano e das doenças.

Benfares, Idrissi e Hamid (2018) propõem uma arquitetura de sistema de acompanhamento de pacientes, que proporciona o acompanhamento do estado psíquico do paciente. Os autores usam ontologias e sistemas de recomendação para proporcionar aos pacientes o acompanhamento em tempo real. Foram desenvolvidas rotinas para recuperar informações de rastreamento dos pacientes com câncer, depressão e transtornos de ansiedade. A pesquisa foi realizada na unidade de Psicologia, no Centro de Oncologia e Hematologia do Centro Hospitalar Universitário de Marrakech.

Yamada *et al.* (2018) propõem a ontologia OSM (Ontologia de Saúde Mental) desenvolvida para apoiar a integração de dados para saúde mental, a partir de dados mantidos por sistemas em saúde mental.

Rosenberg e Hastings (2020) discutem a importância de estruturar e padronizar a semântica da sintomatologia em saúde mental (ou, a compreensão de sinais e sintomas do paciente). Os autores formalizam uma ontologia usando ontologias na área biomédica, tal como a MF. Os autores abordam o esforço de padronização com

o método de ontologia aplicada, valendo-se de recursos de código aberto existentes na *Open Biomedical Ontologies* (OBO).

De acordo com Dooley *et al.* (2017) a plataforma *Genomic Epidemiology Entity Mart* (GEEM) é um portal web projetado para fornecer a especialistas de domínio uma interface orientada por ontologia para examinar padrões de dados em sequências genômicas. O GEEM possui o GenEpiO, um módulo que fornece um conjunto de termos modelados por uma ontologia de aplicação, e que podem ser associados a dados epidemiológicos. A ontologia reutiliza terminologias encontradas na OBO, e tem sido construída usando uma abordagem *bottom-up*, centrada na padronização dos termos necessários para executar os processos relativos aos fluxos de trabalho laboratorial, clínico e epidemiológico.

Abordagens de modelagem *top-down* e *bottom-up* podem ser usadas para construção de ontologias de domínio. De acordo com Kung, Kung e Gardiner (2012, p. 12) a abordagem *top-down* é definida como:

a obtenção do conhecimento a partir da identificação das entidades, tipos de entidades, e os relacionamentos entre elas. A adoção dessa abordagem, geralmente, inicia-se com um conjunto de requisitos de alto nível, tal como uma narrativa. Esses requisitos descritos iniciam um processo de identificação das entidades necessárias para representar os dados, bem como as propriedades deles e como podem estar organizados em bases de dados.

Em contraste, Kung, Kung e Gardiner (2012, p. 12) ainda definem a abordagem *bottom-up* como: “a aquisição do conhecimento pela compreensão das coleções de dados, avançando para uma análise inicial de classes de conceitos de nível inferior, como atributos e relacionamentos, e então, procedendo em direção a um modelo conceitual aceitável”.

De acordo com Matteo e Cuel (2005, p. 64 apud DING; FOO, 2002), essas abordagens possuem propósitos específicos para modelar ontologias de domínio:

- *Top-down*: da generalização à especificação do conhecimento. Conceitos complementares podem ser reutilizados de outras ontologias;
- *Bottom-up*: da especificação à generalização do conhecimento. Visa elicitar os conceitos por meio da análise ou processamento de dados estruturados ou não;
- *Middle-out*: inicia-se dos conceitos mais importantes da generalização à especialização.

De acordo com El Ghosh *et al.* (2016) a *middle-out* é uma estratégia que pode combinar as outras duas abordagens em atividades executadas para modelar a

ontologia em módulos. Ainda, de acordo com El Gosh *et al.* (2016, p. 2), “a reutilização de outras ontologias, que capturam conhecimentos semelhantes ou complementares de ontologias de alto nível, pode contribuir na abordagem *top-down*”. Os autores destacam a importância das ontologias de alto nível no reuso do conhecimento. Conforme abordado por Francesconi *et al.* (2010), as abordagens *top-down* e *bottom-up* podem se complementar em projetos que necessitam compreender genericamente o domínio, mas, adicionalmente, explora o conhecimento de diferentes conjuntos de dados para identificar novos conceitos visando compor uma ontologia.

2.4 Análise qualitativa e quantitativa

De acordo com Freitas e Moscarola (2000) a busca por informações adequadas e validadas não deve se basear somente nos dados estruturados, dito puramente quantitativos. De acordo com os autores, dados estruturados encontram-se em planilhas, relatórios volumosos, números, percentuais e gráficos. No entanto, é necessário recorrer aos dados de natureza qualitativa, como textos, discursos, entrevistas, trechos de livros ou reportagens.

Os dados qualitativos são um desafio aos pesquisadores de diferentes áreas. Eles “escondem” informações, opiniões, perfis, que exigem uma leitura atenta. O uso de ferramentas de software ou técnicas de análise qualitativa, possibilitam a obtenção de informações que possam responder ou orientar as análises de dados quantitativos (FREITAS e MOSCAROLA, 2000; CALDIERARO *et al.*, 1998).

Aplicações de software permitem que dados quantitativos (dado elaborado, métricas, indicadores, resultados indicadores) sejam analisados por meio relatórios e *dashboards*¹². De acordo com Freitas e Moscarola (2000), ao analisar dados quantitativos deseja-se ir do dado bruto ao dado elaborado, por meio de interpretação e síntese; com objetivo de compreender o dado elaborado e como pode ser usado para tomada de decisões.

Freitas e Moscarola (2000) apresentam um exemplo com a utilização de uma coleção de dados que representa o gênero dos respondentes, e na qual há 3 não-respostas. Neste caso, para apoiar a interpretação dos números tabulados, bastaria

¹² *Dashboard* é o mesmo que um painel para exibição das informações mais importantes necessárias para alcançar um ou mais objetivos; consolidados e organizados em uma única tela para que o as informações podem ser monitoradas facilmente (FEW, 2006).

clicar em cima do número 3 e se teria uma lista dos 3 questionários correspondentes. Em seguida, bastaria clicar num botão para corrigir o número e não-respostas para realizar a alteração. Por outro lado, imagine-se ainda um gráfico de correlação onde um ponto desponha do comportamento de todos os outros: “do que se trata este comportamento”? Deste modo, bastaria clicar em cima do ponto para saber dos detalhes.

A análise qualitativa é usada para análise temática de tratamentos psicoterápicos. De acordo com Braun e Clarke (2006) a análise temática é um método analítico qualitativo amplamente utilizado dentro da psicologia. Os autores descrevem que a análise temática está relacionada a outros métodos analíticos qualitativos que buscam temas ou padrões, e em relação a diferentes posições epistemológicas e ontológicas. Os autores defendem ainda a análise temática como um método útil e flexível para a pesquisa qualitativa dentro e fora da psicologia.

Análises qualitativas podem ser apoiadas por softwares (CAQDAS – *Computer Assisted Qualitative Data Analysis*). A tarefa de *coding* permite definir ou categorizar os dados (textuais) que são analisados para diferentes objetivos de pesquisas. Essa tarefa utiliza de procedimentos que revelam temas embutidos nos dados (WILLIAMS e MOSER, 2019). CAQDASs organizam os *codings* em estruturas baseadas em desenhos qualitativos, que servem de base para codificar (anotar) os textos dos tratamentos psicoterápicos.

2.5 Grafos de Conhecimento

De acordo com Pan *et al.* (2017) um grafo de conhecimento (GC) possui restrições em sua estrutura impostas por meio de ontologias. Possibilitam a inferência de novos fatos enriquecendo o conhecimento e o seu compartilhamento. Os grafos de conhecimento são frequentemente constituídos a partir de várias fontes de dados (datasets) dependendo das técnicas que permitem extrair e ingerir as fontes de dados para geração dos grafos. O resultado é a diversificação de termos e relações entre os dados.

A manipulação da diversidade de dados utiliza as representações de esquema que contém as identificações e o contexto das entidades presentes no grafo de conhecimento. Esse esquema define uma estrutura de alto nível, onde a identidade dos objetos denota quais nós no grafo se referem à mesma entidade do mundo real,

enquanto o contexto especifica os objetos considerados verdadeiros para o domínio específico. Também são considerados como grafos de dados, destinados a acumular e transmitir conhecimento do mundo real, cujos nós representam entidades de interesse e cujas bordas representam relações entre essas entidades (HOGAN *et al.*, 2020). Hogan *et al.* (2020) destaca a importância das ontologias como modelos conceituais utilizados para anotar conjuntos de dados. Esta anotação permite gerar fragmentos de conhecimento do domínio que podem ser representados por grafos de conhecimento.

Bax e Silva (2020) abordam o uso de ontologias para anotação de dados com vistas a harmonizar e normalizar variáveis de estudos científicos. Propõem que os dados estejam organizados em grafos RDF¹³ para serem explorados por consultas a fim de darem origem a outros arquivos de dados.

2.6 Dicionário Semântico de Dados

Rashid *et. al* (2017) usam padrões de metadados para configurar a anotação semântica por um SDD. A anotação semântica proposta por Rashid *et. al* (2017) utiliza os seguintes documentos:

- *InfoSheet*: referências para descrição dos SDDs;
- *Dictionary Mapping*: anotação semântica das colunas das coleções de dados;
- *CodeBook*: códigos correspondentes a conceitos de ontologia;
- *Code Mapping*: mapeamento de termos dos datasets que correspondem a conceitos existentes na ontologia;
- *TimeLine*: anotação de intervalos temporais;
- *Properties Table*: para fins de customizar a descrição por outras ontologias de topo.

A ferramenta `sdd2rdf` interpreta o SDD e processa os dados, formando um grafo RDF. Para acessar os dados anotados, o `sdd2rdf` cria consultas no formato SPARQL¹⁴. São geradas também regras SWRL¹⁵ que auxiliam em novas inferências.

A técnica SDD é utilizada neste artigo visando explicitar os conceitos sobre os dados para gerar grafos de conhecimento dos dados anotados semanticamente. Os

¹³ *Resource Description Framework*

¹⁴ *SPARQL Protocol and RDF Query Language*

¹⁵ *Semantic Web Rule Language*

grafos serão utilizados para encontrar relações entre doenças, sintomas e tratamentos, servindo para auxiliar na definição dos níveis de cuidados com o paciente.

2.7 Trabalhos Correlatos

Hocker *et al.* (2021a) utilizaram a QualiCO para anotar dados qualitativos usando templates de metadados. Os autores avaliaram quantitativamente o reuso da QualiCO visando melhorar as tarefas de codificação. No entanto, apesar da anotação semântica pela QualiCO, os autores não abordaram como os dados anotados poderiam estar representados em GCs.

Miñarro-Giménez *et al.* (2018) utilizaram a ontologia SNOMED CT (SNOMED *Clinical Terms*) para codificar dados qualitativos de textos clínicos, a fim de identificar as fontes de divergências com o uso da SNOMED CT. A prevalência dos conceitos anotados contribuiu para revelar as divergências e o nível de compreensão da SNOMED CT para codificar os termos. Apesar dos autores terem usado a codificação e análise quantitativa, não abordaram GCs para apoiar na análise das divergências sobre os dados anotados.

Charlotte Siefridt *et al.* (2020) desenvolveram um extrator multiterminológico para anotar automaticamente textos de consultas médicas usando a SNOMED CT. A acurácia dos termos anotados foi avaliada por meio de cálculos de precisão (com o uso das funções recall e *F-measure*). Os autores consideraram a anotação eficaz no domínio de medicina familiar. Porém, não representaram os dados anotados em GCs, ou como a SNOMED CT poderia ser utilizada como estrutura para popular dados do estudo apresentado.

Pfaff, Neubig e Krcmar (2018) modelaram uma ontologia de serviços de Tecnologia da Informação (TI) usando o conhecimento sobre dados de questionários submetidos aos representantes de organizações de TI. Uma rotina em NLP foi implementada para acessar os conceitos da ontologia e anotar textos contidos nas respostas dos questionários. Além dos textos, indicadores (KPIs) usados para avaliar os serviços de TI foram anotados como dados quantitativos. Com os dados armazenados em uma base dados semânticos, consultas SPARQL foram elaboradas para gerar fragmentos de GCs a fim de serem analisados para responder questões de competência. No entanto, para anotar as fontes de dados (oriundas dos

questionários) foi necessário um mapeamento de metadado específico para cada formato de dado.

Iglesias *et al.* (2017) criaram um modelo de *linked data* para análise de sentimentos e emoções em redes sociais. Os termos da análise de sentimentos foram extraídos automaticamente de textos por meio do processamento de vocabulários baseado na linguagem de marcação EmotionML . Ao extrair os termos, a rotina de processamento fez a anotação semântica deles usando a ontologia DBpedia. Complementarmente, a prevalência dos conceitos analisados foi tabulada e anotada. Ao final, os autores demonstraram fragmentos de GCs para permitir a análise de sentimentos com dados quali-quantitativos. Porém, os autores não detalharam como a ferramenta foi construída, dificultando o entendimento mais específico sobre como a anotação foi realizada.

Heibi e Peroni (2021) aplicaram a técnica de citação aberta como metodologia para coletar e anotar os dados de artigo científico. Na esteira da web semântica, os autores utilizaram a ontologia *Citation Typing Ontology* (CiTO) e o recurso COCI (*Crossref open DOI-to-DOI*) para definir metadados que descrevem citações em publicações científicas, capturando o objetivo de um autor ao se referir a outro trabalho acadêmico. Os autores realizaram a anotação manual das citações de acordo com conceitos existentes na CiTO. Do ponto de vista quantitativo os autores agruparam áreas temáticas conforme conceitos presente na ontologia, para analisar a prevalência das diferentes áreas que citaram o artigo científico. Apesar de terem utilizado dados anotados semanticamente em análises quali-quantitativas, não representaram os dados anotados em GCs ou como as análises poderiam ser recuperadas por eles.

Tietz e Sack (2019) apresentaram como os princípios de *linked data* e as técnicas de NLP podem ser usados para estruturar, publicar e analisar texto em linguagem natural para pesquisas sociológicas. Os autores descreveram a categorização de textos para análise como problema de pesquisa qualitativa para desenvolver novas teorias ou testar as existentes. Eles propuseram a implementação da anotação semântica com o uso de ontologias, que explicitamente estruturam o conhecimento e definem relacionamentos entre conceitos e indivíduos. Os autores usaram a anotação semântica para o processo de análise de conteúdo em sociologia.

Amdouni e Gibaud (2018) usaram tecnologias semânticas para aprimorar a recuperação de dados de biomarcadores e aumentar a interoperabilidade no campo biomédico. Representaram semanticamente os conceitos no domínio de

biomarcadores, que se fundamentam na articulação de três eixos: qualidade medida, ferramenta de medição e a ferramenta de decisão. A ontologia Imaging *Biomarker Ontology* (IBO) foi desenvolvida para suportar a anotação dos dados coletados dos biomarcadores. Os autores ilustraram a utilidade do IBO para anotar dados quantitativos e qualitativos da coleção TCGA (*The Cancer Genome Atlas*) e como os dados podem ser recuperados por GCs.

3 ESPECIFICAÇÃO DE UM MÉTODO FUNDAMENTADO EM MODELAGEM ONTOLÓGICA PARA ANÁLISE QUALI-QUANTI

O método QuaLonto especifica as etapas para fundamentar a modelagem ontológica e anotação semântica para aprimorar a elaboração de análises quali-quantitativas. As etapas do QuaLonto adotam características de um processo iterativo e incremental. Ao executar o método, torna-se possível a construção de uma ontologia para anotar *datasets* de interesse por especialistas do domínio, e para explorar (ou inferir) conhecimento sobre um determinado domínio de aplicação. O método foi elaborado com o objetivo de questões que envolvam a necessidade de explicitar o conhecimento presente nos *datasets*.

A ideia que fundamenta um processo iterativo é gerar artefatos (ou aplicações) de forma incremental, permitindo que especialistas aproveitem o que foi aprendido durante o desenvolvimento de versões anteriores e incrementais de entrega do artefato. As etapas do processo começam com uma implementação simples de um subconjunto dos requisitos para aprimorar iterativamente a sequência de versões em evolução até que a aplicação completa seja implementada. A cada iteração, modificações no projeto podem ocorrer a partir da adição de novas funcionalidades (BASILI; TURNER, 1975 apud LARMAN; BASILI, 2003).

A utilização do QuaLonto é motivada por um problema que pode ser resolvido a partir da explicitação do conhecimento e anotação semântica de *datasets*. A proposta se baseia em uma abordagem “*bottom-up*”, usando dados e SDDs a partir de uma ontologia existente (ou servindo de base para criar uma ontologia), sendo possível integrar de forma complementar a uma abordagem “*top-down*”, onde a sua aplicação é definida na próxima seção. As seções seguintes descrevem as etapas do método QuaLonto que contemplam as técnicas citadas acima.

3.1. Identificar problemas

Esta etapa é centrada na identificação do problema a ser tratado na iteração, e que abrirá caminho para definir os objetivos para implementar a solução. Para identificar o problema é necessária a descrição de um minimundo que define de forma sucinta o domínio de aplicação. Por meio do minimundo é possível dimensionar a complexidade em termos de tipos de dados ou até mesmo o processo de aquisição deles. Além disso, essa definição ajuda na compreensão dos dados a serem anotados e deve contar com a participação de especialistas de domínio para contribuir e validar o minimundo. Para ampliar o entendimento desta etapa, artefatos devem ser contemplados a fim de caracterizar os dados qualitativos e quantitativos. Por exemplo, questionários, textos, áudios ou até mesmo vídeos.

O resultado das definições servirá para construir uma primeira versão do modelo da ontologia, possibilitando a aplicação de uma abordagem top-down. Ao compreender melhor o domínio com a definição do minimundo, ontologistas podem pesquisar outras ontologias para que possam ser reutilizadas a fim de enriquecer o conhecimento sendo modelado.

Outras atividades fazem parte desta etapa. Especialistas de domínio devem estabelecer questões de competência que objetivam detalhar o problema geral em problemas menores. As questões de competência norteiam a análise dos dados e motivam a anotação semântica. Além disso, o minimundo pode ser redefinido ao executar as etapas posteriores do método. Assim, é possível que novos conceitos, dados ou outras questões de competência sejam formuladas. Neste caso, as demandas por redefinição do minimundo devem ser priorizadas na Lista de Pendências, para que possa ser discutido se devem ou não ser utilizadas.

3.2 Definir Objetivos

Esta etapa define os objetivos da solução visando orientar a etapa de Concepção e Desenvolvimento. A lista de objetivos tem como base o minimundo, as questões de competências, artefatos especificados na etapa anterior, e que instigam o desenvolvimento das etapas posteriores.

Os objetivos podem determinar as propriedades e informações necessárias para gerar os GCs, e que visam orientar as respostas das questões de competência. Alguns

dos objetivos definidos podem ser adiados e acrescentados à Lista de pendências. A definição dos objetivos conduz a anotação semântica, pois, ao ser realizada, é necessário preencher os templates de metadados com informações inerentes ao domínio definido.

Ao realizar a definição dos objetivos é possível que o entendimento sobre o domínio possa ser verificado, e, com isso, os atores envolvidos nesta etapa podem suscitar a alteração do minimundo descrito na etapa anterior, havendo a necessidade de executar novamente as etapas iniciais especificadas pelo método.

3.3 Concepção e Desenvolvimento

A etapa de Concepção e Desenvolvimento utiliza o minimundo para iniciar a modelagem da ontologia com a estruturação de conceitos, classes e propriedades para responder às questões de competência. Essas atividades conduzem o design da solução em conjunto com a anotação semântica.

Editores de ontologias devem suportar a modelagem ontológica. Sugere-se, ainda, o uso de tecnologias para anotação semântica, tais como: planilhas eletrônicas ou editores de textos. Os metadados preenchidos consolidam a anotação semântica com o uso da técnica SDD, e são considerados artefatos produzidos nesta etapa.

Datasets devem ser preparados para serem anotados semanticamente. A anotação semântica proposta neste método suporta a geração de GCs no formato RDF. Sugere-se a ingestão dos GCs em repositórios de dados semânticos, para que possam ser acessados por instruções a fim de recuperar informações e orientar os especialistas do domínio na análise dos dados.

O método QuaLonto determina, nesta etapa, um processo fundamentado em modelagem ontológica para compreender e preparar dados para análises quali-quantitativas por meio de GCs. O processo organiza seis atividades definidas da forma que segue:

- Atividade 1 - Modelagem da ontologia: Visa compreender o domínio por meio da modelagem ontológica. O minimundo é usado para fundamentar termos (ou conceitos e significados), representados por sujeitos ou respectivos complementos verbais dentro de uma oração. A elicitación dos termos deve suportar o entendimento e a criação das classes de conceitos e predicados (*objects properties*) da ontologia.
- Atividade 2 – Codificação: Um esquema de codificação é estruturado, nesta atividade, para análise qualitativa dos dados. A estrutura hierárquica da ontologia

fornece conceitos para que sejam utilizados como códigos (*codes* para o *coding*), formando um *codebook* dentro da ferramenta de CAQDAS. A codificação é usada como estratégia para permitir a análise de dados qualitativos com o uso de códigos classificados em categorias, para associá-los aos dados analisados.

- Atividade 3 – Exportação da Codificação: A codificação é exportada pela ferramenta de CQDAS. Os dados qualitativos codificados são selecionados e transpostos de linhas para colunas. Ao transpor os dados, os códigos são caracterizados por variáveis (colunas) no dataset, permitindo que estejam organizados para executar as atividades posteriores. Esta atividade expõe a necessidade de unificar os dados qualitativos em um formato tabular e compatível com tabelas de dados originados de datasets quantitativos. Conforme sugerido por Driscoll *et al.* (2007) o método misto inclui projetos que modificam uma forma de dados para outra, sendo necessária a definição do procedimento de transformação de dados qualitativos.
- Atividade 4 – Identificação do objeto de estudo: O objeto do estudo deve ser identificado univocamente por uma variável em ambos os datasets qualitativos e quantitativos. Por exemplo, se o objeto do estudo for um paciente a variável é o identificador do paciente e que deve ser anotada semanticamente por SDDs quali-quantitativos. Esta atividade identifica o objeto de estudo com o objetivo de combinar ou integrar as coleções de dados quali-quantitativos (DRISCOLL *et al.*, 2007).
- Atividade 5 – Anotação Semântica: A técnica SDD é proposta nesta atividade. Os templates de metadados são preenchidos com o apoio dos especialistas de domínio. A ontologia definida na Atividade 1 é utilizada para anotar semanticamente os dados preparados nas atividades iniciais (2, 3 e 4). As questões de competência servem de base para o design da solução e execução da técnica SDD.
- Atividade 6 – Geração do Grafo de Conhecimento: Os GCs são gerados em formato RDF por meio da execução do script *sdd2rdf* proposto por Rashid *et al.* (2017), tendo como entradas os metadados e os datasets.

Finalmente, esta etapa de Concepção e Desenvolvimento prevê a validação da ontologia. A cada iteração deve ocorrer a validação por meio de uma ontologia de topo, com objetivo de legitimar o modelo ontológico gerado a partir da explicitação do conhecimento sobre os datasets.

3.4 Demonstração

Com os GCs gerados na etapa anterior é necessário demonstrar o atendimento às questões de competência, apresentando aos envolvidos o resultado da iteração. É possível demonstrar o atendimento às questões de competência por meio de consultas ao grafo. Outras análises realizadas pelos cientistas de dados ou por ferramentas criadas pelos desenvolvedores podem complementar a demonstração.

A partir de um repositório de dados semânticos, os GCs podem ser extraídos em pequenas e grandes quantidades de dados. Essa extração de dados pode permitir que fragmentos do conhecimento sejam analisados com o objetivo de orientar os especialistas de domínio na inferência de novos conhecimentos. É possível que os grafos analisados sejam criticados, motivando os especialistas na reformulação do conhecimento modelado e explicitado nas etapas de identificação de problemas, definição de objetivos e, concepção e desenvolvimento. Daí a importância de manter o controle da evolução e da matriz de rastreabilidade dos artefatos produzidos para análises dos dados anotados semanticamente.

3.5 Avaliação

Nesta etapa, verifica-se se as etapas anteriormente especificadas estruturam procedimentos e artefatos que visam responder à investigação da pesquisa e questões de competência.

O grafo pode ser validado com o uso de consultas, a partir de visualizações dos dados incorporados, com o tratamento do grafo, ferramentas de armazenamento e processamento. A própria capacidade do método em explicitar o conhecimento sobre *datasets* e representá-los conceitualmente por meio de ontologias e grafos é um indicador da validade da abordagem proposta, sobretudo se tal capacidade for validada no contexto de diferentes estudos ou pesquisas. Avalia-se aqui, não somente a capacidade de anotar semanticamente *datasets*, mas também, a possibilidade de permitir que o método, ao ser executado, possa orientar especialistas de domínio na melhoria da análise de *datasets*.

A concepção e desenvolvimento da solução devem atender às questões de competência elaboradas pelos especialistas de domínio. Esse procedimento pode ser considerado uma forma de avaliação, e que pode motivar a reestruturação das atividades ou artefatos previstos para execução do método em questão. Assim, esta

etapa pode realimentar a próxima iteração, influenciando na identificação do problema, na definição do objetivo e no design, além de poder gerar tarefas para a Lista de pendências.

3.6 Divulgação

De acordo com Gonçalves (2020) a divulgação é um subprocesso que representa o estágio final de uma iteração, no qual as descobertas são compartilhadas com os interessados nas análises dos dados anotados semanticamente. A divulgação pode ser realizada por meio de dashboards publicados na internet, por exemplo, onde os dados são combinados em gráficos usando filtros de campos que possibilitem as análises de interesse sobre os dados anotados.

Outros tipos de divulgação podem ser acadêmicos por meio de revistas científicas. Esta etapa não ocorre em todas as iterações, pois depende do estágio de evolução da integração semântica. Nesta etapa, os problemas, os objetivos da solução e o design podem ser revistos de acordo com o feedback do público.

3.7 Controle

Nesta etapa é sugerido o controle da evolução da execução do QuaLonto. A cada iteração os atores envolvidos produzem diferentes artefatos em cada etapa do método, tornando necessário o controle da produção dos artefatos. O controle pode ser por meio de técnicas de versionamento e configuração de software, ou até mesmo o uso de matriz de rastreabilidade.

4 CONSIDERAÇÕES FINAIS

O método especificado é fundamentado em modelagem ontológica e estabelece etapas para compreensão (com uso de ontologias) e preparação de dados para análise quali-quantitativa. O método formaliza o uso de ontologias para enriquecer o conhecimento, com o objetivo de auxiliar na compreensão dos dados. O método é independente de uma ontologia específica, tanto para fornecer os *codes* da análise qualitativa, quanto para anotação semântica. No entanto, o método propõe atividades iniciais para a modelagem ontológica do conhecimento para criar uma primeira versão

de ontologia. A preparação de dados, apoiada por SDD, reutiliza o conhecimento disponível na ontologia a fim mapear os dados por meio de templates de metadados. Argumentou-se neste trabalho que a geração de grafos de conhecimento, com o uso da anotação semântica pelo SDD, pode melhorar a compreensão dos dados podendo aprimorar análises quali-quanti.

Como trabalhos futuros, dados reais serão tabulados, a partir da aplicação do questionário de estratificação de risco em ansiedade e tratamentos psicoterápicos em pacientes, visando validar o método especificado.

Agradecimentos

Agradecimentos ao Centro de Pesquisa e Extensão do UniAcademia e Curso de Graduação em Psicologia do UniAcademia.

ABSTRACT

Quali-quantitative analysis is a research technique that requires integrating qualitative and quantitative data under the same objects of study and produced in different research tracks. Using various forms of representation of this data makes it possible to clarify more complex relationships and better assimilate the research information. Understanding and Preparation Data are initial activities of quali-quantitative research that use techniques to explain knowledge about data. Ontologies are commonly used to structure the knowledge embedded in datasets and can facilitate extracting information, disambiguating terms, and preserving the semantics of variables. In this work, a systematic method based on semantic data annotation is specified to enable the data, information, and knowledge generated by research to be explored through graphs. A semantic data dictionaryization technique is suggested to create graphs from ontologies, metadata, and data. It allows domain experts to explore the information and knowledge investigated by explaining them from the data. In this way, the method proposes activities that generate graphs to organize semantically annotated data to explore and filter information to answer research questions of interest. They facilitate the graphical visualization of data and have the potential to improve quali-quantitative analyses. As future work, it proposes a case study in psychology to illustrate and validate the proposed method in quali-quantitative research that explores non-synthetic psychotherapeutic data.

Keywords: Psychotherapeutic Treatment. Knowledge Graphs. Quali-quantitative analysis. Ontology.

REFERÊNCIAS

AMDOUNI, Emna; GIBAUD, Bernard. Imaging Biomarker Ontology (IBO): A Biomedical Ontology to Annotate and Share Imaging Biomarker Data. *Journal on Data Semantics* 7, no. 4 (2018): 223-236.

BAX, M. P. ; SILVA, Evaldo de Oliveira . Dicionários Semânticos de Dados para Integrar Dados de Prontuários Eletrônicos de Pacientes (Semantic Data Dictionary to Integrate Data of Electronic Medical Records). In: *Ontobras, 2020, Vitória. Proceedings of the XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS 2020)*. Aachen: CEUR, 2020. v. 2728. p. 298-303.

BENFARES, Chaymae; EL IDRISSEI, Younès El Bouzekri; HAMID, Karim. Personalized healthcare system based on ontologies. In: *International Conference on Advanced Intelligent Systems for Sustainable Development*. Springer, Cham, 2018. p. 185-196.

BERTHOLD, M.R., Borgelt, C., Höppner, F., Klawonn, F.: *Guide to intelligent data analysis: how to intelligently make sense of real data*. Springer (2010).

BRAUN, Virginia, e CLARKE, Victoria. "Using thematic analysis in psychology." *Qualitative research in psychology* 3, no. 2 (2006): 77-101.

BRENAS, Jon Hael; SHIN, Eun Kyong; SHABAN-NEJAD, Arash. Adverse childhood experiences ontology for mental health surveillance, research, and evaluation: advanced knowledge representation and semantic web techniques. *JMIR mental health*, v. 6, n. 5, p. e13498, 2019.

CALDIERARO, Fábio; FREITAS, Henrique Freitas; PETRINI, Maira e POZZEBON, Marlei. "Sistemas de informação de marketing: uma aplicação e seus resultados." *PPGA/UFRGS: Série Documentos para Estudo* 3 (1998): 1-17.

CEUSTERS, Werner; SMITH, Barry. Foundations for a realist ontology of mental disease. *Journal of biomedical semantics*, v. 1, n. 1, p. 1-23, 2010.

CHARLOTTE, Siefriid; GROSJEAN, Julien; LEFEBVRE, Tatiana; ROLLIN, Laetitia; DARMONI, Stefan; SCHUERS, Matthieu. Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations. *International Journal of Medical Informatics* 133 (2020): 104009.

DOOLEY, Damion M; GRIFFITHS, Emma J.; GOSAL, Gurinder; BRINKMAN, Fiona SL; HSIAO, William WL. The Genomic Epidemiology Ontology and GEEM Ontology Reusability Platform. In *JOWO*. 2017.

DRISCOLL, David L; APPIAH-YEBOAH, Afua; Philip Salib; RUPERT, Douglas J. "Merging qualitative and quantitative data in mixed methods research: How to and why not." (2007).

EL GHOSH, Mirna; NAJA, H.; ABDULRAB, H.; KHALIL, M. Towards a middle-out approach for building legal domain reference ontology. *International Journal of Knowledge Engineering* 2, no. 3 (2016): 109-114.

FREITAS, H. M. D., & MOSCAROLA, J. (2000). Análise de dados quantitativos e qualitativos: casos aplicados. Porto Alegre: Sphinx Consultoria.

FRANCESCONI, Enrico; MONTEMAGNI, Simonetta; PETERS, Wim; TISCORNIA, Daniela. Integrating a bottom-up and top-down methodology for building semantic resources for the multilingual legal domain. Springer Berlin Heidelberg, 2010.

GONÇALVES, José Eugênio de Assis et al. Método ágil de integração semântica de dados científicos baseado em ontologias. Tese de Doutorado. 2020.

GUARINO, Nicola (Ed.). Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy. IOS press, 1998.

HAN, Dong, and STOLFFEL, Kilianl. "Ontology based qualitative case studies for sustainability research." In Proceedings of the AI for an Intelligent Planet, pp. 1-8. 2011.

HASTINGS, Janna et al. Representing mental functioning: Ontologies for mental health and disease. 2012.

HEIBI, Ivan; PERONI, Silvio. A qualitative and quantitative analysis of open citations to retracted articles: the Wakefield 1998 et al.'s case. Scientometrics 126, no. 10 (2021): 8433-8470.

HOGAN, Aidan, BLOMQUIST, Eva, COCHEZ, Michael, D'AMATO, Claudia, MELO Gerard de, GUTIERREZ, Claudio, GAYO, José Emilio Labra, KIRRANE, Sabrina, NEUMAIER, Sebastian, POLLERES, Axel, NAVIGLI, Roberto, NGOMO, Axel-Cyrille Ngonga, RASHID, Sabbir M., RULA, Anisa, SCHMELZEISEN, Lukas, SEQUEDA, Juan, STAAB, Steffen, ZIMMERMANN, Antoine. Knowledge Graphs. arXiv preprint arXiv:2003.02320, 2020.

HOCKER, J., BIPAT, T., MCDONALD, D. W., & Zachry, M. (2021). Developing an Ontology for Qualitative Coding Schemas - QualiCO. Disponível em: <https://eacer.de/eacer-programmes/conference/26/contribution/50889/>. Acesso em 28 de jun de 2022.

IGLESIAS, Carlos A.; SANCHEZ-RADA, J Fernando; VULCU, Gabriela; BUITELAAR, Paul. Linked data models for sentiment and emotion analysis in social networks. In Sentiment Analysis in Social Networks, pp. 49-69. Morgan Kaufmann, 2017.

KUNG, Hsiang-Jui; KUNG, L.; GARDINER, Adrian. Comparing top-down with bottom-up approaches: teaching data modeling. In Proceedings of the Information Systems Educators Conference ISSN, vol. 2167, p. 1435. 2012.

LARMAN, Craig; BASILI, Victor R. Iterative and incremental developments. a brief history. Computer, v. 36, n. 6, p. 47-56, 2003.

MATTEO, Cristani; CUEL, Roberta. A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems (IJSWIS)* 1, no. 2 (2005): 49-69.

MIÑARRO-GIMÉNEZ, José Antonio; MARTÍNEZ-COSTA, Catalina; KARLSSON Daniel; SCHULZ Stefan; GØEG, Kirstine Rosenbeck. Qualitative analysis of manual annotations of clinical text with SNOMED CT. *Plos one* 13, no. 12 (2018): e0209547.

PAN, Jeff Z. et al. (Ed.). *Exploiting linked data and knowledge graphs in large organisations*. Heidelberg: Springer, 2017.

PFAFF, Matthias; NEUBIG, Stefan; KRCMAR, Helmut. Ontology for semantic data integration in the domain of IT benchmarking. *Journal on data semantics* 7 (2018): 29-46.

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

ROSENBERG, Rasmus Larsen; HASTINGS, Janna. Mapping the patient's experience: An applied ontological framework for phenomenological psychopathology. *Phenomenology and Mind* 18 (2020): 200-219.

SOMEKH, Bridget; LEWIN, Cathy, eds. *Research methods in the social sciences*. Sage, 2005.

TIETZ, Tabea; SACK, Harald. Linked Data Supported Content Analysis for Sociology. In *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings* 15, pp. 34-49. Springer International Publishing, 2019.

YAMADA, Diego Bettiol et al. Ontology-Based Inference for Supporting Clinical Decisions in Mental Health. In: *International Conference on Computational Science*. Springer, Cham, 2020. p. 363-375.

WILLIAMS, M., & MOSER, T. (2019). The art of coding and thematic exploration in qualitative research. *International Management Review*, 15(1), 45-55.