UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO

JOSÉ EUGÊNIO DE ASSIS GONÇALVES

Método Ágil de Integração Semântica de Dados Científicos Baseado em Ontologias

Belo Horizonte
2020

JOSÉ EUGÊNIO DE ASSIS GONÇALVES

Método Ágil de Integração Semântica de Dados Científicos Baseado em Ontologias

Tese apresentada ao Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Escola de Ciência da Informação da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de doutor em Ciência da Informação, área de concentração Ciência da Informação.

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida pelo autor e aprovada pela Comissão julgadora.

Linha de Pesquisa: Gestão e Tecnologia

Orientador: Marcello Peixoto Bax

BELO HORIZONTE

2020

G635m Gonçalves, José Eugênio de Assis

Método ágil de integração semântica de dados científicos baseado em ontologias [recurso eletrônico] . / José Eugênio de Assis Gonçalves. - 2020.
1 recurso eletrônico. (108f. : il., color): pdf.

Orientador: Marcello Peixoto Bax
Tese (Doutorado) - Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 103-108.
Exigências do sistema: Adobe Acrobat Reader.

1. Ciência da Informação - Teses. 2. Ontologias (Recuperação da informação) - Teses. 3. Web semântica - Teses. I. Título. II. Bax, Marcello Peixoto. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4.03

Ficha Catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

RESUMO

Integrar dados gerados por pesquisas científicas é uma atividade cada vez mais importante para a evolução da Ciência de Dados. Tal integração pode ser realizada com o auxílio de esquemas de dados (modelos), que definem como estes devem ser compreendidos, relacionados e formatados, determinando como são organizados. Contudo, se por um lado esquemas de dados relacionais pré-definidos possam favorecer a sua integração, compartilhamento e reúso pelos membros de uma comunidade científica, por outro, retiram a flexibilidade de representação dos dados pelo pesquisador, já que este deve respeitar o esquema pré-definido, caso intencione compartilhar seus dados com a comunidade. A pesquisa visa explorar e propor uma forma de integrar dados que não se prende à rigidez de esquemas relacionais pré-definidos. Propõe-se a utilização de ontologias para permitir que cada estudo científico utilize um desenho conceitual próprio, e ainda mantenha a capacidade de integração e reúso dos dados coletados pelo estudo. A integração é obtida a partir dos conceitos comuns aos estudos, definidos formalmente por ontologias. Espera-se que o uso de ontologias contribua para a interoperabilidade de dados e sistemas. Ao invés de esquemas relacionais rígidos, utiliza-se estruturas canônicas em formato de triplas: "sujeito", "predicado" e "objeto", interligadas e constituindo um grafo. O objetivo da pesquisa é desenvolver um método iterativo para facilitar a realização do processo de integração semântica de dados produzidos durante pesquisas científicas. O método permite que o pesquisador conceba a ontologia de domínio (que integra os dados) em ciclos curtos de desenvolvimento, ao longo da pesquisa. Esta é a principal contribuição do método proposto. Ele dispensa o pesquisador de ter que desenvolver a ontologia de integração, para somente depois integrar os dados. Fundamentado na *Agile Design Science Research Methodology*, ele permite integrar os dados e evoluir a ontologia a cada ciclo, com a participação de todos os atores envolvidos. Durante a fase de validação dos resultados desta pesquisa, notou-se que a colaboração entre todos os envolvidos foi facilitada com o uso do método proposto, e as decisões puderam ser tomadas mais prontamente em vista do acesso precoce dos mesmos aos dados e informações integradas semanticamente, cuja análise foi realizada com o auxílio de artefatos elaborados para esta finalidade. O método foi validado, utilizando-se uma pesquisa que integrou dados socioeconômicos e ambientais com informações sobre casos de dengue e esquistossomose no Brasil.

Palavras-chave: Integração Semântica de Dados. Ontologias. Grafos de Conhecimento. Web Semântica. Métodos Ágeis. Ciência de Dados. Dados Epidemiológicos.

ABSTRACT

Integrating data generated by scientific research is an increasingly important activity for the evolution of Data Science. Such integration can be accomplished with the aid of data schemes (models), which define how they should be understood, related and formatted, determining how they are organized. However, if, on the one hand, predefined relational data schemes can favor their integration, sharing and reuse by members of a scientific community, on the other hand, they remove the flexibility of data representation by the researcher, since he must respect the pre-defined scheme if he intends to share your data with the community. The research aims to explore and propose a way to integrate data without the rigidity of pre-defined relational schemes. It is proposed to use ontologies to allow each scientific study to use its own conceptual design, and still maintain the ability to integrate and reuse the data collected by the study. The Integration is obtained from concepts common to studies, formally defined by ontologies. The use of ontologies is expected to contribute to the interoperability of data and systems. Instead of rigid relational schemes, canonical structures in the form of triples of “subject”, “predicate” and “object” are used, interconnected and constituting a graph. The objective of the research is to develop an iterative method to facilitate the realization of the process of semantic integration of data produced during scientific research. The method allows the researcher to design the domain ontology (which integrates the data) in short development cycles, throughout the research. This is the main contribution of the proposed method. It frees the researcher from having to develop the integration ontology, only to later integrate the data. Based on the Agile Design Science Research Methodology, it allows integrating data and evolving ontology with each cycle, with the participation of all the actors involved. During the validation phase of the results of this research, it was noted that collaboration between all involved was facilitated with the use of the proposed method, and decisions could be made more readily in view of their early access to data and semantically integrated information, whose analysis was performed with the aid of artifacts designed for this purpose. The method was validated, using a survey that integrated socioeconomic and environmental data with information on cases of dengue and schistosomiasis in Brazil.

Keywords: Semantic Data Integration. Ontologies. Knowledge Graphs. Semantic Web. Agile Methods. Data Science. Epidemiological data.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABox valores associados ao modelo conceitual. 26, 27, 46

ADSRM *Agile Design Science Research Methodology*. 54–57, 82, 107, 108, 110

CHEAR *Children's Health Exposure Analysis Resource*. 59

CI Ciência da Informação. 17, 50

CSV arquivos de texto com valores separados por vírgula. 6, 31, 32, 35, 62, 63, 66–68, 70, 72, 75–77, 101, 111

DFG German Research Foundation. 21

DOI Digital Object Identifiers. 21

DSR Design Science Research. 18, 49–52, 54, 56, 108

FAIR Localização (*Findability*), Acessibilidade, Interoperabilidade e Reutilização. 24

Fiocruz Fundação Oswaldo Cruz. 58, 62, 84, 89, 109

HADatAc *Human-Aware Data Acquisition Framework*. 18, 19, 36, 46, 47, 60, 66–69, 73, 74, 79, 81–85, 90, 91, 94, 96, 100, 107–112

HAScO *Human-Aware Science Ontology*. 6, 36–38, 40, 42–44, 46, 47, 63, 64

JSON-LD *Javascript Object Notation for Linked Data*. 102

LOD Linked Open Data. 15

OAS Especificação de Acesso a Objetos. 7, 19, 72, 75–80, 83, 88, 110

OBDM *Ontology-Based Linking Open Database*. 103

omics Estudos relacionados ao genoma, as proteínas e ao metabolismo. 24

OWL Ontology Web Language. 28, 40, 41

R2RML RDB2RDF *Mapping-Language*. 32, 33, 104, 105

RDF *Resource Description Framework* . 6, 19, 27–29, 31–34, 46, 47, 56–63, 66, 67, 69–73, 75, 77, 82, 96, 103, 104, 106, 112

RPI Rensselaer Polytechnic Institute. 16

SDD Dicionário Semântico de Dados. 7, 19, 73–80, 83, 87, 88, 90, 93, 94, 101, 104, 110, 111

SIO *Semanticscience Integrated Ontology* . 33, 38–40, 63

SPARQL Protocolo e linguagem de consulta RDF SPARQL. 29, 31, 60, 82, 99, 100, 104, 105

SSD Design Semântico de Estudo. 7, 19, 69, 70, 75–80, 83, 88, 90, 91, 101, 110

TBox definição de classes e propriedades como um vocabulário de domínio. 26, 27, 46, 111

TWC Tetherless World Constellation. 16, 84

UFMG Universidade Federal de Minas Gerais. 17, 84

URI Identificador Uniforme de Recursos. 21, 28, 29, 40, 46, 63, 66, 70–73, 75, 77, 78

URL Localizador Padrão de Recursos. 28, 63

W3C *World Wide Web Consortium*. 14, 28, 33, 42, 104

SUMÁRIO

1 INTRODUÇÃO

Para os autores Fox e Hendler (2009), a coleta e a análise de dados são essenciais para toda a Ciência, mas são especialmente importantes no contexto do novo paradigma da "Ciência de Dados", onde as descobertas científicas são intensivamente dependentes da gestão adequada dos dados coletados. O entendimento da relação entre variáveis socioeconômicas e ambientais com uma determinada doença, tal como a dengue, por exemplo, envolve o tratamento de um grande volume de dados. Em um estudo como esse, os pesquisadores poderiam examinar informações sobre como a incidência da dengue na população pode ser afetada por essas variáveis.

Em muitos casos, tais informações são provenientes de diferentes estudos, os quais precisariam ser integrados para este tipo de análise. À medida que se avança no caminho da medicina personalizada, por exemplo, a compreensão dos dados coletados dos ensaios clínicos e de outros estudos é crucial para a pesquisa. Infelizmente, a coleta, organização e disseminação de dados estão se tornando cada vez mais difíceis de realizar com eficiência. Entender um fenômeno e provar uma hipótese requer grandes quantidades de dados, e nem todos os pesquisadores dispõem dos recursos e meios para coletar e tratar esses dados.

Os ensaios clínicos custam caro, exigem recursos não triviais e podem levar anos para serem realizados, dependendo do estudo. Assim, as informações provenientes de tais estudos, normalmente não estão prontamente disponíveis para todos os pesquisadores. Esta é uma das razões pelas quais os campos das ciências da saúde (medicina genética, biomedicina, etc) estão caminhando para um mais amplo compartilhamento desses dados por meio de repositórios públicos ou voltados para grupos privados e disponíveis aos pesquisadores. Ao compartilhar esses dados, os pesquisadores podem aprofundar seu entendimento e fazer conexões com vários conjuntos de dados.

No entanto, os pesquisadores enfrentam atualmente vários problemas com essa abordagem. Primeiro, cada pesquisador ou laboratório descreve os fenômenos observados na realidade de forma idiossincrática. Uma consequência óbvia é que poucos ou nenhum dos conjuntos de dados gerados serão conceitualmente "harmônicos" entre si; em geral eles são relacionados, mas não são estruturados e formatados da mesma maneira, portanto, não podem ser integrados facilmente para análise. Segundo, o pesquisador principal (coordenador da pesquisa) geralmente sabe muito mais detalhes sobre os dados coletados do que consegue transmitir de maneira explícita aos outros membros de sua equipe. Organizar e registrar estes detalhes, de maneira explícita, estruturada e sistemática (mas também de forma consistente conceitualmente e coesa), representaria um ganho potencial considerável para o ciclo da pesquisa científica.

Finalmente, se os conjuntos de dados forem por demais extensos, os processos automatizados podem ser a única maneira de gerar análises abrangentes. É claro que, se os dados não forem "legíveis" por máquina de alguma forma, não será possível definir tais processos.

O conceito de integração semântica de dados (MEEHAN *et al.*, 2017) aparece neste contexto, procurando solucionar ou mitigar alguns destes problemas ao reestruturar os dados em um formato acessível, anotado por metadados e formal (legível por máquina).

Várias organizações de pesquisa estão no processo de desenvolvimento que incorporarão a integração semântica de dados em seus repositórios (UCITELI; KIRSTEN, 2015; BIFFL *et al.*, 2013). Como esse é um fenômeno relativamente recente, não há um método acordado para isso, portanto, existem diversas proposições de como realizar a tarefa (Capítulo 8). Anotando-se com metadados, acrescenta-se valor aos dados, garantindo novas possibilidades na realização de consultas e facilitando a sua integração (VERSTICHEL *et al.*, 2011). Dessa maneira, pesquisadores podem explorar e cotejar vários conjuntos de dados para resultados de interesse e chegar a novas intuições e discernimentos (*insights*) científicos.

A pesquisa científica pode obter diversos benefícios técnicos oriundos da integração semântica de dados. A nossa visão é a de que uma melhor gestão da organização conceitual e da disponibilidade de dados para as pesquisas científicas, está relacionada com o montante de conhecimentos que pode ser extraído desses dados. É por isso que os grupos de pesquisa estão começando a desenvolver ferramentas para estruturar melhor esses dados, para que possam ser analisados mais prontamente. Disponibilizar dados de tal forma também diminui as barreiras inerentes à realização de pesquisas científicas, permitindo que pessoas utilizem dados que não foram coletados por elas. Isso também abre portas para estudos cruzados, complementares e colaboração entre domínios, uma tendência que vem aumentando constantemente nos últimos anos (DARAIO *et al.*, 2016).

A organização conceitual formal dos dados científicos facilita a sua análise, e vem se tornando uma atividade inseparável da prática científica atual (HEY *et al.*, 2009). Além disso, a curadoria de dados para a ciência colaborativa interdisciplinar requer uma nova abordagem online baseada na web que integre o conhecimento de múltiplos domínios e possibilite a colaboração iterativa entre fornecedores de dados, especialistas de domínio e analistas de dados (MCCUSKER *et al.*, 2017). As pesquisas atuais, entretanto, ainda enfrentam problemas tais como uma estrutura de diretórios limitada ou o uso de grandes aquivos para representar a informação (VALLE, 2018). O que se vê ainda, é o acesso a arquivos específicos, armazenados em um notebook do pesquisador, o que torna mais complexo o processo de acesso a dados. Nesse caso, podemos ter uma situação que nenhum acesso é feito a esses dados e metadados, que ficam nos laboratórios, nos notebooks dos cientistas e até na mente dos pesquisadores.

Enriquecer semanticamente o material científico com metadados pode, portanto, trazer oportunidades, tanto para a representação quanto para o armazenamento e recuperação mais granular das coleções de dados. Uma delas é a possibilidade de reúso de dados existentes para novos estudos.

O reúso dos dados tem sido facilitado com o esforço da comunidade científica em padronizar os formatos de dados a serem utilizados por todos, permitindo uma maior interoperabilidade entre projetos de pesquisa. Do ponto de vista estritamente técnico e de forma bastante rudimentar essa interoperabilidade já existe com os padrões definidos pelo *World Wide*

Web Consortium (W3C)[1]. Estes padrões permitem criar ontologias (Seção 2.2.1) que definam o conhecimento existente sobre um determinado domínio, com dados e metadados. Grande parte dessas ontologias são abertas e se interligam formando a Linked Open Data (LOD)[2].

Atualmente, algumas ontologias possuem propriedades, tais como o volume de dados armazenados, que as caracterizam como "grafos de conhecimento" [3]. Cita-se, por exemplo, a DBPedia[4], o Wikidata[5] e outros. Em teoria, uma ontologia, mesmo que não seja tão abrangente como essas, pode ser considerada um grafo de conhecimento ao atender certos critérios que serão discutidos na Seção 2.2.1.2. Porém, na prática, um grafo de conhecimento possui foco na representação das instâncias da ontologia. Outro aspecto interessante das ontologias é a possibilidade de gerar metadados para validar a procedência dos dados, o que melhora a confiabilidade da informação, pois torna possível verificar quem foi o criador da informação. Além disso, a interligação entre as bases de dados é uma forma de adicionar credibilidade à informação.

1.1 Síntese do problema

A utilização de recursos para processar, formatar e integrar dados e informações de diversos estudos, gerando conhecimento em um formato acessível ao processamento de máquina é uma necessidade para o desenvolvimento da pesquisa científica (WILKINSON *et al.*, 2016; MOORE, 2001). Nesse sentido, o emprego de ontologias associadas a ferramentas para a criação de grafos de conhecimento, não somente pode permitir a integração semântica, como fornecer meios para se lidar com grandes volumes de dados. Entretanto, criar uma ontologia não é uma tarefa simples, podendo demandar meses ou anos de trabalho de investigação científica de um domínio.

Além disso, em geral, uma ontologia permanece sempre em evolução, pois ela representa o estado de conhecimento de um domínio em um dado momento no tempo. A criação de uma ontologia é, portanto, um processo inerentemente iterativo (NOY; MCGUINNESS *et al.*, 2001), onde a cada etapa temos uma versão aprimorada da mesma.

Assim, a hipótese subjacente a este trabalho foi a suposição de que o processo de criação do grafo de conhecimento a partir de ontologias, com o intuito de integrar semanticamente dados científicos, pode ser realizado de forma ágil, com a utilização de um método que discipline o processo de evolução da modelagem e que envolva a participação de todos os

[1] https://www.w3.org/
[2] https://lod-cloud.net/
[3] Um grafo é uma que representação abstrata de um conjunto de objetos e das relações existentes entre eles. É definido por um conjunto de nós ou vértices, e pelas ligações ou arestas, que ligam pares de nós. Uma grande variedade de estruturas do mundo real podem ser representadas abstratamente através de grafos, inclusive o conhecimento humano. Os grafos de conhecimento são uma abstração útil para organizar o conhecimento do mundo de forma estruturada pela Internet, capturando relacionamentos entre as principais entidades de interesse; são também uma maneira de integrar informações extraídas de várias fontes de dados. Os grafos de conhecimento desempenham um papel central no aprendizado de máquina e no processamento de linguagem natural como um forma de incorporar o conhecimento contextual sobre os objetos do mundo e explicar o que está sendo aprendido.
[4] http://pt.dbpedia.org/
[5] http://wikidata.org

interessados (SEQUEDA; MIRANKER, 2017). Dessa forma, neste trabalho, foi proposto um método ágil original (inspirado de Conboy, Gleasure e Cullina (2015)) para solucionar o seguinte problema de pesquisa:

Integrar semanticamente dados científicos heterogêneos.

Segundo o método, a evolução do processo de integração se dá por meio de iterações curtas, onde, a cada iteração a ontologia de domínio que modela os dados e o grafo de conhecimento, são avaliados identificando-se problemas e lacunas de modelagem a serem solucionados em futuras iterações.

Como em nossa investigação a integração de dados é planejada para ocorrer no âmbito da execução de um projeto de pesquisa científica, são as perguntas de pesquisa elaboradas pelo(s) pesquisador(es) responsável(eis) que orientam a evolução do processo de integração.

1.2 Proposta da pesquisa

Como revelou a revisão de literatura, embora existam alguns métodos de anotação semântica de dados visando a sua integração no domínio de determinado estudo científico, pode-se afirmar que nenhum deles atingiu maturidade suficiente para ser utilizado em larga escala. Um dos problemas com a maioria dos métodos atuais é que eles requerem que a ontologia de domínio, fonte da anotação, já se encontre madura, pronta ou pelo menos estável o suficiente para ser utilizada.

Assim nestes métodos, para que a ontologia possa ser utilizada no processo de anotação, esta não pode sofrer alterações constantes após o início do mesmo. Tal requisito representa um entrave prático ao processo, uma vez que raramente será encontrada a ontologia adequada ou ideal para marcar um determinado conjunto de dados. Com efeito, na prática, o que tende a ocorrer é que a ontologia de domínio vai sendo desenvolvida conforme evolui a geração e/ou o reúso dos dados no escopo do estudo científico.

A proposta desta tese foi construir um método que seguisse os princípios ágeis e que permitisse que a ontologia de domínio pudesse evoluir em paralelo ao desenvolvimento da pesquisa científica em si. Denominamos o método "Odin" - *Ontology-based Data Integration*. Utilizou-se uma ferramenta para o processo de integração semântica de dados implementada por um *framework*[6] em desenvolvimento pela equipe coordenada por Deborah McGuiness no laboratório Tetherless World Constellation (TWC)[7] do Rensselaer Polytechnic Institute (RPI)[8] no estado de Nova Iorque, com o qual estabeleceu-se parceria para esta investigação. A parceria viabilizou-se graças ao estágio pós-doutoral do orientador científico da presente tese.

A fim de caracterizar os resultados esperados da aplicação do método para realizar a integração semântica de dados, elaboramos as hipóteses a seguir.

[6] hadatac.org
[7] https://tw.rpi.edu/
[8] https://www.rpi.edu/

1.3 Hipóteses

A partir de uma avaliação prévia das características desejáveis do método (seus requisitos) para a integração de dados foram elaboradas hipóteses relacionadas às expectativas da aplicação do mesmo. Assim, considerou-se que:

- Adotar uma metodologia ágil garante a obtenção dos resultados da integração semântica de dados mais rapidamente, facilitando a correção precoce dos problemas encontrados e motivando o(s) pesquisador(res) envolvido(s) na exploração dos dados.
- Utilizar ontologias enriquece semanticamente os dados oriundos de fontes diversas e/ou de vários estudos, para além de meramente integrá-los sintaticamente.
- Adotar a abordagem ontológica de modelagem, explicada na Seção 2.2, auxilia o processo de integração de dados heterogêneos.
- Desenvolver ferramentas que permitam lidar com a ontologia gerada durante o processo de integração garante a utilização e busca das informações desta ontologia.

1.4 Justificativa e relevância

Durante pesquisas científicas de natureza empírica, os dados científicos coletados precisam ser transformados, descritos, integrados, preservados, recuperados e analisados. Estes procedimentos envolvem significativos esforços dos pesquisadores. Principalmente quando estes necessitam lidar com grandes volumes de dados, muitas vezes heterogêneos e não harmonizados em suas unidades de medida, escalas, índices e dimensões, além de oriundos de fontes diferentes.

O método Odin baseia-se na utilização de ontologias, que por hipótese de nossa pesquisa tem potencial para facilitar o processo, permitindo aos pesquisadores realizar a integração semântica de dados agregando conteúdo e formalismo lógico aos mesmos.

Outro aspecto que justificou a proposição do método foi a necessidade de propiciar aos pesquisadores de domínio a obtenção de resultados mais rapidamente, com a participação contínua dos mesmos em cada etapa do processo.

A proposição de um método para organizar e integrar o conhecimento científico no contexto de uma investigação, favorecendo o seu gerenciamento, bem como a recuperação de informações se situa no contexto da curadoria digital de dados científicos (SAYÃO; SALES, 2012). Um trabalho alinhado com o esforço realizado na área da Ciência da Informação (CI) para entender e avançar no tratamento dos desafios recentes impostos pela necessidade de curadoria de dados científicos foi desenvolvido recentemente na Escola de Ciência da Informação da Universidade Federal de Minas Gerais (UFMG) (RESENDE, 2019).

A integração semântica de dados, sistematizada pela utilização do método conforme projetado, favorece a obtenção de conhecimento útil de forma automática a partir dos dados

científicos. O método, por sua vez, ao se mostrar efetivo e de fácil utilização, pode garantir maiores aplicações dessa integração na pesquisa científica.

1.5 Objetivos

O objetivo geral foi desenvolver um método ágil de integração semântica de dados científicos que pudesse ser utilizado para grandes volumes de dados e que fosse de fácil operacionalização para os especialistas de domínio, um público de pesquisadores. O método deveria ser capaz de integrar dados de diferentes fontes (estudos científicos), baseados em estruturas conceituais e formatos heterogêneos.

O método foi avaliado na sua capacidade de preparar os dados para análise [9], integrado-os por meio de ontologias, gerando um grafo de conhecimento.Foram verificados os fatores que contribuíram para a gestão e uso mais efetivo da informação contida no grafo, de maneira que o mesmo pudesse auxiliar no trabalho de exploração e análise dos dados, eventualmente propiciando a obtenção de repostas para questionamentos dos especialistas de domínio.

Para realizar tal intento, aplicou-se o que preconiza a Design Science Research (DSR) (WIERINGA, 2009; BAX, 2015), para derivar os objetivos específicos a seguir. Tais objetivos se configuram como de ordem tanto teórica quanto prática, podendo os resultados serem verificados empiricamente.

- Realizar ciclos de integração de dados utilizando a plataforma de software *Human-Aware Data Acquisition Framework* (HADatAc) (descrita na Seção 3.4) por meio de experimentos. Sendo o principal deles, integrar os dados de uma pesquisa epidemiológica realizada por pesquisadores parceiros.
- A partir dos conhecimentos gerados por meio dos experimentos, propor o método e fazer evoluir a ontologia de domínio em acordo com o método proposto, ou seja, a ontologia deverá poder se aprimorada a cada ciclo de desenvolvimento da integração dos dados.
- Executar consultas aos dados integrados e "ingeridos" com o auxílio da plataforma de software (HADatAc), utilizando a sua interface que expõe e disponibiliza os dados por meio de busca facetada.
- Realizar outros experimentos com o método proposto, e construído nesta pesquisa, para validá-lo em conjunto com pesquisadores de domínios diferentes.

1.6 Contribuições

A principal contribuição do trabalho é a criação do método, que foi construído com a auxílio da plataforma de software HADatAc, apresentado na Seção 3.4.

[9] O processo de preparação envolve transformar, descrever, integrar, preservar e recuperar os dados

A capacidade de fazer evoluir a ontologia de domínio, de forma incremental e cíclica, por meio dos princípios ágeis, constitui também uma contribuição significativa. Tal abordagem permitiu uma interação maior com os pesquisadores da equipe, que tiveram a oportunidade de revisar precocemente a ontologia de domínio a cada ciclo da pesquisa, eliminando a necessidade de uma definição acabada e final da mesma para se obter os primeiros resultados da exploração dos dados.

Com a utilização do método, os especialistas de domínio realizaram a integração semântica de dados. Esta integração representou possibilidades de reúso de dados, ao integrar dados de diversas fontes através de dimensões comuns. Deste modo, foi possível trabalhar com resultados obtidos de vários estudos relacionados.

Os experimentos com a modelagem ontológica utilizando o HADatAc permitiram validar o que ressalta Rector *et al.*, que estabelece que os modelos ou esquemas de dados têm funções diferentes da representação do conhecimento de um domínio, ou seja, as declarações sobre o domínio são diferentes do modelo definido para o armazenamento destas declarações. O modelo pode conter metadados que não pertencem ao domínio, mas que fornecem elementos para auxiliar na estruturação dos dados do domínio, tais como a obrigatoriedade de um determinado campo. Deste modo, segundo Rector *et al.*, as "entidades de informação" sobre estruturas de dados deveriam ser distintas da base de conhecimento sobre o domínio, separando assim, o "conhecimento do domínio", do "modelo das estruturas de dados e informação" que será utilizado para persistir/armazenar o conhecimento do domínio. No HADatAc os templates utilizados para o mapeamento dos dados no grafo RDF definem claramente esta separação. Neste caso, temos três templates principais o Design Semântico de Estudo (SSD), a Especificação de Acesso a Objetos (OAS) e o Dicionário Semântico de Dados (SDD), os quais foram detalhados na Seção 5.2. O SSD e a OAS definem a estrutura da informação, enquanto o SDD captura o conhecimento do domínio.

1.7 Organização geral da tese

Além deste capítulo de introdução, que apresentou a proposta de pesquisa, o problema e as hipóteses levantadas, a questão de pesquisa, os objetivos e as contribuições esperadas, esta tese é composta dos seguintes capítulos:

O Capítulo 2 apresenta o referencial teórico, relatando e definindo conceitos para o entendimento do trabalho. As ontologias e ferramentas utilizadas no método são apresentadas no Capítulo 3. O Capítulo 4 descreve a metodologia empregada na pesquisa, destacando os passos seguidos para a elaboração da mesma. A integração semântica de dados é descrita no Capítulo 5. O Capítulo 6 apresenta experimentos realizados com a aplicação do método. O Capítulo 7 analisa, avalia e sintetiza os resultados obtidos a partir destes experimentos. O Capítulo 8 descreve e discute alguns trabalhos correlatos, ou seja, outros esforços relacionados com a investigação do problema aqui proposto. Finalmente, no Capítulo 9 apresenta-se as considerações finais, avaliando-se a aplicação do método e sua adequação ao atendimento do problema proposto.

2 MODELAGEM ONTOLÓGICA DE DADOS CIENTÍFICOS

Apresenta-se, neste capítulo, os principais conceitos relacionados com a organização de dados e informação por meio de modelos ontológicos de representação semântica. Mais do que organizar dados e informações em uma representação semântica, o que se pretende com a pesquisa é definir um método capaz de integrar dados científicos provenientes de diversas fontes usando ontologias para isso. Biffl *et al.* (2013) discutem necessidades e soluções para o gerenciamento de dados visando a sua reutilização. Os autores avaliam abordagens ontológicas para integrar dados heterogêneos de estudos científicos.

A integração visa organizar dados de diferentes estudos para que possam ser analisados em conjunto. Por exemplo, dados de estudos sobre a taxa de mortalidade por um vírus, realizados de forma independente, podem ser integrados e harmonizados semanticamente, implicando em maiores chances de reutilização dos dados para avanços nas pesquisas sobre o tema e a sua avaliação a partir de amostras mais abrangentes.

As seções seguintes detalham os tópicos que serviram para elucidar as bases teóricas do presente trabalho.

2.1 Ciência de Dados e dados abertos

Ciência de Dados (*data science*) é uma área interdisciplinar voltada para o estudo e a análise de dados, estruturados ou não, que visa a extração de conhecimento ou *insigths* para possíveis tomadas de decisão. Trata-se de um campo que já existe há 30 anos, porém ganhou mais destaque nos últimos anos devido a alguns fatores como o surgimento e popularização do *Big Data* e o desenvolvimento de áreas como o *machine learning* (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

O uso de técnicas para obter informações dos dados melhora com a maior disponibilidade dos mesmos em suas diversas fontes. O acesso aberto a publicações e a abertura de dados científicos define o conceito de dados abertos, que colaboram para a reprodutibilidade das pesquisas, para o aumento a velocidade de divulgação e reúso dos dados, além da maior transparência do financiamento público. Segundo Woelfle, Olliaro e Todd (2011) estas são as características que definem a “Ciência Aberta”. Santos *et al.* (2017) em seu livro “Ciência Aberta e Dados Abertos” sistematizaram um mapeamento seguido de análise do processo de implantação da Ciência Aberta em oito países e na União Europeia, com ênfase nas políticas e na infraestrutura de dados abertos. Segundo os autores, os dados científicos abertos vão permitir uma ciência cada vez mais acessível.

Fox e Hendler (2009) afirmam que enquanto a imprensa levou mil anos para se desenvolver, o uso de computadores para entender os dados criados e armazenados provavelmente levará décadas ou menos. Os autores explicam este novo paradigma a partir de uma variedade de perspectivas disciplinares. Para eles, em alguns casos, a ciência está

atrasada em relação ao mundo comercial em inferir significado a partir de dados e agir com base nesse significado. Os autores acreditam que em breve veremos um momento em que os dados disponibilizados em alguma mídia de arquivamento, estarão publicamente acessíveis na "nuvem" para a sua exploração para humanos e máquinas.

2.1.1 Dados, metadados e proveniência

Segundo Floridi (2005), um metadado fornece indicações sobre a natureza de algum outro dado (usualmente primário). Estas indicações descrevem propriedades como a localização, o formato e outras propriedades adicionais dos dados originais. Um metadado pode ser, por exemplo, um indicador da codificação dos caracteres em um texto, neste caso, a falta do mesmo, poderia levar a uma interpretação errada de caracteres acentuados. Sayão e Sales (2013) ressaltam a importância dos metadados, os quais, segundo eles, cumprem um papel de ponte para o futuro nas estratégias de preservação de dados, além de auxiliar na validação da integridade e autenticidade dos dados digitais de pesquisa. Ainda segundo Sayão e Sales, a qualidade dos esquemas de metadados adotados e o rigor da sua aplicação favorecem a acessibilidade dos dados.

A proveniência de dados é representada por metadados que se referem ao seu histórico de derivação, a partir de suas fontes originais. Pesquisas podem gerar terabytes de dados, exigindo metadados descritivos que auferem sentido aos dados, facilitando o seu reúso. Dentre as diferentes formas de proveniência para vários propósitos, para os dados que registram resultados experimentais, o meio usual de se obtê-la são os artigos científicos (SIMMHAN; PLALE; GANNON, 2005). Entretanto, a utilização de identificadores do tipo Digital Object Identifiers (DOI) estão sendo cada vez mais utilizados como forma de se referenciar dados utilizados em experimentos.

Brase (2004), relata sobre o uso de metadados no registro de dados primários científicos. Tornar os dados referenciáveis, como trabalho independente e não apenas como parte de uma publicação, é uma questão relevante para as mais novas bibliotecas digitais. No contexto do projeto "Publicação e Citação de Dados Primários Científicos", fundado pela German Research Foundation (DFG), a biblioteca nacional alemã de ciência e tecnologia tornou-se a primeira agência de registro em todo o mundo para dados primários. Os conjuntos de dados recebem DOIs e Identificadores Uniformes de Recursos (URIs) exclusivos como identificadores citáveis e todas as informações relevantes de metadados são armazenadas na biblioteca.

2.1.2 Dados científicos

"Dados científicos" são aqueles que possuem grandeza escalar e contém dimensionalidade, tal como definido pelo *Common Data Format* (CDF)[1]. Um conjunto de dados científicos deve possuir dimensões, variáveis e atributos (REW; DAVIS, 1990). Outras definições do CDF, como a possibilidade de manipulação de arquivos superiores a 2Gb, não fazem parte da definição de dado científico para o presente trabalho.

[1] https://cdf.gsfc.nasa.gov/

Segundo Bowers (2012), os dados científicos possuem peculiaridades: (1) são coletados empiricamente por indivíduos ou instituições; (2) a estrutura que permite a sua observação é definida pelo método de coleta; (3) termos e conceitos usados para descrevê-los não são padronizados, variam conforme a área ou seus pesquisadores.

Os dados observacionais são adquiridos na forma de amostras e contextualizados pelos estudos. Portanto, um estudo contém várias amostras e um estudo pode fazer parte de outros estudos, ou seja, dados podem ser compartilhados entre estudos. Deste modo a possibilidade de reúso de dados tem importância para este compartilhamento. Quando estudos heterogêneos podem compartilhar dados comuns e quanto maior o nível de compartilhamento dos mesmos, estes estudos passam a ter seus dados "integrados".

Estudos possuem participantes, amostras e equipamentos. Os dados obtidos de equipamentos, por exemplo, um sensor, são em geral acompanhados de metadados que descrevem o equipamento e sua calibragem. Outros metadados podem adicionar conteúdo semântico aos dados científicos. Num estudo sobre uma dada doença, por exemplo, a informação de uma mudança no protocolo de notificação de casos da mesma pode estar ausente nos arquivos de dados, mas pode ser inserida em um metadado.

Um passo além destes metadados pode ser obtido com a conversão dos dados científicos de um estudo em uma ontologia, que não somente adiciona metadados como facilita a integração semântica dos mesmos. Conforme será descrito na Seção 2.2.2.2, na integração semântica os dados científicos são convertidos de sua representação original em uma representação capaz de representar "conhecimento", formalizada por ontologias que agregam informações adicionais na forma de metadados (MEEHAN *et al.*, 2017).

Os dados científicos, ao serem integrados semanticamente, não são apenas formalizados por ontologias, mas também enriquecidos com estes metadados adicionais (PAN *et al.*, 2017b). Além disso a formalização dos dados científicos em uma ontologia os torna acessíveis e legíveis por máquinas.

2.1.3 O ciclo de vida dos dados na pesquisa científica

O ciclo de vida dos dados, conforme o dataONE[2], é composto por etapas que permitem o seu gerenciamento e a preservação para utilização e re-utilização posterior. Apresenta-se a seguir uma adaptação deste ciclo para os fins do método definido na pesquisa.

- Planejar: descrever os dados que serão compilados e como eles serão gerenciados para torná-los acessíveis ao longo de seu tempo de vida.
- Aquisição de dados:
  - Coletar: fazer as observações manualmente ou através de sensores ou outros instrumentos colocando os dados em formato digital.
  - Assegurar: garantir a qualidade dos dados através de checagens e inspeções.

[2] https://www.dataone.org/data-life-cycle

- Preparação de dados:
  - Transformar: realizar operações para ajustar os dados, por exemplo, convertendo valores ou modificando a estrutura dos arquivos.
  - Descrever: associar aos dados metadados que documentem o que foi coletado.
  - Integrar: nesse momento os dados de fontes diferentes são combinados para formar um conjunto homogêneo de dados que pode ser analisado.
  - Preservar: é o processo de armazenar os dados em um arquivo de longa duração como um data-center.
  - Descobrir: nesta fase dados potencialmente úteis são localizados e recuperados, juntamente com os metadados.
- Analisar: Finalmente os dados são analisados e reinicia-se o ciclo.

2.1.4 Preparação de dados

A preparação de dados envolve a transformação, a descrição, a integração, a preservação e a descoberta de dados.

A etapa de transformação é necessária em diversas áreas, como no reconhecimento de padrões, na recuperação da informação, no aprendizado de máquina, na mineração de dados, na inteligência da Web, etc. Os dados brutos precisam ser formatados para o futuro processamento dos mesmos. Esta etapa pode demandar muito esforço, sendo essencial para o sucesso no tratamento da informação (ZHANG; ZHANG; YANG, 2003). Ela está relacionada com a fase de descrição das informações, onde são identificados os artefatos a serem preservados. Nesse momento pode ser constatada a necessidade de realizar transformações nos dados.

A descrição dos dados, pode ser realizada com a modelagem ontológica (Seção 2.2) que adiciona metadados aos dados de entrada. A integração de dados será detalhada na Seção 2.1.6. Já a preservação e a descoberta de dados são definidas de acordo com a infraestrutura utilizada para o armazenamento e gerenciamento de dados (Seção 3.4).

2.1.5 Princípios para o gerenciamento de dados

Como pôde ser compreendido até aqui, é amplamente aceito e reconhecido que a coleta, preparação e análise de dados são essenciais para o desenvolvimento da ciência em seu esforço empírico de evidenciação de proposições e juízos. Conforme relatado no Capítulo 1, a coleta, organização e disseminação de dados estão se tornando cada vez mais difíceis de realizar com eficiência. Deste modo, esforços no compartilhamento de dados por repositórios públicos ou privados, pelo pesquisadores, podem significar avanços.

Infelizmente, nem todas as iniciativas de compartilhamento de dados são bem-sucedidas. O gerenciamento adequado de dados é “o canal principal que leva à descoberta e à inovação do conhecimento”, promovendo o compartilhamento e a reutilização de dados em comunidades científicas conforme podemos ver em Wilkinson *et al.* (2016). Estes autores

ressaltam que é importante definir um conjunto de princípios comuns, que estabeleçam o que seria um "bom" gerenciamento de dados. Estes princípios, são Localização (*Findability*), Acessibilidade, Interoperabilidade e Reutilização (FAIR). Os princípios FAIR estão sendo cada vez mais citados como uma referência para todos os repositórios de compartilhamento de dados e são características utilizadas para destacar o sucesso de certas iniciativas. Por exemplo, o Immune Epitope Database, o DisGeNET Platform, o BioSharing Portal e o Omics Discovery Index possuem publicações que discutem sua adesão aos princípios do FAIR como forma de ilustrar seu compromisso de facilitar o compartilhamento de dados em suas respectivas comunidades (VITA *et al.*, 2018; PIÑERO *et al.*, 2016; MCQUILTON *et al.*, 2016; PEREZ-RIVEROL *et al.*, 2017).

Existem vários problemas que impedem que os dados estejam adequados aos princípios FAIR. Nem nem todos os conjuntos de dados são harmoniosos entre si, assim, apesar de intimamente relacionados, não são formatados da mesma maneira, dificultando a sua integração. Além disso, o pesquisador principal geralmente é quem mais sabe sobre os dados coletados, mas não consegue transmitir de forma coesa essa informação suplementar com os dados em si. Finalmente, se os dados forem suficientemente grandes, os métodos automatizados podem ser a única maneira de gerar análises abrangentes. Para viabilizar esses métodos automatizados, no entanto, é preciso que os dados sejam legíveis por máquina de alguma forma. É aqui que entra o conceito de integração de dados.

2.1.6 Integração de dados

A integração de dados combina processos técnicos e de negócios utilizados para interligar dados de fontes diferentes em informações significativas e valiosas[3]. Uma solução completa de integração de dados fornece dados confiáveis de uma variedade de fontes. Ela permite aos pesquisadores combinar dados existentes em diferentes locais. Em uma pesquisa sobre saúde, por exemplo, os dados combinados podem gerar novas hipóteses sobre a incidência de uma doença. Como uma estratégia, a integração é o primeiro passo no sentido de transformar dados em informação significativa e de valor.

No artigo "Integração de dados na era do Estudos relacionados ao genoma, as proteínas e ao metabolismo (omics): desafios atuais e futuros" (GOMEZ-CABRERO *et al.*, 2014), os autores ressaltam que a integração de dados é uma ferramenta muito utilizada na pesquisa em ciências da vida. Os autores relatam que, em 2006, 1.062 artigos mencionavam explicitamente "integração de dados" em seu resumo ou título, enquanto esse número mais que dobrou em 2013 (2.365). No entanto, ainda não existe uma definição unificada de integração de dados, nem taxonomia para metodologias de integração de dados, apesar de alguns esforços recentes sobre este tema (BAIROCH; COHEN-BOULAKIA; FROIDEVAUX, 2008; GOBLE; STEVENS, 2008; PHILIPPI, 2008; STEIN, 2002; COUNCIL *et al.*, 2010).

[3] https://www.ibm.com/analytics/data-integration

2.2 Abordagens de modelagem conceitual

Entender como os objetos de interesse, presentes em um estudo científico, “se manifestam” em um dado universo de discurso é uma preocupação central em muitas atividades científicas. Uma caracterização destes objetos e suas interações pode ser obtida pelo reconhecimento de suas relações. Parte significativa do conhecimento científico advém da análise de dados, para os quais os pesquisadores especificam relações mais diversas e granulares e a necessidade de profundidade investigativa. Uma descrição adequada dos dados (com metadados) pode contribuir para a sua análise. A descrição dos dados por metadados, além de proporcionar maior longevidade e exposição para os mesmos, facilita o reúso por outros experimentos, bem como a reprodutibilidade do experimento original.

Para além dos metadados, outro aspecto a ser considerado é a forma com que os dados são modelados conceitualmente, que pode ser feito de duas maneiras: (1) no processo tradicional de desenvolvimento de um sistema de informação, o modelo de dados (ou esquema de dados) é criado pelos analistas e faz parte da aplicação desenvolvida; cada dado inserido no sistema possui então o seu lugar na estrutura do banco de dados de forma pré-determinada pelo esquema criado previamente como parte do processo de análise e especificação do sistema de informação. Por outro lado, (2) no processo de modelagem ontológica, os esquemas serão definidos de acordo com os dados de entrada levantados pelos próprios pesquisadores do domínio; não há uma estrutura de metadados prévia a ser seguida pelo pesquisador, caso este pretenda publicar seus dados, dando-lhes maior visibilidade. Estas duas maneiras de modelagem de dados são apresentadas por West (2011), que as classifica de acordo com o tipo de abordagem utilizada pelo analista:

1. Abordagem por normalização - O analista procura por anomalias e padrões repetidos nos dados, eliminando-os com a criação de entidade adicionais, o que geralmente é chamado de normalização. Neste caso, temos a definição prévia de um modelo normalizado.

2. Abordagem ontológica - O analista utiliza as próprias entidades sobre as quais deseja estruturar os dados como base para a modelagem. Deste modo temos a estrutura das entidades definindo como será a estrutura do modelo sem a necessidade de um esquema pré-estabelecido.

A diferença entre a abordagem ontológica e a normalização para modelagem de dados, segundo West, está nas perguntas que são feitas ao se fazer a análise. Na abordagem da normalização, procura-se extrair grupos repetidos da estrutura subjacente. Já na abordagem ontológica, analisa-se os dados buscando responder o que eles representam, estruturando o modelo em torno disso.

Assim, no processo de modelagem conceitual proposto neste trabalho não são os dados que aderem à um modelo ou esquema prévio, mas o modelo é criado à demanda, conforme a estrutura de dados necessária à cada estudo em particular. Dessa forma, a partir de uma análise dos dados de interesse no domínio da pesquisa, estabelece-se o modelo

conceitual para organizar os dados (de um ou mais estudos) a serem ingeridos na base de dados.

Este processo é similar àquele de construção de uma ontologia e é realizado atribuindo-se classes, relações e indivíduos para representar cada um dos objetos (ou fenômenos) investigados. Isso significa que é possível modelar os dados de forma particular a cada estudo, atribuindo informações adicionais (metadados) que os tornem acessíveis a um rol mais amplo de análises no domínio científico da pesquisa. Além disso, dados estruturados por modelos conceituais diferentes podem ser integrados ao compartilharem características comuns.

A integração semântica de dados é realizada conforme a abordagem ontológica, onde os objetos do estudo, instanciados pelos dados coletados durante a realização do mesmo, são identificados e qualificados por suas classes, atributos e relações, independentemente de um modelo pré-estabelecido. Essa identificação dos objetos adiciona significado aos dados científicos. Assim, estes ganham conteúdo semântico que facilitam o reúso e a reprodução dos resultados do estudo por parte de outros pesquisadores.

2.2.1 Ontologias

O dicionário Caldas Aulete digital[4] apresenta quatro definições para Ontologia: (1) Parte da filosofia que estuda a natureza dos seres, o ser enquanto ser; (2) Doutrina sobre o ser; (3) Doutrina segundo a qual os fenômenos patológicos têm existência própria, não tendo relação com fenômenos fisiológicos; (4) Campo da informática que trata de conceitualizar de forma explícita e formal (portanto processável por máquina e compartilhável) conceitos e restrições relacionados a certo domínio de interesses. No contexto da ciência da informação é esta última acepção que nos interessa. Ao estabelecer os conceitos e relações a um dado domínio de interesse.

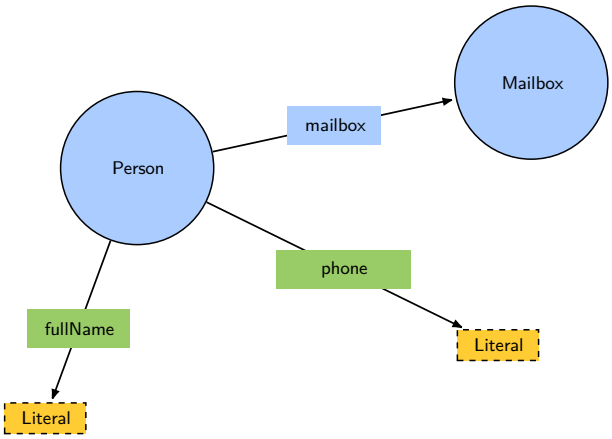
Uma ontologia de domínio representa conceitos que pertencem a um domínio de conhecimento específico, como medicina ou engenharia. À medida que os sistemas que dependem de ontologias de domínio se expandem, eles geralmente precisam interligar conceitos de diferentes ontologias. Deste modo, para a definição de uma ontologia, é necessária a participação dos especialistas de domínio. De fato, não se pode esperar uma ontologia de domínio de qualidade sem a atuação destes profissionais (KAIYA; SAEKI, 2006).

As ontologias representam dois tipos de informação, a definição de classes e propriedades como um vocabulário de domínio (TBox), ou seja, o modelo conceitual dos objetos e os valores associados ao modelo conceitual (ABox), cada TBox pode ter vários objetos ABox correspondentes. O TBox[5] é um "componente terminológico um conceito associado a um conjunto de fatos, conhecidos como ABox".

Os termos ABox e TBox são utilizados para descrever dois diferentes tipos de afirmações nas ontologias. As estruturas TBox descrevem classes e propriedades para cada elemento. Os ABox são ligados aos TBox fornecendo dados sobre os elementos que podem

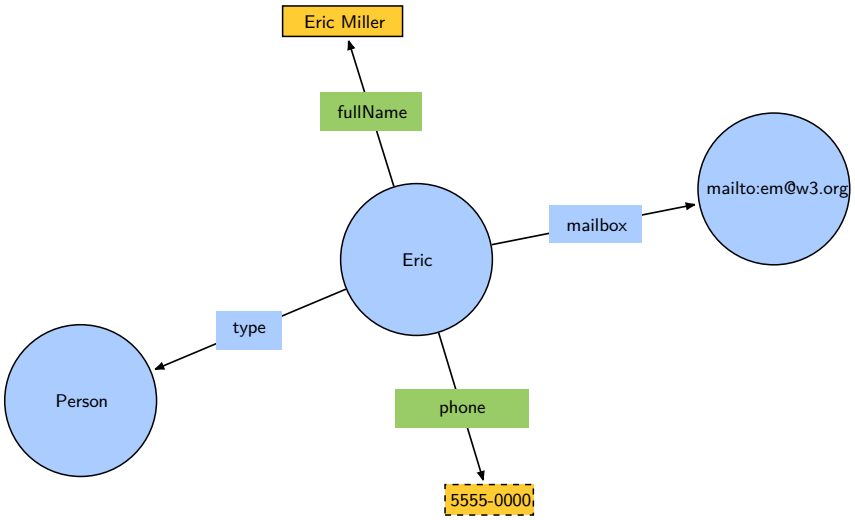
[4] http://www.aulete.com.br/ontologia
[5] https://en.wikipedia.org/wiki/Tbox

Figura 1 – TBox - Esquema para os dados



Fonte: Adaptada de (SHADBOLT; BERNERS-LEE; HALL, 2006)

Figura 2 – ABox - Dados armazenados conforme o esquema TBox



Fonte: Adaptada de (SHADBOLT; BERNERS-LEE; HALL, 2006)

ser representados pelos conceitos estabelecidos pelo TBox. o TBox de uma entidade “pessoa” (*Person*), por exemplo, pode definir que esta pode possuir atributos como, por exemplo, nome completo (*fullName*), telefone (*phone*) e e-mail (*mbox*), conforme mostra a Figura 1. A Figura 2 apresenta o ABox com uma instância específica desse TBox.

O ABox representa um conjunto de valores para cada entidade, ou seja, a relação modelada pelo TBox com uma determinada configuração de dados. Um TBox define um modelo para as instâncias ABox. Assim, no exemplo apresentado, diversas instâncias da entidade “pessoa” podem ser representadas conforme o modelo da Figura 1 (GRUBER, 1993).

O padrão que representa os elementos de uma ontologia em uma base de dados ou arquivo é o RDF. Ele é formado por triplas, formadas por sujeito, predicado e objeto, que podem ser representadas conforme ilustrado na Figura 1 e na Figura 2 (KLYNE; CARROLL,

2006).

Nesse modelo, temos um conjunto de triplas formando um grafo orientado com arestas e vértices rotulados, o grafo RDF. Os sujeitos são representados por um número finito de vértices, ligados através de um número finito de arestas rotuladas com os predicados a um conjunto finito de vértices representando os objetos. Além disso, qualquer vértice pode representar o papel de sujeito, de objeto ou de ambos. No entanto, a utilização de um grafo como forma de armazenamento ainda apresenta problemas de desempenho e escalabilidade (POKORNỲ, 2015). Maiores detalhes das regras podem ser vistos na página sobre os conceitos do RDF do W3C[6].

Qualquer um dos rótulos para sujeito, predicado ou objeto, pode ser representado por um Identificador Uniforme de Recursos (URI). Os rótulos identificados como URIs referenciam objetos que podem ter sido definidos localmente ou em ontologias externas. Deste modo, uma ontologia se torna parte da Web de dados ligados, se integrando a um conjunto de várias ontologias a partir dos URIs externos. O modelo RDF é estendido por uma série de outras linguagens e especificações como, por exemplo, a Ontology Web Language (OWL). Que é um *framework* para representar a informação, definindo uma série de regras para a representação de dados em triplas.

Um URI serve como uma chave que identifica um determinado recurso. Este recurso pode ser um Localizador Padrão de Recursos (URL), neste caso ele representará uma forma de acessar o recurso a partir da Internet. O URI deve ser único para cada recurso referenciado. Esta é a ideia por trás dos dados ligados, ou seja, temos vários identificadores na forma de URIs que levam a recursos de ontologias distintas. Por exemplo o URI de uma pessoa poderia ser: http://dbpedia.org/resource/Albert_Einstein. Como descrevemos, o URI não precisa necessariamente ser representada por um link existente na Web, podendo ser um número ou uma sequência de caracteres.

Alguns exemplos de URIs e dados brutos onde representamos os mesmos no formato *Turtle* podem ser vistos a seguir. O formato *Turtle*, resumidamente, consiste em um conjunto de triplas separadas por ‘;’ e ‘.’, com cada elemento da tripla sendo separado por espaços. Para o formato *Turtle*, as declarações das triplas podem ser encadeadas, ou seja, um conjunto de declarações podem se referir a um mesmo sujeito ao serem separadas por ‘;’. Neste caso, a primeira declaração deve ser uma formação completa de sujeito, predicado e objeto. As declarações seguintes utilizam o mesmo sujeito e definem apenas o predicado e o objeto. O encadeamento é terminado numa declaração finalizada com um ponto final ‘.’. Outros detalhes do formato *Turtle*, incluindo outras estruturas e regras, podem ser vistos na página da recomendação da linguagem do W3C[7].

```
<http://dbpedia.org/resource/Albert_Einstein> birthDate ‘‘14/03/1879’’;
    award <http://dbpedia.org/resource/Nobel_Prize_in_Physics>.
<http://dbpedia.org/resource/Richard_Dawkins> birthDate‘‘26/03/1941’’;
    award <http://dbpedia.org/resource/Michael_Faraday_Prize>.
```

[6] https://www.w3.org/TR/rdf-concepts/
[7] https://www.w3.org/TR/turtle/

No exemplo apresentado, são definidas quatro triplas. Nelas temos dois sujeitos http://dbpedia.org/resource/Albert_Einstein e http://dbpedia.org/resource/Richard_Dawkins, cada um deles têm dois predicados (*award* e *birthDate* que relacionam, respectivamente, um prêmio recebido e a data de nascimento de cada sujeito. Não há nenhuma limitação no número de vezes que um mesmo predicado seja utilizado por um determinado sujeito, deste modo, no exemplo acima, é possível adicionar outras premiações a um determinado sujeito.

No RDF é possível definir *namespaces* (espaços de nomes)[8], que seriam vocabulários específicos para designar uma URI completa de forma abreviada. Pode-se, por exemplo, estabelecer que "http://dbpedia.org/resource/" seja substituído por "dbp:", de forma que se tenha URIs abreviadas. Neste caso um URI como `dbp:Albert_Einstein` seria um sinônimo para o URI http://dbpedia.org/resource/Albert_Einstein.

2.2.1.1 O protocolo e linguagem de consulta SPARQL

Consultas a uma ontologia podem gerar respostas, no formato de tabelas. Por exemplo, considerando as triplas representadas no formato *Turtle* da seção anterior, podemos buscar a data de nascimento de Albert Einstein avaliando os seus valores. Para dados armazenados em RDF uma forma de consultar dados em uma ontologia é através do Protocolo e linguagem de consulta RDF SPARQL (SPARQL)[9]. Assim, se desejarmos buscar a data de nascimento de Albert Einstein a partir das triplas apresentadas anteriormente a consulta SPARQL poderia ser a seguinte:

```
PREFIX dbp: <http://dbpedia.org/resource/>
SELECT ?birth
    WHERE { dbp:Albert_Einstein birthDate ?birth .}
```

A primeira linha da consulta define o prefixo "dbp:" que identifica o espaço de nomes http://dbpedia.org/resource/. Em consultas mais elaboradas utiliza-se diversas declarações como esta, definindo um conjunto de espaços de nomes para indicar as ontologias utilizadas.

Pode-se notar que a cláusula *where* acima contém uma tripla no formato *Turtle*. Esta tripla indica a intenção de selecionar os valores que correspondam ao sujeito Albert Einstein, com o predicado *birthDate*. O campo " "?birth " define uma variável, que neste caso, é a nossa variável de saída e corresponderá a todos os valores que forem encontrados para o sujeito e predicado escolhidos.

A linguagem SPARQL conta com outros elementos que adicionam muitos outros recursos. A cláusula *where*, por exemplo, pode conter várias triplas e utilizar operadores especiais, que propiciam maior poder e flexibilidade às consultas. Além disso, podem ser agrupadas informações de outras ontologias. É possível realizar consultas bem complexas com a linguagem SPARQL.

[8] https://www.w3.org/TR/1999/REC-xml-names-19990114/#NT-QName
[9] https://www.w3.org/TR/rdf-sparql-query/

2.2.1.2 Grafos de conhecimento

O termo “grafo de conhecimento” tem sido utilizado com frequência em pesquisa e negócios, geralmente em estreita associação com tecnologias da Web semântica, dados ligados, dados em grande escala, análise e computação em nuvem. Sua popularidade foi influenciada pela introdução do *Knowledge Graph* do Google em 2012, desde então, o termo tem sido utilizado sem uma definição precisa. Uma grande variedade de interpretações tem dificultado a evolução de um entendimento comum do termo. Diversos trabalhos de pesquisa se referem ao *Knowledge Graph* do Google (EDER, 2012), embora não haja documentação oficial sobre os métodos usados.

O pré-requisito para a adoção generalizada do conceito, tanto acadêmica como comercialmente, é um entendimento comum do mesmo (EHRLINGER; WÖSS, 2016). Entretanto, segundo McCusker *et al.* (2018), atualmente este conceito pode se referir a uma ampla gama de grafos os quais podem conter referências ambíguas e não precisas. No trabalho desenvolvido por McCusker *et al.*, um grafo de conhecimento é definido como “Um grafo, composto por um conjunto de asserções (arestas marcadas com relações) que são expressas entre entidades (vértices), onde o significado do grafo é codificado em sua estrutura, as relações e entidades são inequivocamente identificadas, um conjunto limitado de relações é usado para rotular as arestas, e o grafo codifica a proveniência, justificativa e atribuição das asserções (MCCUSKER *et al.*, 2018).”

Um grafo de conhecimento, de acordo com a definição acima, precisa cumprir certos requisitos como, por exemplo, a proveniência. No entanto, uma ontologia pode ser considerada um grafo de conhecimento sem a necessidade das restrições estabelecidas por McCusker *et al.*. Assim, Färber *et al.* (2016) e Pan *et al.* (2017a) consideram que uma ontologia é um grafo de conhecimento se possui um volume de dados compatível com o *big data* e um conjunto de algoritmos para busca e gestão destes dados.

Deste modo, de acordo com estes autores, se a ontologia fornecer recursos adicionais através de artefatos que permitam pesquisar e gerenciar a informação do grafo e a infraestrutura tiver escalabilidade suficiente para manipular grafos com um número de vértices e arestas considerável, esta ontologia é um grafo de conhecimento.

O método de integração semântica de dados proposto no Capítulo 5 cria uma ontologia, que neste trabalho é armazenada e gerenciada com a utilização da infraestrutura de um *framework* que fornece os recursos para esta seja considerada um grafo de conhecimento de acordo com esta definição.

2.2.1.3 Especificando ontologias com questões de competência

A criação de uma ontologia ocorre em torno de requisitos de uso, os quais podem ser determinados através de questões de competência. Segundo Bezerra, Freitas e Santana (2013), as questões de competência consistem em um conjunto de perguntas declaradas em linguagem natural, de modo que a ontologia possa respondê-las corretamente. Elas são uma forma de levantar os requisitos da ontologia. Dado um conjunto de cenários relacionados a um

domínio do discurso, os desenvolvedores elaboram um conjunto de perguntas que representem as demandas do usuário e limitem seu escopo. Essas questões apoiam o processo de desenvolvimento de duas maneiras:

- Auxiliam na identificação dos principais elementos e seus relacionamentos para criar o vocabulário de ontologia (terminologia) e;
- Fornecem um meio simples de verificar a satisfatibilidade dos requisitos.

Além disso, definir as questões de competência é uma maneira de priorização das tarefas, pois estas podem ser listadas em ordem de importância (AZZAOUI *et al.*, 2013). Isto é particularmente útil no desenvolvimento de ontologias, onde os pesquisadores contam com diversas opções de criação, ao definir as principais questões. No método proposto, essa priorização servirá para definir os objetivos a serem atendidos em cada iteração, conforme será descrito no Capítulo 5.

As questões de competência tornam mais fácil a participação dos especialistas de domínio, os quais normalmente não estão familiarizados com as linguagens de criação tais como a RDF. Após a sua implementação em uma ontologia, estas podem ser testadas automaticamente com a utilização de consultas ao grafo, o que pode ser feito, por exemplo, com a utilização da SPARQL (REN *et al.*, 2014).

A criação de uma ontologia a partir de questões de competência pode ser feita a partir de um *dataset*, normalmente em um formato tabular. A ontologia a ser criada, neste caso, consiste em um mapeamento do *dataset* em um grafo RDF. Existem diversas propostas para tal mapeamento, a seguir serão apresentadas algumas delas, destacando-se características que distinguem as abordagens utilizadas.

2.2.2 Mapeamento de dados em grafos RDF

No mapeamento de dados em grafos RDF, um conjunto de dados e de uma série de regras de mapeamento são utilizados para gerar o grafo RDF. Este mapeamento é normalmente realizado a partir de bancos de dados relacionais, no entanto é possível gerar o grafo a partir de qualquer formato de dados. No presente trabalho, utilizou-se arquivos de texto com valores separados por vírgula (CSV). Algumas implicações da utilização deste formato de arquivo de dados são apresentadas a seguir.

2.2.2.1 Utilização de arquivos CSV

Arquivos com valores separados por vírgula (CSV), representam um formato de dados amplamente utilizado na ciência e na indústria. Um arquivo CSV é composto por um certo número de linhas e um número fixo de colunas. Normalmente, cada registro é codificado como uma única linha no arquivo, com exceção da primeira linha que pode conter cabeçalhos de coluna. Em alguns casos podemos ter uma organização diferente, por exemplo, com mais de um registro representado em colunas da mesma linha do arquivo. Dependendo da configuração do arquivo CSV, podem ser necessárias transformações em sua estrutura.

Figura 3 – Exemplo de dados tabulares que podem ser representados no formato CSV

Id	Taxa dengue	Temperatura
20001102	35	29
20011132	31	27
32323410	19	24

Fonte: Elaborada pelo autor

Supondo que um arquivo CSV contenha atributos de uma única entidade, sua estrutura poderia ser mapeada em uma tabela que representa a entidade, como em um modelo relacional. Neste caso, observamos que cada linha poderá corresponder a um registro da tabela e cada cabeçalho de coluna corresponderá a uma propriedade (ou atributo) da entidade. Estas mesmas suposições não valem quando uma linha representa atributos de mais de uma entidade, ou quando algumas das propriedades são metadados.

A utilização de arquivos CSV, inclui dois desafios: **construção de um modelo para os dados** (que pode ser reutilizado para arquivos de dados futuros que seguem o mesmo esquema) e **mapeamento dos dados para o modelo proposto**. Esses desafios se evidenciam quando se descobre que os cabeçalhos de coluna nas tabelas de dados de origem não revelam com precisão o significado dos dados alí presentes. Um único registro em um conjunto de dados pode conter dados que descrevem várias entidades, por exemplo, taxa anual de casos de dengue na população e temperatura média anual, para um município, conforme ilustrado na Figura 3.

Além disso, ao contrário do que ocorre com dados de tabelas em um banco de dados, esses arquivos não definem os tipos de dados e outras restrições de integridade, fazendo com que um aquivo CSV possa ser criado com dados inválidos. Explorando ainda mais arquivos CSV, observamos alguns problemas comuns relacionados à interpretação dos dados:

- Devido à falta de restrições, já que os arquivos são texto livre, podemos ter, por exemplo, valores textuais inseridos em campos numéricos;
- Registros com dados de vários objetos na mesma linha podem gerar dúvidas sobre que coluna qualifica que objeto;
- Registros podem conter dados e metadados, sem qualquer evidência explícita que permita a sua diferenciação.

Apesar das limitações apresentadas, os arquivos formatados como CSV podem ser convertidos em grafos RDF utilizando os métodos da seção seguinte.

2.2.2.2 Métodos para conversão de dados em grafos RDF

Sequeda (2013) apresentou um trabalho com resultados iniciais de um estudo formal da RDB2RDF *Mapping-Language* (R2RML) com a definição de sua semântica utilizando

o Datalog, uma linguagem de consulta não procedural. A R2RML é uma recomendação do W3C[10] para conversão de dados de bancos relacionais para RDF. Sequeda formaliza a R2RML, estabelecendo que a entrada de um mapeamento *M* é um esquema de dados relacional *R* e uma instância *I* de *R* e que a saída é o grafo *G*, ou seja, *G* é uma função de *M*,*I* e *R*. Assim, o grafo RDF gerado depende do mapeamento, dos dados de entrada e do esquema dos dados.

Para a conversão de dados para RDF, é necessário considerar pelo menos três métodos: (1) "mapeamento direto", (2) "*lifting* de dados"[11] (ou enriquecimento semântico) e (3) "integração semântica de dados" que inclui, além da descrição dos dados dentro da estrutura de uma atividade científica, a "harmonização" desses dados com dados de outras atividades científicas. Podemos afirmar que apenas o terceiro método, que inclui a harmonização de dados, abarca todos os requisitos necessários para a integração de dados de estudos científicos. A seguir serão detalhados cada um destes métodos:

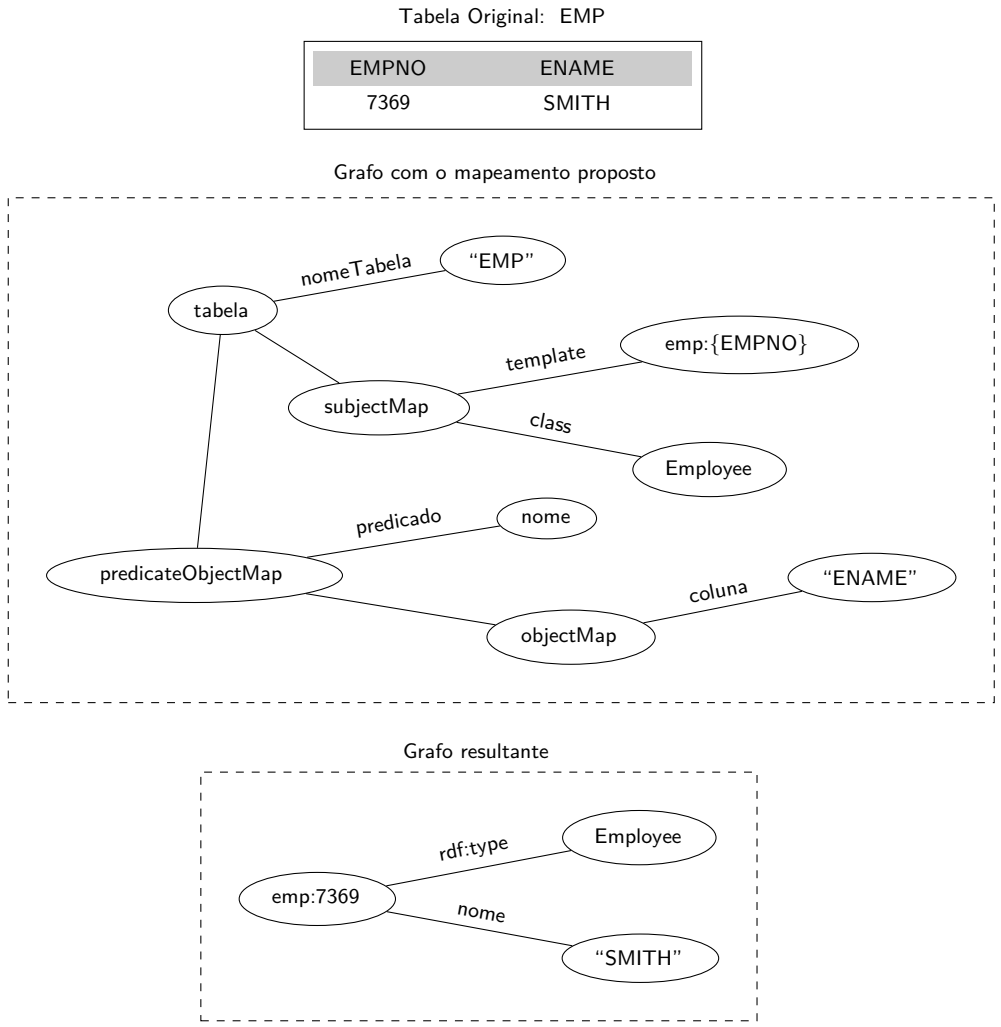
- Mapeamento Direto: de acordo com Pan *et al.* (2017b), o mapeamento direto é uma maneira eficiente de obter uma rápida conversão de dados tabulares para RDF. No entanto, em muitos cenários do mundo real, o mapeamento direto pode não ser suficiente. Um exemplo desta abordagem é apresentado na Figura 4 e utiliza a R2RML como forma de mapeamento. No exemplo da figura temos uma tabela simples, com duas colunas "EMPNO" e "ENAME", onde a primeira é o identificador do registro e a segunda o nome.

  Podemos ver na Figura 4 um grafo do mapeamento proposto, este mapeamento tem uma proposição para o mapeamento dos sujeitos ("subjectMap") e outra para o mapeamento dos predicados e objetos ("predicateObjectMap"). O mapeamento é relativamente simples e trata da conversão de uma tabela em um sujeito, representado pelo seu identificador e seus predicados, formados pelas demais colunas da tabela. Além disso podemos ver que possíveis mapeamentos em ontologias podem ser feitos nesse processo. Nesse caso, por exemplo, o predicado *rdf:type* foi utilizado para identificar o tipo do objeto criado. Apesar do exemplo se basear em um banco de dados relacional, este tipo de conversão pode ser aplicado a diversos tipos de entrada.

- *Lifiting* de Dados: o *lifiting* de dados não é apenas uma transformação de um formato para outro (tabular para RDF), mas pode ser considerado "uma elevação de informações do nível de dados para o nível do conhecimento *legível por máquina*" (PAN *et al.*, 2017b). Neste caso são agregadas informações de ontologias científicas que permitam identificar e relacionar os dados às mesmas. A motivação é de que usar uma ontologia científica de domínio específico melhora a visibilidade da base de conhecimento do estudo e facilita a integração e o reúso dos dados, ou a sua vinculação com dados de outros estudos. Portanto, trata-se não somente em converter dados de um estudo para um grafo RDF, mas também de adicionar ao grafo ligações à ontologias conhecidas na literatura especializada, tais como a *Semanticscience Integrated Ontology* (SIO) (DUMONTIER *et al.*, 2014), a Ontologia de Observação Extensível (OBOE) (MADIN *et al.*, 2007), e outras.

[10] https://www.w3.org/TR/r2rml/
[11] Operação que incorpora informações de ontologias aos dados.

Figura 4 – Mapeamento direto de dados em RDF



Fonte: elaborada pelo autor

- Integração Semântica de Dados: a integração semântica de dados, além de agregar informações de ontologias de domínio, requer uma conversão personalizada que as abordagens descritas anteriormente não podem fornecer (PAN *et al.*, 2017b; PINHEIRO *et al.*, 2018b). Integrar e comparar dados de diferentes estudos exige que os valores assumidos por uma variável sejam comensuráveis. Neste caso, cada estudo pode ter escalas ou propriedades diferentes, o que exige a harmonização dos dados dos estudos.

  A integração Semântica de Dados gera dados representados em RDF, de uma forma integrada e harmonizada, permitindo que as informações sejam identificadas, desambiguadas e interconectadas por vários sistemas para ler, analisar e agir. Uma solução completa deve fornecer ainda uma interface para consulta, a qualdeve ser amigável e com os valores de dados harmonizados entre os estudos. Neste método é necessário um tratamento mais detalhado dos dados, o que exige um processo de conversão mais complexo do que os anteriores.

O presente trabalho propõe a realização da integração semântica de dados científicos representados em formato CSV, o método de integração será descrito no Capítulo 5 sendo realizado durante a etapa de preparação de dados do ciclo de vida da Informação.

3 ONTOLOGIAS E PLATAFORMA DE SOFTWARE USADAS

Descrevem-se aqui as ontologias que fundamentam a integração semântica de dados. Estas ontologias enriquecem os dados com metadados, gerando um grafo com as informações científicas contextualizadas. Será descrito o *framework* HADatAc, utilizado como ferramenta de ingestão de dados e gerenciamento do grafo. De acordo com Meehan *et al.*, a ingestão de dados é o processo que converte os dados originais, armazenando-os no sistema de banco de dados destino da maneira mais eficiente e correta possível.

Ao longo do texto, foi adotado o seguinte padrão: as propriedades ou predicados dos objetos, como por exemplo *sio:Age*, foram escritas em itálico. Outros termos, como sio:Human, são classes de ontologias e foram destacados na fonte *Teletype*.

A primeira ontologia que será descrita é a *Human-Aware Science Ontology* (HAScO), que formaliza as definições específicas necessárias para a codificação de metadados que descrevem dados gerados no âmbito de estudos científicos.

3.1 A ontologia HAScO

Os processos de aquisição, preparação e organização de dados são centrais para o avanço científico. A HAScO (PINHEIRO *et al.*, 2018a) foi desenhada com o propósito específico de codificar metadados de dados científicos. Ela foi concebida especificamente para codificar metadados relevantes nas fases mencionadas. Fases do ciclo da Ciência de Dados, que culmina com a análise estatística por meio da aplicação de diversos algoritmos de análise de dados.

A HAScO reutiliza ontologias existentes, já testadas e aprovadas pela comunidade. Esta ontologia pode ser utilizada para modelar experimentos em domínios diversos, para anotação de dados, para enriquecer semanticamente consultas aos dados e para produzir visões orientadas a dados para grupos específicos de analistas/pesquisadores.

Um estudo científico empírico precisa ter alguns objetivos específicos, um plano bem definido de coleta, seleção, limpeza, organização e preservação dos dados; além de muitos outros componentes, como um líder, uma fonte de financiamento, entre outros.

A descrição detalhada dos objetos relacionados aos estudos adiciona informações úteis, tais como a descrição de sujeitos/participantes, amostras, locais de amostragem, períodos de amostragem e unidades de medida utilizadas para a coleta e representação dos dados.

Na medição de valores fisiológicos, por exemplo, do nível de vitamina D no sangue, uma única amostra temporal representa a informação necessária para definir se o sujeito tem ou não deficiência nesta vitamina. Em outros casos, o nível de inter-relações entre objetos de estudo é tão complexo que as relações precisam ser explicitamente descritas. Por exemplo, um estudo pode envolver dados biométricos e amostras de sangue extraídos da mãe e do filho. Nesse caso, é essencial que a atividade de aquisição de dados descreva como mães, filhos,

amostras de mães e amostras de crianças estão relacionadas.

O mesmo tipo de relacionamento complexo pode ocorrer com relação a dependências temporais e espaciais entre os dados/objetos dos estudos.

A HAScO foi originalmente criada para anotar dados de origem ambiental captados através de sensores (MCGUINNESS *et al.*, 2014). A utilização da HAScO em projetos de saúde humana introduziu a necessidade de registrar dados de humanos (sujeitos) e a ontologia evoluiu para anotar dados epidemiológicos de questionários sobre os seres humanos. Assim, os questionários também foram considerados instrumentos de captura de dados. Os dados passaram a envolver a medição de indicadores fisiológicos, como pressão arterial ou frequência cardíaca, bem como a elicitação de alguns atributos qualitativos, como hábitos tabágicos.

Com a utilização de amostras e hábitos de seres humanos, surgiu a necessidade de conectar os dados dos sujeitos pesquisados às amostras e ao conteúdo dos questionários relacionados, assim, a ontologia passou a contemplar recursos para atender a essa necessidade.

Em síntese, a ontologia HAScO procura atender aos seguintes requisitos da atividade de anotação de dados (PINHEIRO *et al.*, 2018a):

- Permitir o tratamento de dados resultantes da realização de diferentes tipos de estudo, tais como observações ou experimentos.
- Identificar os objetos do estudo e suas inter-relações.
- Gerenciar a qualidade dos dados do estudo.
- Fornecer suporte à descrição temporal e espacial dos dados.
- Permitir a utilização de diversos tipos de fontes de dados como questionários, documentos, simulações, sensores e dados de laboratório.
- Identificar a proveniência dos dados de acordo com a fonte de origem.
- Representar metadados de atividades de aquisição de dados, suportando metadados de estudo e metadados da infra-estrutura de detecção.
- Apoiar a organização e integração de dados científicos heterogêneos e diversificados.
- Permitir que os ontologistas a estendam, agrupando livremente os conceitos para suportar visualizações facetadas que não precisam ser baseadas nas classificações lógicas existentes dentro dela ou em qualquer outra ontologia importada ou referenciada pela mesma.

3.2 Ontologias científicas de suporte à descrição de dados

Para cumprir os requisitos estabelecidos para a HAScO, foi selecionado um conjunto de ontologias fundacionais (ontologias de topo) apropriado para uso na modelagem de dados

científicos. Uma visão geral destas ontologias de suporte será apresentada nas seções seguintes. Essas ontologias foram alinhadas com a HAScO, fornecendo um vocabulário comum de alto nível de abstração e genérico o suficiente para seu uso em vários tipos de estudos científicos. Utilizar essas entidades, atributos, unidades de medida e escalas comuns entre os estudos, tem permitido uma efetiva integração dos dados, conforme atesta a literatura científica da área (MCGUINNESS *et al.*, 2009; RASHID *et al.*, 2017; DUMONTIER *et al.*, 2014).

As ontologias de suporte reutilizadas pela HAScO são:

- *Semanticscience Integrated Ontology* (SIO): A anotação semântica dos conceitos científicos na HAScO é baseada nesta ontologia que define os tipos e relações usados atualmente para objetos, atributos, processos e tempo (DUMONTIER *et al.*, 2014). Ela fornece a estrutura integrada a partir da qual a HAScO está enraizada. O uso da SIO em conjunto com ontologias de domínio permite que os cientistas caracterizem o conjunto de entidades e atributos que são objetos de estudo em domínios científicos mais específicos.
- VSTO-I ("*Virtual Solar-Terrestrial Ontology - Instrument model*"): Uma ontologia que contém conceitos que descrevem entidades que coletam dados, como sensores (FOX *et al.*, 2009).
- UO ("*Units Ontology* "): Representa unidades de medida do sistema internacional (GKOUTOS; SCHOFIELD; HOEHNDORF, 2012).
- PROV ("*Provenance Ontology* "): Fornece um conjunto de classes, propriedades e restrições que podem ser usadas para representar e intercambiar informações de proveniência geradas em diferentes sistemas e sob diferentes contextos (LEBO *et al.*, 2013).

As ontologias de suporte serão detalhadas nas seções seguintes.

3.2.1 Semanticscience Integrated Ontology

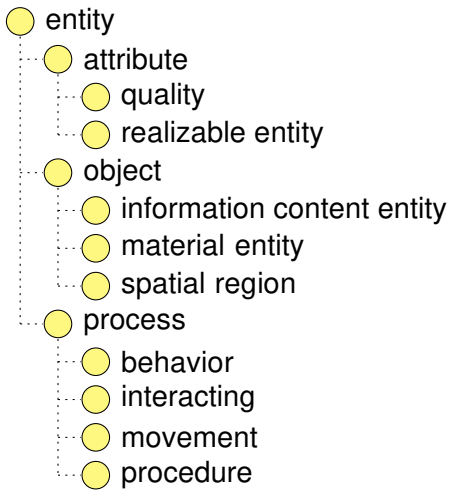
A SIO[1] foi desenvolvida com o objetivo de facilitar a descoberta do conhecimento pela integração de dados baseados em ontologias (DUMONTIER *et al.*, 2014). Esta ontologia e a HAScO (Seção 3) são as principais ontologias utilizadas no processo de ingestão utilizado nesta pesquisa.

A SIO oferece classes e relações para descrever e relacionar objetos, processos e seus atributos com extensões científicas com foco no domínio biomédico, porém não limitado a ele. Suas relações cobrem aspectos do raciocínio qualitativo espacial e temporal, incluindo a localização, a contenção, a sobreposição e topologia; participação e agência, linguística e representação simbólica, bem como comparações e outras relações orientadas à informação.

Embora o desenvolvimento da SIO tenha sido motivado por necessidades do domínio biomédico, ela pode ser aplicada a um conjunto mais amplo de domínios. Ela é uma ontologia de tipos básicos e relações e pode ser utilizada para capturar uma ampla extensão do conhecimento através de um conjunto de domínios emergentes.

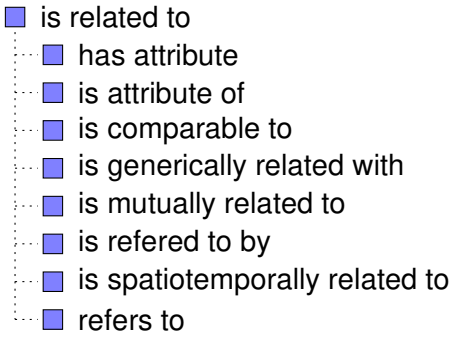
[1] http://sio.semanticscience.org

Figura 5 – Nível principal da hierarquia de classes na SIO.



Fonte: Adaptada de (DUMONTIER *et al.*, 2014)

Figura 6 – Nível principal da hierarquia de propriedades de objetos na SIO.



Fonte: Adaptada de (DUMONTIER *et al.*, 2014)

Na SIO os rótulos em inglês são fornecidos pelo uso da propriedade *rdfs:label* quando forem legíveis, definições na língua inglesa são fornecidas utilizando o termo de metadados Dublin Core (dc:) *dc:description*. Nas Figuras 5 e 6 temos uma amostra da hierarquia de classes e da hierarquia das propriedades de objeto em que'entity' é a classe de nível superior e *'is related to'* é a propriedade de objeto de nível superior. Essa ontologia adere a uma visão de mundo quadridimensional que é familiar para a maioria dos cientistas, distinguindo entre os processos e os objetos que participam deles. Nela, um 'objeto' é uma 'entity' que ocupa espaço e é totalmente identificável por suas características a qualquer momento em que existe. Por outro lado, um 'processo' é uma'entity' que se desdobra no tempo e tem partes temporais. Enquanto uma entidade 'Está localizada em' e 'existe em' algum espaço e tempo, que não precisam ser espaço real ou em tempo real, mas podem ocorrer em um ambiente hipotético (proposicional), virtual (eletrônico), ou configuração fictícia (trabalho criativo). Uma 'qualidade' (atributo intrínseco), 'capacidade' (especificação de ação) ou 'papel' (comportamento, direito e obrigação) podem existir em alguma entidade que a suporta, mas 'é

realizada em' um processo no qual ela desempenha um papel crítico. O valor de uma entidade informativa, como um 'valor de medição' ('quantidade' ou 'posição') é representado utilizando a propriedade *'has value'*.

A *Semanticscience Integrated Ontology* está disponível livremente sob a licença *Creative Commons*[2]. As entidades SIO são identificadas utilizando URIs resolvíveis, inicialmente formuladas como um identificador alfanumérico, por exemplo, sio:SIO_000001, mas é alternativamente acessível utilizando um identificador baseado em rótulos, por exemplo, *sio:is-related-to*. Estes e outros subconjuntos gerados estão disponíveis na Internet[3].

3.2.2 *Virtual Solar-Terrestrial Ontology - Instrument model*

A *Virtual Solar-Terrestrial Ontology - Instrument model* (VSTO-I) (FOX *et al.*, 2009) é uma ontologia que contém conceitos que descrevem entidades capazes de coletar dados (por exemplo, instrumentos, detectores e plataformas) e atividades relacionadas a essas entidades, tais como a implantação de um instrumento em uma plataforma.

O desenvolvimento da ontologia VSTO-I foi conduzido pelo Observatório de Alta Altitude do Centro Nacional de Pesquisa Atmosférica com a colaboração da McGuinness Associates. Ela tem sido utilizada e refinada por várias organizações, incluindo a colaboração com o BCODMO do Instituto Oceanográfico Woods Hole. Algumas das classes da VSTO-I utilizadas pela HAScO serão vistas na Seção 3.3.

3.2.3 Units Ontology

A *Units Ontology* (UO) é uma ontologia que é utilizada atualmente em muitos recursos científicos para a descrição padronizada de unidades de medida. (GKOUTOS; SCHOFIELD; HOEHNDORF, 2012).

A versão inicial da UO foi desenvolvida manualmente utilizando o editor da *Open Biomedical Ontology* (OBO) (DAY-RICHTER *et al.*, 2007). A UO foi refinada e populada com uma combinação de unidades de pesquisa baseadas em: anotações de medições existentes, ensaios, comunicação pessoal com usuários e com o o conhecimento de domínio dos desenvolvedores da ontologia. Essa ontologia contém definições textuais para todos os seus termos e, sempre que possível, fornece links para a fonte da definição.

A UO é mantida em um repositório do *subversion* e é disponibilizada através do *OBO registry* e do site do projeto[4]. Além disso, possui um rastreador de solicitações de termos[5] e uma lista de discussão[6] onde os usuários sugerem alterações e solicitam novos recursos. Ela está disponível no formato OBO e na linguagem OWL (GRAU *et al.*, 2008).

Seu desenvolvimento seguiu os princípios da OBO (SMITH *et al.*, 2007), o que a torna parte do conjunto de ontologias da OBO. Ela tem sido amplamente adotada na comunidade

[2] http://semanticscience.org/ontology/sio.owl
[3] http://goo.gl/0LgN8
[4] http://unit-ontology.googlecode.com
[5] http://code.google.com/p/unit-ontology/issues/list
[6] https://lists.sourceforge.net/lists/listinfo/obo-unit

biomédica por um grande número de ontologias, linguagens de marcação, bancos de dados, iniciativas de padrões, projetos de pesquisa e aplicações, desempenhando um papel central no fornecimento de acesso padronizado a dados biomédicos. A UO forma uma estrutura que facilita a padronização e formalização de unidades e é crucial para o intercâmbio, processamento e integração de dados quantitativos

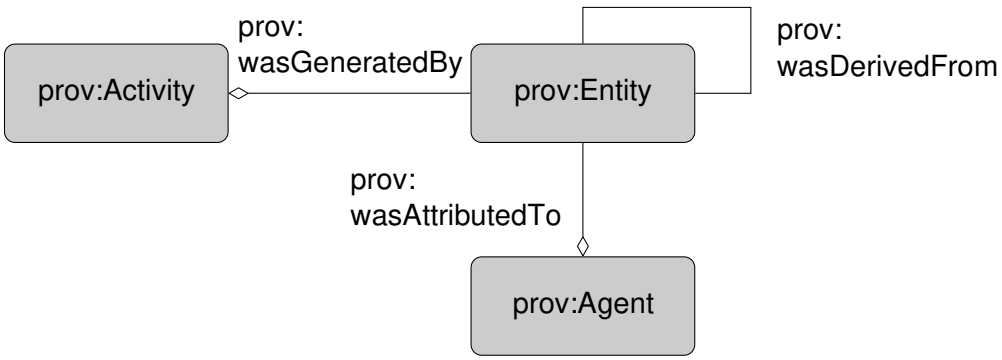
3.2.4 Provenance Ontology

A *Provenance Ontology* (PROV-O) é uma ontologia criada utilizando a OWL (GRAU *et al.*, 2008). Ela fornece um conjunto de classes, propriedades e restrições que podem ser utilizadas para representar e intercambiar informações de proveniência geradas em diferentes sistemas e sob diferentes contextos. Também pode ser especializada para criar novas classes e propriedades para modelar informações de proveniência para diferentes aplicativos e domínios.

A Proveniência é definida a partir de informações sobre entidades, atividades e pessoas envolvidas na produção de um dado ou artefato. Ela pode ser utilizada para formar avaliações sobre a qualidade, confiabilidade e autenticidade. A PROV define um modelo, serializações correspondentes e outras definições de suporte para o intercâmbio de informações de proveniência em ambientes heterogêneos, como a Web. As principais classes da PROV-O são:

- prov:Entity é um tipo físico, digital, conceitual ou qualquer outro artefato com alguns aspectos fixos; podem ser reais ou imaginárias.
- prov:Activity ocorre durante um período de tempo e age sobre ou com umprov:Entity; pode incluir o consumo, processamento, transformação, modificação, realocação, uso ou geração de um prov:Entity .
- prov:Agenté responsável de alguma forma por uma prov:Activity , pela existência de uma prov:Activity ou pela atividade de outro prov:Agent

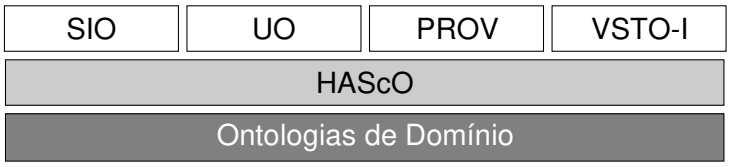
Figura 7 – Diagrama com algumas das classes da PROV-O



Fonte: elaborado pelo autor

A Figura 7 ilustra as entidades apresentadas anteriormente com algumas das ações possíveis para as mesmas, as quais são:

Figura 8 – A HAScO e suas ontologias de suporte



Fonte: (PINHEIRO *et al.*, 2018a)

- *prov:wasGeneratedBy*: Define que uma `prov:Entity` foi produzida por uma `prov:Activity`.
- *prov:wasDerivedFrom*: Define que uma `prov:Entity` foi derivada de outra. Pode ser uma atualização de uma `prov:Entity` resultando em uma nova ou a construção de uma nova `prov:Entity` baseada em uma outra pré-existente.
- *prov:wasAttributedTo*: Indica que uma determinada `prov:Entity` foi atribuída a um `prov:Agent` A proveniência de uma `prov:Entity` é estabelecida a partir de relações como esta que ligam a mesma a um `prov:Agent`

3.3 Implementação dos principais conceitos da HAScO

A Figura 8 mostra a ontologia HAScO e suas ontologias de suporte, representando ainda a utilização de ontologias específicas para o domínio. Os principais conceitos da HAScO estão organizados em três categorias: atividades científicas, instrumentos para aquisição de dados e organização de dados, sendo esta última subdividida em objetos de estudo e esquema de dados.

3.3.1 Atividades científicas

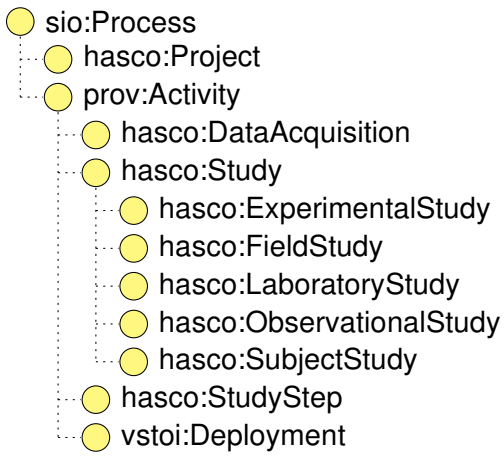
A HAScO usa a visão de que "a ciência é um conhecimento organizado" a partir de um conjunto de eventos. Um dos seus objetivos é a identificação e categorização desses eventos, vistos como atividades científicas, além de aprimorar a representação dos mesmos e suas interdependências. Facilitando as consultas e a integração entre eventos inter-relacionados.

Temos três atividades científicas essenciais definidas na HAScO: `Study`, `DataAcquisition` e `Deployment`. A Figura 9 apresenta a hierarquia de classes que inclui estas três atividades. Conforme ilustra a figura, essas atividades são subclasses da `prov:Activity` da W3C PROV[7].

`Study`: Na HAScO, um estudo pode ser especializado em cinco categorias: `ExperimentalStudy`, `FieldStudy`, `LaboratoryStudy`, `ObservationalStudy` e `SubjectStudy`. Cada estudo pode ser composto de vários passos (`StudyStep`). A ontologia HAScO fornece uma classificação de alto nível dos estudos que pode ser expandida se necessário.

[7] https://www.w3.org/TR/prov-o/

Figura 9 – Parte da hierarquia de classes da HAScO e ontologias de suporte



Fonte: Adaptada de (PINHEIRO *et al.*, 2018a)

Os estudos são classificados como observacionais quando nenhuma variável no estudo é controlada e experimentais quando pelo menos uma variável é controlada. Nas observações, os conjuntos de dados anotados pela HAScO podem ser representados como uma única aquisição de dados se nenhum controle, como calibração de instrumentos, é levado em consideração.

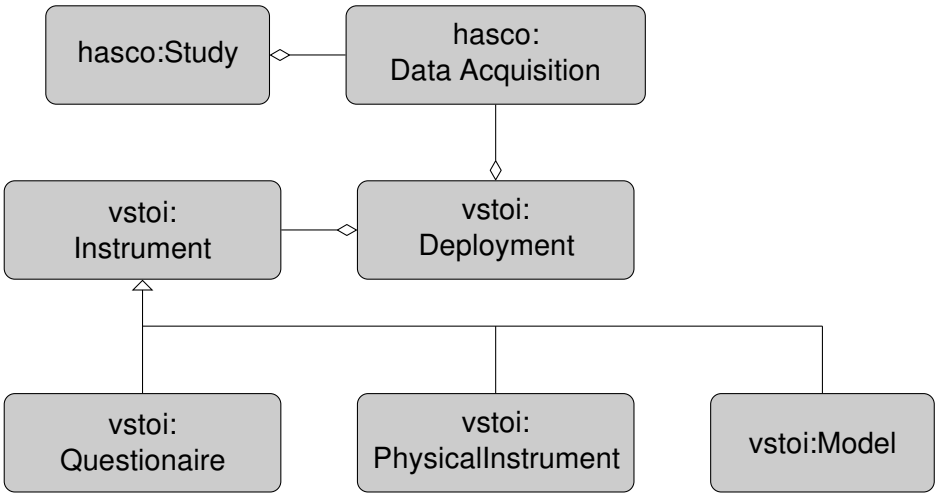
Em experimentos, ter o conjunto de dados dividido em aquisições de dados é uma maneira eficaz de descrever o controle de variáveis. Por exemplo, se um experimento está medindo os efeitos da luz em seres humanos, cada aquisição de dados pode ser caracterizada por eventos como ligar as luzes ou desligar. As subclasses de um estudo não são disjuntivas. Um estudo que é um FieldStudy requer dados de gerenciamento de instrumentos, como condições de implantação e configuração definidas. Um estudo que não é um estudo de laboratório pode não ter gerenciamento de incerteza, ou seja, precisão de computação (limite de detecção), resolução ou confiabilidade de detectores.

Data Acquisitions: na HAScO DataAcquisition é tanto um evento utilizando um instrumento para aquisição de dados, quanto uma coleta global de valores de dados adquiridos pelo instrumento, onde todos os pontos pertencentes à coleção têm exatamente a mesma qualidade. A ontologia define a qualidade de dados como o conjunto de configurações e o conjunto de propriedades dos instrumentos que foram usados para adquirir os dados. No caso de instrumentos físicos, as propriedades do instrumento tais como sua precisão são fundamentais para definir a qualidade dos dados.

Um ponto que deve ser considerado na integração de dados é a garantia de que os dados provenientes dos instrumentos de pesquisa, através de cada aquisição de dados, tenham propriedades que garantam que a análise de cada aquisição de dados seja comensurável. Em outras palavras a integração de dados somente é possível se cada aquisição de dados estiver baseada nos mesmos parâmetros.

Um estudo, em termos de dados, é caracterizado por suas aquisições de dados

Figura 10 – Classes da HAScO relacionadas à aquisição de dados



Fonte: Adaptada de (PINHEIRO *et al.*, 2018a)

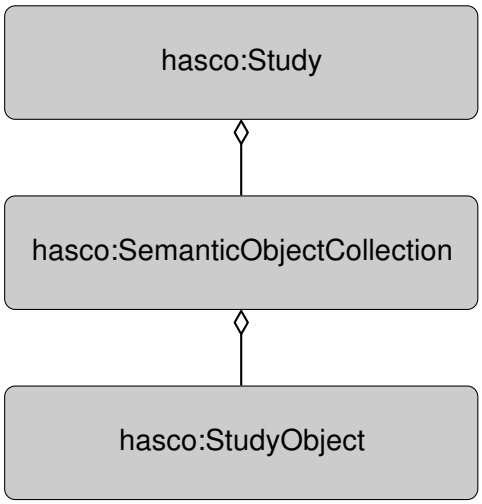
associadas. Podemos determinar quais são as variáveis do estudo ou quais amostras componentes do mesmo simplesmente verificando as variáveis de uma determinada DataAcquisition

Conforme mostrado na Figura 10, cada aquisição de dados é associada a uma única implantação (vstoi:Deployment). As aquisições de dados só existem no contexto de implantações. Isso significa que o momento de início de uma aquisição de dados não pode ocorrer antes da implantação associada. Da mesma maneira o final de uma implantação também determina o término de qualquer aquisição de dados associada à mesma.

Um vstoi:Deployment equivale ao posicionamento de um conjunto de instrumentos em uma plataforma para a aquisição de dados. A qualquer momento, mais de um instrumento pode ser colocado em uma plataforma. além disso, muitas implantações podem ocorrer em uma plataforma a qualquer momento. Uma implantação deve ter uma hora de início e pode ter um momento de parada. Se uma implantação não tiver tempo de parada, presume-se que ela esteja em andamento. Um evento de acionamento indica uma alteração na configuração de implantação, que pode ser uma alteração no próprio instrumento. Qualquer alteração na configuração durante uma implantação em andamento significa que, dentro da implantação, os dados adquiridos antes da alteração só devem ser comparados ou analisados em relação aos dados adquiridos após o evento, se houver um claro entendimento e consideração com relação à manutenção da qualidade dos dados adquiridos. Deste modo, se um instrumento tiver uma maior precisão após a nova implantação ou outra propriedade que impacte nos dados obtidos, não podemos comparar os dados entre as implantações.

Cada vstoi:Deployment possui um conjunto de vstoi:Instrument que podem ser instrumentos de medida físicos, tais como um conjunto de sensores. A ontologia prevê ainda o uso de instrumentos não físicos, cada qual associado a uma "implantação" (vstoi:Deployment). A Figura 10 mostra três especializações do vstoi:Instrument são elas: vstoi:PhysicalInstrument, vstoi:Questionnaire, e vstoi:Model, sendo estes dois últimos questionários e modelos, respectivamente. A generalização de instrumentos não físicos, em questionários e

Figura 11 – Objetos de estudo e coleções de objetos semânticos



Fonte: Adaptada de (PINHEIRO *et al.*, 2018a)

modelos, caracteriza os dados de forma uniforme, no sentido de que cada valor, independentemente de sua proveniência, será adquirido por um instrumento implantado. As características da implantação e as configurações do instrumento irão definir a qualidade dos dados.

Os questionários (vstoi:Questionaire) podem ser vistos como instrumentos para extrair conhecimento humano. Enquanto que os modelos (vstoi:Model) são gerados a partir de máquinas, como por exemplo, através de simulações. Nestes casos o conceito de implantação pode ser abstraído para determinar uma dada pesquisa de campo ou uma execução de um programa de simulação.

3.3.2 Organização da informação em um estudo científico

No decorrer da atividade de pesquisa temos vários experimentos que resultam em dados adquiridos para cada estudo realizado. Estes estudos têm sua estrutura e design descritos através da modelagem dos objetos relacionados aos mesmos. Identificadores gerenciados pelo investigador e relações com objetos de outros estudos definem o conjunto de objetos de cada estudo.

A ontologia prevê uma forma de armazenar os objetos independentemente dos identificadores originais. Isto é feito adicionando cada objeto de estudo à um conjunto, a Coleção de Objetos Semânticos (SOC). Deste modo pode-se trabalhar com os objetos de cada estudo acessando essa coleção. Cada estudo é composto de um ou mais SOCs. A Figura 11 ilustra a hierarquia das classes que formam a definição de cada um destes itens.

3.4 *Framework* para a aquisição e organização e armazenagem de dados - HADatAc

Nesta seção será apresentado o HADatAc, um *framework* que pode ser utilizado para integrar e harmonizar dados de múltiplos estudos científicos. Sua arquitetura se apoia na HAScO e suas ontologias de suporte.

Sintetizando, os três principais objetivos do HADatAc são:

- Extrair valor de dados relevantes de arquivos gerados por instrumentos e mover esses valores para repositórios (de conteúdo).
- Extrair metadados relevantes de documentos gerados por cientistas.
- Anotar semanticamente esses valores de forma que todo o conteúdo seja logicamente vinculado e harmonizado (isto é, compartilhando uma representação unificada) de acordo com ontologias científicas bem estabelecidas.

As implantações existentes do *framework* são construídas utilizando-se as ontologias de fundamentação apresentadas neste capítulo, juntamente com outras ontologias especializadas para o domínio de interesse. Essas ontologias especializadas são importadas a partir da ontologia Base que interliga essas ontologias ao grafo estendendo suas definições (MCCUSKER *et al.*, 2017).

Principais características do HADatAc

- Modelagem conceitual "sem esquema"[8] (*schema-free*): o *framework* utiliza tecnologias semânticas, ontologias, bases de dados de grafos e bases de dados não relacionais para gerir metadados e dados de estudos científicos relevantes. Ele é livre de esquemas, pois o conteúdo desses arquivos de estudo é armazenado sem uma estrutura predefinida e fixa. Isso permite que este inclua objetos, por exemplo, sujeitos, amostras, locais, em seus repositórios à medida que são apresentados, incluindo atributos de objetos considerados relevantes para os estudos.
- Evolutivo: o grafo de conhecimento do *framework* evolui importando as ontologias principais e de domínio ou definindo conceitos e relações que não podem ser encontrados ou reutilizados a partir de ontologias existentes. As ontologias e o grafo subjacente (TBox), assim como os dados científicos carregados como instâncias no grafo (ABox), são gerenciados por um banco de dados armazenado em um grafo RDF (Blazegraph[9]) e por um repositório de documentos Apache SOLR[10]. O repositório SOLR se interliga ao Blazegraph através de URIs. O repositório Blazegraph é usado principalmente para gerenciar a parte de terminologia (TBox) do grafo de conhecimento HADatAc.

[8] Em um banco de dados sem esquema os dados podem ser armazenados sem uma estrutura conceitual anteriormente definida.
[9] https://www.blazegraph.com/
[10] https://lucene.apache.org/solr/

- Escalável: as plataformas de gerenciamento de dados científicos devem lidar com volumes de dados crescentes. O repositório SOLR de backend do HADatAc fornece a escalabilidade necessária para repositórios de dados muito grandes. Apesar de existirem diversas soluções para bases de dados RDF no mercado atual, estas opções ainda não apresentam um bom desempenho. Assim a utilização do SOLR associada ao Blazegraph é uma alternativa que dá maior escalabilidade e desempenho (BELLINI; NESI, 2018).

- Suporte a Proveniência: o *framework* foi desenvolvido para trabalhar com uma ampla gama de fontes de dados, e a infraestrutura captura e preserva a proveniência dos valores de dados adquiridos pelos diversos tipos de instrumentos. O HADatAc classifica as fontes de dados de acordo com os instrumentos (e detectores) utilizados para adquirir os dados, nas três estratégias de aquisição de dados previstas: (a) *medição empírica* utilizando instrumentos físicos como sensores; (b) dados e conhecimento de humanos através de questionários; (c) *geração de dados por computador* utilizando modelos de simulação.

O HADatAc é um *framework* e também uma aplicação *Web*. A arquitetura do HADatAc é esquematizada na Figura 12, O quadro principal no centro da figura representa o núcleo do sistema, que é conectado a seis sub-sistemas satélites. O subsistema API é composto por uma coleção de serviços REST [11] com acesso programático ao conteúdo do HADatAc. O componente principal tem os elementos necessários para dar suporte aos subsistemas satélites incluindo os repositórios SOLR e Blazegraph, uma API Java codificando os conceitos da ontologia HAScO como classes POJO [12] e os sub-sistemas responsáveis pela extração, anotação e armazenamento dos dados e metadados no SOLR e no Blazegraph (PINHEIRO *et al.*, 2018c).

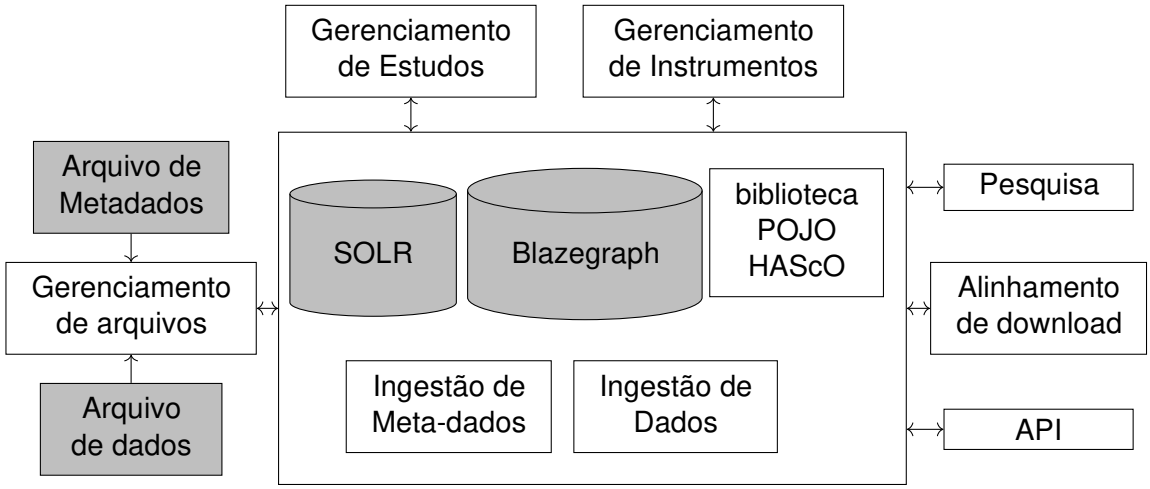
As classes HAScO POJO são utilizadas para construir e manter o grafo de conhecimento do HADatAc. Os dados são adicionados no repositório pela análise dos arquivos carregados através do Subsistema de Gerenciamento de Arquivos, sendo apresentados aos usuários por meio do subsistema de pesquisa, podendo ser baixados por meio do Subsistema de Alinhamento de Objeto e do Subsistema da API.

Com a utilização do HADatAc para processar a ingestão de dados o resultado final do processo constrói um grafo de conhecimento (Seção 2.2.1.2). Tem-se não apenas recursos para o gerenciamento, como também a escalabilidade necessária para tornar possível o tratamento de um grande volume de informação.

[11] *Representational State Transfer* (REST), é um estilo de arquitetura de software que define um conjunto de restrições a serem usadas para a criação de serviços web.

[12] *Plain Old Java Objects* (POJO) são classes Java que possuem apenas atributos privados e métodos para acesso e alteração dos mesmos

Figura 12 – Arquitetura do HADatAc, incluindo repositórios de conteúdo e sub-sistemas



Fonte: (PINHEIRO *et al.*, 2018c).

4 METODOLOGIA

4.1 Etapas do desenvolvimento do trabalho

Como estabelecido no Capítulo 1, o objetivo geral deste trabalho foi desenvolver um método de integração semântica de dados científicos que apresentasse uma solução aprimorada em relação aos trabalhos correlatos atuais. De acordo com o que foi estabelecido, buscou-se uma solução que permitisse evoluir um grafo de conhecimento de forma ágil e participativa.

A partir do que foi proposto, estabeleceu-se um marco teórico, onde foram definidos os conceitos básicos relacionados à integração de dados utilizando grafos de conhecimento. Esses conceitos estão relacionados no Capítulo 2.

O método de integração foi especificado com o apoio da Design Science Research (DSR), que estabelece como princípios a construção de artefatos para obter conhecimentos a partir de um conjunto de instrumentos de pesquisa. A escolha da DSR se justifica pois será criado um artefato conforme descrito na Seção 4.2.

A integração semântica de dados de acordo com o método está descrita no Capítulo 5. Os experimentos que foram realizados são apresentados no Capítulo 6.Os recursos utilizados para a avaliação do método e os resultados obtidos, bem como os esforços na divulgação do mesmo estão apresentados no Capítulo 7.

Os trabalhos correlatos foram detalhados no Capítulo 8. A análise destes trabalhos aprimorou o entendimento de como os pesquisadores têm tratado o tema, permitindo explorar pesquisas sobre a integração semântica de dados científicos, sua motivação e importância, bem como suas aplicações no contexto da ciência de forma mais ampla. Finalmente, no Capítulo 9, temos as considerações finais com relação ao método, relacionando sua aplicação às hipóteses e objetivos apresentados, além das contribuições e limitações.

O uso da DSR e sua aplicação ao problema de pesquisa do presente trabalho será descrito nas próximas seções.

4.2 Utilização da metodologia *Design Science*

A DSR estabelece como princípios a construção de artefatos, que podem ser modelos, construtos, métodos e instanciações (MARCH; SMITH, 1995). Os artefatos são criados e utilizados por um grupo de atores para obter conhecimentos a partir de um conjunto de instrumentos de pesquisa. Conforme define Bax:

> “A DSR é uma metateoria que investiga a geração de conhecimento no processo de concepção de artefatos, i.e., sobre como métodos de design podem constituir pesquisa de caráter científico.”(BAX, 2015).

A metodologia de pesquisa utilizando a DSR na CI está ligada ao seu uso como forma de buscar o aprimoramento do acesso a informação conforme afirmou Vakkari (1994):

> "O propósito para o qual a ciência da informação é concebida é facilitar o acesso à informação desejada. [. . .] Ela é DSR, cuja missão é fornecer, com ajuda da pesquisa, as diretrizes pelas quais o acesso ao informação pode ser melhorada"

A Estratégia de pesquisa da DSR é capaz de orientar tanto a construção do conhecimento, quanto aprimorar as práticas em várias disciplinas. No paradigma DSR, o conhecimento e a compreensão de um domínio do problema e sua solução são alcançados graças à construção e aplicação de um artefato projetado. Segundo Gregor e Hevner (2013), a DSR apostou seu fundamento como um importante e legítimo paradigma de pesquisa de sistemas na CI.

Gregor e Hevner afirmam que, para a DSR atingir todo o seu potencial no desenvolvimento e uso de sistemas de informação, algumas lacunas na compreensão e aplicação dos conceitos da DSR e métodos devem ser preenchidas. Propondo um esquema de comunicação da DSR com semelhanças a publicações mais convencionais, substituindo, no entanto, a seção tradicional de resultados com a descrição dos artefatos gerados pela DSR. No presente trabalho, o enfoque nesta descrição dos artefatos será levado em conta, mas os resultados obtidos com a utilização dos mesmos serão trazidos como uma forma de avaliação da sua qualidade.

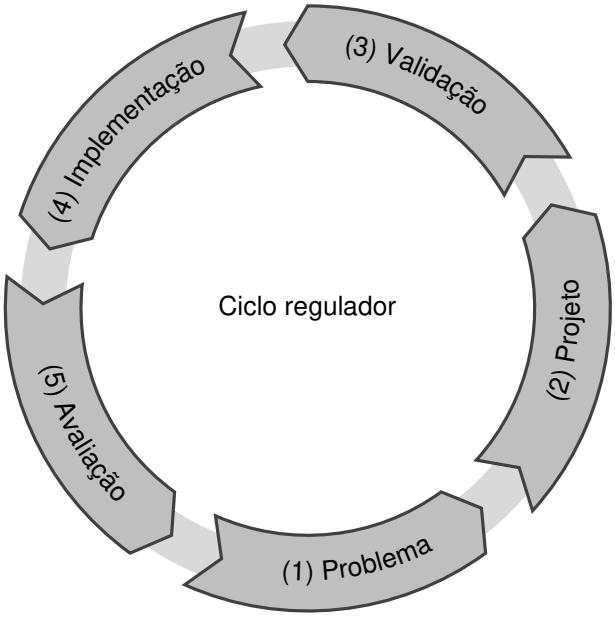
As motivações da DSR partem do fato que o conhecimento necessário para o design é muitas vezes distinto do conhecimento gerado pela pesquisa tradicional, no entanto, o uso da DSR permite satisfazer aos requisitos de design sem deixar de atender aos objetivos teóricos da pesquisa. Ao se envolver mais ativamente no design de artefatos inovadores, a pesquisa pode solucionar problemas mais próximos da vanguarda da prática industrial (HEVNER, 2007).

O processo de construção ou identificação dos artefatos a serem utilizados obedece a um ciclo regulador proposto por Wieringa (2009). As etapas do ciclo estão ilustradas na Figura 13. Cada uma das etapas do ciclo pode representar a busca da solução de um problema prático ou a geração de novos conhecimentos através de respostas a problemas teóricos. O processo se inicia com a investigação de um problema prático. Passa-se então à fase de projeto dos artefatos para a solução do problema, em seguida os projetos são validados. Caso os projetos sejam aprovados, os artefatos são implementados. Finalmente temos a avaliação da implementação, a qual pode gerar novos problemas, nesse caso inicia-se um novo ciclo.

A primeira etapa, a investigação do problema prático, é uma questão de conhecimento porque envolve a descrição, a explicação e até mesmo a predição das possíveis soluções. Essa investigação requer, portanto, um melhor entendimento do problema. Nesta etapa os requisitos da solução estabelecidos devem ser levados em consideração.

A próxima etapa, o projeto da solução, é uma questão prática, uma vez que esse projeto deve definir uma solução prática para o problema da etapa anterior. Ao final desta etapa, o projeto obtido não representa necessariamente uma solução definitiva, pois frequentemente melhorias no projeto podem ocorrer nas fases de validação e implementação. Ressaltamos

Figura 13 – Ciclo regulador de Wieringa



Fonte: Adaptado de (WIERINGA, 2009)

ainda que, conforme define Wieringa, a solução definida nesta etapa pode ser um artefato preliminar, o qual será aprimorado em outros ciclos.

Em seguida tem-se a realização da validação do projeto, onde se verifica se o mesmo atende aos objetivos estabelecidos na definição inicial. Esta etapa busca responder questões de conhecimento relativas à satisfação dos critérios identificados na investigação do problema, à existência de outras soluções que possam satisfazer aos critérios e à abrangência da solução em outros contextos.

Seguindo o ciclo, chega-se à etapa de implementação, onde o que foi projetado nas etapas anteriores será executado, ou seja, trata-se de uma etapa prática na qual os artefatos são gerados.

Finalmente, os artefatos são validados na última etapa do ciclo, que pode ser reiniciado com novos problemas a serem investigados. Tais problemas podem ser relacionados a melhorias nos artefatos criados ou na necessidade de novos artefatos para complementar a solução.

Os problemas práticos a serem investigados no início de cada ciclo podem ter níveis diferentes de complexidade. O projeto de um novo artefato, por exemplo, demanda um esforço bem maior do que melhorias em um artefato existente. A solução destes problemas práticos, independentemente do nível de complexidade, pode ser facilitada com a decomposição dos mesmos. Esta decomposição pode ser feita com o método DSR de Wieringa, que define uma estrutura aninhada do problema que auxilia o pesquisador a analisar (no sentido cartesiano), compreender (no sentido fenomenológico) de forma aprofundada o problema que está tratando e definir os objetivos específicos que precisa alcançar com vistas ao atingimento do objetivo geral. Com esta estrutura o problema é decomposto em problemas práticos (P) e problemas

teóricos ou questões de conhecimento (K). Assim, projeta-se um fluxo da solução do problema de pesquisa com um conjunto de subproblemas, os quais são classificados em práticos e teóricos. Explicita-se, deste modo, quais serão os problemas teóricos e quais são os problemas práticos a serem tratados. O método de Wieringa prevê ainda, a definição de subtipos para os problemas assim definidos:

- Problemas práticos (P):
  - especificação: especificação e desenvolvimento de uma solução;
  - participação: busca de soluções com a participação da equipe;
  - discussão: apresentação dos artefatos, reuniões com os membros envolvidos;
- Problemas teóricos ou questões de conhecimento (K):
  - descrição: descoberta de informações necessárias à investigação.
  - avaliação: observação e diagnóstico dos fatos.
  - predição: estimativa dos efeitos de uma solução.
  - validação: validação de soluções e comparação com critérios.
  - reflexões: questionamentos sobre as lições aprendidas e a geração de conhecimento.

A seção seguinte apresenta como esta metodologia foi aplicada ao problema de pesquisa do presente trabalho.

4.3 Aplicação da DSR ao problema de pesquisa

A partir do problema de pesquisa, foi aplicado o ciclo regulador de Wieringa, projetando uma solução que pudesse atacar o mesmo. Esta solução foi projetada, validada, implementada e avaliada segundo os critérios preconizados pela DSR.

Foi utilizada a estrutura aninhada e o ciclo regulador de Wieringa relatados na Seção 4.2. A estrutura aninhada utilizada pela pesquisa é apresentada na Figura 14 e contém a decomposição do problema de pesquisa em subproblemas práticos e teóricos.

Na Figura 14, o quadro superior denota o problema de pesquisa. Os demais quadros representam os subproblemas quebrados de acordo com a estrutura aninhada, onde os quadros marcados em cinza representam os subproblemas principais, com os quadros a direita representando a decomposição destes subproblemas em problemas menores.

Em cada quadro está estabelecido se temos um problema de conhecimento (K) ou um problema prático (P). A estrutura aninhada é executada de acordo com o ciclo regulador, o que significa que os problemas podem ser refinados a cada execução do ciclo.

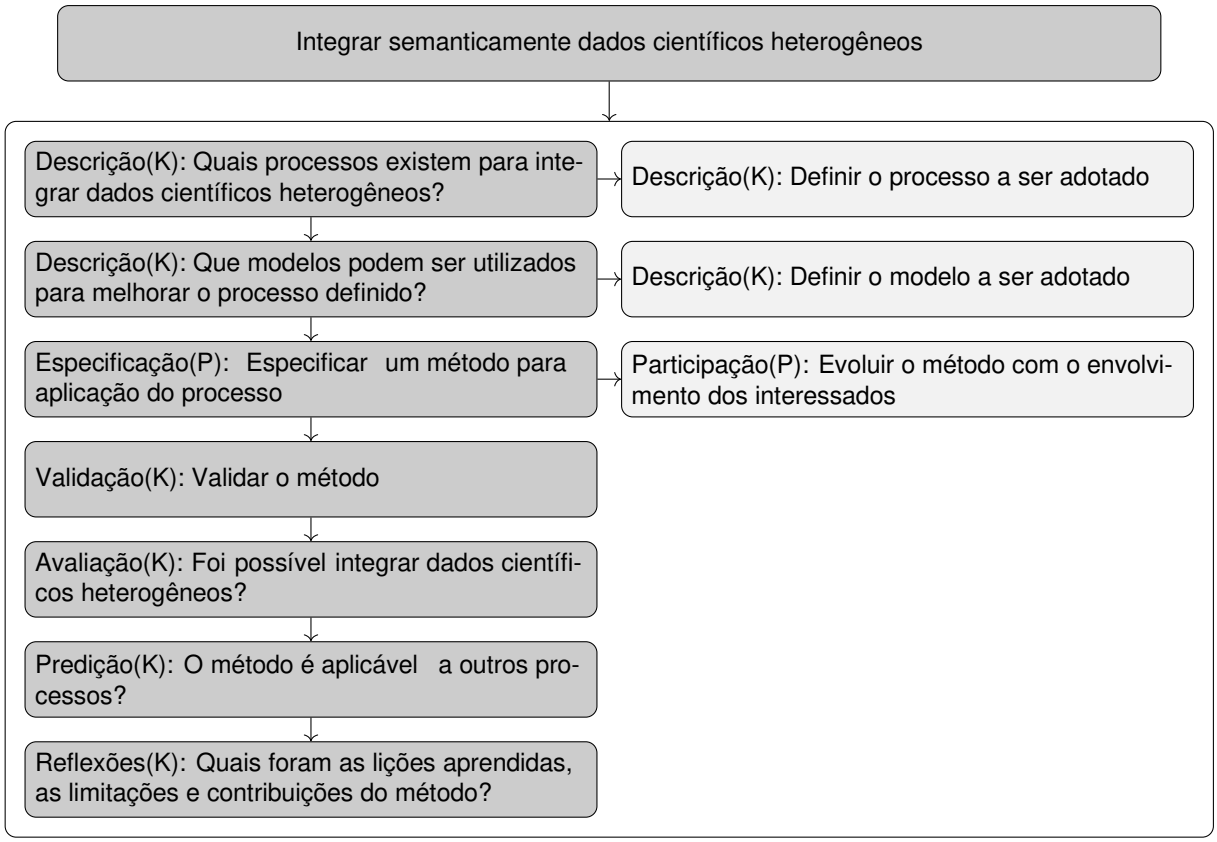
Assim, a partir do problema de pesquisa estabelecido:

Integrar semanticamente dados científicos heterogêneos.

Foram definidos os subproblemas ilustrados na Figura 14 os quais estão detalhados a seguir:

- Quais processos existem para integrar dados científicos heterogêneos?
  - O levantamento dos processos foi feito com a revisão da literatura onde buscou-se identificar trabalhos que apresentaram soluções existentes para a integração de dados (Capítulo 8). Em seguida resolveu-se o subproblema de definição do processo a ser adotado, o qual por sua vez deveria abordar os subproblemas relacionados à viabilidade de executar o processo.
- Que modelos podem ser utilizados para melhorar o processo definido?
  - A partir de uma pesquisa de modelos que possam ser aplicados para uma melhoria do processo escolhido, definiu-se um modelo a ser aplicado.
- Especificação de um método para aplicação do processo.
  - Foi especificado um método segundo o modelo definido e o processo escolhido com a participação de todos os envolvidos. A solução deste subproblema foi a etapa que consumiu a maior parte dos esforços para uma definição satisfatória do método.
- Validação do método.

Figura 14 – Estrutura aninhada do problema de acordo com a DSR de Wieringa



Fonte: Elaborada pelo autor

  - O método foi validado, verificando se os resultados obtidos estavam de acordo com o esperado.

- Foi possível integrar semanticamente dados científicos heterogêneos?
  - Este subproblema procura avaliar se o artefato construído, ou seja, o método, atendeu ao que se estabeleceu com a pergunta de pesquisa.

- O método é aplicável a outros processos?
  - Determinou-se, a partir deste questionamento, se o método pode ser aplicado a outros processos, levantando-se as características necessárias para que um processo alternativo seja utilizado.

- Quais foram as lições aprendidas, as limitações e contribuições do método?
  - Este subproblema foi resolvido com reflexões obtidas das soluções e problemas anteriores. Estas reflexões permitiram extrair o conhecimento que foi gerado a partir da criação do método e estão relacionadas no Capítulo 9.

Com a utilização do método de Wieringa foram analisados os problemas a serem tratados, identificando-se os problemas práticos e os de conhecimento. Esta análise facilitou a reflexão acerca dos objetivos específicos da pesquisa que podem ser entendidos como subprodutos do objetivo geral (WAZLAWICK, 2017). Além disso, a divisão em subproblemas dentro da estrutura aninhada definiu uma direção a ser seguida para a execução do trabalho.

Com base na estrutura aninhada da Figura 14 e na aplicação da metodologia DSR apresentada aqui, o método para a integração semântica de dados com a utilização de ontologias foi definido.

4.4 *Agile Design Science Research*

A metodologia DSR estabelece como princípios a construção de artefatos, que podem ser modelos, construtos, métodos e instanciações (MARCH; SMITH, 1995). Os artefatos são criados e utilizados por um grupo de atores para obter conhecimentos a partir de um conjunto de instrumentos de pesquisa (BAX, 2015). A DSR apresenta-se como uma metodologia de pesquisa, na qual os problemas são vistos como entradas a priori do processo de design. No entanto, muitas vezes, o espaço do problema é visto como emergente e evoluindo em conjunto com o espaço da solução.

Para capturar essa emergência e evolução de problemas, adicionando características da metodologia ágil ao modelo de processo DSR, Conboy, Gleasure e Cullina (2015) apresentaram a *Agile Design Science Research Methodology* (ADSRM), que acrescenta características da metodologia ágil ao mesmo. Segundo os autores a ADSRM (e a filosofia ágil que a sustenta) dá aos pesquisadores licença para reformular o problema, dando maior flexibilidade criativa aos mesmos. Essa reformulação é realizada sem o comprometimento dos padrões acadêmicos, nem da capacidade de posicionamento do trabalho frente a literatura existente.

O problema de pesquisa é tratado na ADSRM de forma iterativa. A cada iteração, o problema é revisto e um conjunto de tarefas e executado, gerando artefatos que são demonstrados e posteriormente avaliados. Ao final de uma iteração, os resultados obtidos podem gerar comunicações científicas formais dos resultados obtidos. A metodologia estabelece que as tarefas a executar, no sentido de endereçar o problema, sejam relacionadas em um documento, a Lista de pendências, a partir da qual são selecionadas aquelas que serão realizadas a cada iteração.

Essa metodologia preconiza ainda, a adição de uma iteração de homologação após um certo número de iterações, com o objetivo de aumentar o rigor que pode estar faltando durante iterações regulares. A frequência destas iterações de homologação é baseada no contexto e dependente da necessidade de rigor ou da falta de rigor percebida. Uma iteração de homologação é diferenciada por orientar-se segundo os seguintes mecanismos chave, que visam aumentar o rigor:

1. Congelar o problema. Embora a capacidade de lidar com a mudança proporcionada pela agilidade esteja presente nos ambientes de design complexos atuais, um nível de rigor pode ser aplicado ao utilizar-se uma fase em que não é permitida turbulência, dinamismo ou improvisação.

2. Congelar o processo. A preocupação do ponto de vista da regulação é que, para se ter mais rigor, há momentos em que o processo deve ser valorizado sobre as pessoas. Novamente, uma única iteração que requer uma adesão cuidadosa ao procedimento, verificações de conformidade e ausência de improvisação pode ser útil para assegurar e manter o rigor.

3. Adicionar artefatos de controle. Adicionar ao processo artefatos orientados pelo rigor. Um exemplo disso seria adicionar ou alterar validações ao conduzir a fase de avaliação.

Com as iterações de homologação é possível validar os artefatos já construídos, permitindo que a solução atual do problema seja analisada sem as interferências que poderiam ocorrer durante uma iteração regular.

5 INTEGRAÇÃO SEMÂNTICA DE DADOS

O método de integração semântica de dados utiliza a modelagem ontológica tal como explicada na Seção 2.2. A partir de uma análise dos fenômenos representados pelos objetos de interesse no domínio de um estudo científico, estabelece-se um modelo conceitual, no formato de uma ontologia de domínio, para organizar os dados (de um ou mais estudos) a serem ingeridos na base de dados. Com a modelagem ontológica, os objetos do estudo, instanciados pelos dados coletados, são organizados, identificados e qualificados por suas classes, propriedades e relações.

Para integrar dados científicos oriundos de pesquisas utilizando o método é necessário elaborar alguns arquivos contendo gabaritos de metadados que descrevem os dados (usaremos os termos “templates” ou “artefatos” para designar tais gabaritos). Tais metadados mapeiam os dados para conceitos de ontologias. Dados, metadados e ontologias são então submetidos a uma plataforma de software, apresentada na Seção 5.2, responsável pela integração por meio de um grafo de conhecimento no padrão RDF.

Uma vez mapeados para as ontologias, os dados adquirem semântica formal o que, por hipótese, facilitaria o reúso e a reprodução dos resultados do estudo por parte de outros pesquisadores. Ao compartilharem características comuns, seria possível ainda integrar dados de diferentes estudos, estruturados por modelos conceituais diferentes.

5.1 O método Odin

5.1.1 Aspecto ágil do método

O método inspira-se na *Agile Design Science Research Methodology* (ADSRM), detalhada na Seção 4.4, que incorpora princípios ágeis à DSR (CONBOY; GLEASURE; CULLINA, 2015) e segue um processo iterativo, organizado por ciclos ou iterações. Tais princípios são expressos no Manifesto Ágil (FOWLER; HIGHSMITH *et al.*, 2001), cuja motivação histórica se deu no campo da Engenharia de Software. Os mais relevantes, já adaptados para o nosso contexto, são:

- Priorizar a satisfação dos pesquisadores através da obtenção precoce e contínua dos artefatos utilizados no processo de integração de dados.
- Incentivar a mudança, tratando novas questões de competência a qualquer tempo.
- Obter os artefatos para integração o mais cedo possível.
- Fomentar o trabalho conjunto dos atores durante todo o projeto.
- Medir o progresso através de entregas de artefatos funcionando.
- Refletir, a intervalos regulares, sobre como melhorar a eficácia, sintonizando e ajustando as decisões de acordo com o que se constatar de novo a cada ciclo.

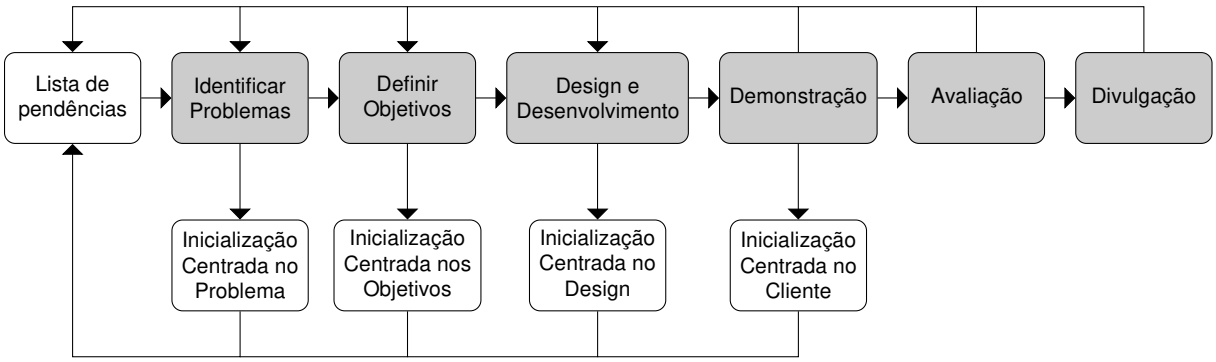
Com base nestes princípios e na ADSRM de Conboy, Gleasure e Cullina, o método orienta a produção dos artefatos de forma iterativa. As tarefas a realizar em cada iteração são selecionadas a partir da Lista de pendências. As iterações compõem-se de uma série de atividades voltadas para a execução, em um intervalo de tempo, de um conjunto específico de tarefas. A cada iteração, cria-se uma versão aprimorada do grafo RDF, validada pelos atores envolvidos.

5.1.2 Atores envolvidos na execução do método

Quatro tipos de atores estão envolvidos e representam papeis específicos, conforme mostra a Figura 15:

- **Especialistas de Domínio** (pesquisadores): são os pesquisadores interessados em realizar a integração e analisar os dados de pesquisa de um dado domínio.
- **Ontologistas**: profissionais que possuem o conhecimento na criação, evolução e pesquisa de ontologias.
- **Desenvolvedores**: colaboram no processo com o seu conhecimento em tecnologia da informação, através da criação de pequenos programas ou até mesmo de soluções mais complexas.
- **Cientistas de Dados**: especialistas em ciências de dados (Seção 2.1), que poderão auxiliar na obtenção de informações a partir dos dados ingeridos, utilizando técnicas estatísticas, de aprendizado de máquina e outros recursos associados ao tratamento de grandes volumes de dados (*Big Data*).

Figura 15 – Etapas do método



Fonte: Adaptação de (CONBOY; GLEASURE; CULLINA, 2015)

5.1.3 Etapas do método

As etapas são executadas com a participação dos atores relacionados na Seção 5.1.2. O fluxo de uma iteração (ou ciclo) tem como objetivo a geração do grafo RDF que integra os dados por meio dos artefatos de integração. Cada iteração se desenvolve através de

um conjunto de etapas que estão representadas pelas caixas em cinza da Figura 15. Conforme mostra a figura, determinadas etapas podem não serem finalizada, retornando o fluxo para uma etapa anterior ou para o exame da Lista de pendências.

Assim, algumas tarefas inciadas em uma dada iteração não percorrem todas as etapas sequencialmente, mas voltam a uma etapa anterior ou são inseridas na Lista de pendências. Neste último caso, as tarefas inseridas na Lista de pendências devem ser tratadas em uma outra iteração.

O método foi criado e suas etapas foram validadas por meio de um experimento que integrou *datasets* no domínio da Epidemiologia oriundos de pesquisas realizadas por parceiros na Fundação Oswaldo Cruz (Fiocruz) (SOUSA *et al.*, 2014; SOUSA; GARBAYO; BARCELLOS, 2017; SOUSA *et al.*, 2018).

5.1.3.1 Identificação do problema

Nesta etapa o problema a ser tratado na iteração é identificado e orientará a concepção dos artefatos. Na primeira iteração esta etapa consiste na definição da primeira versão dos artefatos, que irão fornecer subsídios para a criação do grafo RDF inicial. Assim, a primeira iteração envolve a participação direta dos especialistas de domínio e dos ontologistas. Nas demais iterações o problema será definido a partir da priorização das tarefas da Lista de pendências e da escolha do conjunto de tarefas que farão parte do problema a ser tratado. A maioria destas tarefas será definida a partir das questões de competência elencadas pelos especialistas de domínio, que definirão a modelagem ontológica do grafo RDF conforme estabelecido na Seção 2.2.1.3. A etapa de identificação do problema, pode gerar tarefas para a Lista de pendências, conforme mostra a Figura 15 como "*Inicialização Centrada no Problema*", isso pode ocorrer quando um conjunto de tarefas do problema for postergado, voltando para a Lista de pendências.

É nesta etapa, da primeira interação, que cada um dos cabeçalhos do dataset (nomes de colunas dos arquivos de dados) deve ser vinculado à conceitos das ontologias relacionadas ao domínio do estudo. Contudo, o método pressupõe a necessidade de criação de uma ontologia de domínio específica para a anotação dos dados. Em sua primeira versão, tal ontologia de "Base" é formada por conceitos correspondentes aos termos usados nos cabeçalhos dos arquivos de dados (do estudo científico em questão, cujos dados estão sendo anotados e integrados). Na primeira etapa do ciclo inicial (i.e., na primeira iteração do método), pressupõe-se que tais conceitos ainda não foram encontrados e mapeados para outras ontologias pré-existentes. Na ontologia Base o pesquisador poderá qualificar certos conceitos para serem considerados indicadores científicos (ou seja, atributos que apresentam interesse especial para a análise dos dados integrados).

5.1.3.2 Definição dos objetivos da solução

Nesta etapa serão definidos os objetivos que irão determinar as características da solução que abordará o problema de design definido. O método não se concentra apenas em objetivos específicos, mas também quebra os requisitos em histórias de usuário mais

detalhadas. Deste modo, teremos um objetivo geral de alto nível e sub-objetivos de nível inferior de prioridades e granularidades variadas.

Os principais interessados na definição dos objetivos são os especialistas pesquisadores de domínio, que buscam a comprovação de suas hipóteses de pesquisa.

Para a integração semântica pode-se determinar, nesse momento, desde as características do grafo desejado até opções particulares da própria integração, aspectos que estarão ligados às questões de competência que serão tratadas na iteração. Alguns dos objetivos definidos, também nesta etapa, podem ser adiados e acrescentados à Lista de pendências (*Inicialização Centrada no Objetivo*).

5.1.3.3 Design e desenvolvimento

A partir do que foi definido nas etapas anteriores, deve-se especificar e implementar os artefatos que atenderão à solução. Assim, a partir dos objetivos e do problema, estes artefatos serão desenvolvidos, com a participação de todos os atores da seguinte maneira:

- Especialistas de domínio e os ontologistas: definem e implementam a ontologia Base e os artefatos de mapeamento dos dados no grafo RDF;
- Cientistas de dados: trabalham com os dados, caso já tenham sido ingeridos em ciclos anteriores, utilizando recursos computacionais para obter informações que permitam responder às questões de competência;
- Desenvolvedores: corrigem *bugs*, modificam e criam ferramentas de acordo com as tarefas da Lista de pendências. Podendo ainda auxiliar em transformações necessárias nos arquivos de dados de entrada.

A implementação da ontologia de Base e dos artefatos de mapeamento é a tarefa que demanda o maior tempo da iteração e envolve as seguintes atividades:

- Revisar e compreender as ontologias de fundamentação, definidas no Capítulo 3. Adicionar vínculos dos dados para os conceitos de outras ontologias, visando complementar a ontologia Base. A primeira versão da ontologia Base será construída exclusivamente a partir das ontologias de fundamentação. Posteriormente, serão acrescentados vínculos para ontologias adicionais.
- Construir a ontologia Base (MCCUSKER *et al.*, 2017): O processo é incomum, no sentido de que a ontologia, na iteração inicial, é criada com o propósito de detalhar o *dataset* descrevendo o modelo conceitual que ele representa, sendo aprimorada nas iterações posteriores. Uma forma semelhante de tratar o problema é descrita por McCusker *et al.*, que relata um processo de desenvolvimento específico: “Uma abordagem sob demanda para o desenvolvimento de uma Ontologia baseada em um conjunto de projetos piloto representativos no *Children’s Health Exposure Analysis Resource* (CHEAR)”. Assim, a ontologia Base é desenvolvida sob demanda, a partir de requisitos de uso ou de questões de competência dos conjuntos de dados (Seção 2.2.1.3).

- Gerar os grafos de conhecimento: a geração dos grafos resulta da ingestão semântica de dados, que deve ser realizada com uma ferramenta específica que propicie a recriação deste grafo a cada iteração. No presente trabalho foi utilizado o HADatAc, apresentado na Seção 3.4.

Nesta etapa, novamente o atendimento às questões de competência servirá como base para o design, definindo escolhas de projeto que possam atender às mesmas. Mais uma vez, durante esta etapa, podem ser geradas tarefas para a Lista de pendências, conforme mostrado no diagrama da Figura 15.

5.1.3.4 Demonstração

Com os dados ingeridos na etapa da Seção 5.1.3.3, pode-se apresentar aos envolvidos o resultado da iteração. É possível demonstrar o atendimento às questões de competência, por meio de consultas SPARQL ao grafo. Outras análises realizadas pelos cientistas de dados ou por ferramentas criadas pelos desenvolvedores podem complementar a demonstração.

A implementação precoce e frequente da ingestão de dados, permite que se identifique mais rapidamente possíveis problemas e limitações, que vão gerar novas tarefas para a Lista de pendências (cf. destacado na Figura 15 como *Inicialização Centrada no Cliente*). Nesta etapa é possível testar a estabilidade da solução frente ao problema.

Uma vantagem da demonstração precoce da solução é que os especialistas de domínio poderão visualizar e analisar os dados ingeridos mais cedo, podendo sugerir alterações para as próximas iterações.

5.1.3.5 Avaliação

Nesta etapa verifica-se se a solução de integração de dados construída atende à investigação do problema especificado pelas questões de competência. Isso corresponde aproximadamente ao componente de testes teóricos da pesquisa descritiva ou explanatória tradicional (CRESWELL, 2013), embora a ênfase esteja na utilidade do design.

O grafo RDF pode ser validado com o uso de consultas utilizando o SPARQL, a partir de visualizações dos dados ingeridos, com o tratamento do grafo com ferramentas de processamento. Além disso, a própria capacidade do processo em modelar e representar um conjunto de dados científicos é um indicador da validade da abordagem proposta, sobretudo se tal capacidade for validada no contexto de diferentes estudos científicos. Ou seja, procura-se avaliar não somente a ingestão de dados, mas também a eventual integração dos mesmos com outros estudos previamente ingeridos.

O *design* da solução deve ser revisto, avaliando-se o atendimento às questões de competência elaboradas pelos especialistas de domínio. As questões que não puderam ser respondidas podem então gerar a necessidade de alterar os artefatos de ingestão ou a ontologia Base. 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.

5.1.3.6 Divulgação

O subprocesso de divulgação representa o estágio final de uma iteração, no qual as descobertas são compartilhadas com públicos relevantes por meio de publicações, tanto acadêmicas quanto profissionais. Esta etapa não ocorre em todas 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.

Conforme citado, o método prevê a existência de um documento onde são relacionadas as diversas tarefas a executar, a Lista de pendências.

5.1.4 Lista de pendências

Na Lista de pendências estão contidas todas as tarefas que foram identificadas, no início do problema ou durante a execução das etapas da Seção 5.1.3. Observa-se na Figura 15 que, em cada etapa do processo, é possível ser gerada uma tarefa para a Lista de pendências, a qual será tratada em futuras iterações. As questões de competência (Seção 2.2.1.3) são um caso particular de tarefas da Lista de pendências geradas pelos especialistas de domínio e que norteiam o desenvolvimento do grafo RDF. A Lista de pendências contém tarefas para os diversos atores citados na Seção 5.1.2, as quais podem ter sido geradas por qualquer um dos outros atores.

No caso das tarefas geradas a partir das questões de competência, por exemplo, os especialistas de domínio, os ontologistas, desenvolvedores e cientistas de dados trabalharão em conjunto para encontrar uma solução.

Outras tarefas mais específicas, como uma alteração em um elemento de interface com o usuário, por exemplo, devem ser direcionadas à equipe de desenvolvedores. A cada iteração, na etapa de Identificação do problema, as tarefas da Lista de pendências a serem tratadas são selecionadas e direcionadas aos responsáveis pelo seu tratamento. No caso das questões de competência, esta seleção obedece à priorização estabelecida pelos especialistas de domínio. Assim, define-se um conjunto de tarefas que fará parte da iteração, adiando-se as demais tarefas para iterações posteriores.

5.1.5 Iteração de homologação

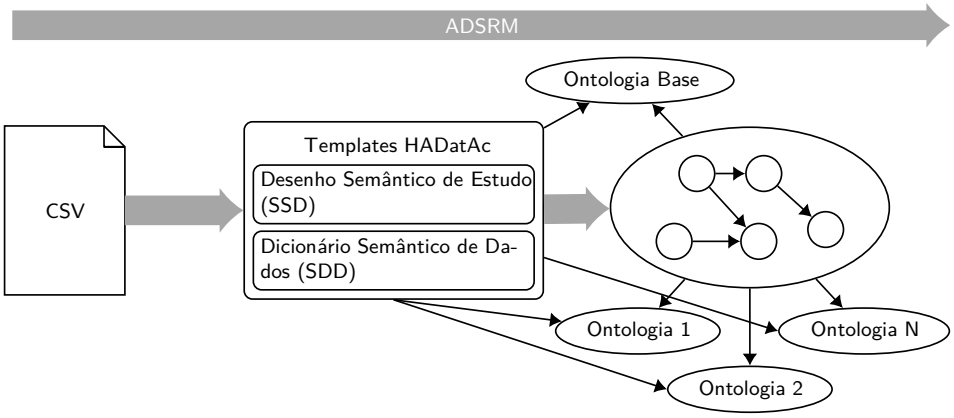
O desafio de introduzir agilidade sem comprometer o rigor processual tem sido explorado em diversos trabalhos (FITZGERALD *et al.*, 2013). No método proposto, uma iteração de homologação é adicionada após um certo número de iterações, com o objetivo de aumentar o rigor que pode estar faltando durante iterações regulares.

Nas iterações de homologação não é feita uma revisão do problema a ser tratado e os artefatos construídos até o momento são mantidos inalterados, exceto por possíveis artefatos

de controle. Tais artefatos de controle podem ser utilizados, por exemplo, para verificar se o grafo RDF fornece respostas satisfatórias às questões de competência.

5.1.6 Ingestão semântica de dados

Figura 16 – Processo de ingestão semântica proposto



Fonte: elaborada pelo autor

A ingestão semântica dos dados concentra-se na fase do trabalho científico onde dados heterogêneos originados de múltiplas fontes são semanticamente anotados de acordo com um modelo conceitual especificado por uma ou mais ontologias. Este modelo conceitual pode evoluir conforme amadurece o entendimento do pesquisador sobre o domínio de conhecimento representado. Com essas anotações semânticas os dados são integrados, podendo ser posteriormente consultados pelos especialistas de domínio, usando como critérios de seleção os metadados descritivos (FOX; HENDLER, 2009). As anotações podem servir também para guiar o trabalho analítico desses especialistas.

Esta seção descreve o processo de ingestão semântica de dados utilizado neste trabalho. Nesta seção e na Seção 5.2 serão apresentados exemplos baseados em experimentos realizados com parte dos *datasets* fornecidos pelos pesquisadores da Fiocruz. Nos exemplos, teremos um recorte dos *datasets* e, em alguns casos, modificações em seu *layout*.

O processo de ingestão (Figura 16) é o objetivo principal de uma iteração e consiste basicamente na definição do mapeamento, na criação/edição da ontologia Base e na interligação de conceitos à eventuais outras ontologias de domínio. Ele se fundamenta no uso das ontologias apresentadas no Capítulo 3 e na ontologia Base. A cada iteração, um *dataset* (i.e., um ou mais arquivos CSV), é convertido em um grafo RDF a partir de um mapeamento que o interliga à ontologia Base e às demais ontologias. O grafo RDF gerado ao final da iteração terá, portanto, ligações com cada uma das ontologias utilizadas.

De acordo com o método proposto, como já explicado acima, os problemas selecionados para cada iteração são tratados por meio de modificações no mapeamento e na ontologia Base. A modelagem ontológica da solução é realizada, portanto, por meio destas modificações. Assim, os especialistas de domínio e ontologistas vão definindo a evolução do

grafo RDF, aprimorando a modelagem a cada iteração.

Em alguns casos, entretanto, o processo de ingestão pode não ser realizado em uma iteração, por exemplo, se esta iteração for executada apenas para analisar informações já existentes no grafo de conhecimento. A seguir será descrito o processo de ingestão e como essa operação funciona com o uso da ontologia Base e das ontologias de fundamentação descritas no Capítulo 3.

5.1.6.1 Fluxo de trabalho da ingestão semântica de dados

O fluxo de trabalho da ingestão implica em validar a sintaxe dos arquivos de dados e de metadados individuais e em seguida rotear os dados para o seu destino correto no grafo sendo construído. A ingestão segue um algoritmo utilizado pela ferramenta que implementa o processo de ingestão para adicionar metadados aos dados e construir o grafo.

Formalmente, o grafo (G) é gerado a partir dos arquivos CSV de entrada (C), de um conjunto de ontologias (O) e de um mapeamento (M), ou seja, G é uma função de C,O e M. O conjunto de ontologias (O) é subdividido em ontologias de suporte (apresentadas no Capitulo 3) e a ontologia de domínio que descreve os dados em C. Tanto M, quanto a ontologia de domínio em O são definidos a partir dos dados de entrada C. A ontologia de domínio em O é inicialmente denominada ontologia Base. Esses conjuntos evoluem a cada iteração, com possíveis ligações à novas ontologias e com a revisão da ontologia Base.

As ontologias utilizadas no processo de modelagem conceitual incluem aquelas especificadas no Capítulo 3 e a ontologia Base, que será descrita na Seção 5.1.6.2. Nos exemplos que serão apresentados a seguir, os URIs associados a cada uma destas ontologias serão abreviados na forma de "*prefixo*:identificador". Os prefixos determinam o "espaço de nomes" da ontologia, conforme a Tabela 1. O "identificador", é uma sequência alfanumérica que representa um elemento do grafo, vértice ou aresta, da ontologia indicada pelo prefixo. Pode-se verificar que, para a ontologia Base, o prefixo é vazio, o que denota que qualquer URI na forma ":identificador" representa um elemento desta ontologia. Os prefixos e identificadores geram o URI definitivo de acordo com o espaço de nomes associado ao prefixo. Um URI definido como `hasco:namedTime`, por exemplo, equivale ao URL http://hadatac.org/ont/hasco/namedTime.

Tabela 1 – *Namespaces* das ontologias utilizadas

Prefixo	*NameSpace*	Ontologia
hasco:	http://hadatac.org/ont/hasco/	*Human-Aware Science Ontology* (HAScO)
sio:	http://semanticscience.org/resource/	*Semanticscience Integrated Ontology* (SIO)
:	http://ws1.assis.bhz.br/gbd#	Ontologia Base

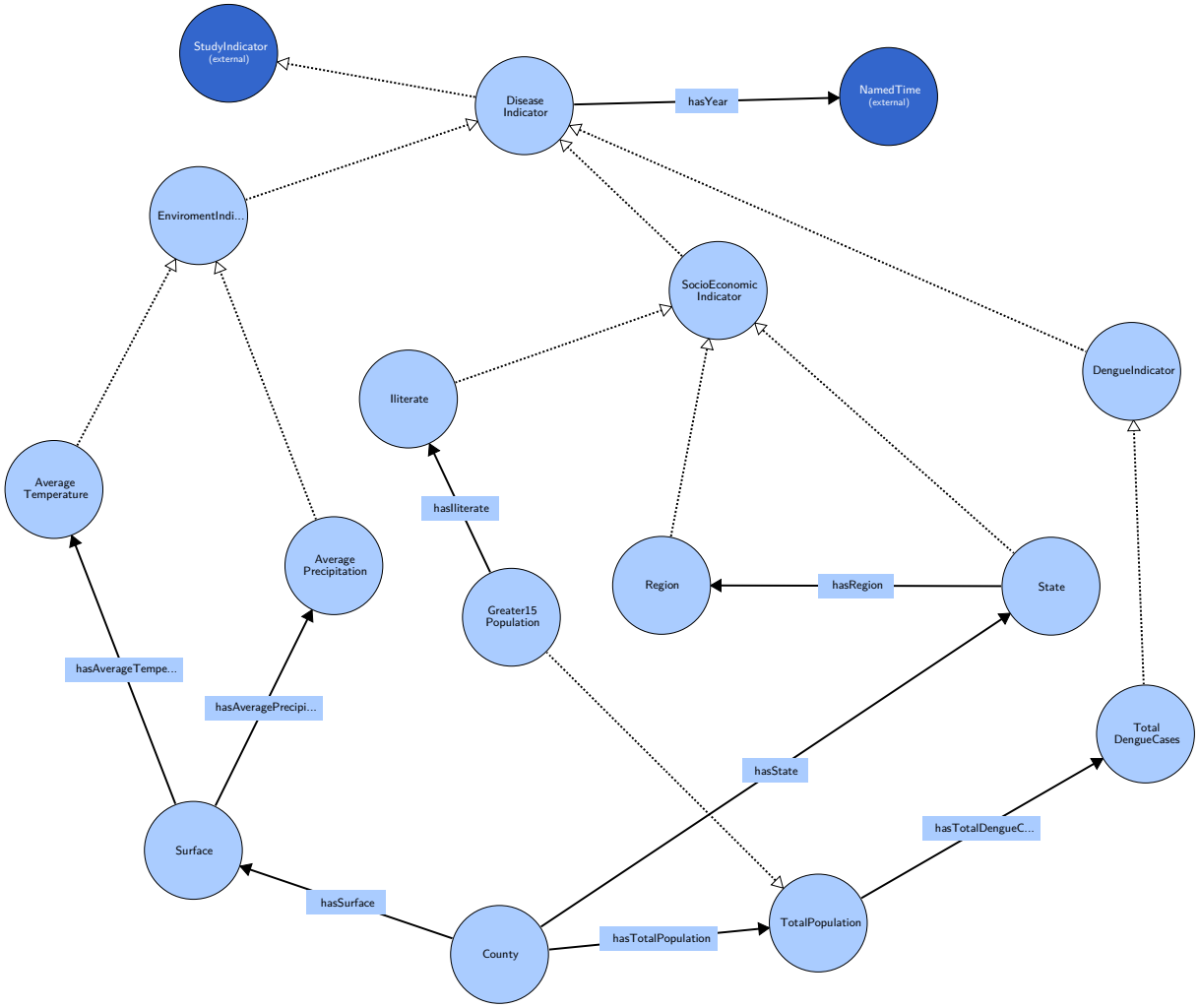
Fonte: elaborada pelo autor

5.1.6.2 Ontologia Base

Conforme relatado na Seção 5.1.3, o processo de ingestão prevê a existência de uma ontologia criada especificamente para a anotação do estudo, a ontologia "Base". Na

Figura 17, pode-se verificar um subgrafo ilustrativo dessa ontologia, tal como foi utilizada em um experimento realizado para validar o método (estudo sobre a incidência da Dengue apresentado na Seção 6.3). Já na Figura 18 podem ser observadas a hierarquia de classes e a hierarquia de propriedades da ontologia.

Figura 17 – Ontologia Base



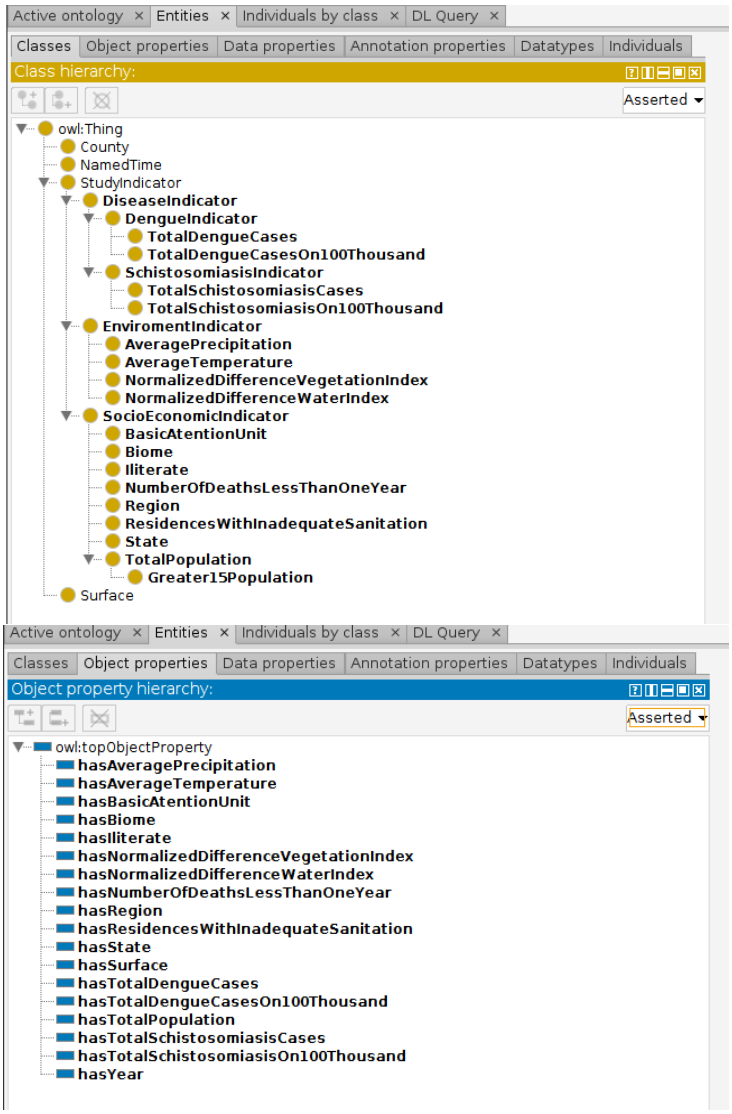
Fonte: elaborada pelo autor

A ontologia Base, em sua primeira versão, é um modelo conceitual que descreve os dados do domínio do estudo de forma inicial e simplificada. Isto é feito porque, normalmente, no início do processo ainda não existem definições precisas para os conceitos empregados ou estes últimos não foram procurados em uma ontologia de referência disponível publicamente.

A cada iteração, esta ontologia é então revista, com o acréscimo de ligações (mapeamentos) para conceitos encontrados em ontologias específicas do domínio do estudo. Tal revisão deve contar com a participação dos especialistas do domínio. Estes, com o auxílio do ontologista, poderão identificar as melhores opções para a interligação do estudo a ontologias existentes.

Usando a ontologia HAScO, apresentada na Seção 3, o pesquisador pode fazer a

Figura 18 – Ontologia Base - hierarquia de classes e propriedades



Fonte: Importação da ontologia no Protégé https://protege.stanford.edu/

qualificação de certos dados da ontologia Base para que sejam considerados como indicadores científicos, os quais podem ser utilizados para explorar as hipóteses elaboradas para o estudo em questão. Isto é feito declarando certos conceitos da ontologia Base como subclasses do conceito hasco:StudyIndicator.

Pode-se verificar no grafo da ontologia Base (Figura 17) que a principal classe de interesse: :Disease, é uma subclasse da hasco:StudyIndicator. Esta classe, possui duas subclasses: :Socioeconomicindicatore :Enviroment-Indicator. Desta forma, uma classe como :Average-Temperatureestá associada à classe hasco:StudyIndicatoratravés da classe :Envoment-Indicator A classe hasco:StudyIndicatortorna as sub-classes dela derivadas indicadores que podem ser explorados no grafo gerado. As demais classes da figura completam o exemplo, com outros indicadores do estudo. Especificar uma classe como sendo um Indicador informa as ferramentas de análise sobre variáveis de interesse que podem ser

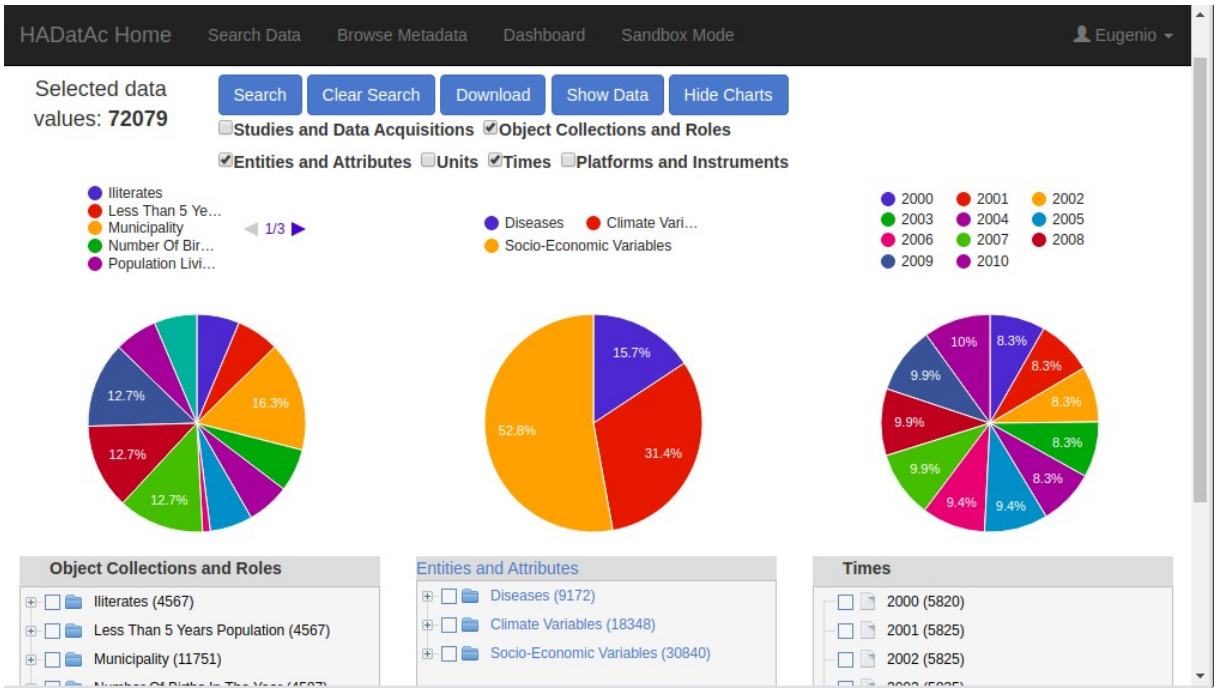
utilizadas, por exemplo, para gerar visualizações facetadas dos dados inseridos no grafo RDF. A Figura 19 apresenta de maneira facetada, na caixa *"Entities and Attibutes"*, os indicadores presentes na ontologia Base.

A criação da ontologia Base e de um mapeamento (representado pelo conjunto de templates de metadados) determinam o grafo resultante da integração semântica dos dados de um conjunto de arquivos CSV. A Seção 5.2 detalha uma forma específica de realizar este mapeamento (ou "ingestão"), que foi utilizada para validar o método utilizando o *framework* HADatAc.

5.2 Ferramenta utilizada para "ingerir" e apresentar os dados

O HADatAc (Seção 3.4), implementa um processo de mapeamento para criar o grafo RDF. O mapeamento define metadados que especificam como os dados de entrada, oriundos de um ou mais estudos científicos, devem ser processados para construir o grafo de conhecimento correspondente e interligá-los às ontologias definidas.

Figura 19 – Tela de busca facetada do HADatAc



Fonte: Elaborada pelo autor

O processo de ingestão semântica de dados implementado no HADatAc implica nos três passos seguintes: (1) codificar os dados e metadados de interesse do estudo científico com base em ontologia(s); (2) mover os dados de arquivos CSVs para bancos de dados do tipo grafo para que possam ser consultados; e (3) garantir que os dados e metadados sejam conectados através de URIs em um grafo RDF.

O mapeamento dos dados para ingestão por meio do *framework* HADatAc é feito com a criação de templates específicos, processados no decorrer das seguintes fases:

- Processamento do "*Semantic Study Design*" (SSD), que prepara a estrutura principal do grafo sendo construído, pela criação das coleções que organizarão os objetos (indivíduos da ontologia) e o relacionamento entre eles e a respectiva coleção;
- Processamento da "*Object Access Specification*" (OAS), que estabelece como os dados são roteados para instanciar os objetos nas coleções definidas na primeira fase do processo, conforme os dados de entrada são lidos, além de outras informações como as permissões de acesso aos dados;
- Processamento do "*Semantic Data Dictionary* " (SDD) que finaliza a anotação semântica dos dados, ou seja, faz o mapeamento dos valores dos objetos instanciados para as ontologias.

O HADatAc acrescenta os requisitos para que o grafo RDF gerado seja considerado um grafo de conhecimento. Tais requisitos (cf. Seção 2.2.1.2), se referem à possibilidade de trabalhar com um volume de dados compatível com o *big data* e à existência de um conjunto de componentes para busca e gestão destes dados (FÄRBER *et al.*, 2016; PAN *et al.*, 2017a).

Um exemplo de um componente do HADatAc para a gestão e busca no grafo RDF pode ser visto na Figura 19. Nesta tela do *framework* é possível visualizar diversas facetas do grafo, selecionando de forma granular um subconjunto do grafo para análise. Esta seleção fornece o acesso a recortes do grafo que correspondam a uma determinada configuração das facetas.

Na Figura 19 pode-se escolher, por exemplo, as coleções de objetos, um conjunto de indicadores definidos na ontologia Base e valores para a dimensão temporal. O exemplo da Figura 19 ilustra uma funcionalidade do HADatAc para a gestão do grafo RDF. O *framework* possui, no entanto, recursos variados para trabalhar com o grafo. Estes recursos estão descritos no manual do usuário[1].

O processo de ingestão no HADatAc está baseado na utilização de dados e metadados armazenados em arquivos CSV. Esse tipo de arquivo foi o assunto da Seção 2.2.2.1. Na seção seguinte relacionamos algumas configurações possíveis dos arquivos de dados a serem ingeridos. Tais configurações tem implicações na forma como os templates devem ser definidos na ferramenta.

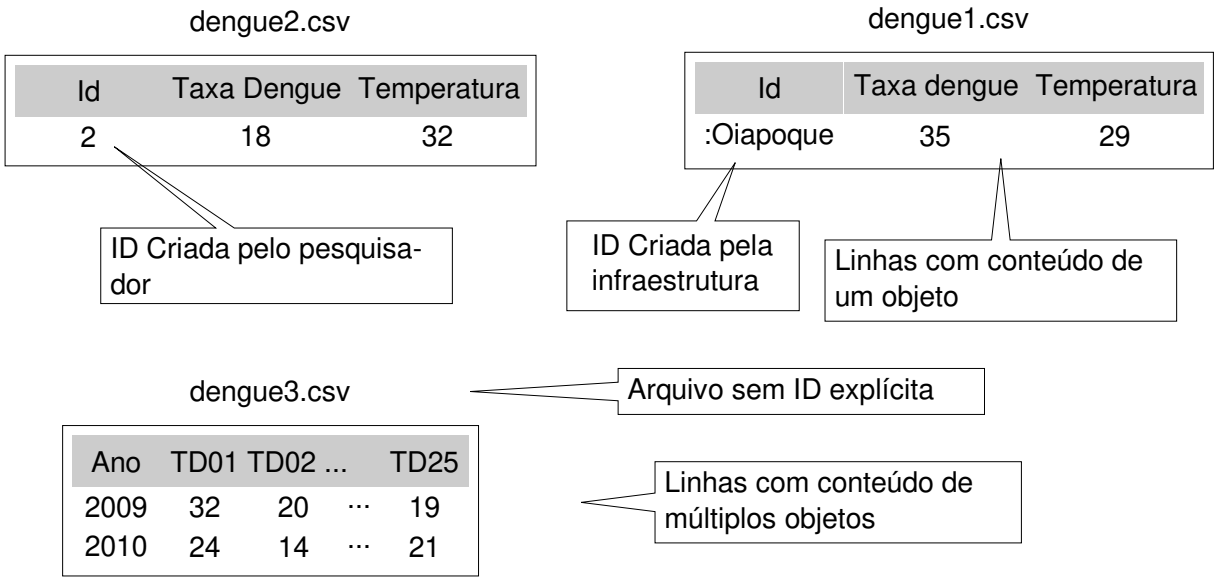
5.2.1 Sobre a diversidade de formatos de arquivos de dados CSV

A entrada do processo de ingestão é composta por arquivos de dados CSV (dados da pesquisa) e arquivos templates (metadados) também em CSV. A Figura 20 apresenta exemplos de arquivos de dados oriundos da pequisa que foi objeto de um experimento de integração usando o método proposto e a ferramenta Hadatac. Esses arquivos ilustram a diversidade de formatos CSV que podem ocorrer.

Por exemplo, nos arquivos dengue1.csv e dengue2.csv, alguns objetos são explicitamente identificados pela coluna "id", que identifica um município. Por outro lado, nenhum

[1] https://github.com/paulopinheiro1234/hadatac/wiki/HADatAc-User-Guide

Figura 20 – Diferentes organizações possíveis de dados em arquivos CSV



Fonte: Adaptada de (PINHEIRO *et al.*, 2018b)

identificador é explicitamente descrito em dengue3.csv. Em vez disso, está implícito como cada coluna caracteriza o objeto descrito. Assim, no arquivo dengue3.csv, temos 26 colunas. A primeira delas, *Ano*, contém os anos em que cada valor da taxa de casos dengue foi levantado. As demais 25 colunas, TD01 a TD25, contém o valor da taxa em 25 municípios diferentes, cuja identificação deve ser estabelecida pelo pesquisador.

Os três exemplos da Figura 20 resumem características que fazem com que os mesmos sejam tratáveis pelo HADatAc, ou seja, arquivos com "ids" explícitas ou não, mas com cabeçalhos com valores únicos para cada coluna. Caso um arquivo de entrada não possua estas características ele poderá ainda ser utilizado no processo de ingestão do HADatAc desde que se realize um processo prévio de transformação (Seção 5.2.5).

Os arquivos dengue1.csv e dengue2.csv mostram outro exemplo de conhecimento implícito, onde o atributo "Taxa Dengue (de casos)" não é uma propriedade direta do objeto identificado pelo *id*, ou seja, o município, mas sim da população daquele município[2]. Ou seja, na medida do interesse da pesquisa pode-se explicitar, no modelo conceitual ontológico, objetos muitas vezes implícitos nos dados. Nesse caso da Dengue, pode-se explicitar que é a "Taxa Dengue" é uma propriedade da população do município, e não do município.

Muitas vezes, o conhecimento implícito sobre relações de objeto e como os valores estão associados aos objetos são codificados de maneiras *ad-hoc* em arquivos de metadados e até em bancos de dados auxiliares. Nesses casos, apenas os pesquisadores que criaram esses arquivos e bancos de dados entendem como seu conteúdo deve ser usado para interpretá-los.

Finalmente, às vezes, observa-se ainda que certos arquivos são inteiramente sobre

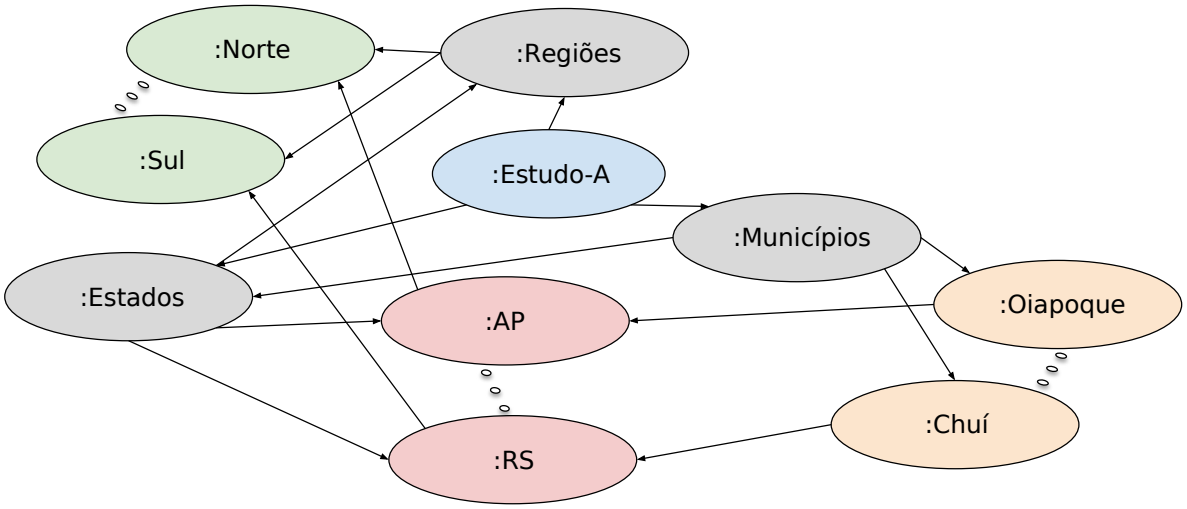
[2] O cálculo da taxa de casos de dengue, para o estudo em questão, equivale ao número de casos da doença para cada 100 mil habitantes

dados de um único objeto, e o conhecimento implícito nesses casos é sobre que arquivo descreve propriedades de que objeto.

A seguir serão apresentados os três principais templates de metadados utilizados para a ingestão de dados no HADatAc. O tratamento de cada um será discutido por meio de exemplos.

5.2.2 Design Semântico de Estudo (SSD)

Figura 21 – Design semântico de estudo



Fonte: elaborada pelo autor

Além de organizar os arquivos de dados relacionados ao estudo, o Design Semântico de Estudo (SSD) descreve um estudo científico em termos de alguns conjuntos de objetos denominados "coleções semânticas de objetos". A Figura 21[3] ilustra um Design Semântico de Estudo (SSD) descrevendo dados de um certo :Estudo-A, em termos de seus objetos, durante a preparação para a ingestão. Neste caso, o pesquisador principal especificou a coleção de objetos semânticos :Municípios, englobando os 5571 municípios brasileiros; uma coleção de estados :Estados e outra de regiões :Regiões. Cada uma contendo as instâncias de municípios, estados e regiões brasileiras.

As coleções se relacionam por meio do conceito de "escopo" (cf., coluna *hasScope*, na Tabela 2). A coleção Estados, está no escopo de Regiões e Municípios no escopo de Estados. Assim, o processamento do SSD irá gerar o grafo RDF mostrado na Figura 21, com cada município associado a seu estado e cada estado à sua região. Todas estas coleções estão vinculadas a coleção principal, que representa o estudo, no caso :Estudo-A.

Os SSDs são expressos em formato tabular, onde cada linha descreve uma coleção de objetos. Um exemplo de SSD é mostrado na Tabela 2. As coleções mostradas na Figura 21

[3] As propriedades relativas as arestas do grafo não foram explicitadas para não complicar em demasia a figura.

Tabela 2 – Especificação do *Design* Semântico de Estudo para o Estudo A

id	**isMemberOf**	**hasScope**	**hasTimeScope**	**cardinality**
:Estudo-A				1
:Municípios	:Estudo-A	:Estados		5571
:Estados	:Estudo-A	:Regiões		26
:Regiões	:Estudo-A			5

Fonte: elaborada pelo autor

Tabela 3 – Especificação do relacionamento entre coleções do SSD da Tabela 2

originalID	**type**	**scopeID**
:Oiapoque	:Município	:AP
:Chui	:Município	:RS

Fonte: elaborada pelo autor

estão especificadas nesta tabela, onde se pode observar o relacionamento entre elas. A coleção de municípios, por exemplo, tem seu escopo definido para a coleção de estados (*hasScope*).

A coluna "cardinalidade" (*cardinality*) especifica quantos objetos serão criados em cada coleção. O vínculo que materializa o relacionamento entre os objetos das coleções é expresso em tabelas específicas para cada uma delas. A tabela 3 ilustra esse vínculo para a coleção de :Municípios. Cada município é identificado pela propriedade hasco:originalID , e está associado a um estado identificado pela propriedade hasco:scopeID

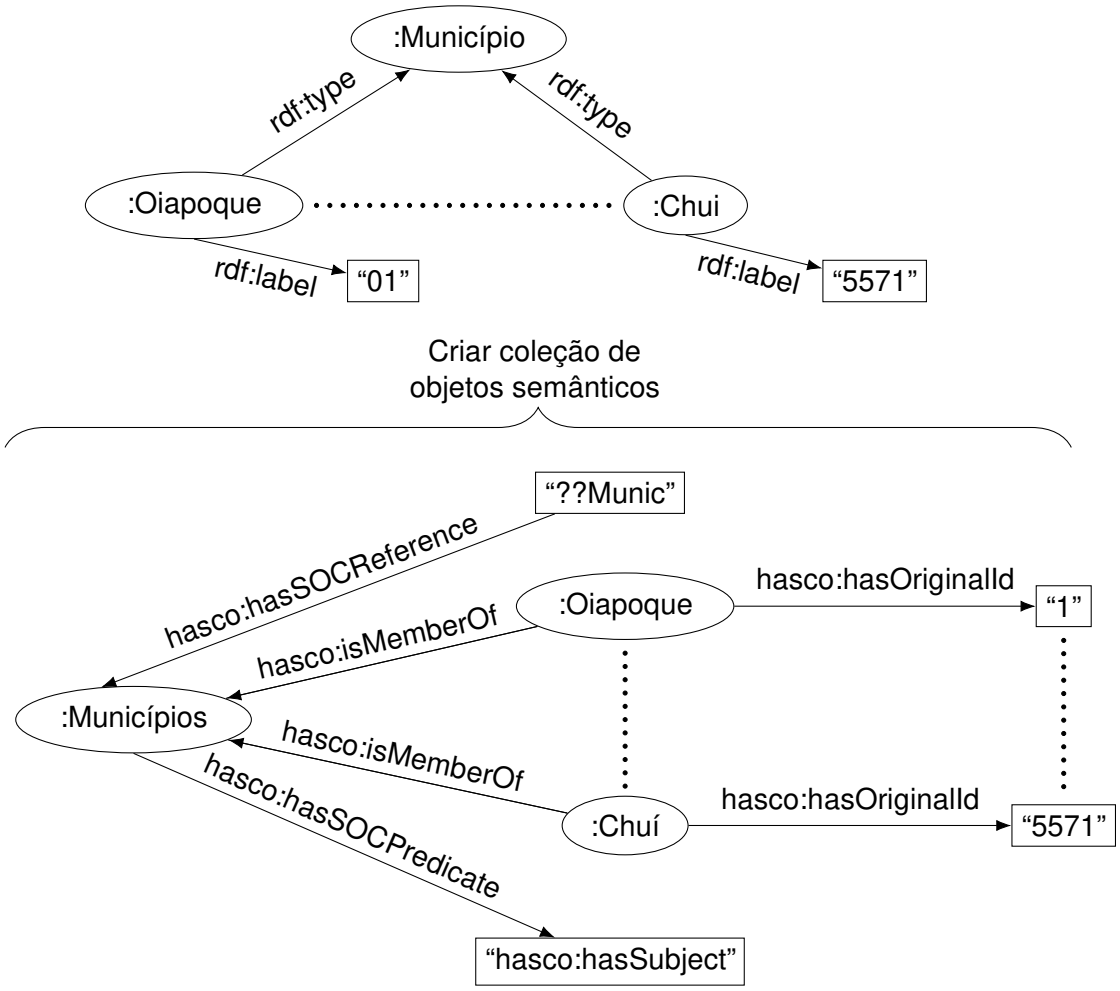
5.2.2.1 Como identificar objetos no grafo?

Um problema relacionado à atribuição de valores a atributos de objetos é a recuperação, no grafo, do objeto correto que terá seus atributos instanciados pelos dados. O exemplo em dengue1.csv na Figura 20 representa o cenário mais simples, pois o objeto já foi identificado usando o URI no grafo RDF (:Oiapoque).

Contudo, quando se trata de dados de estudos científicos é praticamente impossível forçar o pesquisador a utilizar uma forma padrão de identificação dos objetos. É importante que o pesquisador tenha total independência na definição de como os objetos serão identificados. Vale notar que, na prática, o uso de URIs ainda não é considerado um padrão "de fato" para identificar objetos. Em dengue2.csv da Figura 20, por exemplo, os sujeitos são identificados por inteiros armazenados na coluna *id*. Neste caso, quem gerou os dados foi responsável por atribuir e assegurar que os identificadores sejam usados de maneira consistente em vários arquivos CSV.

Qualquer identificador, que não seja definido por um URI, é sempre relativo à existência de um espaço de nomes que forneça o escopo do identificador. Este espaço de nomes é definido através da utilização de uma *coleção de objetos semânticos*, que é um recurso RDF, caracterizado pelo seguinte:

Figura 22 – Criação de uma coleção de objetos semânticos



Fonte: Adaptada de (PINHEIRO *et al.*, 2018b)

- É um recurso RDF do tipo `hasco:ObjectCollection`;
- Ele está conectado a qualquer número de recursos RDF através da propriedade *hasco:isMemberOf* - os membros SOC são chamados *Objetos de estudo*;
- *Objetos de Estudo* de um dado SOC compartilham um tipo comum diferente de `rdf:Thing`;
- *Objetos de Estudo* têm uma propriedade literal *hasco:hasOriginalId* opcional.

Durante o processo de atribuição de entidades a um SOC dentro de uma determinada base de conhecimento, é necessário verificar se os valores `hasco:originalId` da entidade são exclusivos dentro da coleção de membros deste SOC.

A abordagem para agrupar recursos RDF existentes em SOCs consiste no seguinte:

- criar um novo SOC no grafo RDF com um URI conhecido para representar a coleção (por exemplo, `:Municípios`). Esse novo objeto conterá todas as instâncias existentes na coleção que serão identificadas por um determinado valor;

- usar *hasco:isMemberOf* para adicionar os objetos exclusivos ao SOC recém-criado; e
- para os casos em que o identificador do objeto não é um URI, adicionar um atributo *hasco:hasOriginalId* ao objeto. O valor desta propriedade será o *id* original do objeto.

A Figura 22 mostra as instâncias RDF existentes do tipo :Município. A figura também mostra o resultado da criação de um SOC com o URI :Municípios que agrupa os sujeitos existentes. Neste caso, ao ter acesso ao SOC, pode-se percorrer os objetos que pertencem a ele e procurar por um determinado objeto, inspecionando o *hasco:originalID* de cada objeto membro.

A ligação de um SOC a outros objetos é feita pela propriedade *hasco:hasSOCPredicate*. Para ilustrar o uso do *hasco:hasSOCPredicate*, suponha que *hasco:hasSubject* seja um valor para *hasco:hasSOCPredicate* de :Municípios. Neste caso, todos os sujeitos do SOC são conectados ao objeto de estudo através do predicado *hasco:hasSubject*. Além disso, assumindo que o estudo irá analisar dados da população dos municípios, o SOC da mesma pode ser um conjunto de objetos :População Estes SOC pode então ser ligado a cada entidade por meio de uma propriedade de objeto *:hasPopulation*.

O *hasco:hasSOCReference* é um atributo utilizado para se referir a objetos dentro de SOCs. Por exemplo, o valor "??Munic" relaciona os membros do SOC de :Municípios no Dicionário Semântico de Dados (RASHID *et al.*, 2017) que será descrito na Seção 5.2.3.

Objetos sem identificadores explícitos

O único requisito de um arquivo CSV é que os valores tenham um separador (em geral vírgulas). Não se exige que os objetos possuam algum identificador explícito. O arquivo dengue3.csv na Figura 20, replicado na Tabela 4, mostra um exemplo de um arquivo de dados com a taxa da dengue de 25 municípios.

Tabela 4 – Múltiplos objetos na mesma linha

Ano	TD01	TD02	...	TD25
2009	32	20	...	19
2010	24	14	...	21

Fonte: elaborada pelo autor

Nesse caso, está implícito nos rótulos das colunas que "TD01" corresponde a um determinado município, "TD02" a outro município e assim por diante. A falta de identificadores de objetos torna o entendimento de quaisquer conjuntos de dados totalmente dependentes de quem os gerou, uma vez que a estrutura dos dados não está explícita no arquivo de dados. O desafio é encontrar uma forma de descrever explicitamente tal estrutura usando metadados.

A solução para o problema da identificação dos objetos no grafo será apresentada na Seção 5.2.4, que descreve o template Especificação de Acesso a Objetos (OAS). O OAS fornece meios para relacionar os identificadores a objetos cujos atributos espanham-se por colunas ou linhas dos arquivos de dados a serem ingeridos.

Antes de apresentar esta solução, será exemplificado o uso do SDD. Outro template utilizado no processo de ingestão de dados do HADatAc, sendo o principal ponto de interligação para o mapeamento dos dados de entrada nas ontologias selecionadas.

5.2.3 Dicionário Semântico de Dados (SDD)

O Dicionário Semântico de Dados (SDD), introduzido em (RASHID *et al.*, 2017), é uma abordagem para anotar semanticamente os dados do domínio de conhecimento da pesquisa presentes no *dataset* a ser integrado. O SDD mapeia esses dados para os objetos, seus relacionamentos e atributos de uma ontologia ou conjunto de ontologias. Com essa abordagem pode-se expressar dados em relação a ontologias abertas e conhecidas, orientando a criação de um grafo RDF onde os dados ingeridos caracterizam as entidades e seus relacionamentos.

Um SDD contém informações sobre as entidades (conjunto de objetos) e atributos referidos nas colunas dos arquivos de dados CSV, utilizando URIs de ontologias para transmitir essas informações de forma não ambígua e legível por máquina. Ele é composto por 4 planilhas (ou tabelas):

- *Infosheet*: contém informações sobre o estudo e referências às demais tabelas do SDD.
- *Dictionary Mapping*: lista os atributos dos objetos que serão instanciados, informações básicas para o mapeamento.
- *Codebook* : lista faixas de valores permitidos e mapeamentos de códigos (por exemplo: 1=Masculino, 2=Feminino).
- *Timeline*: informações temporais a respeito dos dados.

Tabela 5 – *Dataset* origem para o SDD

Id	Nome	Precipitação	Temperatura	Saneamento	TaxaDengue
2903508	Belo Campo	512	31.8	59.2	6.6
2927408	Salvador	1579	28.6	0.7	227.8

Fonte: elaborada pelo autor

Toma-se como exemplo, a parte de um *dataset* exemplificado na Tabela 5, com 6 colunas, e linhas referindo dados ambientais, socioeconômicas e da taxa de Dengue dos municípios identificados pela coluna “Id”. Um mapeamento deste *dataset* é definido pela Tabela 6.

Na Tabela 6, “??Munic” representa a classe de objetos que tem as propriedades *hasco:originalID* e *:Nome*. Cada uma destas propriedades é mapeada nas colunas “Id” e “Nome” do *dataset* (da Tabela 5). Os valores da coluna “Id” do tipo *hasco:originalID* fazem a ligação com os objetos da Coleção de Objetos Semânticos :Municípios.

A Tabela 5 tem ainda as colunas “Precipitação”, “Temperatura”, “Saneamento” e “TaxaDengue”, que correspondem a propriedades indiretas. Estas propriedades se referem a

Tabela 6 – SDD para a tabela 5 - mapeamento dos atributos

Rótulo	Atributo	*isAttributeOf*
Id	hasco:originalID	??Munic
Nome	:NomeMunicípio	??Munic
Precipitação	:Precipitação	??Superfície
Temperatura	:Temperatura	??Superfície
Saneamento	:Saneamento	??Residências
TaxaDengue	:Taxa-Dengue	??População

Fonte: elaborada pelo autor

Tabela 7 – SDD para a tabela 5 - relacionamento entre os objetos

Rótulo	Entidade	Relação	*inRelationTo*
??Munic	:Município		
??Superfície	:Superfície	sio:inRelationTo	??Munic
??Residência	:Residências	sio:inRelationTo	??Munic
??População	:População	sio:inRelationTo	??Munic

Fonte: elaborada pelo autor

atributos das classes de objetos “??Superfície”, “??Residência” e “??População”. As classes “??Superfície”, “??Residência” e “??População” são propriedades da classe “??Munic” e não podem ser definidas apenas com os dados da Tabela 6.

Para resolver a questão adiciona-se uma segunda tabela que faz o relacionamento entre as classes de objetos, a Tabela 7. Esta tabela identifica as classes de objetos “??Superfície”, “??Residência” e “??População”, que são objetos implícitos identificados no *dataset* de entrada. Estes objetos implícitos são relacionados à classe “??Munic” através do predicado *sio:inRelationTo*. Deste modo, o predicado *sio:inRelationTo* estabelece a conexão entre os objetos implícitos e os objetos da classe “??Munic”.

As tabelas 6 e 7 compõem o *Dictionary mapping* do SDD para o exemplo, nomeado como “SDD-Dengue”. Este exemplo representa um SDD simplificado. Conversões mais complexas podem ser obtidas visto que um SDD para o *framework* HADatAc contém metadados adicionais no próprio *Dictionary mapping* e nos demais artefatos listados anteriormente.

A partir do processamento de dados da Tabela 5 e usando as informações das Tabelas 6 e 7, a ferramenta gera o grafo, serializado no formato *Turtle* (Seção 2.2.1):

```
:MUNIC01
        hasco:originalID      "2903508";
        :NomeMunicípio        "Belo Campo";
        sio:inRelationTo      :SUPERFÍCIE01;
        sio:inRelationTo      :RESIDENCIAS01;
        sio:inRelationTo      :POPUL01.
```

```
:SUPERFÍCIE01
        :Precipitação       512;
        hasco:namedTime     2009;
        :Temperatura        31.8.
:RESIDENCIAS01
        :Saneamento         59.2.
:POPUL01
        :Taxa-Dengue        6.6.
:MUNIC02
        hasco:originalID    "2927408";
        :NomeMunicípio      "Salvador";
        sio:inRelationTo    :SUPERFÍCIE02;
        sio:inRelationTo    :RESIDENCIAS02;
        sio:inRelationTo    :POPUL02.
:SUPERFÍCIE02
        :Precipitação       1579;
        :Temperatura        28.6.
:RESIDENCIAS02
        :Saneamento         0.7.
:POPUL02
        :Taxa-Dengue        227.8.
```

A seguir descreve-se o template que permite estabelecer a interligação entre o SDD e o SSD quando esta não pode ser definida diretamente. Com isto é possível dar uma solução para o tratamento de arquivos CSV como os da Tabela 4, que não definem um identificador explícito.

5.2.4 Especificação de Acesso a Objetos (OAS)

A Especificação de Acesso a Objetos (OAS) mapeia o conteúdo de arquivos de dados CSV para objetos que compõem o grafo. Um detalhe adicional é que alguns objetos especificados pelo SSD irão pré-existir no grafo (já que o SSD é ingerido antes do SDD) e, nesse caso, precisam ser recuperados para que os valores possam ser adicionados às suas propriedades.

Supondo que os arquivos CSV sejam compostos de valores de propriedades de objetos, critérios para identificação desses objetos precisam ser determinados. Conforme a discussão sobre arquivos CSV ilustrada na Figura 20, alguns objetos podem conter um identificador explícito que pode já existir no grafo RDF criado. Tal identificador será o URI do objeto no grafo. No caso de existir um identificador explícito, o arquivo será considerado pronto para ser armazenado no grafo RDF.

Não existe um método padrão para identificar objetos em um grafo RDF a partir de valores em arquivos CSV. A tarefa de criação de novos objetos requer conhecer uma série de informações sobre: (1) como os objetos são identificados em arquivos CSV; (2) como

identificadores são usados para recuperar objetos, se tais identificadores existirem; (3) como os objetos existentes podem ser recuperados se não tiverem identificadores nos arquivos CSV; (4) como procurar objetos existentes nas coleções (por exemplo, procure uma pessoa na coleção de pessoas usando o seu CPF).

Tabela 8 – Exemplo de Especificação de Acesso a Objetos (OAS)

Atributo	Valor	Descrição
máscara de arquivos	dengue*	máscara para os arquivos de dados
dicionário de dados	SDD-dengue	SDD para interpretação de dados
escopo de linha		escopo no nível de linha
escopo de célula		escopo no nível de célula
proprietário-email	joe@example.org	proprietário dos dados
url de permissão	http://example#team	guia de acesso a dados

Fonte: elaborada pelo autor

Uma Especificação de Acesso a Objetos (OAS) (p.ex., Tabela 8) especifica uma solução possível para representar as informações acima. Ela indica como processar os dados de um ou mais arquivos CSV. Uma OAS determina:

- uma máscara de nome de arquivo usada para selecionar os arquivos de dados de entrada (atributo “máscara de arquivos” na Tabela 8);
- o Dicionário Semântico de Dados (SDD) que deverá ser usado para processar o(s) arquivo(s) (atributo “dicionário de dados” na Tabela 8). O SDD descreve como os objetos serão mapeados no grafo resultante (detalhado na Seção 5.2.3);
- o escopo dos objetos, que é composto de “escopo de linha” e “escopo de célula”, que indicam respectivamente se os dados de cada linha do aquivo correspondem à um único objeto ou se correspondem a dados de diversos objetos presentes na mesma linha (o escopo dos objetos é detalhado na Seção 5.2.4);
- o endereço de e-mail da pessoa que é o proprietário real dos dados; e a política de permissão do item para acessar os dados.

A Tabela 8 mostra um exemplo de especificação OAS para processar o arquivo `dengue.csv`da Figura 20; especifica-se que joe@example.org é o e-mail do proprietário do conteúdo (dos dados) que formarão o grafo resultante; que qualquer pessoa com permissão de acesso à http://example#team terá poderá visualizar o conteúdo gerado.

A especificação OAS da tabela determina ainda, que cada arquivo de nome “dengue*” será processado com a ajuda do “dicionário semântico de dados” “SDD-dengue”, apresentado na Seção 5.2.3.

Neste exemplo, o SDD define uma das colunas do arquivo da Figura 20 como o identificador do objeto município interligando-o, assim, a uma coleção do SSD. A OAS permite que esta interligação entre o SSD e o SSD seja feita sem que um identificador esteja explicitado

no SDD. Isso é possível com a definição do escopo dos objetos que será descrito na próxima seção.

Tabela 9 – Exemplo de arquivo de dados com identificadores URI

Id	Nome	Precipitação	Temperatura	Saneamento
:MUNIC01	Belo Campo	512	31.8	59.2
:MUNIC02	Salvador	1579	28.6	0.7

Fonte: elaborada pelo autor

Escopo de objetos

Um CSV de entrada como na Tabela 9, deve ser mapeado para objetos existentes no grafo RDF. No caso da Tabela 9 fica claro qual é o objeto pelo identificador relacionado na coluna "Id", que é o URI do objeto. Neste caso, a OAS deve processar todas as colunas de uma linha do arquivo CSV como atributos de um único tipo de objeto (onde cada linha do arquivo corresponde a uma instância desse tipo de objeto). Estes atributos podem ser diretos ou indiretos. Atributos indiretos de um objeto são atributos de objetos (implícitos) que, por sua vez são atributos deste. Neste exemplo, todos os atributos são diretos e o objeto está definido pelo URI. A forma de tratamento de atributos indiretos, via anotação de objetos implícitos, foi descrita na Seção 5.2.3.

Tabela 10 – Exemplo de arquivo de dados com identificadores numéricos

Id	Nome	Precipitação	Temperatura	Saneamento
2903508	Belo Campo	512	31.8	59.2
2927408	Salvador	1579	28.6	0.7

Fonte: elaborada pelo autor

Por outro lado, na Tabela 10, que possui uma estrutura análoga à da Tabela 9, os identificadores são numéricos e não referenciam um objeto já existente no grafo. A Tabela 10 pode ser processada pela OAS da Tabela 8, mas para tanto será necessário definir o escopo de linha como :Municípios. Com o escopo de linha definido, a OAS orienta a criação correta dos objetos definidos pelo SDD e sua interligação ao SSD.

Tabela 11 – Exemplo de configuração que pode ser tratado com o escopo de célula

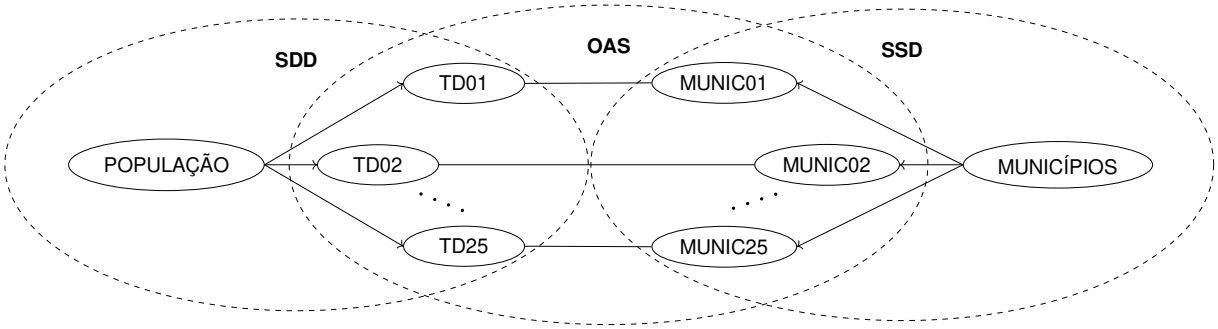
Ano	TD01	TD02	...	TD25
2009	30	27	...	25
2010	23	25	...	32

Fonte: elaborada pelo autor

Escopo de células

Uma outra situação ocorre quando temos valores de objetos distintos na mesma linha. Neste caso define-se um escopo de célula. Isto pode ser verificado na Tabela 11, onde "TD01" é o cabeçalho da coluna para a taxa de casos de dengue no município :MUNIC01 e "TD02" o cabeçalho da coluna para a taxa de casos de dengue no município :MUNIC02 e assim sucessivamente.

Figura 23 – Utilização do escopo de célula na OAS



Fonte: Elaborada pelo autor

Neste caso várias colunas são relativas a um conjunto diferente de municípios, sem associação entre eles. A solução é definir um escopo de linha vazio e um escopo de célula como: "<<TD01, :MUNIC01>, <TD02, :MUNIC02>,..., <TD25, :MUNIC25>>", que consiste em uma lista de nomes de coluna e URIs dos objetos que serão mapeados. Esta definição de escopo permite interligar os valores mapeados pelo SDD à coleção de objetos semânticos do SSD, conforme ilustrado na Figura 23. Com esta definição de escopo de célula na OAS, as instâncias do objeto MUNICIPIO que representarão o escopo para cada célula (colunas "TD01" a "TD25"), são determinadas.

Deste modo, conforme ilustrado na figura, o escopo das colunas "TD01" a "TD25" será dado pelos municípios "MUNIC01" a "MUNIC25", permitindo que a OAS faça a ligação entre a coleção de municípios do SSD com a taxa de casos de dengue da população definida no SDD[4].

Na OAS existem outras formas de definir o escopo de célula, por exemplo, pode-se indicar que todas as colunas mapeadas por um SDD possuem o mesmo escopo. Isto é feito indicando-se o caractere "*" como nome da coluna. Este foi o recurso utilizado no experimento apresentado na Seção 6.2, onde cada SDD tratava dados relativos a sujeitos diferentes.

A possibilidade de indicar o escopo na OAS configura, portanto, uma outra forma de interligar os objetos definidos no SDD às coleções de objetos semânticos do SSD, tornando possível determinar o "Id" (*hasco:originalID*) através da OAS. Desta forma é possível criar o SDD sem a indicação de um "Id" no mesmo.

[4] População é um objeto implícito no SDD que está relacionado aos municípios determinados pela OAS

5.2.5 Eventuais transformações de dados necessárias

Conforme estabelecido na Seção 2.1.4, em certos casos, é necessário realizar uma transformação nos dados antes do seu tratamento. Para o *framework* utilizado, um arquivo como o da Tabela 12 não pode ser ingerido sem que seu formato seja alterado.

Tabela 12 – Taxa de casos da dengue - vários anos na mesma linha

Id	Nome	2007	2008	2009	2010
2903508	Belo Campo	0.0	0.0	6.6	12.5
2927408	Salvador	49.1	84.1	227.8	230.3

Fonte: elaborada pelo autor

A Tabela 12 contém valores para a taxa de casos de dengue, por município e por ano, de 2007 a 2010. O intervalo de anos na primeira linha pode variar, não sendo conveniente definir cada ano como um atributo no SDD. Deve ser incluída uma etapa de transformação para converter o arquivo para outro formato, por exemplo, como o que é apresentado na Tabela 13. Nesta nova tabela a primeira linha contém apenas rótulos, 'Id', 'Nome', 'Ano' e 'Taxa-Dengue', podendo ser processada a partir da criação de um SDD, conforme definido na Seção 5.2.3.

Tabela 13 – Taxa de casos da dengue - após preparação

Id	Nome	Ano	Taxa-Dengue
2903508	Belo Campo	2007	0.0
2903508	Belo Campo	2008	0.0
2903508	Belo Campo	2009	6.6
2903508	Belo Campo	2010	12.5
2927408	Salvador	2007	49.1
2927408	Salvador	2008	84.1
2927408	Salvador	2009	227.8
2927408	Salvador	2010	230.3

Fonte: elaborada pelo autor

5.2.6 Visão geral da ingestão de dados com o HADatAc

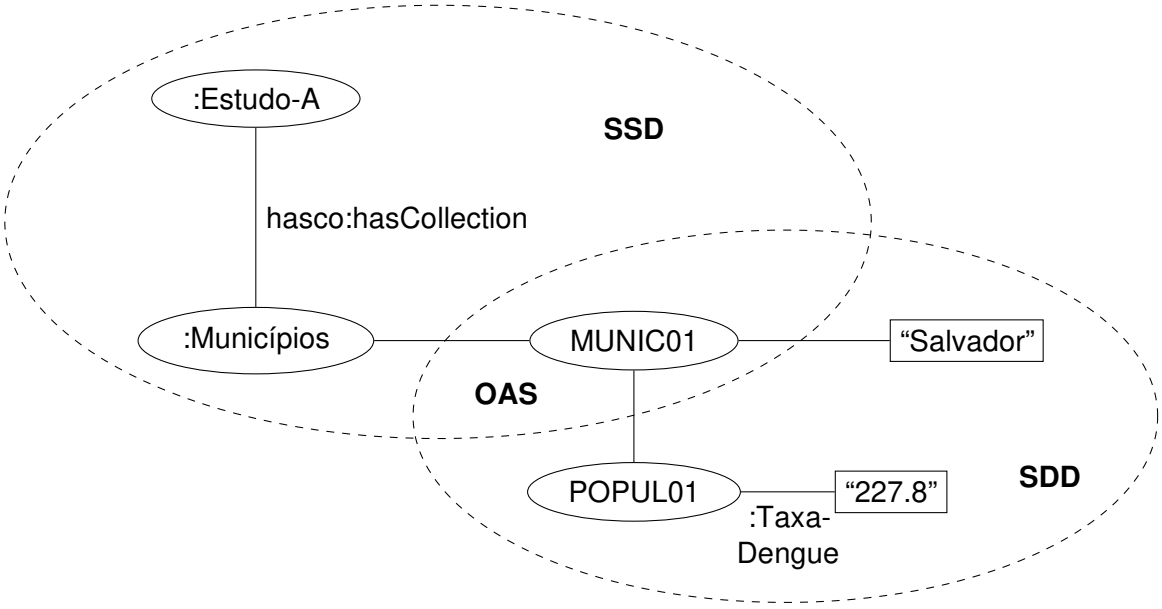
Foi apresentado o processo de Ingestão de Dados implementado no HADatAc (Versão 1.1.12). Como visto, o processo combina o uso de três especificações que, juntas, estabelecem as regras para ingerir o conteúdo dos arquivos de dados de um ou mais estudos científicos para construir o grafo de conhecimento correspondente. Tais especificações são o Design Semântico de Estudo (SSD), o Dicionário Semântico de Dados (SDD) e a Especificação de Acesso a Objetos (OAS).

A um estudo científico podem estar associados mais de um arquivo de dados que deverão ser processados para integração. A ingestão desses arquivos de dados requer especificar o SSD, o SDD e a OAS.

Em síntese, as funções de cada especificação na ingestão de arquivos de dados são as seguintes:

1. O SSD descreve um estudo em termos de seus objetos e coleções de objetos;
2. O SDD descreve o conteúdo de um arquivo de dados em termos de objetos e suas propriedades e pode ser reutilizado para processar vários arquivos de dados;
3. A OAS descreve como os dados de entrada devem ser usados para instanciar objetos descritos semanticamente por um SDD e possíveis mapeamentos para os objetos organizados em coleções de objetos pelo SSD.

Figura 24 – Visão geral da ingestão de dados. SSD, SDD e OAS trabalhando juntos



Fonte: Adaptada de (PINHEIRO *et al.*, 2018b)

A Figura 24 mostra o SSD, o SDD e a OAS trabalhando juntos. O SSD define o ":Estudo-A" e a coleção de objetos semânticos ":Municípios". O SDD, por sua vez, define a criação de instâncias dos objetos :Município e :População Um exemplo de instâncias destes objetos está representado na figura por "MUNIC01" e "POPUL01", correspondendo a uma instância de :Município e :Populaçãorespectivamente. A ligação entre o SSD e o SDD é feita pela criação de instâncias de objetos da coleção :Municípios, ou seja, instâncias de objetos do tipo :Município. Esta criação pode ocorrer no processamento do SDD ou pela definição deste objeto no escopo de célula da OAS.

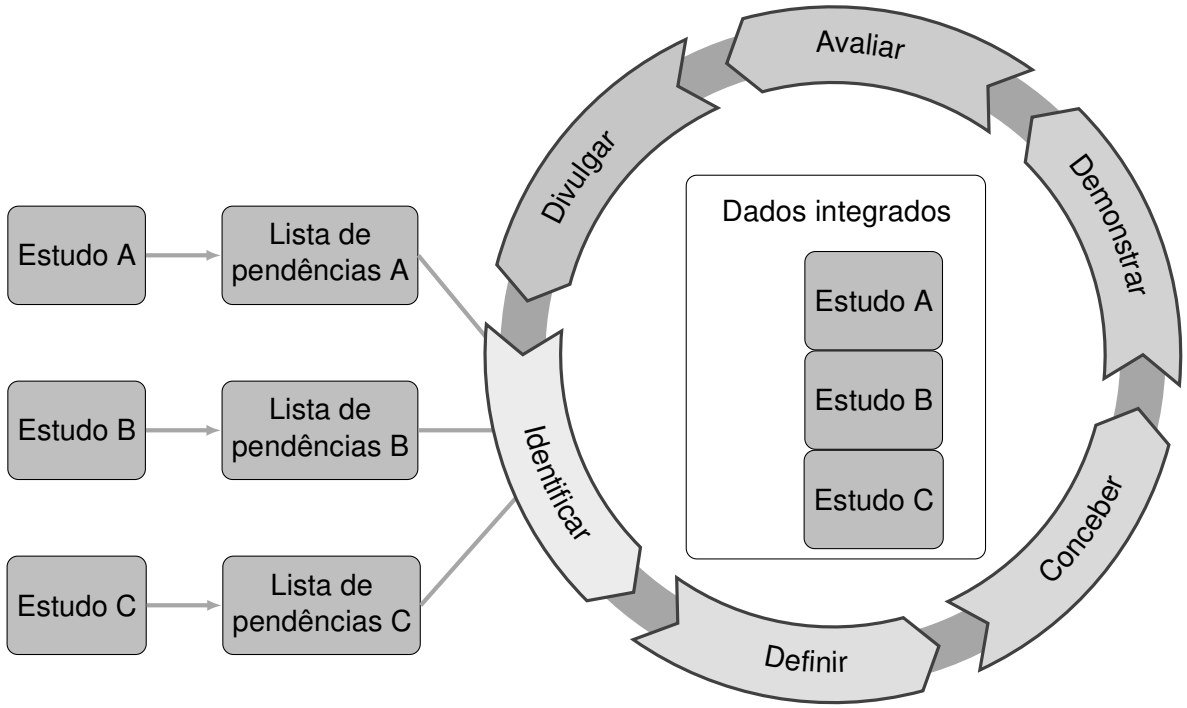
Os objetos podem ser criados no grafo em dois momentos: (1) durante a fase de preparação para ingestão; executada antes de um arquivo de dados ser ingerido (criando coleções de objetos) a partir do SSD e outros arquivos de suporte; e (2) durante o processo de ingestão propriamente dito, que é realizado com o uso do SDD.

5.3 Síntese da integração semântica de dados conforme o método proposto

Uma vez definida a configuração para a integração de dados, o estudo estará pronto para ser preservado no grafo de conhecimento. Tem-se então a possibilidade de investigar os dados com o auxílio dos metadados, que definem "facetas" para a sua visualização e recuperação.

Finalmente é possível analisar os dados obtidos utilizando ferramentas apropriadas. O grafo de conhecimento pode sofrer alterações em iterações subsequentes. Neste caso, os dados do estudo são removidos e reinseridos a cada nova iteração.

Figura 25 – Integração semântica de dados de diferentes estudos



Fonte: elaborada pelo autor

Com o processo de ingestão de dados apresentado na Seção 5.1.6 as informações de diferentes estudos podem ser integradas em um grafo de conhecimento comum. Isto é possível mesmo que o modelo dos dados de entrada seja diferente, desde que exista um mapeamento do mesmo para as ontologias escolhidas. Neste caso as ontologias propostas incluem as ontologias apresentadas no Capítulo 3. Durante o tratamento de um estudo, em sucessivas iterações, apenas os dados deste estudo podem ser removidos e reinseridos, caso a ferramenta utilizada permita, como é o caso do HADatAc. Deste modo, é possível avaliar a integração com outros estudos sem a necessidade de reprocessar os dados destes últimos.

A Figura 25 mostra como ocorre a integração de dados de diferentes estudos de acordo com as diversas etapas do método apresentadas na Seção 5.1.3. A figura apresenta um exemplo de 3 estudos diferentes (A, B e C). Cada um destes estudos possui uma lista de

pendências. No processo de integração, a proposta é executar as iterações isoladamente, a partir da Lista de pendências de cada estudo. Isto é possível porque os subgrafos de cada estudo podem ser tratados de forma independente. Após um conjunto de iterações de todos os estudos, os dados de cada um deles farão parte do grafo de conhecimento gerado. Esta configuração com os dados de todos estudos no grafo de conhecimento é particularmente útil nas iterações de homologação, quando uma versão estável do mapeamento pode ser utilizada. As etapas para a execução das iterações seguem o que foi exposto na Seção 5.1.3.

A escolha da sequência dos estudos a serem tratados pode ser definida pelas características dos mesmos. Se temos uma questão de competência que envolve dados de mais de um estudo, estes precisam ser ingeridos para respondê-la. Por exemplo, se temos uma questão referente à relação da dengue com variáveis ambientais, o *dataset* contendo os dados ambientais deve ser ingerido juntamente com o que contenha os dados sobre a dengue.

5.4 Discussão

O método Odin, apresentado neste capítulo, é a contribuição principal do presente trabalho. Conforme foi descrito nas seções anteriores. Ele tem como base a ADSRM (CONBOY; GLEASURE; CULLINA, 2015), permitindo a evolução do grafo de conhecimento, com a entrega precoce de uma nova versão do grafo a cada iteração.

Além dos especialistas de domínio e dos ontologistas, foram identificados outros atores: os desenvolvedores e os cientistas de dados. O método descreve os papeis de cada um destes atores ao longo do processo. Cada iteração do método é composta por um conjunto de etapas, definidas na Seção 5.1.3. O método se mostrou adequado para evoluir o grafo RDF, sem perder o rigor científico, o qual foi reforçado pela adição das iterações de homologação.

O processo de ingestão de dados envolve além da criação do mapeamento dos dados de entrada no grafo RDF, a evolução da ontologia de domínio. O método prevê a criação de uma ontologia de domínio específica para a anotação do estudo, denominada ontologia Base. A revisão do mapeamento e da ontologia Base a cada iteração define o grafo RDF resultante. Os especialistas de domínio e os ontologistas são os principais responsáveis pela criação e alteração do mapeamento e da ontologia Base.

Sendo que os especialistas estão particularmente interessados em estabelecer as características do grafo que permitam que este possa trazer respostas às questões de competência elaboradas. Tais questões são instrumentos para a modelagem ontológica (Seção 2.2.1.3). Nas etapas de demonstração e avaliação é possível validar o grafo de conhecimento criado a partir de consultas SPARQL ou mesmo com ferramentas ou programas específicos criadas com a participação dos desenvolvedores e cientistas de dados. Esta validação envolve a verificação de que o grafo de conhecimento apresenta respostas corretas às questões de competência.

O método pode ser utilizado em processos de ingestão de dados que permitam a execução iterativa, com a reconstrução do grafo RDF a cada iteração. Sua utilização foi validada para o processo de ingestão utilizando o HADatAc, apresentado na Seção 5.2.

Foram apresentados os principais templates que configuram o mapeamento dos

dados de entrada em uma ontologia para o HADatAc. Estes templates definem a representação da “estrutura da informação” (SSD e OAS) e da “estrutura do conhecimento do domínio” (SDD) de forma separada. Com efeito, as declarações sobre o domínio são de natureza diferentes das declarações sobre como armazenar informações sobre o domínio. Noções como “dado ausente” ou “campo obrigatório” só fazem sentido em relação às estruturas de informação (e não ao domínio) (RECTOR *et al.*, 2019). Deste modo, o SSD define metadados que podem não pertencer propriamente ao domínio, mas à estruturação dos dados conforme os objetivos e tarefas de uma determinada pesquisa científica. O SSD define, por exemplo, a cardinalidade das coleções de objetos, que podem indicar ao pesquisador um dado ausente.

Esta separação obtida pelo uso dos diferentes templates SSD e OAS para a estrutura da informação e SDD para a estrutura do conhecimento do domínio é necessária, como ressaltam Rector *et al.*, e permite isolar o “modelo do conhecimento do domínio”, do “modelo das estruturas de dados e informação” que será utilizado para persistir/armazenar o conhecimento do domínio.

6 APLICAÇÃO DO MÉTODO PARA INTEGRAÇÃO DE DADOS

Os Templates utilizados em cada seção descrita neste caapítulo, correspondendo as diversas iterações bem como a ontologia base estão disponíveis no GitHub[1].

Após decidir por utilizar o HADatAc como ferramenta de ingestão de dados que permitiria validar o método Odin, realizamos dois experimentos. Foram enfrentadas dificuldades devido a detalhes no formato dos templates de metadados apresentados no capítulo anterior. No entanto, com a ajuda dos responsáveis pelo HADatAc no TWC, foi possível ingerir os dados dos primeiros *datasets*. Para isso foi necessário preparar o ambiente, instalando e configurando o HADatAc, conforme descrito a seguir.

6.1 Preparação do ambiente

Foi feito o *download* do código fonte executando-se o processo de instalação de uma versão de desenvolvimento, conforme o guia do usuário [2]. Em seguida, os primeiros experimentos foram executados utilizando-se arquivos de exemplo. Após algumas experiências preliminares e um melhor entendimento do HADatAc, implantou-se uma versão de produção do mesmo em um servidor da UFMG.

Um primeiro experimento serviu como piloto, não somente do processo de ingestão, como também do próprio método. O experimento integrou dados de pulseiras inteligentes e está descrito na próxima seção.

A Integração dos dados das pulseiras coletoras permitiu a criação de uma versão para demonstrar a utilização do método. Além disso, foi possível testar e aprimorar o método aplicado a um processo de ingestão inicialmente mais simples, com poucos dados. Com este experimento definiu-se a abordagem para a utilização do HADatAc de forma iterativa.

6.2 Integração de dados de pulseiras inteligentes

Inicialmente, o método foi aplicado para integrar dados de *datasets* obtidos de duas pulseiras "inteligentes": Mi Band 3[3] e Fit Bit Charge 2 [4]. As decisões de design e os resultados alcançados são apresentados nesta seção. Posteriormente, buscou-se validar mais profundamente o método com a integração de dados epidemiológicos, quando houve o envolvimento de pesquisadores da Fiocruz. Os resultados desta fase estão descritos na Seção 6.3.

Após as primeiras avaliações do HADatAc, estabeleceu-se que o método deveria prever a definição e evolução dos templates e da ontologia "Base" a cada iteração, com a

[1] https://ws1.assis.bhz.br/
[2] https://github.com/paulopinheiro1234/hadatac/wiki/HADatAc-User-Guide
[3] https://www.mi.com/br/mi-band-3/
[4] https://www.fitbit.com/in/charge2

recriação total do grafo. Isto evitaria problemas de consistência da modelagem dos dados. Para esta recriação optou-se por utilizar uma pequena amostra dos *datasets* nas iterações regulares, com a ingestão dos *datasets* completos realizada apenas nas iterações de homologação.

As etapas para a ingestão de dados utilizando o HADatAc, neste caso, deveriam obedecer a esta sequência:

- Criação ou alteração dos templates (SDD, SSD, OAS);
- Criação ou alteração da ontologia Base;
- Remoção do repositório do HADatAc do *dataset* anteriormente ingerido, caso houvesse;
- Remoção, caso necessário, dos templates anteriormente ingeridos, na ordem inversa à utilizada para a ingestão, ou seja, OAS, SDD e SSD;
- Atualização da ontologia Base no HADatAc;
- Atualização dos templates no HADatAc, na sequência normal, ou seja, SSD, SDD, OAS;
- Ingestão do *dataset* na ferramenta, integrando semanticamente os dados pela construção do grafo de conhecimento.

Este primeiro experimento piloto percorreu doze iterações, com a evolução dos templates de ingestão do HADatAc e da ontologia Base. Da primeira à décima iteração utilizou-se os dados da pulseira *Fit Bit Charge 2*. Em seguida, com a realização de duas iterações, adicionou-se ao grafo os dados da *Mi Band 3* (Figura 26).

A facilidade em integrar dados de duas pulseiras diferentes, que forneciam *datasets* em formatos próprios, revelou os primeiros indícios da utilidade do método para integração semântica de dados. Com apenas duas iterações, apesar do formato e da diferença entre os dados gerados por cada uma das pulseiras, foi possível criar um grafo integrando dados obtidos de cada uma delas.

Deste modo foi possível, por exemplo, obter informações comuns à cada pulseira, por exemplo, o número de passos e batimentos cardíacos do usuário em um determinado período de tempo. Além disto, o objetivo que era de testar o método e obter uma compreensão inicial de suas potencialidades, bem como gerar caminhos para aprimorar o processo de ingestão foi atendido.

Nesta demonstração, um objeto vstoi:Deployment (Seção 3.3.1) foi criado para cada pulseira. Com este objeto e sua hierarquia de classes e propriedades foi possível relacionar metadados relativos à cada uma das pulseiras.

Dessa maneira agrega-se, ao grafo final, metadados específicos tais como os sensores existentes nas pulseiras, além de características com modelo da pulseira e outros parâmetros específicos.

[5] Disponível em: https://www.flickr.com/photos Acesso em 04/03/2020

Figura 26 – Pulseiras *Fit Bit Charge 2* e *Mi Band 3*

Fonte: Flickr[5]

A definição dos objetos vstoi:Deploymentdas pulseiras, com as triplas ligadas aos mesmos, está representada na Tabela 14 onde o espaço de nomes padrão édemo-kb

Tabela 14 – Definição dovstoi:Deploymentdas pulseiras utilizadas

Sujeito	Predicado	Objeto
:jose-miband	rdf:type	vstoi:Deployment
	vstoi:hasPlatform	:jose
	hasco:hasInstrument	:miband
:marcello-fitbit	rdf:type	vstoi:Deployment
	vstoi:hasPlatform	:marcello
	hasco:hasInstrument	:fitbit
:miband	rdf:type	vstoi:Instrument
	rdfs:label	Mi Band 3
	hasco:hasDetector	:photoelectric-heart-rate-sensor
	hasco:hasDetector	:adi-ultra-low-power-aceleration
:fitbit	rdf:type	vstoi:Instrument
	rdfs:label	Fit Bit Charge 2
	hasco:hasDetector	:gyroscope
	hasco:hasDetector	:3-axis-accelerometer
	hasco:hasDetector	:compass
	hasco:hasDetector	:ambient-light-sensor
	hasco:hasDetector	:optical-heartbeat-sensor
:jose	rdf:type	sio:Human
:marcello	rdf:type	sio:Human
:photoelectric-heart-rate-sensor	rdf:type	vstoi:Detector
:adi-ultra-low-power-aceleration	rdf:type	vstoi:Detector
:gyroscope	rdf:type	vstoi:Detector
:3-axis-accelerometer	rdf:type	vstoi:Detector
:compass	rdf:type	vstoi:Detector
:ambient-light-sensor	rdf:type	vstoi:Detector
:optical-heartbeat-sensor	rdf:type	vstoi:Detector

Fonte: elaborada pelo autor

Uma amostra dos *datasets* utilizados nesse experimento é representada na Tabela 15 e na Tabela 16, onde estão listados valores obtidos em dois dias de medição de dados de cada pulseira. Essas tabelas foram transpostas com as colunas representando as linhas da tabela original.

Tabela 15 – Amostra do *dataset* obtido a partir da Mi Band

date	2016-01-02	2018-06-06
lastSyncTime	1528254075	1528340511
steps	9728	10296
distance	6874	7334
runDistance	275	310
calories	209	220

Fonte: elaborada pelo autor

Tabela 16 – Amostra do *dataset* obtido a partir da Fit Bit

activityName	Gym	Gym
calories	9	9
activeDuration	428000	3032000
steps	0	4048
logType	tracker	auto_detected
startTime	07/22/15 23:09:05	02/01/16 21:56:09

Fonte: elaborada pelo autor

A Tabela 17 e a Tabela 18 contêm as principais propriedades do mapeamento final, definido na última iteração, para o *dataset* da *Mi Band*. Enquanto que, a Tabela 19 e a Tabela 20 contêm o mapeamento definido para o *dataset* da *Fit Bit*, respectivamente. Estes mapeamentos fazem parte dos SDDs criados para cada um destes *datasets*.

Tabela 17 – SDD para o *dataset* da Mi Band - mapeamento dos atributos

Column	Attribute	attributeOf
date	demo:date	??Activity
lastSyncTime	demo:lastSyncTime	??Activity
steps	demo:steps	??Activity
distance	demo:distance	??Activity
runDistance	demo:runDistance	??Activity
calories	demo:calories	??Activity

Fonte: elaborada pelo autor

Tabela 18 – SDD para o *dataset* da Mi Band - relacionamento entre os objetos

Column	Entity	Relation	inRelationTo
??subject	demo:person		
??Activity	demo:Activity	sio:isPartOf	??subject

Fonte: elaborada pelo autor

Nenhuma propriedade dos SDDs é do tipo `hasco:originalId` e os *datasets* não contêm de fato uma propriedade que identifique unicamente cada instância de dados. No

Tabela 19 – SDD para o *dataset* da Fit Bit - mapeamento dos atributos

Column	Attribute	attributeOf
activityName	demo:activityName	??Activity
calories	demo:calories	??Activity
activeDuration	demo:activeDuration	??Activity
steps	demo:steps	??Activity
logType	demo:logType	??Activity
startTime	demo:startTime	??Activity

Fonte: elaborada pelo autor

Tabela 20 – SDD para o *dataset* da Fit Bit - relacionamento entre os objetos

Column	Entity	Relation	inRelationTo
??subject	demo:person		
??Activity	demo:Activity	sio:isPartOf	??subject

Fonte: elaborada pelo autor

entanto, cada um dos *datasets*, para este exemplo, foi proveniente de uma pulseira pertencente a um indivíduo diferente. Portanto, a determinação do identificador (proprietário da pulseira) estava relacionada ao *dataset* de entrada.

Neste caso, foi possível utilizar o recurso da OAS apresentado na Seção 5.2.4, definindo o mesmo escopo de célula para todas as propriedades de um *dataset*. Assim, a OAS de uma determinada pulseira teve o campo *CellScope* definido como "<<*, demo-kb:subj01>>". Deste modo, demo-kb:subj01 é o identificador do objeto demo:Person, relacionado ao indivíduo que teve seus dados coletados. Utilizou-se o mesmo recurso na OAS referente ao *dataset* da outra pulseira, estabelecendo demo-kb:subj02 como o identificador.

Todos os atributos dos SDDs representados nas tabelas 19 e 17 são do objeto implícito demo:Activity e se ligam à um objeto demo:Person que identifica o indivíduo que deve seus dados coletados. O SSD foi definido de forma que o estudo continha uma coleção de objetos semânticos do tipo demo:Person. Deste modo, a interligação entre esta coleção do SSD aos dados descritos pelo SDD foi feita através da OAS. Permitindo, nesse caso, que não houvesse a necessidade da definição de um identificador no SDD. Utilizando esta abordagem, podem ser processados *datasets* de vários indivíduos, bastando especificar, nos arquivos OAS correspondentes, o escopo de célula adequado para cada *dataset*. Na Tabela 21 temos a OAS da *Fit Bit Charge 2* onde pode-se verificar a definição das características específicas para a ingestão de dados desta pulseira, como o SDD que deve ser utilizado, restrições de segurança de acesso aos dados e o escopo de célula (*cell scope*). Estas mesmas configurações podem ser vistas na OAS da Tabela 22 para a *Mi Band 3*.

Os procedimentos definidos com o experimento simplificado de integração semântica das pulseiras inteligentes serviram de base para estabelecer as condições mínimas para a realização da ingestão de dados oriundos de um projeto de pesquisa desenvolvido por pesquisa-

Tabela 21 – OAS para o *dataset* da Fit Bit

da name	2019-02-FITBIT
study	2019-02
data dict	2019-02-FITBIT
owner email	eugeniog@gmail.com
permission uri	http://localhost:9000/hadatac
cell scope	<<*,demo-kb:sbj01>>

Fonte: elaborada pelo autor

Tabela 22 – OAS para o *dataset* da Mi Band

da name	2019-02-MIBAND
study	2019-02
data dict	2019-02-MIBAND
owner email	eugeniog@gmail.com
permission uri	http://localhost:9000/hadatac
cell scope	<<*,demo-kb:SBJ-subj02>>

Fonte: elaborada pelo autor

dores da Fiocruz. A Seção 6.3 descreve este trabalho que completou a avaliação e consequente validação do método, com a participação dos diversos atores previstos na Seção 5.1.2.

6.3 Criação de um grafo com dados epidemiológicos

A utilização de dados de pesquisa sobre doenças contou com a colaboração de pesquisadores da Fiocruz, os quais foram também os especialistas de domínio para esta fase. Assim estes pesquisadores não somente forneceram os dados, mas participaram ativamente das iterações, em todas as etapas do método, conforme definidas na Seção 5.1.3. Esta participação englobou a etapa de *design*, quando os especialistas de domínio elaboraram questões de competência (Seção 2.2.1.3) que auxiliaram na definição de características do grafo de conhecimento a ser gerado.

Com o grafo gerado, os pesquisadores tiveram contato com os resultados o mais cedo possível, identificando as necessidades de melhorias evolutivas na ontologia Base (e, em consequência, também no grafo de conhecimento). Isto incluiu a participação dos mesmos nas atividades de busca por conceitos representados em ontologias de domínio, capazes de adicionar semântica formal aos dados da pesquisa.

Inicialmente, trabalhou-se com um *dataset* com dados socioeconômicos e ambientais dos 5571 municípios brasileiros, associado a dados relativos aos casos reportados da dengue nos mesmos. Este *dataset* continha dados dos anos de 2000 a 2010. O modelo representado pelos templates e pela ontologia Base foi criado de forma incremental com a realização das iterações relacionadas na Seção 6.3.1. Nas quais todos os atores participaram, colaborando para a evolução do grafo de conhecimento.

6.3.1 Histórico de execução das iterações

O método apresentado no Capítulo 5 foi aplicado com a execução das etapas estabelecidas no mesmo (Seção 5.1.3). Num primeiro momento, observando as premissas do método, o *dataset* foi ingerido no HADatAc utilizando um modelo conceitual simplificado, porém funcional, i.e., a primeira versão da ontologia Base. A ontologia Base e o grafo de conhecimento foram então evoluindo de acordo com os requisitos estabelecidos pelos especialistas de domínio e ontologistas nas iterações seguintes, incorporando, quando necessário, alterações identificadas pelos desenvolvedores e cientistas de dados. No modelo do *dataset*utilizado, os valores anuais estão definidos para cada município, os quais estão associados a um estado da Federação e a uma região. Um resumo das decisões tomadas em cada iteração será apresentado a seguir.

- **Iteração 0**: Criação do grafo inicial, com um único SOC (Coleção de estados) no SSD.
- **Iteração 1**: Observou-se que a abordagem deveria ser alterada para considerar a maior granularidade do *dataset*, incluindo anos e municípios. Deste modo, adicionou-se ao SOC as coleções de municípios e anos. Para esta abordagem enfrentou-se problemas de infra-estrutura, pois a cardinalidade das coleções aumentou consideravelmente. Como o HADatAc armazena as coleções do SSD no *Blazegraph* este problema foi ainda maior, visto que o mesmo não apresenta um desempenho satisfatório para grandes volumes de dados.
- **Iteração 2**: Utilização de apenas duas coleções: anos (hasco:TimeCollection) e estados para evitar problemas com a infra-estrutura de software.
- **Iteração 3**: Aprimoramento dos templates de ingestão, mantendo as coleções definidas na iteração anterior.
- **Iteração 4**: Adição de uma coleção de regiões ao SSD.
- **Iteração 5**: Revisão do SSD com a adição de uma coleção de municípios, removendo-se a coleção de anos.
- **Iteração 6**: A solução adotada na iteração anterior não foi satisfatória, pois existia a necessidade de identificar os dados temporais. Para re-introduzir a dimensão temporal e manter todas as coleções já definidas no SSD (regiões, estados e municípios), utilizou-se um recurso do SDD que permite definir um dos seus atributos como sendo um valor temporal. Para isso o atributo da coluna que representa o ano foi definido como *hasco:namedTime*. Isto resolveu o problema de infra-estrutura.
- **Iteração 7**: Refinamento do grafo final com a definição de objetos implícitos no *dataset* no SDD (identificados 10 novos objetos).
- **Iteração 8**: Mapeamento da ontologia Base em ontologias de referência. Realizou-se, nessa iteração, um trabalho conjunto com os especialistas de domínio, onde foram identificados algumas ontologias e classes que poderiam ser reutilizadas.

- **Iteração 9**: Na etapa de demonstração, foi detectado um problema que impedia a criação de indicadores por estado e região. Esta é uma limitação do HADatAc, que não permite indicadores para coleções do SSD. Foi criada uma tarefa na Lista de pendências para alterar o *framework*.
- **Iteração 10**: Para contornar o problema da iteração anterior, optou-se por adotar a estratégia de definir identificadores para estado e região no SDD. Deste modo, foi possível obter um grafo final com todos os indicadores necessários para responder às questões de competência.

Numa segunda fase, a integração de dados com o grafo já existente foi realizada a partir de um *dataset* com casos de esquistossomose nos municípios brasileiros. A motivação para trabalhar com este segundo *dataset* foi comprovar a aplicação do método para integrar dados de diferentes estudos. Inicialmente foram criados templates adicionais para tratar o *dataset* com os dados da esquistossomose. Ingerindo o *dataset* com esses novos templates foi possível gerar um sub-grafo integrado ao grafo gerado anteriormente.

Assim, os dados socioeconômicos e ambientais, já existentes no grafo de conhecimento, foram associados ao sub-grafo do novo *dataset*. Posteriormente, em uma nova iteração, optou-se por fazer uma alteração no esquema inicial dividindo-se a ingestão em quatro sub-grafos: dados socioeconômicos, ambientais, dados da dengue e dados da esquistossomose. Com esta divisão, tem-se o isolamento dos sub-grafos que podem ser gerenciados de forma independente.

Tabela 23 – Amostra do *dataset* com dados socioeconômicos e ambientais

UF	BA	BA
REGIAO	NE	NE
CD_GEOCMU	2900801	2927408
NM_MUNICIP	Alcobaça	Salvador
biome	4	4
year	2010	2010
pop	21271	2675656
analf	23.59	3.97
san	33	0.1
temp	29.16	30.22
prec	1284	1772
ndvi	0.721	0.488
ndwi	0.553	0.293

Fonte: elaborada pelo autor

Uma amostra contendo o cabeçalho e duas linhas com as principais propriedades de cada um dos três *datasets* (dados socioeconômicos e ambientais, da dengue e da esquistossomose) pode ser vista nas tabelas 23, 24 e 25. Estas tabelas estão transpostas para uma melhor formatação.

Tabela 24 – Amostra do *dataset* com número de casos de dengue anuais

CD_GEOCMU	2900801	2927408
number	12	6161
year	2010	2010

Fonte: elaborada pelo autor

Tabela 25 – Amostra do *dataset* com número de casos de esquistossomose anuais

CD_GEOCMU	2900801	2927408
number	1	18
year	2007	2010

Fonte: elaborada pelo autor

Tabela 26 – Mapeamento do SDD para o *dataset* com dados socioeconômicos e ambientais

Column	Attribute	attributeOf	inRelationTo
UF	:UF	??state	
REGIAO	:Region	??state	
CD_GEOCMU	hasco:originalID	??municipality	
NM_MUNICIP	:municipalityName	??municipality	
biome	:Biome	??municipality	
year	hasco:namedTime	??timeInterval	
pop	:TotalPopulation	??municipality	
analf	:ratio	??iliterates	??totalPopulation
san	:residencesWithInadequateSanitation	??totalResidences	??municipality
temp	:averageTemperature	??surface	??municipality
prec	:averagePrecipitation	??surface	??municipality
ndvi	:normalizedDifferenceVegetationIndex	??vegetation	??municipality
ndwi	:normalizedDifferenceWaterIndex	??waterBodies	??municipality

Fonte: elaborada pelo autor

Tabela 27 – Relacionamento entre os objetos do SDD para o *dataset* com dados socioeconômicos e ambientais

Column	Entity	Relation	inRelationTo
??municipality	:Municipality	sio:part_of	??state
??totalPopulation	:TotalPopulation	:isTotalPopulationOf	??municipality
??totalResidences	:TotalResidences		??municipality
??surface	:Surface		??municipality
??vegetation	:Vegetation		??municipality
??waterBodies	:WaterBodies		??municipality
??iliterates	:Iliterates	sio:part_of	??totalPopulation
??timeInterval	sio:TimeInterval		
??state	:State		??municipality

Fonte: elaborada pelo autor

Para ilustrar o *Dictionary Mapping* (mapeamento) e o relacionamento entre os objetos implícitos para cada um dos *datasets* no SDD, temos para o *dataset* da Tabela 23, o mapeamento exibido na Tabela 26 e o relacionamento entre os objetos na Tabela 27. Para o *dataset* da Tabela 24 o mapeamento está na Tabelas 28 e o relacionamento na Tabela 29. Finalmente, para o *dataset* da Tabela 25 o mapeamento está na Tabela 30 e o relacionamento na Tabela 31. É interessante observar que, os SDDs apresentados já acrescentam semântica aos dados, definindo mais claramente os atributos e relacionamentos entre eles. Futuras evoluções da ontologia Base permitem adicionar um significado ainda maior, interligando os conceitos à ontologias de domínio. Todas os três *datasets* são vinculados ao mesmo identificador (*hasco:originalID*).

Pode-se observar nos SDDs definidos, que o componente ligado à dimensão temporal (*hasco:namedTime*) é o atributo da coluna *year*, definindo o ano para todas as demais colunas de uma linha. Já o *hasco:originalID* é o atributo da coluna "CD_GEOCMU" que define um código único para cada município. Assim, estas duas colunas determinam um identificador único para os demais atributos.

Tabela 28 – Mapeamento do SDD para o *dataset* com dados da dengue

Column	Attribute	attributeOf
CD_GEOCMU	hasco:originalID	??municipality
number	:totalDengueCases	??totalPopulation
rate	:totalDengueCasesOn100Thousand	??totalPopulation
year	hasco:namedTime	??timeInterval

Fonte: elaborada pelo autor

Tabela 29 – Relacionamento entre os objetos do SDD para o *dataset* com dados da dengue

Column	Entity	Relation	inRelationTo
??municipality	:Municipality		
??totalPopulation	:TotalPopulation	:isTotalPopulationOf	??municipality
??timeInterval	sio:TimeInterval		

Fonte: elaborada pelo autor

Tabela 30 – Mapeamento do SDD para o *dataset* com dados da esquistossomose

Column	Attribute	attributeOf
CD_GEOCMU	hasco:originalID	??municipality
number	:totalSchistosomiasisCases	??totalPopulation
rate	:totalSchistosomiasisOn100Thousand	??totalPopulation
year	hasco:namedTime	??timeInterval

Fonte: elaborada pelo autor

Tabela 31 – Relacionamento entre os objetos do SDD para o *dataset* com dados da esquistossomose

Column	Entity	Relation	inRelationTo
??municipality	:Municipality		
??totalPopulation	:TotalPopulation	:isTotalPopulationOf	??municipality
??timeInterval	sio:TimeInterval		

Fonte: elaborada pelo autor

6.3.2 Integração de novos dados da esquistossomose

Após avaliar os dados relativos à esquistossomose, foi detectado pelos especialistas de domínio, que ocorreu uma mudança no protocolo de notificação da informação dos casos desta doença. O que se verificou foi que, a partir do ano de 2006, tal notificação passou a ser compulsória. Com o objetivo de investigar se ocorreu uma mudança no volume de casos da doença notificados após esta alteração, foi gerado um novo *dataset* pelos especialistas de domínio com dados de casos da doença de 2007 a 2017. Uma amostra deste *dataset* é apresentada na Tabela 32. Observando a configuração da tabela, pode-se constatar, devido à configuração dos dados de entrada, a necessidade de uma transformação nos dados para ingestão no HADatAc. Esta transformação envolveu, além da mudança de formato (Seção 5.2.5), a necessidade de gerar o dígito verificador dos municípios para manter o mesmo padrão dos *datasets* dos experimentos anteriores. Com estas alterações foi possível ingerir os dados com apenas uma iteração.

Para esta ingestão foi necessário adotar uma estratégia para a integração dos dados, pois existia uma interseção entre o *dataset* existente e o novo *dataset* (o primeiro *dataset* continha dados de 2000 a 2010 e o segundo de 2007 a 2017). Deste modo, optou-se por criar no SDD, para este novo *dataset*, um predicado diferente para o total de casos de esquistossomose. Assim *:totalSchistosomiasisCases* foi substituído pelo predicado *:totalSchistosomiasisCasesStudy2* no SDD. Este novo predicado foi definido na ontologia Base como uma sub-classe da propriedade *:totalSchistosomiasisCases* do SDD anterior. A separação dos valores relativos ao total de casos de esquistossomose de cada experimento, através de predicados distintos, permite que sejam feitas análises com os dados obtidos em cada experimento separadamente ou em conjunto utilizando funções de agregação para sumarizar os valores. Este tipo de abordagem mostra a versatilidade de utilizar grafos de conhecimento para a integração de dados científicos.

Tabela 32 – Amostra do *dataset* com número de casos de esquistossomose de 2007 a 2017

Município	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
110011	21	12	16	27	2	3	-	2	5	12	17
110012	15	6	7	2	-	3	-	5	4	1	7

Fonte: elaborada pelo autor

O novo *dataset* não possuía informação da taxa de casos de esquistossomose

(que pode ser calculada a partir da população). Além disso, a partir do ano de 2011, os dados socioeconômicos e ambientais não estavam no grafo de conhecimento. Apesar destas limitações, os dados foram corretamente ingeridos e integrados ao grafo existente. Ressalta-se ainda, que os dados faltantes podem ser inseridos a qualquer momento, estendendo as informações representadas pelo grafo.

7 AVALIAÇÃO DO MÉTODO E DISCUSSÃO DOS RESULTADOS

Neste capítulo serão analisados, avaliados e sintetizados os resultados obtidos com o uso do método a partir dos experimentos apresentados no Capítulo 6.

7.1 Modelagem ontológica com o método proposto

A modelagem ontológica do grafo RDF foi realizada por meio de modificações, organizadas em ciclos iterativos, dos artefatos de mapeamento do HADatAc e da ontologia Base. A cada ciclo, os dados presentes nos arquivos de entrada foram mapeados para conceitos de ontologias (i.e., anotados semanticamente). Segundo Noy, McGuinness *et al.* (2001), a modelagem ontológica segue um conjunto de regras fundamentais:

- Não existe uma maneira correta de modelar um domínio - sempre existem alternativas viáveis.
- O desenvolvimento de ontologias é necessariamente um processo iterativo.
- Os conceitos na ontologia devem corresponder aos conjuntos de objetos e aos seus relacionamentos no seu domínio de interesse. É mais provável que sejam substantivos (objetos) ou verbos (relacionamentos) em frases que descrevem o domínio.

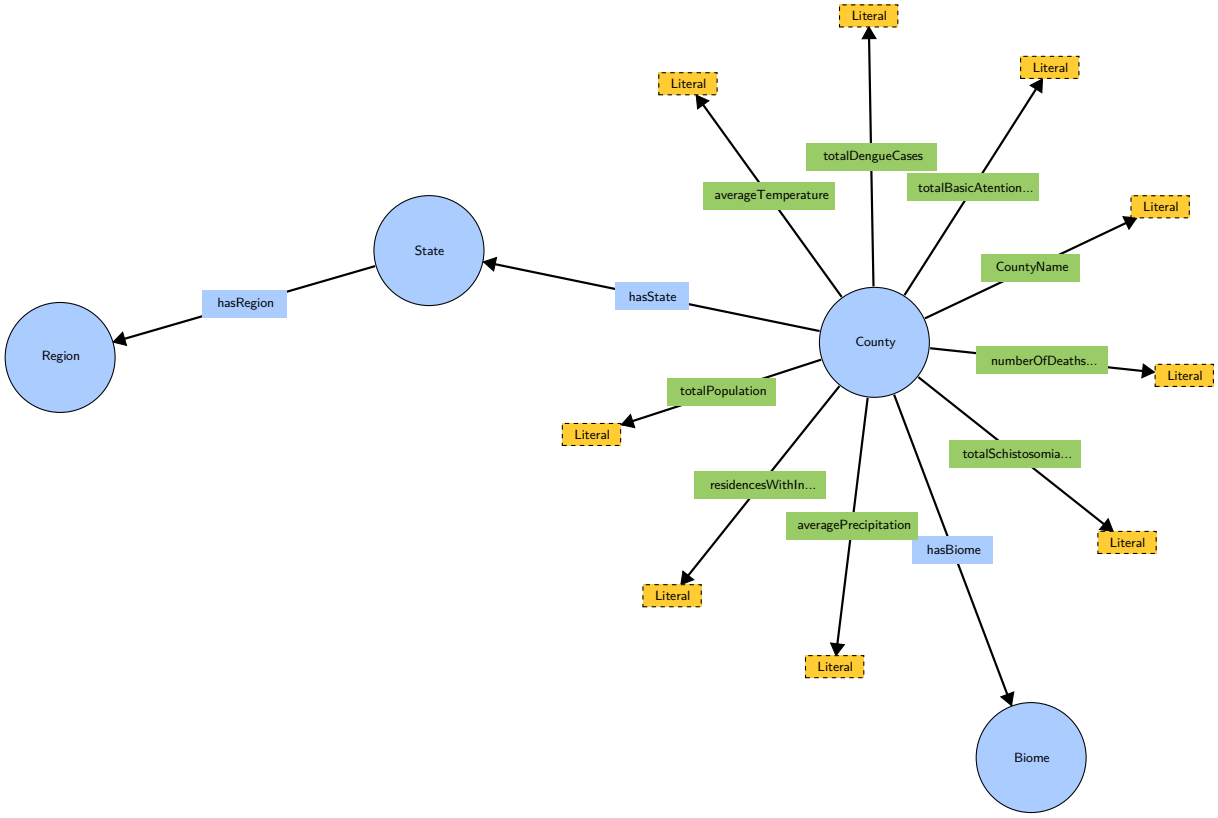
O método sistematiza a modelagem ontológica, permitindo anotar a semântica contida nos conjuntos de dados científicos (*datasets*) que se deseja integrar. As alternativas de modelagem podem ser ajustadas nas iterações previstas no método, onde a cada iteração, o grafo é recriado do início. Já a correspondência da ontologia aos objetos e relacionamentos do domínio é garantida pela participação contínua dos especialistas de domínio.

No sentido de tornar ágil o processo de anotação, considera-se inicialmente que os metadados são os cabeçalhos das colunas dos *datasets* e comporão a versão inicialmente primitiva da ontologia Base, ou seja, a primeira versão da “ontologia” Base é propositadamente bastante simples e rudimentar. Inicialmente, esta ontologia nada mais é do que a lista dos cabeçalhos das colunas dos *datasets*. O método se refere a esta lista como equivalendo à uma ontologia porque ela será enriquecida à medida que os ciclos iterativos forem ocorrendo. Assim, o que inicialmente era uma simples lista de termos, transformar-se-á, no decorrer do processo de anotação, em uma ontologia segundo a acepção plena do conceito, i.e., a “uma conceitualização formal, explícita e compartilhada” (segundo definição clássica de Gruber), contendo classes, relações e instâncias.

Assim, de acordo com o método, os especialistas de domínio e ontologistas vão definindo, a cada iteração, a evolução da ontologia Base, aprimorando a modelagem conceitual a cada iteração. Neste processo de evolução, conceitos de outras ontologias (preferencialmente

reconhecidas) podem ser incorporados à ontologia Base, conforme o princípio de reutilização de conceitos Mireot (COURTOT *et al.*, 2011).

Figura 27 – Versão inicial - ontologia Base

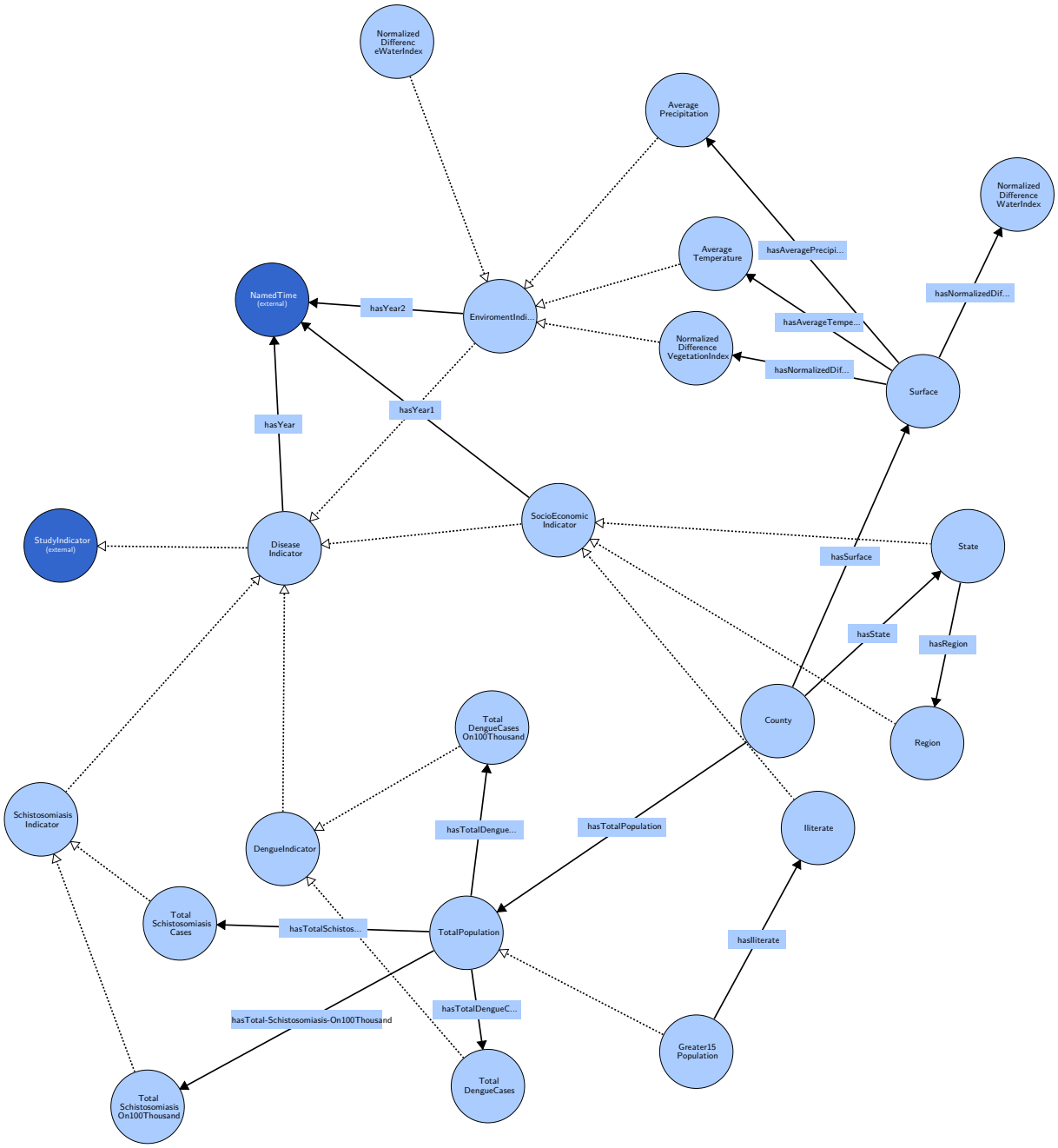


Fonte: elaborada pelo autor

Como descrito acima, de acordo com o método, na fase inicial da modelagem, a ontologia Base foi desenhada para representar os atributos (colunas do *dataset*) do mapeamento de forma simples, sem referenciar conceitos de outras ontologias. Deste modo, uma versão inicial da ontologia pode ser obtida precocemente. A cada iteração, observando-se as premissas estabelecidas por um conjunto de questões de competência, o mapeamento e a ontologia base são revistos, utilizando-se os princípios do método, que determina a manutenção de iterações de curta duração e a entrega precoce de uma nova revisão funcional da ontologia em cada uma dessas iterações.

No experimento realizado, utilizando dados socioeconômicos, ambientais e da incidência anual da dengue e esquistossomose nos municípios brasileiros, a ontologia Base foi desenhada para representar os atributos do mapeamento de forma simples, sem referências à outras ontologias. Os valores anuais foram definidos para cada município, os quais estão associados a um estado da Federação e a uma região. A cada iteração o mapeamento e a ontologia Base foram revistos, gerando-se uma nova versão da ontologia. Na versão inicial da ontologia um município estava relacionado à classe estado que, por sua vez, relacionava-se a uma determinada região. A Figura 27 ilustra estes conceitos a partir de um sub-grafo da ontologia Base, a ontologia foi escrita com os nomes das classes, atributos e relações em

Figura 28 – Versão final da ontologia



Fonte: elaborada pelo autor

inglês. Pode-se observar na figura que cada município (*County*) concentrava todos demais atributos tais como população total (*totalPopulation*), precipitação média (*averagePrecipitation*), temperatura média (*averageTemperature*), casos totais de dengue (*totalDengueCases*), dentre outros.

A evolução da ontologia foi direcionada pelas questões de competência que nortearam a definição das classes, atributos e relações da mesma. Assim, a cada iteração, o conjunto de classes foi se definindo, chegando-se a uma versão mais elaborada da ontologia. Além das decisões de modelagem definidas a partir das questões de competência, a cada iteração

foram sendo definidas características da ontologia que poderiam aprimorá-la e garantir maior flexibilidade na integração de dados adicionais.

Identificou-se, por exemplo, que os dados socioeconômicos e ambientais poderiam ser úteis na análise de outras doenças. Deste modo, o modelo foi revisado colocando-se todos os dados relacionados a indicadores socioeconômicos e ambientais como sub-classes de SocioEconomicIndicatore EnviromentIndicator, respectivamente. Com esta alteração a manutenção da ontologia foi simplificada, facilitando, por exemplo, a adição de dados de outras doenças com a reutilização dos valores de indicadores socioeconômicos e ambientais.

Ao término das iterações obteve-se uma versão final da ontologia. Assim, o modelo com as classes, relacionamentos e atributos, nesta versão da ontologia, deixou de ser um modelo simplificado definido pela ontologia Base. Portanto, a ontologia final define uma hierarquia de classes e atributos mais completa, representando o resultado dos esforços realizados nas diversas iterações para aprimorar a ontologia Base. Um sub-grafo com parte das classes e relacionamentos da ontologia está representado na Figura 28. Os atributos foram omitidos para simplificar a figura. As setas com linhas pontilhadas indicam um relacionamento entre uma classe e sua sub-classe, por exemplo, DengueIndicatoré uma sub-classe de DiseaseIndicator. As classes DiseaseIndicator, SocioEconomicIndicatore EnviromentIndicatorsão superclasses de conjuntos distintos de classes, que agrupam informações das doenças (dengue e esquistossomose), socioeconômicas e ambientais. Conforme relatado na Seção 6.3.1, com essa configuração, é possível gerenciar o grafo manipulando os sub-grafos de forma independente. Assim, a adição de uma nova doença para análise pode ser realizada de forma simples.

7.2 Análises do grafo após a ingestão

A validação das questões de competência foi feita inicialmente com o uso da linguagem de consulta SPARQL, no entanto, o volume de dados exigiu a utilização de artefatos que foram criados pelos desenvolvedores e cientistas de dados. O grafo de conhecimento foi explorado com estes artefatos, gerando-se informações capazes de auxiliar os especialistas de domínio na compreensão do relacionamento entre as diversas variáveis definidas como indicadores na ontologia Base.

7.2.1 Respostas às questões de competência

A validade do método e sua aplicação pôde ser comprovada com o atendimento às questões de competência. Apresenta-se a seguir um exemplo concreto de elucidação de destas questões:

1. Qual a associação da distribuição da taxa da dengue na região Nordeste com o ano?
2. Qual a associação da Dengue com as variáveis temperatura e precipitação nos municípios da região Centro-Oeste em 2010?

Para responder a questão de competência (1) uma consulta SPARQL é suficiente para gerar uma resposta satisfatória. Para a questão (2), a análise é mais complexa, porém ela pode ser respondida com os dados do grafo gerado e o auxílio de técnicas de Ciência de Dados. Estas técnicas podem definir, por exemplo, uma função matemática entre os casos de dengue e as variáveis ano, região, temperatura e precipitação. É possível extrair esta função, com o auxílio de um sistema de aprendizado de máquina supervisionado (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). A relação entre outros indicadores e as variáveis de interesse podem ser obtidas do mesmo modo. Trabalhos futuros utilizando o método Odin podem se beneficiar dos recursos de aprendizado de máquina validando as questões de competência na fase de avaliação.

Vale ressaltar que, resultados obtidos diretamente a partir de funções de aprendizado de máquina, devem contar com a análise crítica dos pesquisadores. Pois, deve-se levar em conta que um algoritmo de aprendizado é suscetível a inconsistências nos dados, que podem ser encontradas pelo ser humano (ATTENBERG; IPEIROTIS; PROVOST, 2011; NUSHI *et al.*, 2017).

7.3 Avaliação do método

Os experimentos realizados mostram que foi possível realizar a integração semântica de dados científicos com o método proposto. Foi gerado um grafo de conhecimento a partir da colaboração dos atores definidos pelo método, nas diversas etapas das iterações, de acordo com o que foi estabelecido na Seção 5.1.3.

Durante a execução do método, foram utilizadas as iterações de homologação. Nestas iterações não se adicionou novas tarefas à Lista de pendências, criando-se uma versão do grafo de conhecimento com o *dataset* completo. Além disso, buscou-se desenvolver ferramentas que auxiliassem na validação do grafo gerado.

A opção por trabalhar com a versão completa do *dataset* somente nas iterações de homologação se relaciona ao tempo de ingestão de uma versão completa, que diminuiria a rapidez nas alterações e testes durante o desenvolvimento de uma iteração regular.

Uma única instalação do HADatAc permite trabalhar com duas instâncias independentes do grafo de conhecimento. Assim, a versão do grafo da última iteração de homologação podia ser mantida enquanto se trabalhava em uma nova iteração. Na interface do HADatAc, estas duas instâncias do grafo são definidas a partir de dois modos de operação: o modo padrão e o modo *sandbox*. Utilizou-se o modo padrão para as iterações de homologação e o *sandbox* para as demais.

Com esta abordagem garantiu-se a existência de uma versão estável e uma versão de trabalho do grafo durante todo o processo. A capacidade de recriar o grafo a cada iteração, aliada à utilização de um subconjunto do *dataset* de trabalho, permitiu executar todas as etapas previstas pelo método. Todos os atores definidos no método participaram, sendo possível certificar a importância de cada um destes no mesmo. Foi possível verificar que a divisão do processo em iterações foi vantajosa, pois os resultados obtidos em cada iteração, mesmo

que preliminares, eram disponibilizados com frequência e permitiam identificar problemas que eram adicionados precocemente à Lista de pendências. Assim, problemas que somente seriam identificados no final da ingestão de dados surgiram em versões iniciais, sendo tratados de acordo com as prioridades dos especialistas de domínio.

7.4 Divulgação

Conforme relatado na Seção 5.1.3, está prevista uma etapa de divulgação no método, que pode acontecer ao final de cada iteração. Esta divulgação é uma exposição formal dos resultados obtidos para a comunidade acadêmica. Para o experimento da Seção 6.3, esta etapa ainda está em desenvolvimento.

A divulgação do presente trabalho, entretanto, já foi realizada com a apresentação do método no 12º Seminário de Pesquisa em Ontologias no Brasil (Ontobrás 2019) (BAX; GONÇALVES, 2019b) e no Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) 2019 (BAX; GONÇALVES, 2019a). No congresso da Ontobrás o objetivo principal foi apresentar o método, enquanto que os resultados preliminares obtidos com a aplicação do método fizeram parte do que foi exposto no ENANCIB.

O artigo com o título: "Método de Integração Semântica Incremental de Dados Científicos Baseado em Ontologias" (BAX; GONÇALVES, 2019b), apresentado no ONTOBRAS 2019, descreveu o mapeamento de um arquivo CSV com a utilização do SSD e do SDD através de um exemplo simplificado. Destacando a originalidade do método de integração incremental proposto. Uma banca avaliou a proposta, indicando direcionamentos para o trabalho. A participação no congresso, por se tratar de um seminário de pesquisas em ontologias, favoreceu o intercâmbio de ideias com pesquisadores envolvidos no assunto.

Quanto ao trabalho exposto no Encontro Nacional de Pesquisa em Ciência da Informação (BAX; GONÇALVES, 2019a), foi apresentado o método proposto e um exemplo baseado nos experimentos já realizados, expondo resultados práticos da utilização e validação do mesmo.

8 TRABALHOS RELACIONADOS

O processo de integração semântica de dados para a geração de grafos de conhecimento tem sido utilizado em diversas áreas. Discute-se aqui as iniciativas consideradas mais fortemente correlacionadas ao esforço de pesquisa desta tese.

8.1 Integração semântica de dados

Chalk (2016) descreve como o movimento por uma ciência global e intrinsecamente ligada à Internet exige a necessidade natural de capturar, armazenar agregar e buscar dados científicos através de grandes corpus de silos de dados heterogêneos. O autor descreve no seu artigo uma forma de capturar os dados utilizando o *Javascript Object Notation for Linked Data* (JSON-LD). Um ponto levantado é que os dados científicos devem ser armazenados desejavelmente anotados com metadados. Os autores afirmam que existe a necessidade de desenvolver abordagens que permitam que os dados científicos fiquem disponíveis em formato aberto e de fácil pesquisa.

O estudo de Fortier *et al.* (2010), apresenta o uso de várias amostras para destrinchar o novelo que caracteriza a análise de doenças. Estas amostras contam com dados de fatores ambientais, estilo de vida e da genética. Tendo em vista esta quantidade de fatores e a necessidade de agrupar dados de diferentes fontes, o estudo busca uma forma padrão de representação do conhecimento dos estudos relacionados. Propondo o DataSHaPER, que é considerado por seus autores uma abordagem científica e uma ferramenta prática.

Um sistema para integrar dados epidemiológicos é apresentado no LIFE (UCITELI; KIRSTEN, 2015), ele é um estudo contendo centenas de entrevistas de moradores de Leipzig. Os dados são heterogêneos e são centralizados e integrados em um repositório comum. A descrição semântica dos dados é feita com a *LIFE Investigation Ontology* (LIO). O objetivo central do *framework* é obter dicas sobre os tipos de dados obtidos na LIFE. São utilizadas consultas ao *framework* para obter dados de pesquisa. Os dados são obtidos por transformações das consultas em SQL *queries*.

Em Rehman *et al.* (2017), ressalta-se que a integração de dados se tornou o mais proeminente aspecto de aplicativos de gerenciamento de dados, especialmente em domínios como ecologia, biologia e geociências. As aplicações científicas complexas de hoje e a ascensão de diversos geradores de dados, dispositivos em domínios científicos (por exemplo, sensores) fizeram a integração de dados uma tarefa desafiadora. Em resposta a esses desafios, aplicativos de gerenciamento de dados estão fornecendo funcionalidades inovadoras que vêm ao preço de alta complexidade. Os autores apresentam uma estrutura de integração semântica de dados que se baseia em ontologias. Explorando um formalismo de lógica de descrição e associando procedimentos de raciocínio, a estrutura é capaz de lidar com formatos heterogêneos e semânticas diferentes. Além de um detalhamento da discussão da capacidade de integração baseada em ontologias, os autores também apresentam uma breve visão da arquitetura do sistema e sua aplicação em um cenário do mundo real a partir da pesquisa.

O estudo de Biffl *et al.* (2013) apresenta uma avaliação de abordagens distintas para a integração de dados heterogêneos. Neste estudo os autores discutem três formas de integração de dados:

- Um repositório central com um modelo de dados fixo, neste caso a entrada de dados deve respeitar o mesmo esquema para todos os estudos;
- Agregação de repositórios locais heterogêneos, aqui os dados dos repositórios são convertidos para o formato do repositório central;
- Integração de eco-sistema empírico. Para este método, os dados são mapeados em uma ontologia central que agrega os dados de cada estudo.

A integração de eco-sistema empírico apresentada por Biffl *et al.* é a forma de integração de dados proposta pelo presente trabalho, que utiliza um conjunto de ontologias e um método próprio (Capítulo 5), para automatizar o processo.

A necessidade de uma base de dados formada por um grafo RDF apoiada por ontologias para a criação de sistemas de informações abertos é defendida por Daraio *et al.* (2016), que ressaltam a possibilidade de uma maior interoperabilidade entre diferentes bancos de dados. Outro ponto levantado é a possibilidade de dimensões adicionais de dados comparando com as dimensões padrões definidos por um Framework. Os autores focaram o seu estudo nas vantagens de um sistema como o *Ontology-Based Linking Open Database* (OBDM), principalmente: a abertura dos dados (liberdade de compartilhamento); abertura dos termos (nomes ou terminologia usada); interoperabilidade e qualidade dos dados. Uma forma para a criação de um grafo RDF a partir de dados de pesquisa é realizar o mapeamento de dados tabulares tradicionais para este grafo. O que será descrito na seção seguinte.

8.2 Métodos de mapeamento de dados em RDF

Existe um grande número de ferramentas para mapear dados de todos os tipos de fontes em um grafo RDF. Reeve e Han (2005) e Uren *et al.* (2006) descrevem a arquitetura e analisam o desempenho de algumas destas ferramentas. Alguns exemplos de ferramentas incluem o MUSE (MAYNARD, 2003), Armadillo (DINGLI; CIRAVEGNA; WILKS, 2003) e KIM (POPOV *et al.*, 2003). No entanto, a maioria dos dados científicos nos principais *data centers* ainda têm seu conteúdo codificado em formato tabular. Pior ainda, o compartilhamento e a consulta de dados são restritos a conjuntos de dados (granulação grossa, arquivo por arquivo), não permitindo a consulta de valores de dados em nível mais detalhado por critérios de busca.

Além disso, pode-se afirmar que o resultado da aplicação da maioria das ferramentas de conversão e técnicas de transformação de dados tabulares para o padrão RDF não resulta em conteúdo adequado para que cientistas possam explorar e baixar dados para analise (NE'EMAN *et al.*, 2019). Esses grafos não seriam capazes de descrever dados em termos de objetos de estudo (por exemplo, sujeitos, amostras, eventos, instrumentos); nem de suportar consultas que resultem em dados harmonizados (ou seja, dados que podem ser comparados entre estudos).

O mapeamento de dados tabulares para uma um grafo RDF pode ser feito com a utilização de uma linguagem de mapeamento. Uma dessa linguagens é a RDB2RDF *Mapping-Language* (R2RML). Diversos trabalhos fornecem uma visão geral de tecnologias individuais de mapeamento (DAS; SUNDARA; CYGANIAK, 2012; SAHOO *et al.*, 2009; MICHEL; MONTAGNAT; FARON-ZUCKER, 2014), ressaltando que tornar dados hospedados em bancos de dados relacionais (RDB) acessíveis para a web semântica tem sido um campo ativo de pesquisa durante a última década. Vários projetos como o Apache Any23[1], o Triplify (AUER *et al.*, 2009), o Open Refine[2] e o Karma (GUPTA *et al.*, 2012) foram motivados pela necessidade de facilitar a transformação de dados tabulares em estruturas de dados semanticamente vinculadas e representadas por grafos (*Linked Data*) (ERMILOV; AUER; STADLER, 2013; WAAL *et al.*, 2014). Por serem genéricas, contudo, a maioria dessas abordagens e ferramentas abstraem o contexto em que são usadas, tornando a sua aplicação menos efetiva no escopo mais específico de dados oriundos de pesquisas científicas, apenas convertendo dados relacionais em RDF ou expondo dados relacionais para que eles possam ser consultado através do SPARQL (HE *et al.*, 2007). Apesar dos provedores dispostos a publicar seus dados em um formato legível por máquina serem um pouco desencorajados pela dificuldade de escolher entre as ferramentas existentes e a linguagem de mapeamento.

A publicação pelo W3C da R2RML ocorreu em setembro de 2012 (BERNERS-LEE, 2011), tornando-se um padrão de linguagem para descrever mapeamentos entre um relacionamento de banco de dados e uma representação RDF equivalente. Isso marcou um novo passo para a atualização da web de dados. A R2RML incentiva os desenvolvedores de ferramentas de mapeamento de bancos de dados relacionais para o RDF a cumprir com uma linguagem de mapeamento padrão. Por outro lado, os provedores de dados devem ser beneficiados pela adoção desta linguagem comum, permitindo-lhes dissociar a integração de dados relacionais de problemas de ferramentas ou abordagens específicas e assegurando a sustentabilidade. No entanto, as escolhas feitas na especificação R2RML implicam em algumas limitações nos tipos de mapeamento que podem ser expressos. Além disso, uma linguagem de mapeamento independente de implementação como a R2RML não aborda algumas das perguntas comuns que ocorrem ao traduzir dados relacionais existentes em RDF, como a escolha de reutilizar os dados existentes e vocabulários ou como os mesmos serão acessados ou consultados.

8.2.1 A abordagem do Dicionário Semântico de Dados

Uma proposta de solução correlata para tratar o problema, no contexto específico de dados científicos, é apresentada em (RASHID *et al.*, 2017). A solução permite a integração de dados de múltiplos domínios utilizando um template formal para anotar as colunas do *dataset*, o Dicionário Semântico de Dados (SDD). Este trabalho de pesquisa se inspira nesta solução para realizar o processo de ingestão de dados aplicando um método próprio, conforme detalhado no Capítulo 5.

[1] https://any23.apache.org
[2] http://openrefine.org

A possibilidade de evoluir o grafo por alterações incrementais, no modelo e nas ontologias relacionadas, representa uma contribuição relevante da presente pesquisa. A seção seguinte referencia um trabalho análogo, que propõe um método iterativo para o problema da integração de dados.

8.2.2 Um método iterativo para integração semântica de dados

Um método iterativo para a integração semântica de dados foi descrito em (SEQUEDA; MIRANKER, 2017), nesta proposta Sequeda e Miranker dividem o procedimento em duas tarefas principais:

- Criação da ontologia: uma ontologia inicial derivada do esquema do banco de dados, também conhecida como ontologia putativa (SEQUEDA *et al.*, 2011; SEQUEDA; ARENAS; MIRANKER, 2012). Esta ontologia putativa é transformada gradualmente na ontologia alvo. Este processo é semelhante ao da proposta deste trabalho (Seção 5.1.3).
- Mapeamento dos dados: com a ontologia putativa criada é possível definir o mapeamento dos dados que pode começar com o mapeamento direto conforme descrito na Seção 2.2.2.

No método proposto por Sequeda e Miranker, as duas tarefas citadas anteriormente, devem ser executadas de forma iterativa. Com um conjunto mínimo de questões de competência (Seção 2.2.1.3) sendo tratado a cada iteração. Os autores relacionam três tipos de atores envolvidos no processo:

- Usuários de negócio: são aqueles que possuem o conhecimento do negócio e que podem priorizar as questões a serem resolvidas.
- Desenvolvedores: tem conhecimento da tecnologia do banco de dados e do modelo de dados utilizado.
- Engenheiros de conhecimento: permitem a comunicação entre os desenvolvedores e os usuários, possuindo conhecimento em modelagem de dados utilizando ontologias.

O método de Sequeda e Miranker prevê uma fase de captura do conhecimento e outra de implementação. Na primeira fase, são descobertos os conceitos e relacionamentos do conjunto de questões selecionadas para serem tratadas e sua conexão com o banco de dados, definindo a ontologia alvo. Na segunda fase, o mapeamento é desenvolvido em R2RML e as questões são implementadas e validadas com a execução de consultas SPARQL.

Pode-se verificar que o método Odin, que foi apresentado no Capítulo 5, é relacionado ao método de Sequeda e Miranker. Especialmente na definição da ontologia Base e sua evolução gradual.

8.3 Discussão

Apresentou-se neste capítulo as iniciativas correlatas mais próximas, propostas na literatura científica para tratar os principais aspectos do problema da integração semântica de

dados. Destaca-se que tais iniciativas buscam também aprimorar fatores como a "encontrabilidade" (*findability*), interoperabilidade e maior possibilidade de acesso e reúso dos dados integrados (Seção 2.1.5).

A transformação de dados em grafos de conhecimento permite a integração de dados, de acordo com o método de eco-sistema empírico (BIFFL *et al.*, 2013), onde os dados são mapeados em uma ontologia central que agrega os dados de cada estudo. Assim, as informações oriundas de diversos estudos passam a compartilhar de um modelo comum definido por esta ontologia central. No presente estudo foi possível validar esta integração ao trabalhar com diferentes conjuntos de dados (*datasets*) (Capítulo 7).

Os esforços revisados acima destacam o interesse no uso de grafos RDF como uma forma de melhorar a interoperabilidade. Compreendeu-se melhor como a integração semântica de dados, processo que adiciona ao grafo RDF informações adicionais, eleva os dados originais ao nível de conhecimento formal e processável por máquinas. Vale reforçar que a presente pesquisa define um método iterativo e participativo para a integração semântica de dados, através de um processo particular. Este método se assemelha ao apresentado na Seção 8.2.2, possuindo, no entanto, características particulares conforme descrito no Capítulo 5.

9 CONSIDERAÇÕES FINAIS

Neste trabalho foi elaborado um método de anotação de dados visando a sua integração semântica no domínio de pesquisas ou estudos científicos. O método se inspira na ADSRM (CONBOY; GLEASURE; CULLINA, 2015) definindo atores e etapas para uma construção incremental de um grafo de conhecimento, fundamentado em uma ontologia de domínio. Nesta construção está prevista a participação de todos os atores de acordo com os princípios da metodologia ágil (Seção 5.1.1). Durante o processo de integração, o método produz, de forma incremental, artefatos para anotar com metadados e gerar o grafo de conhecimento. Estes artefatos são criados no decorrer de iterações (cíclicas e incrementais). A cada iteração, uma versão aprimorada do grafo é construída (ou reconstruída), com a evolução da ontologia de domínio (a ontologia Base).

Ao longo dos experimentos realizados, foi possível criar o grafo de conhecimento com o método, evoluindo a ontologia, com algumas vantagens. Dentre elas está a possibilidade de apresentar, de forma precoce, aos envolvidos no projeto cada versão do grafo gerado. Isto contribuiu para a rápida identificação dos problemas e limitações do design do estudo científico definido na iteração anterior. Assim, refinamentos sucessivos deste design, em busca de uma melhor modelagem conceitual dos dados puderam ser realizados em iterações subsequentes. A comunicação frequente entre os atores (especialistas pesquisadores de domínio, ontologistas, desenvolvedores e cientistas de dados) colaborou para se alcançar um melhor resultado nas iterações. Além disso, novos requisitos de modelagem, necessários para que o pesquisador pudesse responder as perguntas de competência do estudo, foram acrescentados ao design ao longo das etapas de cada iteração. Estes novos requisitos alimentaram a Lista de pendências, sendo priorizados e tratados nas iterações seguintes.

A integração de estudos provenientes de *datasets* com formatos e granularidades diferentes mostrou que o método é uma forma eficiente e eficaz de tratar o problema de integração semântica de dados. Pode-se concluir que o método também se mostrou útil, porque a integração exigiu a execução de um número reduzido de iterações para gerar uma primeira versão satisfatória do grafo[1].

Os especialistas de domínio puderam comprovar a utilidade da criação incremental do grafo de conhecimento, participando da sua construção. Ao tomar contato com o modelo de *templates* do HADatAc, os especialistas puderam entender como os conceitos eram definidos nesses templates.

A participação de ontologistas na criação do grafo de conhecimento contribuiu para tornar o modelo conceitual "mais ontológico". Tais profissionais auxiliaram os especialistas de domínio na definição do mapeamento conceitualmente mais adequado para os *datasets* utilizados. Esse mapeamento, por sua vez, definiu a configuração final do grafo. Outra tarefa desse profissional foi a busca e definição das interligações dos dados com as ontologias de domínio. Tal tarefa exigiu a colaboração dos especialistas de domínio, que puderam apontar

[1] Em analogia com o conceito de "Produto Mínimo Viável", parte da cultura determinada pelos princípios ágeis.

direções com base na sua experiência. Apesar de ter sido executado para os *datasets* de exemplo, o processo de definição das interligações com ontologias de domínio é uma tarefa complexa e teve sua execução restrita a algumas iterações, necessitando maior enfoque em trabalhos futuros.

O método facilitou a contribuição contínua de todos os desenvolvedores nas iterações. Desde a construção de pequenos programas para tratamento de dados até a criação de soluções mais elaboradas para fornecer informações sumarizadas pelos indicadores através de consultas ao grafo.

Verificou-se, ainda, a utilidade do envolvimento dos cientistas de dados. Profissionais que podem aplicar técnicas estatísticas, de aprendizado de máquina e outros recursos para tratar grandes volumes de dados. Com o auxílio destes profissionais, os especialistas de domínio obtiveram respostas para questões de competência não tratadas por consultas diretas aos dados ingeridos. O papel do cientista de dados, apesar de presente no método, poderia ser ainda melhor explorado e aproveitado em trabalhos futuros. O que implica em mais investimentos na evolução das ferramentas de acesso ao grafo para automatizar os processos de análise desenvolvidos por estes profissionais.

9.1 Validação do método

Com a utilização de grafos de conhecimento os dados de pesquisa são organizados semanticamente, podendo ser reutilizados com maior facilidade. Para a avaliação e validação do método foi escolhido um processo específico para a ingestão de dados utilizando o *framework* HADatAc[2]. Esta escolha não limita a solução definida pelo método, que pode ser aplicado a outros processos de ingestão de dados.

Com a avaliação do método respondeu-se a questão de pesquisa, verificando-se as hipóteses elaboradas inicialmente, conforme exposto na seção seguinte.

9.2 Verificação da questão de pesquisa e hipóteses

O método Odin responde ao problema de pesquisa, mostrando que é possível integrar semanticamente dados científicos utilizando ontologias em um processo ágil. A ontologia (juntamente com a ontologia de domínio que o fundamenta) pode ser construída de forma iterativa com a participação contínua dos especialistas de domínio. Além disso, ele prevê três atores adicionais, para além do pesquisador do domínio: desenvolvedores, ontologistas e cientistas de dados.

O grafo de conhecimento é construído de acordo com os princípios da metodologia ágil, que dão ênfase à entrega precoce da melhor solução possível em cada iteração. No encerramento de uma iteração, a solução é avaliada identificando-se tarefas a serem realizadas nas próximas iterações. As etapas que compõem uma iteração se baseiam na metodologia ADSRM que incorpora os princípios ágeis à DSR.

[2] http://hadatac.org

As hipóteses estabelecidas foram verificadas confirmando que:

- A adoção de uma metodologia ágil garantiu a obtenção dos resultados da integração dos dados mais rapidamente, facilitando a correção precoce dos problemas encontrados.
- A utilização de ontologias permitiu não somente a integração sintática de dados heterogêneos, como o enriquecimento semântico dos mesmos.
- A abordagem ontológica de modelagem contribuiu para o processo de integração de dados heterogêneos além de trazer maior flexibilidade e rapidez ao mesmo.
- A utilização do HADatAc como ferramenta para gerenciamento da ontologia gerada pelo processo de integração garantiu a utilização e busca das informações desta ontologia, além de fornecer a base para a realização da ingestão semântica dos dados.

9.3 Atendimento aos objetivos

O objetivo de desenvolver um método ágil de integração semântica de dados científicos heterogêneos que possa ser utilizado para grandes volumes de dados e que seja de fácil utilização para os especialistas de domínio foi alcançado.

- Foram realizados ciclos de integração de dados de estudos provenientes de pesquisas epidemiológicas em parceria com pesquisadores da Fiocruz.
- A evolução da ontologia de domínio foi realizada de acordo com o método, obtendo-se, a cada iteração, uma versão aprimorada da mesma.
- A utilização dos recursos dos HADatAc para consultar os dados ingeridos, tais como a interface de busca facetada, foi apresentada aos atores envolvidos.
- Foi feita a análise e validação dos dados ingeridos no HADatAc em conjunto com os pesquisadores de domínio.

9.4 Contribuições

Conforme exposto no Capítulo 1, a gestão adequada dos dados científicos é essencial para a Ciência, sendo especialmente importante no contexto do novo paradigma da “Ciência de Dados”, onde as descobertas científicas são intensivamente dependentes desta gestão (FOX; HENDLER, 2009). Além disso, o volume e a diversidade de dados a ser gerenciado em uma pesquisa científica é cada vez maior. Nesse cenário, o presente trabalho apresenta um método para integrar dados científicos heterogêneos com o uso de ontologias de forma ágil, permitindo a gestão adequada desses dados. Para isso, utilizou-se o *framework* HADatAc para a ingestão dos dados, cuja infraestrutura garante o gerenciamento de um grande volume de dados conforme relatado na Seção 3.4.

A construção do grafo de conhecimento tem seu foco em atender as expectativas dos pesquisadores, os especialistas de domínio. O método não somente prevê a participação

contínua desses especialistas, como garante que o grafo será construído de acordo com essas expectativas. Assim, o método contribuiu para que os especialistas de domínio pudessem definir, a cada iteração, a modelagem ontológica do grafo. Isto foi feito com a utilização de questões de competência (Seção 2.2.1.3) que orientam a criação do grafo através da elaboração de questionamentos relacionados à pesquisa científica em desenvolvimento.

O método estabelece uma evolução gradual e iterativa do grafo de conhecimento. Nesta evolução, os especialistas de domínio podem verificar se o grafo gerado atende aos requisitos estabelecidos inicialmente. Como o método é baseado em uma metodologia ágil, a ADSRM, a validação do grafo e as possíveis ações de correção ocorrem de forma precoce a cada iteração. Assim, os especialistas de domínio podem revisar o *design* do grafo mais prontamente, sem a necessidade de aguardar por uma versão final do mesmo para isso. Deste modo, o trabalho de preparação de dados é facilitado e mantém o foco em solucionar as questões desses especialistas. Portanto, foi possível evoluir o grafo de conhecimento de forma ágil e atender aos requisitos dos pesquisadores.

Para a preparação dos dados é necessária a colaboração de outros profissionais além dos especialistas de domínio. Neste sentido, a definição dos demais atores e seu papel nas diversas etapas do método foi outra contribuição obtida, permitindo, inclusive, identificar limitações e melhorias futuras conforme será exposto nas próximas seções.

A validação da integração semântica de dados científicos com a utilização de um grafo de conhecimento, trouxe elementos para solidificar a abordagem ontológica como uma solução alternativa não somente para a representação semântica de dados, como também para a integração semântica dos mesmos.

9.4.1 Contribuição para o HADatAc

Como explicado anteriormente, a ingestão de dados pelo HADatAc utiliza dois *templates* de metadados (SSD e SDD) e um para ligar os objetos declarados pelos dois anteriores (OAS). No início da nossa investigação não havia uma compreensão solidificada por parte dos desenvolvedores da Plataforma a respeito da necessidade da existência de todos esses *templates* conjuntamente. Portanto, outra contribuição de nossa pesquisa foi justificar e confirmar a necessidade de separar os templates.

A nossa pesquisa confirmou que separar, no modelo, as entidades de informação das entidades de domínio, como faz o HADatAc é, de fato, uma vantagem. O experimento utilizando os dados da dengue e esquistossomose permitiu observar a utilidade desta separação. Foram identificadas características definidas no SSD, tais como a cardinalidade, que auxiliaram, por exemplo, na identificação de dados ausentes. Neste caso, o total esperado de amostras coletadas (equivalente a dados dos 5571 municípios) pôde ser confrontado com aquele obtido do *dataset*.

O SSD organiza os conceitos do domínio representados pelo SDD de acordo com as necessidades do pesquisador. Um SSD é análogo ao modelo estrutural de uma tarefa de investigação científica, relacionando conceitos como sujeitos, amostras, localização e tempo.

De acordo com o problema em investigação, estes conceitos devem então ser mapeados para os conceitos representados pelo SDD.

9.5 Dificuldades e limitações

Apesar do experimento ter sugerido benefícios da utilização do método para disciplinar o trabalho de integração semântica de dados científicos, não houve tempo suficiente para medir objetivamente os diversos fenômenos que o compõem. Para tal, seria necessário definir os conceitos, variáveis e indicadores mais relevantes que permitissem a observação do desempenho dos especialistas em cenários em que estes utilizam e não utilizam o método (cenário controle). Portanto, é preciso reconhecer que a avaliação realizada neste trabalho se restringe ao desempenho do método. Trabalhos futuros podem avaliá-lo mais profundamente, utilizando experimentos observáveis e mensuráveis.

Foram enfrentadas, também, dificuldades no processo de ingestão de dados. Conforme visto na Seção 2.2.2.1, a utilização de arquivos de entrada no formato CSV fez com que alguns campos tivessem seus valores inconsistentes, como foi constatado posteriormente na etapa de validação. Trata-se, contudo, de um problema comumente enfrentado na fase de preparação de dados e não inerente ao método. Outro problema enfrentado durante o processo de ingestão de dados foi relacionado ao desempenho do mesmo, observado em ocasiões onde o TBox representava um volume muito grande de dados Além disso, o *framework* HADatAc estava em constante evolução funcional ao longo do processo, o que em alguns momentos, trazia a necessidade da revisão dos templates de ingestão de dados. Estes templates, por sua vez, são gerados por arquivos CSV ou planinhas eletrônicas e carecem de recursos para facilitar a sua criação.

A ingestão de dados utilizando o HADatAc, bem como outros processos de ingestão existentes são limitados a um conjunto de formatos de dados de entrada. Isso exige esforços na transformação desses dados, que podem ser complexas, caso o formato da entrada não seja adequado.

Outro problema da representação em forma de grafos é relacionado ao tratamento de dados que possuem um número excessivo de valores de dados. Por exemplo, em dados coletados por extensos períodos temporais. Para estes casos a representação em forma de grafos, mesmo quando adotadas soluções como as do HADatAc, pode enfrentar problemas de escalabilidade.

A avaliação dos trabalhos correlatos mostrou que as soluções atuais para a integração semântica de dados precisam evoluir para que tenham uma maior usabilidade. De toda forma, realizar o mapeamento entre dados e conceitos nunca será uma tarefa trivial, dada a carga semântica inerente. Assim, vencer esta limitação exige um maior envolvimento dos atores especializados que entendam a lógica dos mapeamentos. No futuro, caso o mapeamento possa ser realizado de forma mais simples e intuitiva, os próprios especialistas pesquisadores de domínio poderão realizá-lo.

Finalmente, a extração de informações do grafo gerado foi realizada através de

consultas ao mesmo e com a utilização de alguns artefatos criados especialmente para este fim, complementando as funcionalidades do HADatAc. Apesar de estar fora do escopo deste trabalho, um maior investimento na criação destes artefatos, inclusive com uma utilização mais ampla das técnicas da Ciência de Dados, poderia ter contribuído para aprimorar a validação do grafo de conhecimento.

Ainda é necessário dedicar um tempo à incorporação de conceitos de ontologias já estabelecidas na ontologia Base. Neste sentido são necessárias iterações adicionais com a participação dos ontologistas e especialistas de domínio, as quais podem ser realizadas em trabalhos futuros.

9.6 Trabalhos futuros

O método Odin apresenta-se como uma solução para o problema de geração de grafos de conhecimento. Trabalhos futuros podem utilizar o mesmo em outros processos de ingestão de dados. Esses processos, por sua vez, necessitam de maiores investimentos para que possam ser executados de forma simples pelos próprios especialistas de domínio, desde que instruídos no campo teórico das ontologias. Da mesma maneira, acredita-se que melhorias no processo de mapeamento de dados tabulares para grafos RDF podem ser realizadas, propiciando aos especialistas de domínio uma interface que forneça recursos amigáveis para a geração deste mapeamento.

Trabalhos futuros podem aprimorar a aplicação do método buscando desenvolver um processo de mapeamento mais rápido e fácil pelos especialistas de domínio, bem como a uma maior utilização de recursos da Ciência de Dados. Esta utilização pode ser integrada com a adição de algoritmos à ferramenta de tratamento do grafo. Um exemplo possível seria um algoritmo de aprendizado de máquina. Com este tipo de melhoria inferências que poderiam ser difíceis de se realizar se tornariam mais simples, podendo ser realizadas pelos próprios especialistas de domínio.

Outras melhorias nas ferramentas de ingestão podem, no futuro, criar condições para que os especialistas de domínio possam utilizá-las de forma ainda mais independente. Isso pode ser possível com a criação de soluções para que o método possa ser auxiliado pelas ferramentas permitindo, por exemplo, pesquisar elementos de ontologias de domínio de forma amigável. Assim, parte da experiência dos ontologistas e cientistas de dados estaria embutida nas próprias ferramentas. Além disso, a visualização facetada pode ser incrementada, adicionando-se a exibição de gráficos relacionados. Deste modo, pode-se chegar a uma solução completa para o mapeamento de dados tabulares em um grafo de conhecimento e para o gerenciamento deste grafo utilizando o método.

REFERÊNCIAS

ATTENBERG, J. M.; IPEIROTIS, P. G.; PROVOST, F. Beat the machine: Challenging workers to find the unknown unknowns. *In*: **Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence**. [*s.l.*: *s.n.*], 2011.

AUER, S. *et al.* Triplify: light-weight linked data publication from relational databases. *In*: ACM. **Proceedings of the 18th international conference on World wide web**. [*s.l.*], 2009. p. 621–630.

AZZAOUI, K. *et al.* Scientific competency questions as the basis for semantically enriched open pharmacological space development. **Drug discovery today**, Elsevier, v. 18, n. 17-18, p. 843–852, 2013.

BAIROCH, A.; COHEN-BOULAKIA, S.; FROIDEVAUX, C. **Review of the selected proceedings of the Fifth International Workshop on Data Integration in the Life Sciences 2008**. [*s.l.*]: BioMed Central, 2008.

BAX, M.; GONÇALVES, J. Grafos de conhecimento para preparação e reutilização de dados científicos. *In*: **Encontro Nacional de Pesquisa em Ciência da Informação**. Florianópolis: Universidade Federal de Santa Catarina, 2019.

BAX, M.; GONÇALVES, J. Método de integração semântica incremental de dados científicos baseado em ontologias. *In*: **12º Seminário de Pesquisa em Ontologias no Brasil**. Porto Alegre: Ontobras, 2019.

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

BELLINI, P.; NESI, P. Performance assessment of rdf graph databases for smart city services. **Journal of Visual Languages & Computing**, Elsevier, v. 45, p. 24–38, 2018.

BERNERS-LEE, T. **Design issues: Linked data (2006)** . 2011. http://www.w3.org/DesignIssues/LinkedData.html. 01/07/2018.

BEZERRA, C.; FREITAS, F.; SANTANA, F. Evaluating ontologies with competency questions. *In*: IEEE. **2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)**. [*s.l.*], 2013. v. 3, p. 284–285.

BIFFL, S. *et al.* Replication data management: Needs and solutions–an initial evaluation of conceptual approaches for integrating heterogeneous replication study data. *In*: IEEE. **2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement**. [*s.l.*], 2013. p. 233–242.

BOWERS, S. **Scientific workflow, provenance, and data modeling challenges and approaches**. [*s.l.*]: Springer, 2012.

BRASE, J. Using digital library techniques–registration of scientific primary data. *In*: SPRINGER. **International Conference on Theory and Practice of Digital Libraries** . [*s.l.*], 2004. p. 488–494.

CHALK, S. J. Scidata: a data model and ontology for semantic representation of scientific data. **Journal of cheminformatics**, Springer, v. 8, n. 1, p. 54, 2016.

CONBOY, K.; GLEASURE, R.; CULLINA, E. Agile design science research. *In*: SPRINGER. **International Conference on Design Science Research in Information Systems** . [*s.l.*], 2015. p. 168–180.

COUNCIL, N. R. *et al.* **Steps toward large-scale data integration in the sciences: Summary of a workshop**. [*s.l.*]: National Academies Press, 2010.

COURTOT, M. *et al.* Mireot: The minimum information to reference an external ontology term. **Applied Ontology**, IOS Press, v. 6, n. 1, p. 23–33, 2011.

CRESWELL, J. W. **Research design: Qualitative, quantitative, and mixed methods approaches**. [*s.l.*]: Sage publications, 2013.

DARAIO, C. *et al.* The advantages of an ontology-based data management approach: openness, interoperability and data quality. **Scientometrics**, Springer, v. 108, n. 1, p. 441–455, 2016.

DAS, S.; SUNDARA, S.; CYGANIAK, R. R2rml: Rdb to rdf mapping language, w3c recommendation 27 september 2012. **Cambridge, MA: World Wide Web Consortium (W3C)(www. w3. org/TR/r2rml)**, 2012.

DAY-RICHTER, J. *et al.* Obo-edit—an ontology editor for biologists.**Bioinformatics**, Oxford University Press, v. 23, n. 16, p. 2198–2200, 2007.

DINGLI, A.; CIRAVEGNA, F.; WILKS, Y. Automatic semantic annotation using unsupervised information extraction and integration. *In*: **Proceedings of SemAnnot 2003 Workshop**. [*s.l.*: *s.n.*], 2003.

DUMONTIER, M. *et al.* The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. **Journal of biomedical semantics**, BioMed Central, v. 5, n. 1, p. 14, 2014.

EDER, J. S. **Knowledge graph based search system**. [*s.l.*]: Google Patents, 2012. US Patent App. 13/404,109.

EHRLINGER, L.; WÖSS, W. Towards a definition of knowledge graphs. *In*: **SEMANTiCS (Posters, Demos, SuCCESS)**. [*s.l.*: *s.n.*], 2016.

ERMILOV, I.; AUER, S.; STADLER, C. Csv2rdf: User-driven csv to rdf mass conversion framework. *In*: **Proceedings of the ISEM**. [*s.l.*: *s.n.*], 2013. v. 13, p. 04–06.

FÄRBER, M. *et al.* Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. **Semantic Web**, IOS Press, n. Preprint, p. 1–53, 2016.

FITZGERALD, B. *et al.* Scaling agile methods to regulated environments: An industry case study. *In*: IEEE PRESS. **Proceedings of the 2013 International Conference on Software Engineering**. [*s.l.*], 2013. p. 863–872.

FLORIDI, L. Semantic conceptions of information. 2005.

FORTIER, I. *et al.* Quality, quantity and harmony: the datashaper approach to integrating data across bioclinical studies. **International journal of epidemiology**, Oxford University Press, v. 39, n. 5, p. 1383–1393, 2010.

FOWLER, M.; HIGHSMITH, J. *et al.* The agile manifesto. **Software Development**, [San Francisco, CA: Miller Freeman, Inc., 1993-, v. 9, n. 8, p. 28–35, 2001.

FOX, P.; HENDLER, J. A. Semantic escience: encoding meaning in next-generation digitally enhanced science. **The Fourth Paradigm**, v. 2, 2009.

FOX, P. *et al.* Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience. **Computers & Geosciences**, v. 35, n. 4, p. 724–738, abr. 2009. ISSN 0098-3004.

FRAMEWORK, O. Q. Guidelines for oecd statistical activities. **Version**, v. 1, p. 2003.

GKOUTOS, G. V.; SCHOFIELD, P. N.; HOEHNDORF, R. The Units Ontology: a tool for integrating units of measurement in science. **Database**, v. 2012, jan. 2012.

GOBLE, C.; STEVENS, R. State of the nation in data integration for bioinformatics. **Journal of biomedical informatics**, Elsevier, v. 41, n. 5, p. 687–693, 2008.

GOMEZ-CABRERO, D. *et al.* **Data integration in the era of omics: current and future challenges**. [*s.l.*]: BioMed Central, 2014.

GRAU, B. C. *et al.* Owl 2: The next step for owl. **Web Semantics: Science, Services and Agents on the World Wide Web**, Elsevier, v. 6, n. 4, p. 309–322, 2008.

GREGOR, S.; HEVNER, A. R. Positioning and presenting design science research for maximum impact. **MIS quarterly**, v. 37, n. 2, 2013.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge acquisition**, Elsevier, v. 5, n. 2, p. 199–220, 1993.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing? **International journal of human-computer studies**, Elsevier, v. 43, n. 5-6, p. 907–928, 1995.

GUPTA, S. *et al.* Karma: A system for mapping structured sources into the semantic web. *In*: SPRINGER. **Extended Semantic Web Conference**. [*s.l.*], 2012. p. 430–434.

HE, B. *et al.* Accessing the deep web. **Communications of the ACM**, ACM, v. 50, n. 5, p. 94–101, 2007.

HEVNER, A. R. A three cycle view of design science research. **Scandinavian journal of information systems**, v. 19, n. 2, p. 4, 2007.

HEY, T. *et al.* **The fourth paradigm: data-intensive scientific discovery** . [*s.l.*]: Microsoft research Redmond, WA, 2009. v. 1.

KAIYA, H.; SAEKI, M. Using domain ontology as domain knowledge for requirements elicitation. *In*: IEEE. **14th IEEE International Requirements Engineering Conference (RE'06)**. [*s.l.*], 2006. p. 189–198.

KLYNE, G.; CARROLL, J. J. Resource description framework (rdf): Concepts and abstract syntax. 2006.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning:A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, v. 160, p. 3–24, 2007.

LEBO, T. *et al.* Prov-o: The prov ontology. **W3C recommendation**, W3C, v. 30, 2013.

MADIN, J. *et al.* An ontology for describing and synthesizing ecological observation data. **Ecological informatics**, Elsevier, v. 2, n. 3, p. 279–296, 2007.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. **Decision support systems**, North-Holland, v. 15, n. 4, p. 251–266, 1995.

MAYNARD, D. Multi-source and multilingual information extraction.**Expert Update**, v. 6, n. 3, p. 11–16, 2003.

MCCUSKER, J. P. *et al.* What is a knowledge graph? **Semantic Web**, 2018.

MCCUSKER, J. P. *et al.* Broad, interdisciplinary science in tela: An exposure and child health ontology. *In*: ACM. **Proceedings of the 2017 ACM on Web Science Conference** [*s.l.*], 2017. p. 349–357.

MCGUINNESS, D. *et al.* Semantic escience for ecosystem understanding and monitoring: The jefferson project case study. *In*: **AGU Fall Meeting Abstracts**. [*s.l.*: *s.n.*], 2014. v. 1, p. 3712.

MCGUINNESS, D. L. *et al.* The emerging field of semantic scientific knowledge integration. **IEEE Intelligent Systems**, IEEE, v. 24, n. 1, p. 25–26, 2009.

MCQUILTON, P. *et al.* Biosharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. **Database**, Oxford University Press, v. 2016, 2016.

MEEHAN, J. *et al.* Data ingestion for the connected world. *In*: **CIDR**. [*s.l.*: *s.n.*], 2017.

MICHEL, F.; MONTAGNAT, J.; FARON-ZUCKER, C**A survey of RDB to RDF translation approaches and tools**. Tese (Doutorado) — I3S, 2014.

MOORE, R. W. Data management systems for scientific applications. *In*: **The Architecture of Scientific Software**. [*s.l.*]: Springer, 2001. p. 273–284.

NE'EMAN, Y. *et al.* Understanding the data management capabilities of health science repositories: A survey of nih data centers - student draft. 2019.

NOY, N. F.; MCGUINNESS, D. L. *et al.***Ontology development 101: A guide to creating your first ontology**. [*s.l.*]: Stanford knowledge systems laboratory technical report KSL-01-05 and . . . , 2001.

NUSHI, B. *et al.* On human intellect and machine failures: Troubleshooting integrative machine learning systems. *In*: **Thirty-First AAAI Conference on Artificial Intelligence**. [*s.l.*: *s.n.*], 2017.

PAN, J. Z. *et al.***Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering: 12th International Summer School 2016, Aberdeen, UK, September 5-9, 2016, Tutorial Lectures**. [*s.l.*]: Springer, 2017. v. 9885.

PAN, J. Z. *et al.***Exploiting linked data and knowledge graphs in large organisations**. [*s.l.*]: Springer, 2017.

PEREZ-RIVEROL, Y. *et al.* Discovering and linking public omics data sets using the omics discovery index. **Nature biotechnology**, v. 35, n. 5, p. 406, 2017.

PHILIPPI, S. **Data and knowledge integration in the life sciences**. [*s.l.*]: Oxford University Press, 2008.

PIÑERO, J. *et al.* Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants.**Nucleic acids research**, Oxford University Press, p. gkw943, 2016.

PINHEIRO, P. *et al.* Annotating diverse scientific data with hasco. *In*:**Proceedings of the Seminar on Ontology Research in Brazil**. [*s.l.*: *s.n.*], 2018.

PINHEIRO, P. *et al.* An ontology-based annotation workflow for semantic data ingestion - draft. 2018.

PINHEIRO, P. *et al.* Hadatac: A framework for scientific data integration using ontologies. *In*: **Proceedings of the ISWC**. [*s.l.*: *s.n.*], 2018.

POKORNỲ, J. Graph databases: their power and limitations. *In*: SPRINGER. **IFIP International Conference on Computer Information Systems and Industrial Management** [*s.l.*], 2015. p. 58–69.

POPOV, B. *et al.* Kim–semantic annotation platform. *In*: SPRINGER.**International Semantic Web Conference**. [*s.l.*], 2003. p. 834–849.

RASHID, S. M. *et al.* The semantic data dictionary approach to data annotation & integration. *In*: **SemSci@ ISWC**. [*s.l.*: *s.n.*], 2017. p. 47–54.

RECTOR, A. *et al.* On beyond gruber:“ontologies” in today’s biomedical information systems and the limits of owl. **Journal of Biomedical Informatics: X**, Elsevier, p. 100002, 2019.

REEVE, L.; HAN, H. Survey of semantic annotation platforms. *In*: ACM.**Proceedings of the 2005 ACM symposium on Applied computing**. [*s.l.*], 2005. p. 1634–1638.

REHMAN, M. A. *et al.* Semantic based data integration in scientific workflows.**International Journal Of Advanced Computer Science And Applications**, Science & Information Sai Organization Ltd 19 Bolling Rd, Bradford, West Yorkshire, 00000, England, v. 8, n. 7, p. 314–325, 2017.

REN, Y. *et al.* Towards competency question-driven ontology authoring. *In*: SPRINGER. **European Semantic Web Conference**. [*s.l.*], 2014. p. 752–767.

RESENDE, L. C. de. **A Curadoria de dados científicos na ciência da informação: levantamento do cenário nacional**. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, 2019. http://hdl.handle.net/1843/32413.

REW, R.; DAVIS, G. Netcdf: an interface for scientific data access**IEEE computer graphics and applications**, IEEE, v. 10, n. 4, p. 76–82, 1990.

SAHOO, S. S. *et al.* A survey of current approaches for mapping of relational databases to rdf. **W3C RDB2RDF Incubator Group Report**, v. 1, p. 113–130, 2009.

SANTOS, P. X. d. *et al***Livro Verde-Ciência aberta e dados abertos: mapeamento e análise de políticas, infraestruturas e estratégias em perspectiva nacional e internacional**. [*s.l.*]: Fiocruz, 2017.

SAYÃO, L. F.; SALES, L. F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade**, Universidade Federal da Paraíba-Programa de Pós-Graduação em Ciência da Informação, v. 22, n. 3, 2012.

SAYÃO, L. F.; SALES, L. F. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. 2013.

SEQUEDA, J. F. On the semantics of r2rml and its relationship with the direct mapping. *In*: CITESEER. **International Semantic Web Conference (Posters & Demos)** . [*s.l.*], 2013. v. 2013, p. 193–196.

SEQUEDA, J. F.; ARENAS, M.; MIRANKER, D. P. On directly mapping relational databases to rdf and owl. *In*: ACM. **Proceedings of the 21st international conference on World Wide Web**. [*s.l.*], 2012. p. 649–658.

SEQUEDA, J. F.; MIRANKER, D. P. A pay-as-you-go methodology for ontology-based data access. **IEEE Internet Computing**, IEEE, v. 21, n. 2, p. 92–96, 2017.

SEQUEDA, J. F. *et al.* Survey of directly mapping sql databases to the semantic web. **The Knowledge Engineering Review**, Cambridge University Press, v. 26, n. 4, p. 445–486, 2011.

SHADBOLT, N.; BERNERS-LEE, T.; HALL, W. The semantic web revisited **IEEE intelligent systems**, IEEE, v. 21, n. 3, p. 96–101, 2006.

SIMMHAN, Y. L.; PLALE, B.; GANNON, D. A survey of data provenance in e-science. **ACM Sigmod Record**, ACM, v. 34, n. 3, p. 31–36, 2005.

SMITH, B. *et al.* The obo foundry: coordinated evolution of ontologies to support biomedical data integration. **Nature biotechnology**, Nature Publishing Group, v. 25, n. 11, p. 1251, 2007.

SOUSA, T. *et al.* The environmental burden of diarrhea in young children attributable to inadequate sanitation in brazil. **Journal of Water, Sanitation and Hygiene for Development**, International Water Association, v. 4, n. 3, p. 509–520, 2014.

SOUSA, T.; GARBAYO, L.; BARCELLOS, C. Glocal gdb use of subnational data and attributable risk factors by regions in the context of the united nations millennium objectives: A case study of the gbd estimation in children attributable to wash in brazil. *In*: . [*s.l.*: *s.n.*], 2017.

SOUSA, T. C. M. d. *et al.* Doenças sensíveis ao clima no brasil e no mundo: revisão sistemática. **Revista Panamericana de Salud Pública**, SciELO Public Health, v. 42, p. e85, 2018.

STEIN, L. Creating a bioinformatics nation. **Nature**, Nature Publishing Group, v. 417, n. 6885, p. 119, 2002.

UCITELI, A.; KIRSTEN, T. Ontology-based retrieval of scientific data in life. **Datenbanksysteme für Business, Technologie und Web (BTW 2015)-Workshopband**, Gesellschaft für Informatik eV, 2015.

UREN, V. *et al.* Semantic annotation for knowledge management: Requirements and a survey of the state of the art. **Web Semantics: science, services and agents on the World Wide Web**, Elsevier, v. 4, n. 1, p. 14–28, 2006.

VAKKARI, P. Library and information science: its content and scope. *In*: **Advances in librarianship**. [*s.l.*]: Emerald Group Publishing Limited, 1994. p. 1–55.

VALLE, M. **Scientific Data Management** . 2018. http://mariovalle.name/sdm/scientific-data-management.html. 21/06/2018.

VERSTICHEL, S. *et al.* Efficient data integration in the railway domain through an ontology-based methodology. **Transportation Research Part C: Emerging Technologies**, Elsevier, v. 19, n. 4, p. 617–643, 2011.

VITA, R. *et al.* Fair principles and the iedb: short-term improvements and a long-term vision of obo-foundry mediated machine-actionable interoperability. **Database**, Oxford University Press, v. 2018, 2018.

WAAL, S. van der *et al.* Lifting open data portals to the data web. *In*: **Linked Open Data–Creating Knowledge Out of Interlinked Data**. [*s.l.*]: Springer, 2014. p. 175–195.

WAZLAWICK, R. **Metodologia de pesquisa para ciência da computação**. [*s.l.*]: Elsevier Brasil, 2017. v. 2.

WEST, M. **Developing high quality data models**. [*s.l.*]: Elsevier, 2011.

WIERINGA, R. Design science as nested problem solving. *In*: ACM. **Proceedings of the 4th international conference on design science research in information systems and technology**. [*s.l.*], 2009. p. 8.

WILKINSON, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. **Scientific data**, Nature Publishing Group, v. 3, 2016.

WOELFLE, M.; OLLIARO, P.; TODD, M. H. Open science is a research accelerator**Nature Chemistry**, Nature Publishing Group, v. 3, n. 10, p. 745–748, 2011.

ZHANG, S.; ZHANG, C.; YANG, Q. Data preparation for data mining. **Applied artificial intelligence**, Taylor & Francis, v. 17, n. 5-6, p. 375–381, 2003.