Integração de Dados de Publicações Científicas usando uma Abordagem baseada em Ontologias

João Paulo S. Andrade
jaosantos.andrade@gmail.com
LEDS, Instituto Federal do Espírito Santo
Serra - ES, Brasil

Julio Cesar Nardi
julionardi@ifes.edu.br
Instituto Federal do Espírito Santo
Colatina - ES, Brasil

Fabiano Borges Ruy
fabianoruy@ifes.edu.br
LEDS, Instituto Federal do Espírito Santo
Serra - ES, Brasil

ABSTRACT

Data integration has gained increasing importance with the growing volume of available data (so-called Big Data). It allows to put data from different sources together in an aligned way, so meaningful information can be extracted. Scientific publications are available in several data sources along the Internet. Integrating data from these distributed sources faces a challenge, mainly if we consider the heterogeneous aspects inherent to the structure/schema and the semantics of data. This paper presents a semantic integration initiative for integrating data from three distinct scientific publication sources into a shared repository. The integration process is conducted by a systematic approach that creates and applies a well-founded domain ontology to address the semantic aspects and to map concepts, relations, and data between the different data sources. As a final result, a web tool was built over the integrated repository to support some final user queries.

KEYWORDS

Data Integration, Scientific Publications, Semantics, Ontology

1 INTRODUÇÃO

Com o volume de dados em crescimento constante, iniciativas de integração de dados vêm sendo desenvolvidas há décadas. Isso reflete a importância de se ter dados integrados e o desafio que existe ao lidar com a criação de tais iniciativas [18].

Integrar diferentes bases de dados manualmente é um processo oneroso e muito propenso a erro, levando à necessidade de se buscar automatizar e sistematizar processos e ferramentas de integração de dados [18]. O grande volume de dados disponíveis (rotulado como Big Data) em várias bases de dados espalhadas e disponíveis pela Web traz às iniciativas de integração um desafio em capturar, preparar e trabalhar com os dados, lidando com problemas advindos de heterogeneidades sintáticas, semânticas e de esquema existentes entre tais dados [18].

No contexto acadêmico, publicações científicas são um importante meio de compartilhar informações visando o desenvolvimento tecnológico e científico [8]. Os textos acadêmicos são publicados em uma grande variedade de veículos de publicação (e.g., revistas e anais de conferências). Alguns portais buscam indexar tais artigos (tais como Google Scholar, Semantic Scholar, Scopus, DBPL), facilitando a busca, e conseguem oferecer, além dos artigos em si, dados importantes nesse domínio, como o perfil de autores, relevância dos veículos, fatores de impacto, citações etc. Por usarem diferentes algoritmos de identificação e indexação, cada portal apresenta um conjunto diferente de publicações, e também de informações relacionadas, servindo a diferentes públicos ou propósitos. Para os pesquisadores brasileiros, por exemplo, especialmente os associados a programas de pós-graduação, uma informação relevante é o estrato Qualis [1] de cada artigo. A CAPES fornece uma lista de periódicos e conferências (para a computação) com seus respectivos Qualis, atualizada periodicamente. Algumas ferramentas[1] apoiam a busca pelo estrato Qualis de periódicos e conferências, no entanto, consideram apenas os veículos, não havendo uma associação direta com os artigos de um dado pesquisador, por exemplo. Assim, a contabilização das publicações Qualis de um pesquisador, ou de um Programa de Pós-Graduação, demanda um esforço de contagem considerando o veículo de publicação ao qual o fator de impacto está associado.

Nesse sentido, o domínio de publicações científicas se apresenta como propício a explorar o desafio da integração de dados distribuídos pela Web, além de ser didático e conhecido de modo a favorecer à discussão no âmbito da integração de dados. Este trabalho se propõe a integrar dados de bases de dados distintas de publicações científicas, particularmente autores e seus artigos, associando-os respectivos estratos Qualis. Entretanto, integrar os dados de publicações científicas, a partir de diferentes plataformas com atenção à sua semântica não é uma tarefa simples. Envolve a compreensão de cada plataforma e dos dados nelas contidos, o mapeamento entre esses esquemas de dados, formas de captura, integração e apresentação dos dados envolvidos [2]. Para isso, foi proposta e aplicada uma Abordagem de Integração Semântica, orientada a modelos. A abordagem oferece um suporte a iniciativas de integração para que sejam realizadas de forma sistemática, explorando os aspectos semânticos e favorecendo a captura, armazenamento e apresentação dos dados. Como ponto central, e referência semântica para a integração dos dados, há uma ontologia de domínio. Como resultado da iniciativa de integração de publicações acadêmicas, além da ontologia no domínio de publicações e do repositório de dados, foi produzida uma ferramenta de consulta em que se pode buscar por pesquisador e obter seus respectivos artigos, compilados a partir de duas bases distintas, e classificados segundo o estrato Qualis.

Este artigo está organizado como segue. A Seção 2 discute aspectos de Integração Semântica de Dados. A Seção 3 apresenta o domínio de Publicações Científicas e portais e informações relacionadas. A Seção 4 descreve a iniciativa de integração de publicações científicas e como ela foi realizada segundo a abordagem. A Seção 5 demostra alguns resultados obtidos, incluindo a ferramenta de consulta. A Seção 6 discute trabalhos correlatos e a Seção 7 traz as considerações finais.

[1] e.g., https://ppgcc.github.io/discentesPPGCC/pt-BR/qualis/

2 INTEGRAÇÃO SEMÂNTICA DE DADOS

Integração de dados não é uma área recente na Computação. Em 1975, por recomendação do *National Bureau of Standards and the Association for Computing Machinery*, discutiam-se questões relativas à necessidade de dicionários de dados para facilitar o entendimento e a identificação de dados vindos de diferentes fontes [18].

Do ponto de vista prático, a integração de dados pode ser entendida como o processo de combinar duas ou mais fontes de dados[2]; cada fonte de dados possuindo seu respectivo esquema, o qual descreve como os dados estão estruturados [18].

Tal processo envolve, basicamente, atividades como [18]: (i) mapeamento de esquemas (por meio do alinhamento de conceitos e relações); (ii) implementação do mapeamento de esquemas (por meio regras/funções de transformação); (iii) resolução de entidade (identificação das diferentes instâncias de uma entidade/conceito); (iv) consolidação de entidade (agrupamento e organização das instâncias de uma determinada entidade); e (v) limpeza de dados (atividade de apoio realizada paralelamente às demais).

Fontes de dados são, em geral, construídas de maneira isolada (não integradas) e, por consequência, podem apresentar heterogeneidade sintática (causada pelo uso de diferentes linguagens/representações), semântica (causada por diferentes significados/interpretações dos dados em contextos diferentes) ou de esquema (causada por diferenças na estrutura de organização dos dados) [4].

Ainda que a heterogeneidade sintática e a de esquema sejam importantes, a heterogeneidade semântica tem sido foco de diferentes trabalhos, recentemente apresentando desafios que acompanham o desenrolar do processo de integração [18] . Nesse contexto, ontologias têm sido aplicadas em iniciativas de integração de dados a fim de lidar com problemas semânticos [18] [3] [14] [6] [5]. Basicamente, são aplicadas de três maneiras [21]: (i) *ontologia única*, os esquemas das fontes de dados são diretamente relacionados a uma ontologia global compartilhada; indicada nos casos em que as fontes de dados possuem mesmo nível de granularidade sobre o domínio em abordado; (ii) *múltiplas ontologias*, cada fonte de dados conta com sua própria ontologia local; as ontologias locais são mapeadas, então, entre si; (iii) *abordagem híbrida*, combina as duas anteriores; ontologias locais são criadas para cada fonte de dados, mas não integradas entre si, mas sim à ontologia global compartilhada.

Quanto à finalidade, ontologias podem ser classificadas em [12]: ontologia de referência e ontologia operacional. *Ontologias de referência* são tipicamente usadas de maneira *off-line*, ou seja, para apoiar humanos na tarefa de negociação de significado e estabelecimento de consenso a respeito de conceitos e relações de um dado domínio. Versões especializadas de ontologias de referência podem ser criadas para uso *run-time* sendo, portanto, processáveis por máquinas. Essas versões são chamadas *ontologias operacionais*, as quais sacrificam expressividade e fundamentação teórica em busca de certas propriedades computacionais.

Soluções de integração de dados baseadas em ontologias (*ontology-based data integration - OBDI*) [3] tipicamente são compostas por 3 (três) elementos: (i) representação do domínio de conhecimento fornecida por ontologia(s), (ii) fontes de dados a serem integradas (geralmente heterogêneas e independentes umas das outras) e (iii) mapeamento entre os dados e entidades das fontes de dados com os conceitos e relações da(s) ontologia(s). Por meio dessa tríade, é possível fornecer ao usuário consumidor de informação uma interface que o permita acessar os dados integrados.

[2]Os termos "fonte de dados" (*datasource*) e "conjunto de dados" (*dataset*) são usados na literatura de maneira flexível/indistinta. Neste trabalho, usaremos apenas o termo "fonte da dados", buscando uma noção mais ampla.

3 PUBLICAÇÕES CIENTÍFICAS

A escrita científica é uma etapa fundamental e indissociável do fazer Ciência, pois sem ela não há comunicação e debate sobre a descoberta frente ao que já estava posto [8]. No Brasil, já há alguns anos, tem crescido a discussão acerca da necessidade de não apenas aumentar o volume de publicações científicas, mas, principalmente, da qualidade das publicações [20].

A qualidade da publicação científica tem sido mapeada em uma série de índices associados a veículos de publicação (revistas, conferências), a publicações científicas (artigos, capítulos de livro) e a pesquisadores. Tais índices buscam, mesmo com limitações, traçar algum perfil de qualidade e/ou impacto científico. Alguns dos índices mais conhecidos pela comunidade científica são: fator de impacto JCR (*Journal Citation Reports*), nº citações, índice H e índice i10.

Veículos de publicação de qualidade dão reputação ao conteúdo e ampla divulgação por meio dos melhores indexadores e bases de publicações científicas [20]. As bases de publicações científicas clássicas (e.g., Web of Science, Scopus e ACM Library) juntamente com as máquinas de busca acadêmicas (e.g., Google Scholar e Semantic Scholar) desempenham um importante papel no suporte ao pesquisador, facilitando o acesso a trabalhos científicos relevantes. Nesse contexto, destaca-se o caráter "aberto" das máquinas de busca acadêmicas, uma vez que muitas das bases de publicações científicas clássicas apresentam certas restrições de acesso.

Ainda que haja restrições no acesso a certos veículos de publicação e às publicações relacionadas, há um movimento importante no sentido de compartilhar, livremente, resultados de publicações científicas e índices de qualidade/impacto científico. Tal movimento conta com o livre acesso a certos veículos de publicação (*open access journals*) e a relatórios de índices fornecidos por diferentes instituições como, por exemplo, o Qualis Referência (pela Capes) e o JCR (pelo ISI - *Institute for Scientific Information*).

Esses dados, entretanto, estão acessíveis de maneira descentralizada (em geral, nos sites de cada instituição responsável) cabendo ao pesquisador interessado realizar a busca, o mapeamento e a integração por conta própria. Ademais, tais dados, em geral, estão disponibilizados em formatos distintos (e.g., PDF, XLS), o que gera ainda mais dificuldade para o pesquisador. Assim, o cenário de integração de dados de publicações apresenta desafios comuns a outros cenários de integração de dados (e.g., [18] [22]) que levam à necessidade de se utilizar abordagens sistemáticas e bem fundamentadas.

4 INTEGRANDO DADOS DE PUBLICAÇÕES CIENTÍFICAS

A iniciativa de integração apresentada neste trabalho considera dados de publicações disponíveis no Google Scholar, Semantic Scholar e estratos Qualis fornecidos pela Capes. Ela foi conduzida utilizando uma abordagem que compreende quatro fases: Planejamento, Modelagem, Implementação e Execução. Tais fases abordam desde a

definição do escopo das fontes de dados utilizadas até o acesso aos dados já integrados por parte do usuário final.

4.1 Planejamento

A fase de **Planejamento** visa a definir os objetivos de integração e os limites de escopo de uma iniciativa. Assim, são estabelecidos: (i) o Propósito da Iniciativa; (ii) as Fontes de Dados a serem utilizadas; (iii) a descrição do Escopo de abrangência da iniciativa de integração, considerando quais dados devem ser integrados e quais não serão considerados; e (iv) as Questões de Interesse, que suportam a criação de um modelo comum de integração (e.g., Ontologia) e a definição de consultas sobre o repositório a ser criado. Esta fase é importante para definir o foco da iniciativa de integração, delimitar seu escopo e antecipar questões a serem tratadas ao longo do processo. A Tabela 1 sumariza os itens mencionados.

Table 1: Sumarização do Planejamento

Item	Descrição
`Propósito`	Prover o usuário com dados de publicações científicas de diferentes fontes com seus respectivos índices de qualidade científica.
`Fontes de dados`	Google Scholar, Semantic Scholar e Relatórios Qualis da Capes.
`Descrição do escopo`	Os dados considerados são: (i) dados de pesquisadores (nome, índices bibliométricos, afiliação e áreas de interesse); dados de publicações (título, autores, data de publicação, tipo de publicação, total de citações); dados de veículos de publicação e índice Qualis.
`Questões de interesse`	Q1: Quais as publicações de um dado pesquisador? Q2: Qual o estrato Qualis de cada publicação? Q3: Quais os índices de publicação obtidos pelo autor nos últimos 5 anos?

4.2 Modelagem

A fase de **Modelagem** trata da escavação dos modelos de cada fonte de dados e da criação do modelo de referência de integração. Em seguida, define os mapeamentos semânticos entre esses modelos. Esta fase tem um papel fundamental na iniciativa, pois permite lidar com heterogeneidades (de origem sintática, estrutural ou semântica) que surgem ao se integrar diferentes fontes de dados. Por meio de modelos, torna-se mais explícita a estruturação dos conceitos, relações e instâncias/dados pertencentes ao domínio advindos de cada fonte de dados. Esta fase é composta por três atividades.

(i) *Escavar modelos das fontes de dados.* Aborda a construção do modelo estrutural de dados (e.g., um diagrama de classes) para cada fonte de dados a ser integrada. Os modelos são construídos analisando-se a estrutura/esquema e os dados de cada fonte. Nessa atividade, é necessário entender as estruturas dos dados lidando com as possíveis linguagens de representação (JSON, XML etc.) e formatos (PDF, XLS etc.) nos quais o conjunto de dados de cada fonte é obtido. Diferenças entre essas representações e formatos são tratadas. Como resultado, os modelos escavados passam a ser representados em uma linguagem única (e.g., UML - *Unified Modeling Language*).

(ii) *Criar modelo conceitual de integração.* Aborda a construção de um modelo conceitual comum (e.g., uma ontologia de referência), que represente o domínio de interesse e cujo escopo é aquele definido para a iniciativa de integração. Esse modelo tem a função de apoiar o entendimento do domínio como ele é, bem como dirimir eventuais dúvidas durante o processo de negociação/alinhamento de significado de conceitos e relações.

(iii) *Realizar mapeamento entre os modelos.* Suporta a definição dos mapeamentos semânticos entre os conceitos e relações dos modelos escavados de cada fonte de dados com os conceitos e relações do modelo conceitual de referência. Dependendo do escopo e do propósito da integração, o modelo de referência pode ser reduzido a um modelo mais operacional cujo propósito estará mais voltado à aplicação específica e a aspectos de implementação dos mapeamentos.

Neste trabalho, foram escavados os modelos de cada fonte de dados utilizada (Google Scholar, Semantic Scholar e estratos Qualis), considerando as respectivas linguagens de representação e formatos disponibilizados. Os dados do Google Scholar e Semantic Scholar foram disponibilizados pelas plataformas no formato texto estruturado por tags HTML. Os dados do Qualis foram obtidos em planilhas Excel e em formato texto. Para desenvolver/escavar cada modelo, analisou-se o esquema e o conjunto de dados a fim de identificar conceitos e relações. Os modelos escavados, exemplificando as origens dos atributos, são apresentados pelas Figuras 1, 2 e 3.

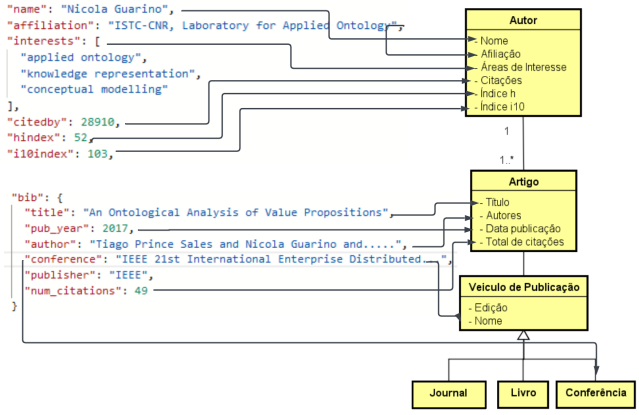

Figure 1: Modelo escavado a partir do Google Scholar.

Por se serem fontes de dados heterogêneas, podem-se notar diferenças entre os modelos escavados. Enquanto o Semantic Scholar traz mais informações sobre métricas individuais de cada artigo, o Google Scholar fornece mais informações sobre métricas do autor. O Qualis, por sua vez, traz informações não disponíveis nas duas outras fontes de dados, tanto para periódicos quanto para eventos.

Uma vez criados os modelos das fontes de dados, desenvolveu-se o modelo conceitual de referência, mais especificamente uma

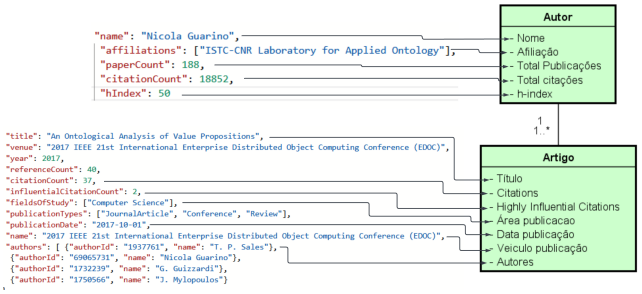

Figure 2: Modelo escavado a partir do Semantic Scholar.

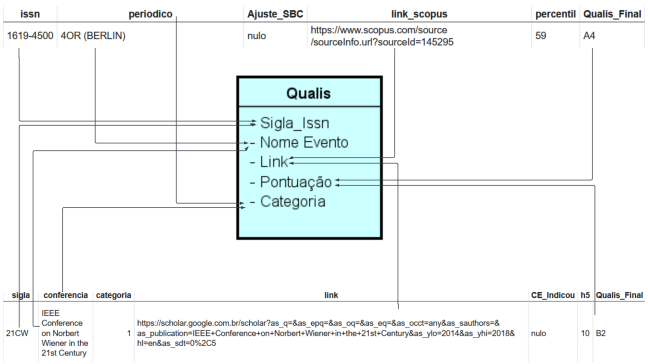

Figure 3: Modelo escavado a partir dos estratos Qualis.

ontologia de referência para o domínio de publicação científica. Tal ontologia tem o propósito de representar o domínio estudado com a maior fidelidade possível, independentemente dos dados a serem trabalhados na iniciativa. Assim, pode ser utilizada como uma interlíngua fidedigna ao domínio de publicações científicas. As Figuras 4, 5 e 6 apresentam a Ontologia de Publicações Científicas. O primeiro diagrama aborda os tipos de publicação, autoria e edição; o segundo foca nos índices de qualidade/impacto científico; e o terceiro aborda a afiliação de um autor a uma instituição e suas áreas de pesquisa.

Na especificação da ontologia de referência, foi utilizada OntoUML [13], uma linguagem de modelagem de ontologias baseada em um perfil UML que incorpora distinções ontológicas, representadas por estereótipos, advindas de UFO (*Unified Foundational Ontology*) [11]. UFO é uma ontologia de fundamentação que oferece distinções ontológicas básicas como eventos, objetos, propriedades e tipos, dentre outros [11]. Tal ontologia que vem sendo amplamente utilizada como fundamentação para a construção de outras ontologias (p.ex., [10] [9] [7] [17]). Para além dos benefícios em se adotar explicitamente distinções ontológicas na construção de ontologias [11], a escolha por OntoUML foi também motivada pela disponibilidade de um ferramental de suporte à engenharia de ontologias que inclui verificação de modelos e transformação para linguagens de implementação como OWL (*Ontology Web Language*).

No modelo apresentado pela Figura 4, **Autor Científico** é uma **Pessoa Física**, quando esta estabelece uma relação de **Autoria** com um determinado **Texto Científico**. **Texto Científico Editado** e **Texto Científico Publicado** são especializações de **Texto Científico**, na medida em que passa pela edição e publicação e toma parte em **Obra Científica Editada** e **Obra Científica Publicada**, respectivamente. Tais obras possuem relação, respectivamente, com **Editor** e com **Publicador**. **Publicação de Obra Científica**, além de relacionar obra com publicador, também a associa com **Veículo de Publicação Científica**, que pode ser **Periódico** ou **Evento**. Textos científicos publicados são especializados em **Resumo Publicado**, **Artigo Publicado** e **Capítulo de Livro Publicado**. Cada um desses tipos possui suas particularidades. Como exemplo, considere **Artigo Publicado**, que pode ser especializado em **Artigo de Evento Publicado** e **Artigo de Periódico Publicado**, cada um sendo membro, respectivamente, de **Anais de Evento** e **Fascículo**.

Pelo modelo da Figura 5, **Texto Científico Publicado**, **Veículos de Publicação Científica** e **Autor Científico** são **Entidades Avaliáveis**. Portanto, tais entidades possuem **Índice de Qualidade Científica**, que é especializado em **Índice de Autor**, **Índice de Veículo de Publicação** e **Índice de Texto Científico Publicado**.

Conforme mostra a Figura 6, um **Autor Científico** possui **Áreas de Pesquisa**. Ademais, ele pode ser um **Autor Científico Independente** ou um **Autor Científico Afiliado**. Neste caso, ele é um **Colaborador** de uma **Instituição**, com a qual possui um vínculo caracterizado, no modelo, pela **Afiliação**.

Uma vez desenvolvida a ontologia de referência, a atenção se volta ao contexto específico da iniciativa de integração, focando-se nos dados disponíveis para integração. Assim, uma ontologia operacional foi desenvolvida, para lidar com aspectos de *design* necessários à solução final de integração. Tais aspectos requerem que sejam incorporadas preocupações relativas a como conceitos e relações são implementados em linguagens de representação que focam mais em questões de implementação do que em expressividade conceitual. Assim, passa-se a se preocupar com a migração do nível conceitual de referência para o nível operacional (de implementação).

A Figura 7 apresenta o modelo da ontologia operacional, que permite a integração dos dados em um formato padronizado, uniformizando o armazenamento e o processamento de consultas. Esse modelo segue os princípios do modelo de referência, mas com algumas adaptações visando otimizar a atividade de implementação. Nesse modelo simplificaram-se algumas relações entre as classes. Por exemplo, Autor Cientifico e Métricas de Autor foram unificadas; o mesmo foi aplicado para Texto Científico Publicado e suas métricas. Ademais, algumas hierarquias de classes na ontologia de referência foram também reduzidas (e.g., a hierarquia de Texto Científico) otimizando a implementação. De todo modo, a ontologia de referência pode ser sempre acessada para dirimir eventuais dúvidas conceituais durante o processo de implementação.

Escavados os modelos das fontes de dados e derivado a ontologia operacional, foi realizado o mapeamento semântico [16] entre o modelo operacional e os modelos escavados, conforme apresenta Tabela 2. Por meio desse mapeamento é possível identificar e associar cada elemento dos modelos das fontes de dados com o modelo operacional que servirá de base para a construção do repositório integrado. Como resultado, se estabelece como os dados obtidos de cada fonte serão ajustados, quando pertinente, e inseridos no repositório, já dirimindo eventuais heterogeneidades existentes.

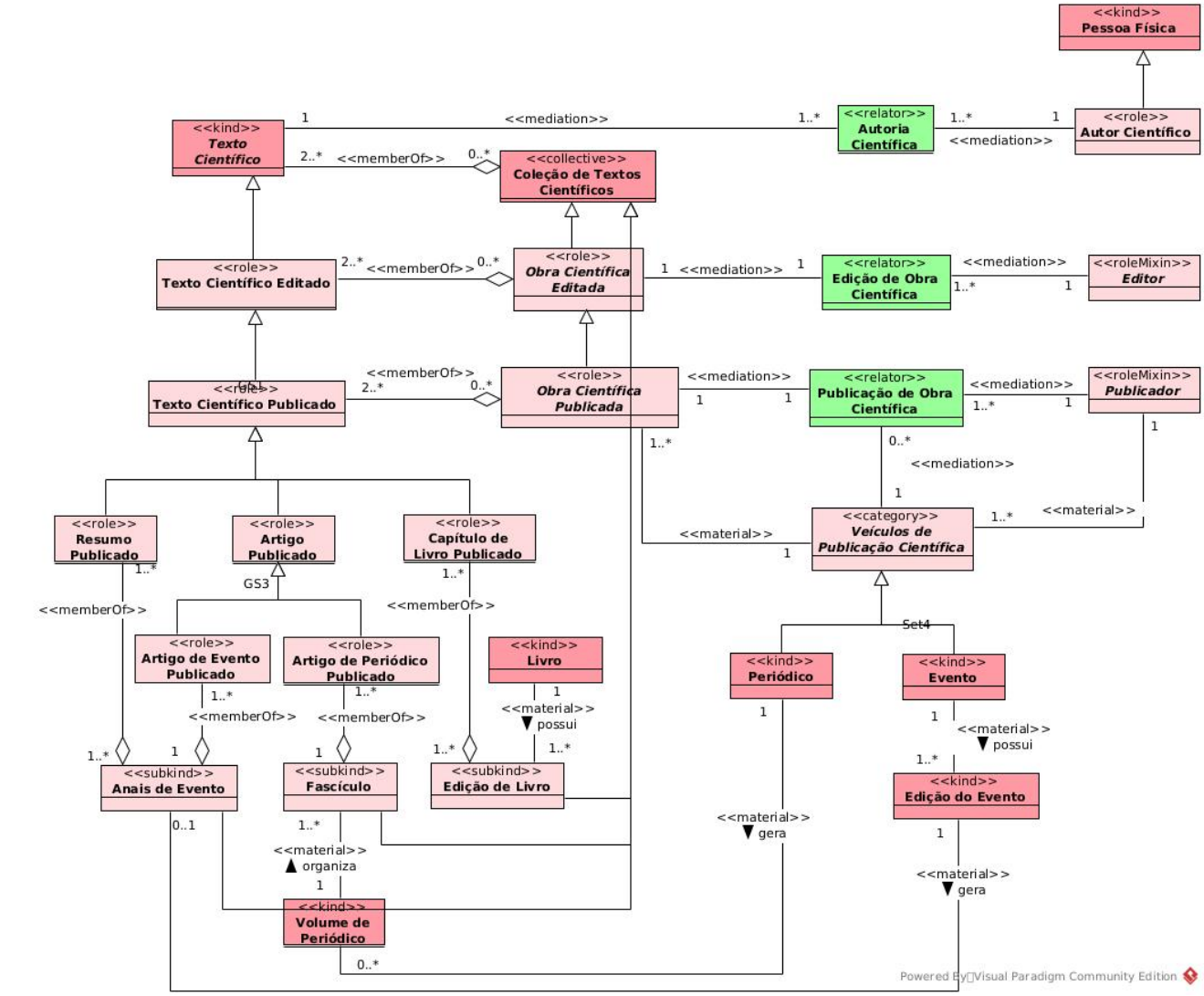

Figure 4: Ontologia de referência: fragmento de publicações.

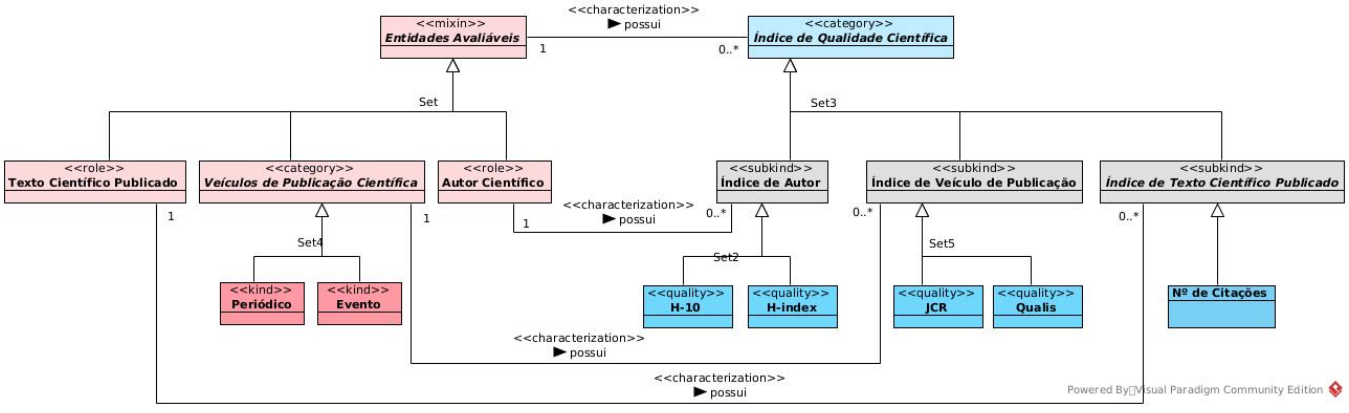

Figure 5: Ontologia de referência: fragmento de índices de qualidade.

Table 2: Mapeamento entre a ontologia operacional e os modelos das fontes de dados

Ontologia	Google Scholar	Semantic Scholar	Qualis
`Autor Cientifico`	Autor	Autor	—
`Autor Científico.indicei10`	Autor.indicei10	Autor.indicei10	—
`Autor Científico.indiceh`	Autor.indiceh	Autor.h-indice	—
`Autor Científico.citations`	Autor.citations	Autor.totalCitações	—
`Área de Pesquisa`	Autor.areasInteresse	—	—
`Instituição`	Autor.afiliacao	Autor.afiliacao	—
`Texto Científico Publicado`	Artigo	Artigo	—
`Texto Científico Publicado.numCitacoes`	Artigo.totalCitações	Artigo.citations	—
`Veículo de Publicação Científica`	Veículo de Publicação	Artigo.veiculoPublicacao	Veículo
`Qualis.extrato`	—	—	Qualis.pontuação

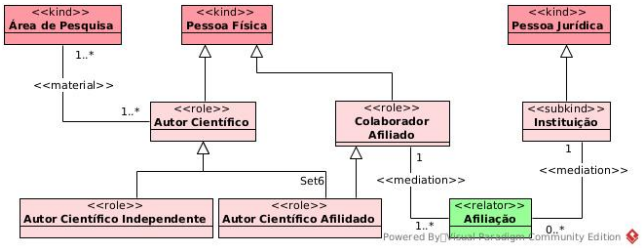

Figure 6: Ontologia de referência: fragmento de afiliação e áreas de pesquisa.

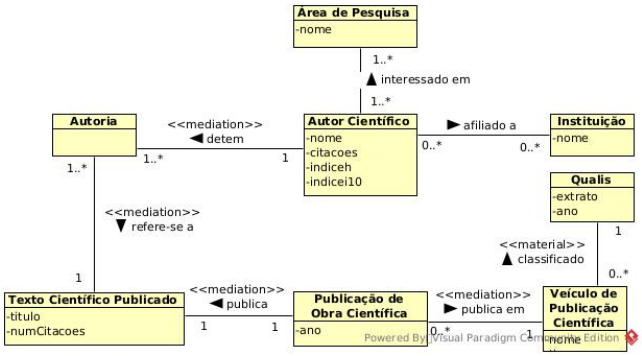

Figure 7: Modelo da ontologia operacional.

4.3 Implementação

Na fase de **Implementação** são desenvolvidos os algoritmos para coleta e transformação dos dados e estabelecido um repositório comum para abrigar os dados integrados. Essa fase é composta por três atividades: (i) Implementar a Obtenção dos Dados, (ii) Desenvolver o Repositório dos Dados, e (iii) Implementar Limpeza e Transformação dos Dados. Uma vez definido o domínio dos dados, as fontes de informação e o modelos de referência e operacional, deve-se então identificar e extrair a informação relevante nas plataformas de publicação. Assim, a fase de implementação aborda todo processo de criação dos algoritmos para a extração de informação relevante nas plataformas selecionadas até a disponibilização dos dados integrados ao usuário final. A Figura 8 ilustra a arquitetura da solução implementada.

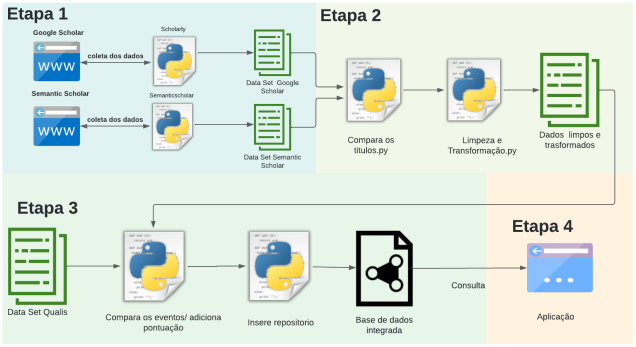

Figure 8: Arquitetura de alto nível da solução.

A fase de implementação iniciou com o desenvolvimento dos algoritmos para a coleta dos dados das fontes. Para tanto, foram utilizadas bibliotecas Python, a saber: (i) *scholarly*, que permite recuperar informações de autores e publicações do Google Scholar; e (ii) *semanticscholar*, que visa a recuperar dados da API Semantic Scholar. Com os algoritmos de coleta criados, a etapa seguinte foi criar os algoritmos para comparar os títulos dos artigos, utilizando a biblioteca *fuzzywuzzy* que fornece o índice de similaridade entre duas strings. Foram considerados iguais os artigos com similaridade superior a 60 nos títulos. No processo de limpeza e transformação, foram removidos espaços em branco e acentos dos títulos para facilitar o uso das informações como instâncias na ontologia. Para coletar os dados referentes ao Qualis, a biblioteca *pandas* permitiu a leitura dos arquivos contendo as informações. A mesma lógica de comparação dos artigos foi aplicada para comparar os eventos/periódicos e atribuir a pontuação Qualis a cada artigo individualmente.

O repositório foi criado com base no modelo operacional, gerando um arquivo OWL (*Ontology Web Language*) e a ferramenta *Protégé* apoiou o suporte a visualização e conferência dos dados. O código de população dos dados no repositório usou a biblioteca *RDFlib*, que permite serializar o OWL na aplicação, para que os dados do repositório (instâncias da ontologia) possam ser criados via código.

4.4 Execução

Por fim, na fase de **Execução**, os algoritmos criados são executados, realizando a coleta das informações, os processos de limpeza e mapeamento dos dados, e a inserção dos dados já integrados e aderentes à ontologia no repositório. A partir de então é possível consultar e apresentar os dados consistentemente.

A execução inicia-se com a coleta das informações de pesquisadores e suas publicações. A coleta inicia com a busca do pesquisador no Google Scholar, utilizando o Scholarly, que retorna uma estrutura no formato *JSON* contendo as informações do pesquisador e os títulos de suas publicações. Para obter as informações detalhadas de cada publicação, há uma busca direcionada, considerando apenas os artigos dos últimos 5 anos. Em seguida, consultou-se a base Semantic Scholar. Nela, as buscas são feitas por "id" do pesquisador, e os dados recuperados de maneira similar.

Tendo as informações referentes ao pesquisador das duas bases de dados, os títulos dos artigos são obtidos e comparados a fim de evitar duplicação na inserção de dados no repositório. Realizada a comparação, a limpeza e transformação dos dados, obtém-se a lista de publicações do pesquisador dos últimos 5 anos. De posse dessas publicações, os eventos e periódicos são comparados para adicionar o respectivo Qualis. Para o Qualis, utilizou-se uma planilha CVS contendo o índice de conferências e periódicos.

Com o repositório já criado, foi populado com base na tabela de mapeamentos e nos dados obtidos das bases de publicações. A Figura 9 apresenta o processo de população do repositório.

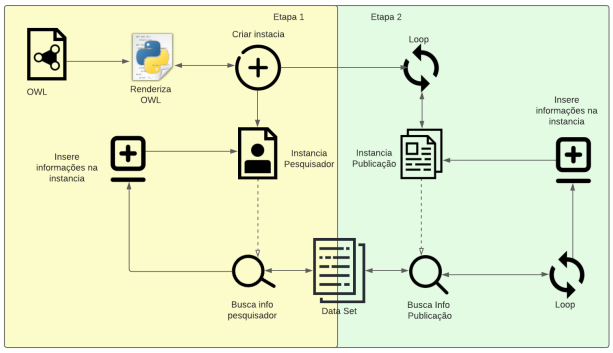

Figure 9: Modelo de comparação de dados.

A população do OWL com as informações do pesquisador e suas publicações é feita da seguinte forma. Primeiro o OWL é renderizado na aplicação e é criada uma instância para o pesquisador, a partir da qual é realizada uma busca para obter as informações que serão inseridas. Em seguida, para popular o OWL com as informações das publicações, é criada uma instância para cada publicação do pesquisador. Por fim são obtidas as informações do artigo, que é associado ao pesquisador instanciado inicialmente.

5 RESULTADOS OBTIDOS

Para demonstrar a utilização do repositório e prover uma ferramenta útil a pesquisadores brasileiros, foi desenvolvido um protótipo em aplicação Web, conforme ilustra a Figura 10. Por meio dele é possível realizar consultas pré-definidas e responder às questões de interesse da iniciativa: Q1. Quais as publicações de um dado pesquisador? Q2. Qual o estrato Qualis de cada publicação? Q3. Quais os índices de publicação obtidos pelo autor nos últimos 5 anos?

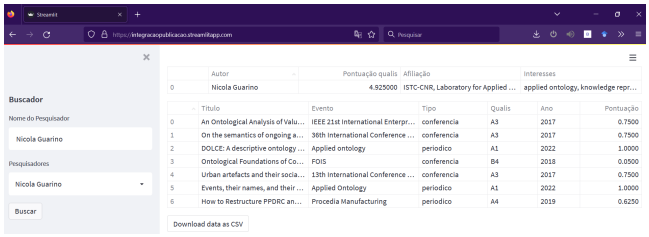

Figure 10: Interface da aplicação web.

Ao acessar a aplicação, o usuário digita o nome do pesquisador e solicita a consulta (barra na lateral direita). Em seguida, o algoritmo busca a lista de autores por similaridade de nome. Com o resultado da consulta, o usuário seleciona o pesquisador desejado. Assim, inicia-se o processo de integração no qual são aplicados os algoritmos para a coleta, limpeza e população de dados no repositório. Por fim, a partir da busca no repositório, é apresentada uma tabela com os dados das publicações do pesquisador com os respectivos Qualis.

A aplicação Web foi criada usando o *framework streamlit*, que é uma estrutura de código aberto e que facilita a criação e implantação de aplicações Web usando Python. Tal aplicação foi implantada/hospedada por meio do *Streamlit Teams* e pode ser acessada no link https://integracaopublicacao.streamlitapp.com/.

6 TRABALHOS CORRELATOS

Em [18], Sagi e colegas apresentam uma iniciativa de integração de dados oceânicos usando um processo de integração organizado em três fases: *discovery*, *merge* e *evaluate/correct*. Os autores advogam pela aplicação de ontologias no suporte à integração e ao acesso aos dados. São utilizados vocabulários e taxonomias para apoiar algoritmos que automatizam partes do processo de integração. Tal como propõe este artigo, Sagi e colegas também preocupam-se em adotar um processo de integração bem estabelecido cujos aspectos semânticos inerentes à integração de fontes de dados heterogêneas são abordados ao longo de processo.

Em [19], Sun e colegas propõem um *framework* - GeoDataOnt - de integração e compartilhamento de dados geoespaciais, composto por três módulos (ontologias essencial, de morfologia e de proveniência) e três níveis (geral, de domínio e de aplicação, em referência aos níveis de generalidade de ontologias), sendo implementado em OWL. Os autores preocupam-se em caracterizar problemas semânticos da integração de dados e estabelecem ontologias como principal artefato para apoiar sua solução. Em relação ao uso de um processo sistemático, o foco está em apresentar as ontologias que compõem o *framework*. Como consequência, aspectos relacionados ao mapeamento semântico entre fontes de dados não são abordados. Ademais, estando voltado à proposição de um *framework* conceitual, o trabalho não foca em apresentar uma aplicação com um repositório de dados integrado para o usuário.

Em [22], Wang e colegas apresentam uma iniciativa de integração no domínio de publicação cientifica, utilizando uma ontologia para mediar o alinhamento semântico entre fontes de dados

heterogêneas. A ontologia é especificada em RDF e os mapeamentos semânticos entre os esquemas das fontes de dados e a ontologia são descritos usando consultas SPARQL. Os autores propõem uma arquitetura em quatro camadas: aplicação, que se comunica com os usuários; mediação, que atua para uniformizar/alinhar as operações sobre as fontes heterogêneas; *wrappers*, onde se encontram os *wrappers* de acesso a cada fonte de dados; e origem, portando os dados. Assim como o presente trabalho, há uma preocupação em se estabelecer os mapeamentos entre as fontes de dados e a ontologia e, então, permitir que o usuário tenha acesso aos dados integrados. Um diferencial está em nossa preocupação nas atividades de cunho conceitual que precedem os mapeamentos e instanciações da ontologia. Advoga-se, portanto, que uma boa análise conceitual suportada por ontologias bem fundamentadas tende a facilitar o trabalho de implementação da solução de integração.

Em [15], Bravo e colegas descrevem uma abordagem para população automatizada de uma ontologia de publicações científicas, para favorecer o acesso e recuperação de informações. Tal abordagem é baseada em uma estratégia de enriquecimento semântico com aplicação de técnicas de cálculo de similaridade entre publicações científicas. Sua arquitetura contempla atividades como extração de dados, população da ontologia, enriquecimento semântico e aplicação de raciocínio. Há uma ontologia de publicações científicas com conceitos como autor, publicação e tópico de pesquisa. De maneira similar à nossa abordagem, os dados são extraídos das fontes, trabalhados e instanciados em uma ontologia, seguindo um processo base mínimo que guia a iniciativa de integração. Neste trabalho, entretanto, focamos em mapear conceitos e relações das fontes de dados de maneira manual, analisando a fundo questões semânticas inerentes aos esquemas das fontes de dados e suas relações com a ontologia. Isso influência, p.ex., no nível de detalhamento e representação da ontologia de referência proposta, em contraponto àquela utilizada por Bravo e colegas. No nível de implementação, enquanto a ontologia de Bravo e colegas é mapeada em um diagrama de classes, neste trabalho optou-se por transformar a ontologia de referência, inicialmente representada em OntoUML, em uma versão operacional especificada em OWL, a partir de onde são realizadas as consultas cujos dados são apresentados ao usuário.

7 CONCLUSÕES

Este trabalho apresenta uma iniciativa de integração de dados de publicações científicas baseada em ontologia, a qual foi conduzida tomando-se como base quatro fases: planejamento, modelagem, implementação e execução. Como contribuições são apresentadas: (i) uma ontologia de referência do domínio de publicações científicas (incluindo índices de qualidade/impacto científico); (ii) um processo composto por quatro fases conduzidas na realização do integração; e (iii) uma aplicação resultante da iniciativa que integra dados de publicações científicas advindos do Google Scholar, do Semantic Scholar e do extrato Qualis da Capes.

Tal integração favorece a busca por publicações de um determinado pesquisador a partir de mais de uma fonte de dados, e a sua classificação e pontuação a partir de um índice de qualidade.

Como trabalhos futuros citam-se o desenvolvimento de uma abordagem sistemática para integração semântica de dados baseada em modelos conceituais, melhorias nos algoritmos de captura e transformação dos dados e da própria aplicação Web construída, e a aplicação dessa abordagem em outros cenários de integração de dados como, por exemplo, de dados de casos e óbitos de Covid.

ACKNOWLEDGMENTS

Os autores agradecem ao apoio da FAPES e CAPES (processo 2021-2S6CD, FAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados) e pela Bolsa de Mestrado (processo 158/2020 - Edital FAPES 14/2019).

REFERENCES

[1] Rita de Cássia Barradas Barata. 2017. Dez coisas que você deveria saber sobre o Qualis. *Boletim Técnico do PPEC* 2, 1 (2017), 17p–17p.
[2] Gustavo Britto, Fabiano B Ruy, and Carlos LB Azevedo. 2020. Um ambiente para integração de dados abertos relativos à despesa pública. *Ontobras* (2020).
[3] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Riccardo Rosati, and Gestionale Antonio Ruberti. 2018. Ontology-Based Data Access and Integration.
[4] Isabel F Cruz, Huiyong Xiao, et al. 2005. The role of ontologies in data integration. *Engineering intelligent systems for electrical engineering and communications* 13, 4 (2005), 245.
[5] Eduardo M da Silva, Filipe W Mutz, and Fabiano B Ruy. 2022. Uso de Ontologias no suporte a aplicação de Machine Learning: um caso no domínio de Evasão Escolar. *Ontobras* (2022).
[6] Eduardo M da Silva, Fabiano B Ruy, and Filipe W Mutz. 2022. Abordagem para Análise de Múltiplas Fontes de Dados de Evasão Escolar. *Anais do Computer on the Beach* 13 (2022), 149–156.
[7] Thomas Derave, Tiago Prince Sales, Frederik Gailly, and Geert Poels. 2021. Comparing digital platform types in the platform economy. In *International Conference on Advanced Information Systems Engineering*. Springer, 417–431.
[8] Alacoque Lorenzini Erdmann. 2011. A importância da publicação científica. *Revista de Enfermagem da UFSM* 1, 2 (2011).
[9] Cristine Griffo, João Paulo A Almeida, and Giancarlo Guizzardi. 2018. Conceptual modeling of legal relations. In *International Conference on Conceptual Modeling*. Springer, 169–183.
[10] Cristine Griffo, João Paulo A Almeida, Giancarlo Guizzardi, and Julio Cesar Nardi. 2021. Service contract modeling in enterprise architecture: An ontology-based approach. *Information Systems* 101 (2021), 101454.
[11] Giancarlo Guizzardi. 2005. Ontological foundations for structural conceptual models. (2005).
[12] Giancarlo Guizzardi. 2007. On ontology, ontologies, conceptualizations, modeling languages. In *and (Meta) Models, Frontiers in Artificial Intelligence and Applications, Databases and Information Systems IV, IOS*. Citeseer.
[13] Giancarlo Guizzardi, Claudenir M Fonseca, Alessander B Benevides, João Paulo A Almeida, Daniele Porello, and Tiago P Sales. 2018. Endurant types in ontology-driven conceptual modeling: Towards OntoUML 2.0. In *Conceptual Modeling: 37th Int. Conference, ER 2018, Xi'an, China, October, 2018, Proc. 37*. Springer, 136–150.
[14] Julio C Nardi, Vitor N Alves, Carlos AR Junior, Fabiano B Ruy, and Robson Prucoli Posse. 2021. Semantic Sensor Network Ontology como modelo de referência em soluções de monitoramento agrícola. In *Anais do XIII Congresso Brasileiro de Agroinformática*. SBC, 273–282.
[15] Robert W Reid, Jacob W Ferrier, and Jeremy J Jay. 2020. Automated gene data integration with Databio. *BMC Research Notes* 13, 1 (2020), 1–5.
[16] Fabiano B Ruy. 2017. *Software Engineering Standards Harmonization: An Ontology-Based Approach.* Ph. D. Dissertation. Doctoral Thesis. Postgraduate Program in Computer Science. Federal University of Espirito Santo.
[17] Fabiano B Ruy, Érica F Souza, Ricardo A Falbo, and Monalessa P Barcellos. 2017. Software Testing Processes in ISO Standards: How to Harmonize Them?. In *Anais do XVI Simpósio Brasileiro de Qualidade de Software*. SBC, 296–310.
[18] Tomer Sagi, Yoav Lehahn, and Koby Bar. 2020. Artificial intelligence for ocean science data integration: current state, gaps, and way forward. *Elementa: Science of the Anthropocene* 8 (2020).
[19] Kai Sun, Yunqiang Zhu, Peng Pan, Zhiwei Hou, Dongxu Wang, Weirong Li, and Jia Song. 2019. Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. *Big Earth Data* 3, 3 (2019), 269–296.
[20] Gilson Luiz Volpato. 2007. Bases teóricas para redação científica. *São Paulo: Cultura Acadêmica* (2007).
[21] Holger Wache, Thomas Voegele, Ubbo Visser, Heiner Stuckenschmidt, Gerhard Schuster, Holger Neumann, and Sebastian Hübner. 2001. Ontology-based integration of information-a survey of existing approaches. In *Ois@ ijcai*.
[22] Jinpeng Wang, Jianjiang Lu, Yafei Zhang, Zhuang Miao, and Bo Zhou. 2009. Integrating Heterogeneous Data Source Using Ontology. *J. Softw.* 4, 8 (2009), 843–850.