Submission date: 20/12/2021
Camera ready submission date: 12/09/2022
1st round notification: 08/08/2022
Available online: 06/10/2022
Publication date: 06/10/2022

Section: extended version
Extended versions of selected papers from the X Brazilian Workshop on Social Network Analysis and Mining

Publicações Sobre COVID-19: Uma Análise Semântica Sobre Artigos Retratados

Publications About COVID-19: A Semantic Analysis On Retracted Articles

Hugo Duca[1], Ingrid Pacheco[1], Giseli Rabello Lopes[1], Maria Luiza M. Campos[1], Jonice Oliveira[1]

[1]Programa de Pós-Graduação em Informática – Instituto de Computação
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

{hugo.duca,mluiza}@ppgi.ufrj.br,
{ingridpacheco,giseli,jonice}@ic.ufrj.br

Abstract. *In a hurry for new discoveries to fight the COVID-19 pandemic, several retracted articles began to appear and be cited as references for other studies. In the present work, social network analysis techniques and linked data were used to assess the influence of retracted articles on those that referred to them. As a result, it was found that the most referred retracted work had 2,745 citations, 7 retracted works refer to others in the same situation, and 82.11% of the references were made after the retraction. In addition, the most used words in the titles were associated with possible treatments for the disease and the countries which had the most referred retracted articles are France and Vietnam.*

Keywords. *COVID-19; Retracted Articles; Social Network Analysis; Semantic Web; Linked Data; Ontologies.*

Resumo. *Com a pressa por novas descobertas para o enfrentamento à pandemia do COVID-19, diversos artigos retratados começaram a surgir e a serem citados como referência para outros estudos. Neste trabalho, são utilizadas técnicas de análise de redes sociais e dados conectados para avaliar a influência dos artigos retratados sobre os que os referenciam. Como resultado foi verificado que o trabalho retratado mais referenciado teve 2.745 citações, 7 retratados referenciam outros na mesma situação, e que 82,11% das referências foram feitas após a retratação. Além disso, as palavras mais usadas nos títulos eram associadas aos possíveis tratamentos da doença e os países com artigos retratados mais citados são França e Vietnã.*

Palavras-Chave. *COVID-19; Artigos Retratados; Análise de Redes Sociais; Web Semântica; Dados Conectados; Ontologias.*

Cite as: Duca, H., Pacheco, I., Lopes, G. R., Campos, M. L. M., Oliveira, J. (2022). Publications About COVID-19: A Semantic Analysis On Retracted Articles (Publicações Sobre COVID-19: Uma Análise Semântica Sobre Artigos Retratados).iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 15(1), 22:1-22:25. DOI:10.5753/isys.2022.2413

1. Introdução

Com o avanço da pandemia da *Corona Virus Disease 2019* (COVID-19), muitas publicações tiveram de ser aceleradas em busca de auxiliar no enfrentamento da crise [Soltani e Patini 2020] e trazer mais conhecimento para o público. Entretanto, devido à velocidade com que os trabalhos eram realizados, por vezes, suas fontes não eram verificadas corretamente ou seus dados eram infundados, o que acarretava em resultados imprecisos e uma consequente *retratação*. Em junho de 2020, em uma pesquisa no *Retraction Database*[1] usando as palavras-chave: "COVID-19", "coronavirus disease 2019", "coronavirus 2019", "SARS-COV-2" e "2019-nCov", foi possível encontrar 26 artigos que haviam sido removidos ou retratados [Soltani e Patini 2020].

De acordo com Sheth e Thaker (2014), uma retratação indica que o trabalho em questão não deveria ter sido publicado, que suas conclusões e dados não devem ser usados para apoio a pesquisas futuras.

A retratação de artigos pode ser definida como um mecanismo para corrigir a literatura e alertar as pessoas sobre artigos que contenham conteúdos ou dados errôneos e que suas descobertas ou conclusões possam ser falhas. O seu principal propósito não é punir os autores, mas garantir a integridade das publicações [Browman et al. 2019].

Diversos podem ser os motivos para um artigo ser retratado, dentre eles: plágio, pesquisa não confiável, dados publicados anteriormente e pesquisa antiética. Quando ocorre uma retratação, significa que ocorreu algum problema tão significativo com a pesquisa que suas descobertas podem ser invalidadas [Sheth e Thaker 2014].

Uma das grandes questões surge quanto à citação desses artigos. Quando usada corretamente, uma citação serve como uma ferramenta valiosa para apoiar uma afirmação, método ou hipótese. Por outro lado, um trabalho que foi oficialmente marcado como retratado pelo editor foi, em essência, riscado na íntegra do registro acadêmico [Silva e Bornemann-Cimenti 2017]. Entretanto, muitos são os casos em que os artigos, mesmo após sofrerem retratação, continuam sendo usados para apoiar novas descobertas científicas, o que pode colocar em risco o processo científico. Ainda assim, muitas vezes, as redes de citações acadêmicas expõem as ligações existentes na literatura entre artigos, autores e projetos de pesquisa [Shotton 2010], sem explicitar a real intenção por trás de cada citação existente.

Diversos são os fatores que podem levar autores a fazer citação de um artigo após a sua retratação, mas um deles é que, algumas vezes, as publicações que sofreram retratação são sinalizadas como retiradas ou retratadas apenas na plataforma do editor, mas não na plataforma dos agregadores de conteúdo. Dessa forma, o banco de dados torna-se desatualizado e os autores não são avisados sobre a questão [Bar-Ilan e Halevi 2017]. No site `retractionwatch.com`, pode-se observar exemplos de artigos que possuem o número de citações ainda maior após a retratação do mesmo, demonstrando o crescente fenômeno da citação pós retratação.

Neste contexto, procurando analisar retratações e citações na temática da COVID-19, são formuladas as seguintes perguntas de pesquisa a serem respondidas pelo presente

[1] http://retractiondatabase.org/RetractionSearch.aspx

trabalho:

- **P1** Quais são os artigos retratados mais referenciados?
- **P2** Quais os principais temas e similaridades entre os artigos retratados?
- **P3** As citações ocorreram antes ou após o aviso de retratação?
- **P4** Existe uma cadeia de artigos retratados (que referenciam outros na mesma condição)?
- **P5** Quais são os artigos que mais referenciam artigos retratados (possivelmente diminuindo significativamente a sua qualidade)?
- **P6** Quais as principais áreas de conhecimento dos artigos retratados sobre COVID-19?
- **P7** Quais os países com mais artigos retratados e mais referências a eles?
- **P8** Quais os autores que mais publicaram artigos retratados?
- **P9** Quais os veículos de publicação que mais publicaram artigos retratados?
- **P10** Quais as intenções de citação aos artigos retratados?

Visando responder as questões de pesquisa elencadas e considerando a falta de um conjunto de dados único com as informações necessárias, o presente trabalho visa realizar este estudo através do uso de técnicas de análises de redes sociais e também uso de princípios de dados conectados para possibilitar análises semânticas sobre os dados bibliográficos, citações e retratações foram construídos, ambos através de *scripts* em *Python*. A fim de responder as questões de **P1** a **P5**, foi criado um primeiro conjunto de dados contendo 5.884 trabalhos que referenciavam 73 artigos retratados na temática de COVID-19. Foi construída uma rede com base no primeiro conjunto de dados e empregadas diferentes técnicas de análise de redes sociais (ARS) [Wasserman e Faust 1994].

Assim, foi possível encontrar os 5 artigos retratados mais referenciados, majoritariamente após o aviso de retratação, dentre os quais um deles é citado por 2.745 artigos e outro ainda referencia 2 trabalhos na mesma situação, demonstrando a existência de uma cadeia de referências de retratados. Ademais, foi possível descobrir que uma grande porção destes mencionava remédios que eram citados como possíveis tratamentos para a COVID-19 no início da pandemia. A partir destes resultados, ficou perceptível o quão recorrente são as citações a artigos retratados e o quanto isto pode impactar na confiabilidade das pesquisas.

Por outro lado, para responder as questões de pesquisa de **P6** a **P10**, era necessário um olhar mais aprofundado sobre os artigos retratados, analisando também os seus autores e até os seus países. Essa imersão permite evidenciar se esta é uma prática comum a certo pesquisador ou até mesmo quais os países que tendem a publicar mais retratações, para que possivelmente revejam seus métodos de avaliação e pesquisa. Tal correlação, entretanto, não se fazia presente nos dados extraídos das bases mencionadas e, para isto, a criação de dados conectados e o uso de ontologias auxiliares foram necessários.

Dados conectados, por sua vez, são dados que apontam para outros dados, ou seja, dados inter-relacionados. Eles foram criados com a finalidade de melhorar a qualidade com as quais dados estavam sendo publicados na Web e padronizar um conjunto de boas práticas [Berners-Lee 2006]. Devido ao seu princípio de incluir *links* para outras fontes de dados, é possível trazer mais semântica para os dados brutos encontrados anterior-

mente, necessitando de um vocabulário auxiliar para conseguir relacionar os dados complementares. Neste ponto, portanto, o presente artigo propõe o uso das ontologias FaBio (*FRBR-aligned Bibliographic Ontology*), uma ontologia para registrar e publicar descrições de referências bibliográficas na Web Semântica, e CiTo (*Citation Typing Ontology*), uma ontologia para a caracterização de citações bibliográficas [Peroni e Shotton 2012].

Dessa forma, o segundo conjunto de dados foi gerado posteriormente e já continha 151 artigos retratados que foram triplificados, possibilitando que buscas mais complexas pudessem ser executadas para explorar as informações semânticas relacionadas. Tal proposta estende trabalho anterior [Duca et al. 2021], utilizando técnicas e ferramentas de dados conectados (*Linked Data*), para realizar análises semanticamente mais aprofundadas sobre a rede de citações dos artigos retratados sobre COVID-19.

O restante deste artigo está organizado da seguinte forma. Na seção 2, são discutidos os trabalhos relacionados a artigos, citações e retratações na temática da COVID-19. Na seção 3, é feita uma fundamentação teórica dos principais conceitos relacionados ao trabalho. Na seção 3.1 são discutidas métricas utilizadas na análise de redes sociais. Na seção 3.2, é feita uma contextualização sobre dados conectados, incluindo uma breve apresentação das ontologias FaBio e CiTo. Na seção 4, são apresentadas as informações sobre a criação dos conjuntos de dados. Na seção 5, são apresentadas as análises realizadas e os resultados obtidos. Por fim, na seção 6, são apresentadas as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção, são discutidos alguns trabalhos relacionados à publicação, citações e retratações de artigos na temática da COVID-19/SARS-CoV-2.

2.1. Publicações e Citações

Diferentes estudos enfatizam o crescimento das publicações sobre COVID-19 ao longo da pandemia. No final de 2020, havia estimativas acerca da publicação de mais de 100.000 artigos e pré-*prints* sobre o assunto, e da possibilidade de este número ultrapassar os 200.000 em dezembro do mesmo ano [Else 2020]. Além disso, dentre as análises apresentadas estão os principais tópicos abordados (modelagem epidemiológica, controle de propagação, saúde pública, diagnósticos e testes, saúde mental e mortalidade em hospitais) e os países dos autores centrais (China e Estados Unidos) [Boschieroa et al. 2021].

Em outro estudo, foram utilizados dados coletados da PubMed[2], entre março e abril de 2020, sendo realizada uma classificação de acordo com o nível de evidência para uma avaliação quantitativa de qualidade metodológica e uma análise narrativa dos pontos fortes e fracos das publicações da COVID-19 [Zdravkovic et al. 2020]. Este mostrou evidências de que a qualidade das publicações sobre o tema nos três periódicos da área médica mais bem classificados (*The New England Journal of Medicine - NEJM, The Journal of the American Medical Association - JAMA, e The Lancet*) está abaixo da média de qualidade desses periódicos. Os autores relataram ainda que o número de publicações

[2]https://pubmed.ncbi.nlm.nih.gov/

sobre COVID-19 é quase igual ao número de publicações em todos os outros tópicos, destacando a importância de fomentar um debate sobre o valor científico, ética e sobrecarga de informações nas pesquisa sobre a temática.

Além disso, o uso de dados abertos conectados tem sido estudado e implementado em diversos segmentos. No domínio das instituições voltadas à pesquisa científica, alguns trabalhos fomentam a estruturação de um processo para utilizar os dados abertos conectados nas atividades de organização, formalização, compartilhamento, relacionamento e exploração de dados bibliométricos/cientométricos [Rautenberg 2017]. Em outra pesquisa, discute-se como profissionais de Ciência da Informação têm lidado com dados bibliográficos no contexto de dados conectados. Foi verificada a necessidade de uma adequação dos dados bibliográficos aos dados conectados, porém apesar das dificuldades encontradas para essa adequação, ela seria vantajosa para as bibliotecas [Jesus e Castro 2019].

Apesar de a publicação durante a pandemia parecer repleta de riscos, é tentadora pelo imenso interesse do público. Em comparação a outros tópicos, os artigos sobre COVID-19 demostraram gerar mais citações (mediana 45 *vs* 2 citações) [Cortegiani et al. 2021]. Um estudo que usou como fonte o site Retraction Watch[3] relatou que houve um aumento na velocidade das revisões em periódicos da área médica e um alarmante crescimento nas retrações de artigos relacionados à COVID-19 (15 pré-*prints* e 24 artigos de periódicos retirados ou retratados até dezembro de 2020) [Else 2020].

2.2. Retratações e Citações

Estudos recentes, feitos em novembro de 2020, apresentam análises preliminares sobre retratações de artigos associados à COVID-19, usando as bases de dados *Retraction Watch* e *PubMed*. De acordo com eles, naquela época já existiam 39 casos de artigos retratados, encontrados tanto em periódicos bem estabelecidos quanto naqueles com fator de impacto mais baixo, tendo seus autores um *h-index* moderadamente alto [Cortegiani et al. 2021].

Dentre os principais motivos de retratações estão desde duplicatas e plágio até questões metodológicas e de má interpretação de dados [Boschieroa et al. 2021]. Duplicação, questões éticas e plágio são mais frequentes em periódicos com baixo indicador de SJR (*SCImago Journal Rank*), enquanto periódicos com alto indicador de SJR têm, em sua maioria, questões metodológicas como motivo de retratação.

De acordo com os levantamentos realizados, pode-se observar que grande parte dos trabalhos relacionados concentra-se em caracterizar publicações e retratações sobre a COVID-19. Em uma investigação dos 200 artigos acadêmicos mais recentes publicados em 2020, mais da metade deles, incluindo os publicados em periódicos conceituados, usaram os artigos retratados para apoiar suas descobertas científicas e não notificaram as retratações [Piller 2021]. Além disso, não foram encontrados estudos mais aprofundados em relação às redes de citações destes artigos retratados.

3. Fundamentação Teórica

Nesta seção, são apresentadas algumas métricas utilizadas para análise de redes sociais. Além disso, o uso de dados conectados é discutido, incluindo a apresentação de algumas

[3]https://retractionwatch.com/

ontologias específicas para descrição de referências bibliográficas e citações.

3.1. Análise de Redes Sociais

Uma rede social é definida por um conjunto finito de atores e as relações definidas sobre eles [Wasserman e Faust 1994]. A presença de informações relacionais é um aspecto crítico e definidor de uma rede social. Na análise de redes sociais, uma rede representa um grupo de atores (nós) que se relacionam. Essas relações, *links* ou vínculos (arestas) se caracterizam por fluxos de informação. Desta maneira, fluxos, nós e vínculos constituem os elementos básicos de uma rede [Ribeiro e Bastos 2011].

Dentre as métricas para análise de redes sociais estão incluídas as de centralidade. O papel da centralidade dentro de uma rede é parte do que geralmente tentamos entender durante a análise. Existem diversas maneiras de se medir a centralidade, sendo que cada uma delas serve para auxiliar em um tipo de entendimento específico [Cherven 2015]. A seguir é feita uma breve introdução sobre estas métricas, em especial, as que são empregadas para as análises apresentadas na seção 5.

- Centralidade local: é calculada como o número total de *links* diretos com os demais nós da rede, consequentemente, um valor elevado de centralidade representa uma posição mais centralizada do nó. Esses nós podem ajudar a facilitar o fluxo de informações de um grupo para o outro dentro de um contexto organizacional [Hatala 2006].
- Centralidade de intermediação (Betweeneess centrality): é calculada como o número de menores caminhos entre quaisquer dois nós da rede que passam por determinado nó. Nós com alto valor de centralidade de intermediação podem oferecer o caminho mais direto a clusters desconectados, sendo denominados como pontes. No entanto, ser uma ponte não é um pré-requisito para ter um alto valor de de centralidade de intermediação, mas é comum que esses nós sejam classificados como criticamente importantes usando essa medida [Cherven 2015].

Além disso, com o propósito de prover uma melhor visualização para grafos, comumente utilizados na representação de redes sociais, alguns algoritmos foram desenvolvidos. Neste artigo optou-se por usar o algoritmo de distribuição *Fruchterman Reingold*, que faz um bom trabalho em distribuir os vértices de maneira uniforme, tornando os comprimentos das bordas homogêneos e refletindo a simetria [Fruchterman e Reingold 1991].

3.2. Dados Conectados e Ontologias

Dados Conectados (*Linked Data*)[4] são uma coleção de conjuntos de dados interrelacionados na Web, que vêm sendo utilizados no contexto da Web Semântica no que se cunhou chamar de Web de Dados. Baseados em representações em RDF[5], e acesso via linguagem SPARQL[6], provêm apoio à interoperabilidade e inferências para geração de novas informações, especialmente quando anotados usando vocabulários e ontologias compartilhadas.

[4]https://www.w3.org/standards/semanticweb/data
[5]https://www.w3.org/TR/rdf11-concepts/
[6]https://www.w3.org/TR/sparql11-query/

Em Computação, nos refere-se a uma ontologia como um tipo especial de objeto de informação ou artefato computacional. Uma ontologia fornece os meios para descrever explicitamente a conceituação por trás de uma base de conhecimento [Swartout et al. 1997], ou seja, as entidades e relações relevantes que emergem de sua observação, que são úteis para os nossos propósitos [Staab e Studer 2009]. As ontologias são hoje uma das formas mais comuns de representação do conhecimento, constituindo um dos pilares da web semântica. O reuso de ontologias ou de pelo menos alguma parte delas, é considerado uma boa prática, podendo acelerar significativamente o tempo de desenvolvimento, além de facilitar interligações futuras dos dados nelas anotados [d'Aquin et al. 2012].

A seguir, as ontologias FaBio e CiTo são brevemente comentadas. Elas foram selecionadas por terem sido empregadas para a triplificação do segundo conjunto de dados construído neste trabalho, conforme apresentado na seção 5. Essas ontologias fazem parte da SPAR[7], a *Semantic Publishing and Referencing Ontologies*, um conjunto de ontologias que permitem a criação de metadados RDF legíveis por máquina abrangentes para todos os aspectos de publicação e referência semântica.

FaBiO é uma ontologia para registro e publicação de descrições de entidades na web semântica que são publicadas ou potencialmente publicáveis e que contêm ou são referências bibliográficas. A FaBiO já importa várias entidades de padrões existentes para descrições de entidades bibliográficas, ou seja, FRBR, DC Terms, PRISM e SKOS [Peroni e Shotton 2012] . Dessa maneira, essa ontologia possibilita representar uma ampla variedade de objetos bibliográficos, incluindo artigos de conferências (*fabio: ConferencePaper*), artigos de periódicos (*fabio: JournalArticle*), edições e volumes de periódicos (*fabio: JournalIssue e fabio: JournalVolume*), usando termos que são comuns à comunidade acadêmica.

Segue um exemplo de referência bibliográfica, descrita em texto simples, extraída de um dos artigos retratados usados no conjunto de dados da seção 4.2.

> *Ahmed Elgazzar, Basma Hany, Shaimaa Abo Youssef et al. Efficacy and Safety of Ivermectin for Treatment and prophylaxis of COVID-19 Pandemic, 13 November 2020, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-100956/v1].*

Dessa referência é possível extrair informações como: ser um artigo acadêmico, nome dos autores, data de publicação, título do artigo, local onde foi publicado, número das páginas onde o artigo foi publicado, DOI, dentre outras informações, que estão disponíveis, porém de uma forma não totalmente estruturada.

Utilizando FaBiO, que também inclui parte dos vocabulários FRBR, DC Terms e PRISM, é possível criar uma descrição das informações desta referência. Uma parte de seu uso pode ser visto na Figura 1 que contem um trecho dos dados do mesmo artigo citado anteriormente.

Complementarmente, a ontologia CiTO possibilita registrar cada citação e a intenção da citação (por exemplo, *cito:extends*, *cito:usesMethodIn,*

[7] http://www.sparontologies.net/

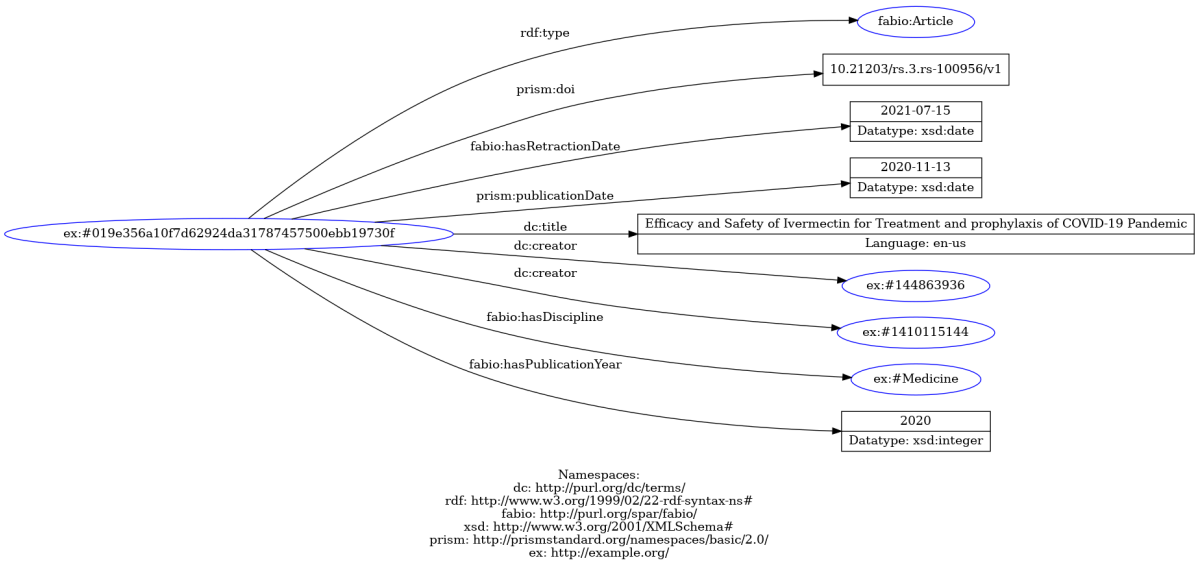

Figura 1. Exemplo de uso da ontologia FaBio

cito:obtainsBackgroundFrom). Em particular, permite criar metadados que descrevem citações que são distintos dos metadados que descrevem as próprias obras citadas, e permite que os motivos de um autor ao se referir a outro documento sejam registrados. A figura 2 contém um exemplo de uso da ontologia, representando as citações de um dos artigos retratados do conjunto de dados usado na seção 4.2.

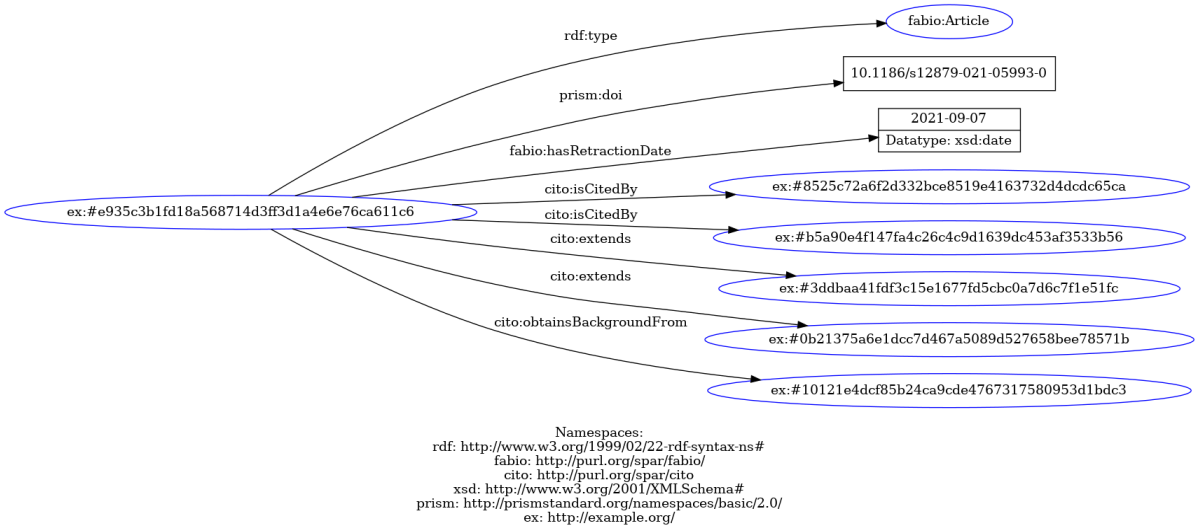

Figura 2. Exemplo de uso da ontologia CiTo

4. Construção dos Conjuntos de Dados

Nesta seção, são descritos os conjuntos de dados gerados para realização das análises. O primeiro conjunto de dados utiliza o formato *GRAPHML*[10] e é explorado através de técnicas de análise de redes sociais. O segundo conjunto de dados, em representação

RDF (*Resource Description Framework*), é adequado para o uso das ferramentas de Web Semântica.

4.1. Conjunto de Dados GRAPHML

Em relação ao primeiro conjunto, foi necessária a construção de um *dataset* que tivesse todas as informações requeridas. Os dados foram coletados de três fontes distintas: *Retraction Watch Database*[3], para obter a listagem dos artigos retratados; *Semantic Scholar*[8], para obter as citações e informações deles; e a *DOI* API[9], para obter a data de publicação dos mesmos. A segunda fonte mencionada ainda foi utilizada para verificar o valor do atributo *isInfluential*, que especifica se o trabalho que fez referência ao retratado foi fortemente influenciado pelo mesmo, dependendo do número de vezes que a referência aparece, assim como a sua localização e contexto.

Portanto, foi utilizado o *Retraction Watch Database* filtrando o campo de título por: *"COVID-19" or "coronavirus disease 2019" or "coronavirus 2019" or "SARS-COV-2" or "2019-nCov"*. Através desta consulta, realizada em 19 de março de 2021, foram obtidos 92 artigos. Destes, 9 não tinham DOI disponível, restando 83 resultados, dos quais 6 estavam aparecendo repetidamente na listagem, portanto, resultando em 77 artigos retratados para análise. Em seguida, a *Semantic Scholar* API foi utilizada para obter mais informações sobre cada artigo retratado e os trabalhos que os citavam. Foi verificado que 4 dos artigos retratados não estavam disponíveis na API, restando 73 artigos nessa situação. Por fim, o *script*, através de requisições para a *DOI* API, obteve a data de publicação dos artigos, um dado importante para a análise e que não estava presente nas requisições iniciais.

O conjunto de dados foi criado em 18 de abril de 2021 no formato *GraphML*[10], contendo uma rede com 5.884 nós (artigos retratados e os que os referenciavam) e 6.623 arestas (citações). Dentre elas, as que representavam citações altamente influentes receberam peso 2 e as demais, peso 1. Além disso, sobre os nós e arestas, foram armazenadas as informações apresentadas na Tabela 1. Por fim, o conjunto de dados e o *script* em *Python* utilizado para o processo de geração do arquivo *GRAPHML* foi disponibilizado para consulta e utilização em futuras pesquisas no *GitHub*[11].

Tabela 1. Informações sobre nós e arestas da rede

Atributo	Para	Descrição
id	nó	DOI do artigo, ou identificador único
retractionDate	nó	Data de retratação, no formato `yyyy-mm-dd`
publicationDate	nó	Data de publicação do artigo, no formato `yyyy-mm-dd`
retracted	nó	*Flag* que identifica se um artigo é retratado
name	nó	Título do artigo
weight	aresta	Peso da aresta (se for artigo influente, aresta tem peso maior)
afterRetraction	aresta	*Flag* para identificar que foi uma citação feita após a retratação
influential	aresta	*Flag* para identificar uma citação influente

[8]https://www.semanticscholar.org/

[9]https://www.doi.org/factsheets/DOIProxy.html#rest-api

[10]http://graphml.graphdrawing.org/

[11]https://github.com/ingridpacheco/RetractedArticlesDatasetGenerator

4.2. Conjunto de Dados RDF

Para análises mais detalhadas sobre os artigos um outro conjunto de dados, desta vez em formato *Resource Description Framework* (RDF), foi gerado realizando os mesmos processos mencionados anteriormente (ver seção 4.1), fazendo requisições para o *Retraction Watch Database*, o *Semantic Scholar* API e a *DOI* API. No total foram obtidos 151 artigos retratados em 17 de junho de 2021, dos quais, depois da remoção daqueles que não tinham DOI disponível e os que não foram encontrados na base do *Semantic Scholar*, restaram 131. Após conseguir as citações, a base de conhecimento totalizou 9.070 artigos.

Após obter todos esses dados, foi necessário realizar o processo de triplificação, ou seja, transformar os dados em triplas (unidade básica da representação em RDF, composta por *sujeito*, *predicado* e *objeto*) para gerar um arquivo no formato RDF e possibilitar a análise dos dados com o uso de ferramentas para dados conectados. Foi necessário o uso de ontologias para expressar os conceitos envolvidos no domínio de forma efetiva (ver seção 3.2). Para tanto, os vocabulários descritos pelas ontologias FaBiO e CiTo foram utilizados para descrever cada artigo retratado (ou artigo que cite um retratado).

- **fabio:Article**: Mapeia a propriedade *paperId* vinda do *Semantic Scholar* para gerar a URI do artigo. Ex: *ex:019e356a10f7d62924da31787457500ebb19730f a fabio:Article;*
- **prism:doi**: Mapeia o DOI do artigo vindo do *Retraction Watch*. Ex: *prism:doi "10.21203/rs.3.rs-100956/v1";*
- **fabio:hasRetractionDate**: Mapeia a data de retratação (*Retraction Date*) do artigo vindo do *Retraction Watch*. Ex: *fabio:hasRetractionDate "2021-07-15"*^^xsd:date;
- **prism:publicationDate**: Mapeia a data de publicação *(timestamp)* do artigo vindo do *DOI* API. Ex: *prism:publicationDate "2021-05-22"*^^xsd:date;
- **fabio:hasPublicationYear**: Mapeia o ano de publicação (*year*) do artigo vindo do *Semantic Scholar*. Ex: *fabio:hasPublicationYear 2021;*
- **dc:abstract**: Mapeia a descrição (*abstract*) do artigo vinda do *Semantic Scholar*. Ex: *dc:abstract "None"@en-us;*
- **dc:title**: Mapeia o título (*title*) do artigo vindo do *Semantic Scholar*. Ex: *dc:title "Efficacy and Safety of Ivermectin for Treatment and prophylaxis of COVID-19 Pandemic"@en-us;*
- **ex:hasRetractionNatureOfChoice**: Mapeia o tipo de retratação (*Nature of Notice*) vindo do *Retraction Watch*. Ex: *ex:hasRetractionNatureOfChoice "Retraction"@en-us;*
- **ex:hasRetractionMotive**: Mapeia os motivos da retratação (*reasons*) vindos do *Retraction Watch*. Ex: *ex:hasRetractionMotive "+Notice - Limited or No Information"@en-us;*
- **dc:publisher**: Mapeia o veículo de publicação do artigo (*venue*) vindo do *Semantic Scholar*. Ex: *dc:publisher "Phytochemistry reviews : proceedings of the Phytochemical Society of Europe"@en-us;*
- **prism:keywords**: Mapeia as palavras-chaves de um artigo (*topic*) vindas do *Semantic Scholar*. Ex: *prism:keywords "Epidemiology"@en-us;*

- **dc:creator**: Mapeia os autores do artigo (*authorId*) vindos do *Semantic Scholar*. Ex: *dc:creator ex:144863936;*
- **dfcore:hasCountry**: Mapeia os países dos autores do artigo (*Countries*) vindos do *Retraction Watch*. Ex: *dfcore:hasCountry "Malta"@en-us;*
- **fabio:hasDiscipline**: Mapeia os campos de estudo do artigo (*fieldsOfStudy*) vindos do *Semantic Scholar*. Ex: *fabio:hasDiscipline ex:Medicine;*
- **cito:isCitedBy**: Mapeia os artigos que citam o artigo principal (sujeito da tripla). Ex: *ex:019e356a10f7d62924da31787457500ebb19730f cito:isCitedBy ex:5cb480110f0388349743232d9e05fcf0309c3c15;*
- **cito:cites**: Mapeia os artigos que o artigo principal (sujeito da tripla) cita. Ex: *ex:5cb480110f0388349743232d9e05fcf0309c3c15 cito:cites ex:019e356a10f7d62924da31787457500ebb19730f;*

Para mapear a intenção de citação usando a ontologia CiTo, foram levados em consideração os atributos *isInfluential* e *intent* do *Semantic Scholar*. Vale ressaltar que, dentre as informações disponibilizadas sobre os artigos através do *Semantic Scholar* API, o atributo *intent* classifica cada citação de acordo com a intenção de quem a citou, podendo possuir três valores: “Cites Results”, “Cites Methods” e “Cites Background”. O mapeamento foi feito da seguinte maneira:

- Quando a citação tinha o atributo *isInfluential* igual a ***true***, foi utilizada a propriedade *cito:extends*.
- Quando o atributo *intent* era igual a ***Cites Background***, a propriedade usada foi *cito:obtainsBackgroundFrom*.
- Quando o atributo *intent* era igual a ***Cites Results***, a propriedade usada foi *cito:usesDataFrom*.
- Por fim, quando *intent* era igual a ***Cites Methods*** a propriedade usada foi *cito:usesMethodIn*.

Os artigos que faziam citação aos retratados também foram descritos através das ontologias. No entanto, devido à limitação de requisições impostas pela API do *Semantic Scholar*, foram usadas apenas as propriedades: *prism:doi, prism:publicationDate, dc:creator, dc:publisher, dc:title, cito:cites, cito:obtainsBackgroundFrom, cito:usesDataFrom, cito:extends, cito:usesMethodIn, fabio:hasPublicationYear* e *fabio:Article*. Os autores e as áreas de estudo dos artigos também foram representados usando as ontologias mencionadas.

Para os autores, as seguintes propriedades foram utilizadas:

- **foaf:Person**: Mapeia que o autor (*authorId*) é do tipo pessoa. Ex: *ex:6080631 a foaf:Person;*
- **foaf:name**: Mapeia o nome do autor. Ex: *foaf:name "A. Castrioto".*

Já em relação as disciplinas, os seguintes mapeamentos foram efetuados:

- **fabio:SubjectDiscipline**: Mapeia o tipo do campo de estudo. Ex: *ex:Medicine a fabio:SubjectDiscipline;*
- **rdfs:label**: Mapeia uma *label* para o nome da disciplina. Ex: *rdfs:label "Medicine"@en-us.*

Todo o processo de geração do conjunto de dados (e triplicação) foi feito através de *script* em *Python*. Ele está, assim como o *dataset* em formato RDF, disponível para consulta e utilização em futuras pesquisas no *Github*[12].

5. Análises e Resultados

Nesta seção, são detalhadas as análises e resultados encontrados ao longo do trabalho, tanto utilizando técnicas de análises de redes sociais [Duca et al. 2021] quanto fazendo uso de ferramentas de web semântica sobre o conjunto de dados em RDF.

5.1. Conjunto de Dados GRAPHML

Baseado no primeiro conjunto de dados gerado, foi utilizada a ferramenta *Gephi*[13] para realizar as análises referentes a algumas das perguntas de pesquisa (especificamente **P1** a **P5**) previamente estabelecidas (ver seção 1).

5.1.1. Caracterização da Rede de Citações

O estudo da estrutura da rede é uma etapa importante para o entendimento do tipo de análise que pode ser realizada, e neste caso, possibilita uma visão geral da problemática da citação de artigos retratados. A rede possui algumas sub-redes ego centradas (*ego-centered network*), onde o nó central (*ego*) de cada sub-rede é um artigo retratado e o conjunto de nós que o cercam (*alter*) são artigos que fazem referência ao ego. Esse tipo de rede pode ser usada em diversas aplicações, como a análise de transmissão de doenças, estudos de suporte social e redes de discussão [Wasserman e Faust 1994].

Outro aspecto interessante é que, dos 5.884 nós da rede, 73 representam artigos que sofreram retratação e os outros 5.811 nós restantes são os artigos que os citam. Em relação às arestas, a rede é formada por 6.623, das quais 423 têm o atributo *influential* igual a 1, o que significa que pelo menos 6,39% dos artigos desta rede foram fortemente influenciados pelos artigos retratados. Além disso, o número médio de citações (arestas) por artigo retratado (nó) é de 90,72.

Por fim, as características principais da rede são: grau médio de 1,126 e coeficiente de clusterização igual a 0,18 (estes valores baixos são esperados, uma vez que esta rede inclui apenas as citações a artigos retratados). Os números de componentes fracamente e fortemente conectados são, respectivamente, 39 e 5.884. Na figura 3 pode-se ver uma representação do grafo da rede completa usando o algoritmo de visualização *Fruchterman-Reingold*. Cada nó da rede representa um artigo, estando os retratados em vermelho e o tamanho de cada nó sendo diretamente proporcional ao seu grau (centralidade local). Por fim, as arestas representam as referências aos artigos retratados, estando em verde as arestas que ligam os artigos que foram fortemente influenciados ao artigo retratado (ver seção 4).

[12]https://github.com/ingridpacheco/RetractedArticlesDatasetGenerator/tree/master/RDF
[13]https://gephi.org/

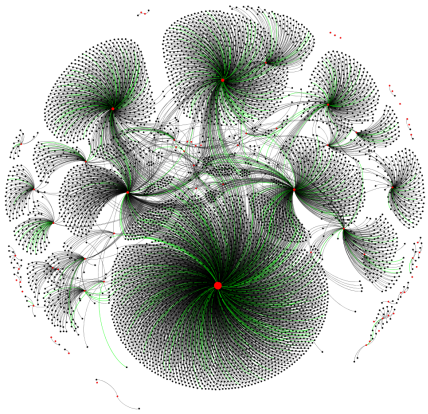

Figura 3. Rede de Citações de Artigos Retratados sobre COVID-19

5.1.2. Artigos Retratados Mais Referenciados

Apesar de todos os artigos retratados terem um grande impacto sobre a validade dos que os referenciam, compreender os mais citados serve para dar uma visão mais concreta e estreita sobre os principais temas que causam retratação dentro da temática da COVID-19, além de elucidar o seu impacto e a forma como a rede está disposta.

Levando em consideração este contexto e para responder a **P1**, uma métrica que se mostrou interessante para esta análise foi a de centralidade local, com a finalidade de encontrar os artigos retratados mais referenciados. Ao aplicar tal métrica sobre a rede, o grafo da figura 4 (a) foi gerado. O nó em destaque, na cor vermelha, representa o artigo com maior valor de centralidade local, correspondente ao artigo retratado mais citado (referenciado por 2.745 artigos), de título *Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial* e DOI 10.1016/j.ijantimicag.2020.105949. Levando em consideração o fato de que o segundo artigo retratado com maior centralidade é citado por 799 artigos, esse valor passa a ser ainda mais significativo, expressando o quanto o artigo é influente dentro da rede.

Estendendo as análises, os 5 artigos retratados mais referenciados foram filtrados, ainda utilizando a métrica de centralidade local, a fim de tentar encontrar similaridades entre eles. A figura 4 (b) representa os 5 principais artigos retratados na cor amarela e suas citações influentes (que possuem o atributo *influential* na aresta de ligação igual a 1) em diferentes cores, dependendo do artigo que referenciam.

A Tabela 2 apresenta informações sobre os top 5 artigos retratados mais citados. Salienta-se que a retratação do primeiro artigo ocorreu por problemas em relação aos seus resultados, o que possivelmente prejudicou os 150 artigos que foram fortemente influenciados por ele, consequentemente, tendo suas validações questionadas. O segundo foi publicado no *National Science Review* e teve como principal motivo de retratação erro no texto, o que também pode acabar afetando a qualidade dos artigos que o referenciam. O terceiro foi publicado no *The New England journal of medicine* e os motivos de sua retratação se expandem, não só tendo incertezas nos resultados como também nos dados

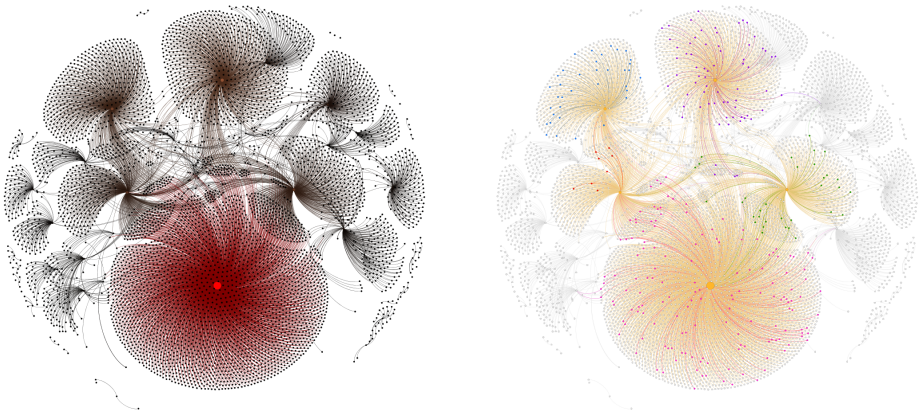

Figura 4. Rede: (a) com a métrica de centralidade; (b) de retratados com seus nós influentes

utilizados, mostrando os problemas que referenciá-lo pode causar. O quarto, publicado no *Engineering*, não teve o seu motivo de retratação muito claro, apenas tendo a indicação de que o artigo foi retirado, sem maiores informações sobre o que o levou a essa situação. Por fim, o quinto artigo, publicado no *The Lancet*, é citado por outro retratado, 10.1016/S0140-6736(20)31174-0, e ainda é um dos 7 retratados que cita outros na mesma condição. Inclusive, dentre eles, este é o único que cita mais de um retratado, neste caso, sendo os artigos 10.1016/j.ijantimicag.2020.105949 e 10.1056/NEJMoa2007621. Devido a este fato, o artigo cria uma ponte entre dois trabalhos retratados muito relevantes na rede, e fica com o maior valor de intermediação (711.833), como visto na figura 5 (a). Ademais, este artigo e suas principais conexões são destacados na figura 5 (b).

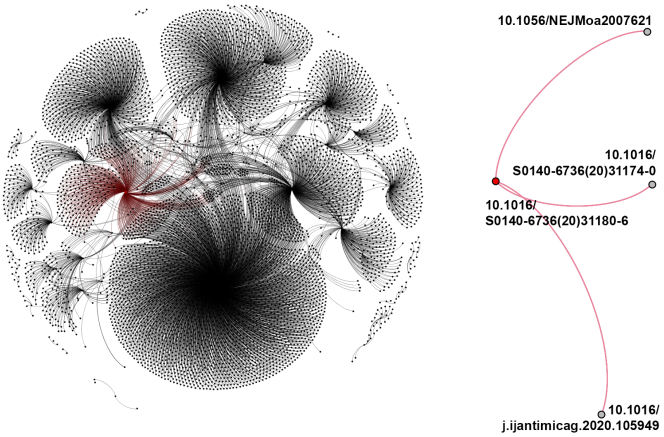

Figura 5. Rede: (a) destacando métrica de intermediação; (b) ego centrada no nó com maior valor de intermediação

Passando para a análise da rede como um todo, um dos primeiros pontos que se pode notar é um tema em comum que aparece nos títulos de 75% deles: *tratamento para a COVID-19*. A fim de responder a **P2**, análises sobre a incidência de termos nos títulos

Tabela 2. Informações sobre citações dos top 5 artigos retratados

Título/DOI	Citações	Citações de Influentes	Data Retratação	Antes	Após
Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial *10.1016/j.ijantimicag.2020.105949*	2.745	150	11/04/2020	260	2485
On the origin and continuing evolution of SARS-CoV-2 *10.1093/nsr/nwaa036*	799	68	19/06/2020	49	750
Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19 *10.1056/NEJMoa2007621*	627	36	04/06/2020	117	510
Experimental Treatment with Favipiravir for COVID-19: An Open-Label Control Study *10.1016/j.eng.2020.03.007*	508	33	01/04/2020	30	478
RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis *10.1016/S0140-6736(20)31180-6*	471	8	04/06/2020	36	435

dos artigos, usando a ferramenta *WordArt.com*[14], foram feitas considerando: (a) todos os artigos da rede e (b) apenas do subgrupo dos 73 artigos retratados. Foram executadas as seguintes etapas de pré-processamento: remoção de *stopwords*, remoção de números e *stemming* [Baeza-Yates e Ribeiro-Neto 2013]. Além destas, também foram removidos os termos utilizados para a coleta dos dados no *Retraction Watch Database* (ver seção 4). Os resultados dessas análises estão sumarizados na Tabela 3. Dentre as palavras-chave mais frequentes nos dois grupos, pode-se evidenciar *Hydroxychloroquine*, *Treatment* e *Drug*. Esses termos se destacam devido ao contexto atual da pandemia da COVID-19, momento em que são divulgados estudos relacionados ao tratamento da mesma.

Além disso, outra análise sobre a rede de citações de artigos retratados foi conduzida com o intuito de considerar o aspecto temporal das citações, verificando a quantidade de referências feitas antes e após os avisos de retratação. A Tabela 2 também apresenta os quantitativos de citações antes e após as retratações para os cinco artigos mais citados na rede. Já a figura 6 ilustra a rede, destacando em amarelo as citações efetuadas (a) antes e (b) após as retratações. Como demonstrado, para os cinco artigos mais citados, o número de citações após a retratação foi expressivamente maior do que as citações anteriores ao aviso de retratação. Desta forma, 82,11% de todas as citações da rede aconteceram após,

[14]https://wordart.com/

Tabela 3. Comparativo entre palavras-chave de todos os artigos da rede versus as palavras-chave dos artigos retratados

Todos os Artigos		x	Artigos Retratados	
Termo	**Peso**		**Termo**	**Peso**
Patient	893	**1°**	Patient	20
Review	717	**2°**	Pandemic	9
Pandemic	663	**3°**	Infect	8
Treatment	629	**4°**	Analysis	7
Hydroxychloroquine	601	**5°**	Clinic	6
Infect	576	**6°**	Effect	6
Clinic	524	**7°**	Treatment	6
Drug	438	**8°**	Hospital	5
Analysis	416	**9°**	Hydroxychloroquine	5
Study	380	**10°**	Mortality	5

e apenas 17,89% ocorreram antes, respondendo assim a **P3**.

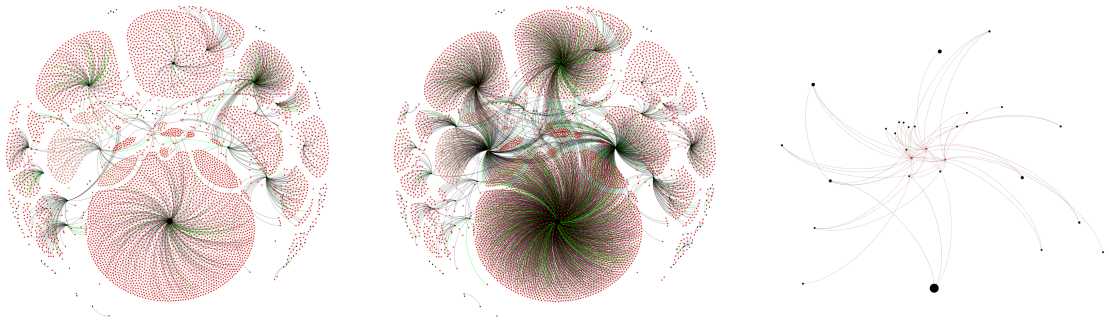

Figura 6. Rede destacando: citações feitas (a) antes e (b) após retratações; (c) artigos que mais referenciam retratados

5.1.3. Artigos Que Mais Referenciam Retratados

Em resposta a **P4**, há 7 retratados que referenciam artigos na mesma situação. Além disso, foi feita uma análise para encontrar quais eram os artigos que mais citavam retratados. Ao analisar a rede procurando por estes nós, filtrando os que referenciavam mais de um retratado, foi possível encontrar 648 artigos (já considerando o quinto artigo retratado mais citado mencionado na seção anterior), desta maneira respondendo a **P5**. Porém, os 3 nós que mais tiveram conexões, destacados em vermelho, mostrados na figura 6 (c), possuem um diferencial em relação aos outros devido à quantidade de retratados que eles citam. Enquanto o quarto lugar referencia apenas 5 artigos, os 3 principais (10.1136/medethics-2020-106494, 10.1007/s11192-020-03661-9 e 10.1016/j.pulmoe.2020.10.011) referenciam uma quantidade significativa de artigos retratados, respectivamente, 19, 17 e 16.

O grande número chama a atenção para o tema em comum que eles tratam, pois todos abordam o assunto de artigos retratados sobre COVID-19 como uma forma de crítica para a velocidade com que as publicações são realizadas, ao invés de priorizar a qualidade

e a busca por informações corretas, portanto, fazendo as referências propositalmente para trazer exemplos que elucidem o problema.

Dentre eles, entretanto, vale mencionar que o artigo de DOI 10.1007/s11192-020-03661-9, que possui o título de *Retracted COVID-19 articles: a side-effect of the hot race to publication* [Soltani e Patini 2020], é o único que cita todos os 5 retratados mencionados antes, evidenciando o quão relevante ele é para o tema e para a rede utilizada.

5.2. Conjunto de Dados RDF

Após todas as análises estabelecidas com o primeiro conjunto de dados terem sido realizadas (ver seção 5.1), o segundo *dataset* é explorado utilizando a ferramenta *GraphDB*[15]. O arquivo RDF gerado é importado e a partir dele são realizadas as análises referentes às perguntas de pesquisa previamente estabelecidas (ver seção 1), especificamente de **P6** a **P10**. Através de consultas SPARQL[16] e a exploração dos grafos gerados, foi possível fazer descobertas acerca dessa rede de citação a artigos retratados sobre COVID-19.

5.2.1. Características da Base de Conhecimento

A base de conhecimento gerada utiliza as ontologias FaBio e CiTo para mapear as publicações científicas e a intenção das referências e citações entre os artigos. A base possui 9.070 artigos, dos quais 131 são artigos retratados, de acordo com o *Retraction Watch*[2], entre os retratados 7 artigos são contados repetidamente, por terem sofrido retratação mais de uma vez. Os motivos de retratação (*fabio:retraction*) são "Correction" com 3 artigos, "Expression of concern" com 11 e "Retraction" com 117 artigos. A consulta *SPARQL* usada obter essas informações está no Apêndice 1.1.

Nesse aspecto, vale ressaltar o artigo "***Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis***"que teve três retratações, sendo a primeira em 06/03/2020, a segunda em 06/04/2020 e a terceira em 30/05/2020, com os motivos de *Correction*, *Expression of Concern* e *Retraction* respectivamente. Outra característica interessante da base de conhecimento é o número de autores, um total de 53.196 dentre os quais, 639 são autores de artigos retratados.

5.2.2. Retratações por Área de Conhecimento

Apesar da temática central dos artigos dessa base de conhecimento ser medicina, é interessante observar que houve outras áreas de conhecimento com artigos retratados sobre COVID-19. Na figura 7, podemos observar que apesar de medicina aparecer em primeiro lugar com 76,2% dos artigos retratados, outras áreas também merecem destaque, tais como: Biologia com 7% dos artigos, além de Geografia e Ciências Políticas com 4,9% dos artigos cada, respondendo a **P6**.

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

[16]https://www.w3.org/TR/sparql11-query/

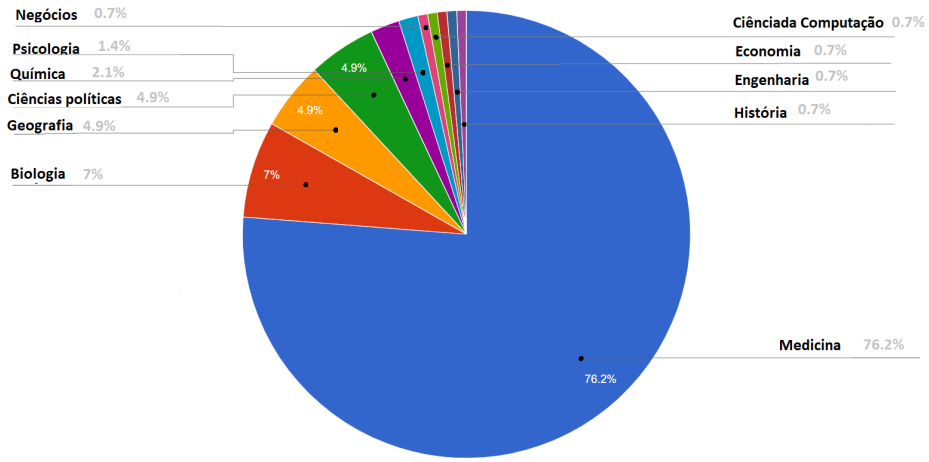

Figura 7. Distribuição dos artigos retratados por área de conhecimento

A Consulta *SPARQL* usada para obtenção dos dados está disponível no Apêndice 1.2.

5.2.3. Citações por País

Ao analisar a nacionalidade dos autores dos artigos, foram verificados 40 países diferentes. O país com mais artigos retratados foi os EUA com 34 artigos, em segundo lugar a China empatada com Malta tendo 27 artigos, seguidos pela Índia com 9 artigos retratados, e finalmente a Espanha e o Reino Unido empatados em quarto lugar com 5 artigos cada, respondendo a **P7**. A Figura 8 mostra a distribuição de todos os artigos retratados, sendo os países com mais artigos em tons de vermelho, e aqueles com menos artigos em tons de verde . A consulta usada para obter esses resultados está disponível no Apêndice 1.3.

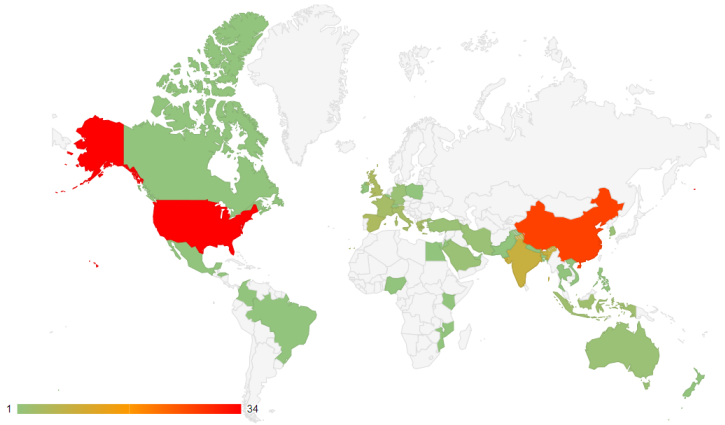

Figura 8. Distribuição dos artigos retratados por país

Estendendo as análises, foi possível verificar os países dos artigos retratados mais referenciados. A Tabela 4 exibe os 5 mais referenciados nessa situação. Cabe ressaltar

que, apesar do grande número de citações, os países França, Vietnã e Suíça não estavam entre os países com mais artigos retratados, sendo o número de artigos retratados deles 4, 1 e 1 respectivamente. Isso demonstra que, apesar de poucos artigos retratados, os autores desses países publicaram artigos que foram muito influentes na temática da COVID-19. A consulta usada para obter esses resultados está disponível no Apêndice 1.4.

Tabela 4. Países com artigos retratados mais citados

País	Total de Citações
França	3.200
Vietnã	3.176
China	2.548
Estados Unidos	1.752
Suíça	532

5.2.4. Autores e Veículos de Publicação

Ao realizar uma consulta pelos autores de artigos retratados outra característica interessante pode ser observada, alguns autores possuem mais de um artigo retratado, um autor chamado "V. Grech" destaca-se, sendo autor de 23 artigos retratados e de outros 9 artigos que fazem parte da rede de citações a artigos retratados. Esse é um número muito expressivo, principalmente quando comparado ao segundo autor com mais artigos retratados, "S. Cuschieri" que possui 6 artigos nessa situação, seguido por "Amit N. Patel", "C. Gauci" e "Mariella Scerri", ambos empatados em terceiro lugar com 5 artigos retratados cada, "Steve Agius" está no quarto lugar com 4 artigos e "G. Victor" em quinto com 3 artigos, respondendo a **P8**.

A Consulta SPARQL utilizada para obter a informação dos autores com mais artigos retratados está no Apêndice 1.5.

Além disso, o número elevado de publicações retratadas chama atenção para a qualidade dos artigos publicados por esses autores. Por esse motivo, foi feita uma consulta nos veículos de publicação com mais artigos retratados. Em primeiro lugar com 26 artigos retratados está o *EARLY HUMAN DEVELOPMENT*, já o segundo colocado foi o *MEDRXIV* com 7 artigos retratados e em terceiro lugar o *CUREUS* com 3 artigos retratados. Empatados em quarto lugar estão o *ASIAN JOURNAL OF PSYCHIATRY*, o *JACC: CASE REPORTS*, o *JOURNAL OF INFECTION* e o *THE LANCET* com 2 artigos retratados cada. Todos os outros se encontram com 1 artigo retratado, respondendo a **P9**. Cabe ressaltar que 10 artigos retratados não tinham identificação do veículo depublicação na base do *Semantic Scholar*. A consulta *SPARQL* feita para obter a informação dos veículos de publicação, está descrita no Apêndice 1.6.

5.2.5. Intenção de Citação a Artigos Retratados

Utilizando as ontologias Fabio e CiTo (ver seção 3.2), foi possível categorizar a intenção da citação ao referenciar um artigo retratado. Através da tabela 5, que contém os 10

artigos retratados com mais citações, é visto que mais de 25% dessas citações usam os artigos como *background* para trabalhos relacionados, respondendo a **P10**.

As colunas *Extends* (*cito:extends*) e *Background* (*cito:obtainsBackgroundFrom*) representam a distribuição das citações de acordo com a intenção obtida através do *Semantic Scholar*, e a coluna *Total de citações* (*cito:isCitedBy*) representa o total de citações do artigo retratado incluindo aquelas que não foram caracterizadas. Cabe ressaltar que não houve citações com intenção *cito:usesMethodIn* e *cito:usesDataFrom* para os artigos apresentados na Tabela 5. As consultas *SPARQL* usada para encontrar a quantidade de citações do tipo *Background* e *Extends* está disponível no Apêndice 1.7.

Tabela 5. Intenção de citação dos 10 artigos retratados mais citados.

Artigo	Extends	Background	Total de citações
Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial	197	1166	3176
On the origin and continuing evolution of SARS-CoV-2	74	407	947
Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19	48	218	742
Experimental Treatment with Favipiravir for COVID-19: An Open-Label Control Study	39	184	633
RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis	20	103	532
Withdrawn: Clinical manifestations and outcome of SARS-CoV-2 infection during pregnancy	24	92	305
RETRACTED ARTICLE: Deep learning system to screen coronavirus disease 2019 pneumonia	12	57	226
SARS-CoV-2 infects T lymphocytes through its spike protein-mediated membrane fusion	11	73	180
Obesity and mortality of COVID-19. Meta-analysis	3	46	140
Effectiveness of Surgical and Cotton Masks in Blocking SARS–CoV-2: A Controlled Comparison in 4 Patients	10	49	133

6. Conclusões

Com o passar do tempo, diversas descobertas foram realizadas e publicá-las se tornou necessário para um avanço natural da ciência. Entretanto, devido à pressa pelo compartilhamento de resultados no menor tempo, problemas com relação à qualidade de artigos e suas referências passaram a ocorrer, levando-os à retratação. Portanto, a adoção de ferramentas para analisar artigos, autores e veículos de publicação, a fim de auxiliar pesquisadores na tarefa de citar um artigo para afirmar suas hipóteses, mostra-se muito

importante, considerando o que as evidências iniciais trazidas por esse trabalho parecem indicar.

Mediante o exposto, é possível afirmar que neste trabalho, através da ARS, foram descobertos os artigos retratados mais referenciados no contexto da COVID-19 e estimadas suas influências em outros artigos pelas métricas de centralidade. Ademais, foi possível identificar pesquisas que sofreram retratações e faziam referência a artigos na mesma situação, além de verificar que, nesta rede, a maior parte das citações foram feitas após o aviso de retratação. Também foi detectado que as palavras mais utilizadas nos títulos destes trabalhos referiam-se a possíveis tratamentos e medicamentos relacionados a doença, deixando em aberto a possibilidade de que essas publicações influenciem de alguma forma a disseminação de notícias falsas. Já considerando os dados conectados, a adoção das ontologias CiTo e FaBio possibilitaram a realização de consultas semânticas a fim de descobrir os países com mais artigos retratados e os que são mais citados, junto com suas intenções, além dos autores e veículos de publicação que mais sofreram retratação. Ademais, este artigo pode ser útil para, para fomentar a discussão sobre a questão das citações e retratações a artigos, levantando a necessidade de mecanismos mais eficientes a fim de evitar as citações pós retratação.

No entanto, durante o desenvolvimento deste trabalho, alguns desafios foram encontrados. O principal foi a falta de um conjunto de dados único com informações sobre as retratações e citações ou referências feitas por esses artigos. Para mitigar este problema, foi necessário o desenvolvimento de *scripts* em *Python* que, coletando dados de três bases distintas, criassem os *datasets* utilizados nas análises. Além disso, este trabalho teve algumas limitações, como a intenção dos autores ao referenciarem os artigos retratados não ter sido levada em consideração no primeiro conjunto de dados (ver seção 4.1). Da mesma forma o segundo conjunto de dados (ver seção 4.2), apesar de usar a intenção de citação fornecida pela API do *Semantic Scholar*, não consegue fazer uma caracterização mais aprofundada, como, por exemplo, verificar se a citação foi realizada de forma positiva ou negativa. Por fim, com os dados adquiridos também não é possível saber se o artigo referencia o retratado como uma fonte ou se ele é um análise, às vezes sobre a retratação, deste, como em alguns casos encontrados.

Como trabalhos futuros, pretende-se ampliar essas análises em uma rede completa sobre a COVID-19, também com as conexões entre artigos não retratados, diferentemente da usada nesse trabalho (*ego-centered*), que foi gerada a partir dos retratados. Desta forma, seria possível verificar mais amplamente o quão influentes realmente são os artigos retratados. Além disso, pretendemos realizar a automatização de parte das análises a fim de evitar possível subjetividade nos resultados. Outra abordagem interessante seria o cruzamento da rede de retratações com uma rede de *fake news* associadas à pandemia da COVID-19, a fim de encontrar possíveis conexões entre as notícias falsas e as publicações retratadas (como suas prováveis fontes). Finalmente, seria interessante o uso de técnidas de inteligência artificial, mineração de textos e processamento de linguagem natural para tentar obter a intenção das citações de maneira mais completa, além das padrões já fornecidas através da API do *Semantic Scholar*.

Referências

Baeza-Yates, R. e Ribeiro-Neto, B. (2013). *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. 2nd edition*. Bookman.

Bar-Ilan, J. e Halevi, G. (2017). Post retraction citations in context: a case study. *Scientometrics*, 113:547–565.

Berners-Lee, T. (2006). Linked data. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 13 de nov. 2021.

Boschieroa, M. N., Carvalho, T. A., e de Lima Marson, F. A. (2021). Retraction in the era of covid-19 and its influence on evidence-based medicine: is science in jeopardy? *Pulmonology Journal*, 27:97–106.

Browman, H., Alexander, J., Fennell, C., Hodgkinson, M., e Tierney, H. (2019). Guidelines for retracting articles. Technical report, Committee on Publication Ethics.

Cherven, K. (2015). *Mastering Gephi network visualization : produce advanced network graphs in Gephi and gain valuable insights into your network datasets*. Community experience distilled. Packt Publishing.

Cortegiani, A., Catalisano, G., Ippolito, M., Giarratano, A., Absalom, A. R., e Einav, S. (2021). Retracted papers on sars-cov-2 and covid-19. *British journal of anaesthesia*, 126:e155–e156.

d'Aquin, M., Aquin, Kronberger, G., e Suárez-Figueroa, M. C. (2012). Combining data mining and ontology engineering to enrich ontologies and linked data. *CEUR Workshop Proceedings*, 868.

Duca, H., Pacheco, I., Lopes, G., Campos, M. L., e Oliveira, J. (2021). Artigos retratados sobre covid-19: Uma análise sobre sua rede de citações. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 13–24, Porto Alegre, RS, Brasil. SBC.

Else, H. (2020). How a torrent of covid science changed research publishing - in seven charts. *Nature*, 588:553.

Fruchterman, T. M. J. e Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.

Hatala, J.-P. (2006). Social network analysis in human resource development: A new methodology. *Human Resource Development Review*, 5(1):45–71.

Jesus, A. F. d. e Castro, F. F. d. (2019). Dados bibliográficos para o linked data: uma revisão sistemática de literatura. *Brazilian Journal of Information Science: research trends*, 13(1):45–55.

Peroni, S. e Shotton, D. (2012). Fabio and cito: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, 17:33–43.

Piller, C. (2021). Disgraced covid-19 studies are still routinely cited. *Science*, 371:331–332.

Rautenberg, S. e. a. (2017). Dados abertos conectados e gestão do conhecimento: estudos de caso cientométricos em uma universidade brasileira. *Perspectivas em Ciência da Informação*, 22:116–142.

Ribeiro, E. M. B. d. A. e Bastos, A. A. V. A.-l. B. (2011). Redes sociais interorganizacionais na efetivação de projetos sociais. *Psicologia & Sociedade*, 23:282 – 292.

Sheth, B. P. e Thaker, V. S. (2014). Scientific retraction: a synonym for pseudoscience? *Acta bioethica*, 20.

Shotton, D. (2010). CiTO, the citation typing ontology. *Springer Science and Business Media LLC*, 1(Suppl 1):S6.

Silva, J. A. T. D. e Bornemann-Cimenti, H. (2017). Why do some retracted papers continue to be cited? *Scientometrics*, 110:365–370.

Soltani, P. e Patini, R. (2020). Retracted covid-19 articles: a side-effect of the hot race to publication. *Scientometrics*, 125.

Staab, S. e Studer, R., editors (2009). *Handbook on Ontologies*. Springer Berlin Heidelberg.

Swartout, B., Patil, R., Knight, K., e Russ, T. (1997). Toward Distributed Use of Large-Scale Ontologies. In *Ontological Engineering, AAAI-97 Spring Symposium Series*, pages 138–148.

Wasserman, S. e Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Zdravkovic, M., Berger-Estilita, J., Zdravkovic, B., e Berger, D. (2020). Scientific quality of covid-19 and sars cov-2 publications in the highest impact medical journals during the early phase of the pandemic: A case control study. *PLoS One*, 15.

1. Apêndices

Nesta seção serão transcritas as consultas SPARQL usadas para análise dos dados.

1.1. Características da Base de Conhecimento

```
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX ex: <http://example.org/>
select distinct ?Motivo (count(?s) as ?total)
where {
    ?s a fabio:Article .
    ?s ex:hasRetractionNatureOfChoice ?Motivo
}
group by ?Motivo
```

1.2. Retratações por Área de Conhecimento

```
select   ?nome  (count(distinct(?s)) as ?total)
where {
    ?s a fabio:Article .
```

```
    ?s fabio:hasRetractionDate ?o .
    ?s fabio:hasDiscipline ?discipline .
    ?discipline rdfs:label ?nome
}
group by  ?nome
```

1.3. Retratados por País

```
select   ?local  (count(distinct(?s)) as ?total)
where {
    ?s a fabio:Article .
    ?s fabio:hasRetractionDate ?o .
    ?s dfcore:hasCountry ?local
}
group by  ?local
order by DESC (?total)
```

1.4. Citações a Artigos Retratados por País

```
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX dfcore: <http://ontology.tno.nl/dfcore#>
PREFIX cito: <http://purl.org/spar/cito>
select ?local (count(distinct(?ref)) as ?total)
where{
?s a fabio:Article.
?s fabio:hasRetractionDate ?o.
?s dfcore:hasCountry ?local.
?s cito:isCitedBy ?ref.
}
group by ?local
order by DESC (?total)
```

1.5. Autores de Artigos Retratados

```
select ?o   (count(distinct(?s)) as ?total)
where {
    ?s a fabio:Article .
    ?s fabio:hasRetractionDate ?retractionDate .
    ?s dc:creator ?o .
}
group by ?o
order by DESC(?total)
```

1.6. Veículos de Publicação dos Artigos Retratados

```
select ?v  (count(distinct(?s)) as ?total)
where {
    ?s a fabio:Article .
```

```
    ?s fabio:hasRetractionDate ?o .
    ?s dc:publisher ?veiculo
    .filter(?veiculo != ""@en-US)
    bind(ucase(?veiculo) as ?v)
    }
group by ?v
order by DESC (?total)
```

1.7. Intenção de Citação a Artigos Retratados

1.7.1. Background

```
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX cito: <http://purl.org/spar/cito>
PREFIX ex: <http://example.org/#>
select
?titulo (count(distinct(?ref)) as ?isCitedBy)
where {
    ?s a fabio:Article;
        fabio:hasRetractionDate ?o;
        dc:title ?titulo;
        cito:isCitedBy ?ref.
    ?ref cito:obtainsBackgroundFrom ?s.
}
group by  ?s ?titulo
order by DESC (?isCitedBy)
limit 10
```

1.7.2. Extends

```
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX cito: <http://purl.org/spar/cito>
PREFIX ex: <http://example.org/#>
select
?titulo (count(distinct(?ref)) as ?isCitedBy)
where {
    ?s a fabio:Article;
        fabio:hasRetractionDate ?o;
        dc:title ?titulo;
        cito:isCitedBy ?ref.
    ?ref cito:extends ?s.
}
group by  ?s ?titulo
order by DESC (?isCitedBy)
```