ARTIGO

Comunicado de imprensa como indicador de atenção social qualificada da ciência: a construção de um banco de dados e suas potencialidades

Press releases as an indicator of qualified social attention of science: the construction of a database and its analytical potential

Germana Barata [a, *]

Monique Oliveira [a]

Thais Peixoto [a]

Carlos Caetano Almeida [b]

Alysson Fernandes Mazoni [c]

Rodrigo Costas Comesana [d]

Juan Pablo Alperin [e]

RESUMO: Na última década, a altmetria contribuiu para abrir espaço para indicadores de atenção e impacto social da ciência e para repaginar a cienciometria. Entretanto, métricas de performance na produção científica ainda são centrais na área, o que leva a uma demanda por outras fontes de dados que confiram contexto e atuem na interface ciência-sociedade de forma mais qualificada. Este trabalho demonstrará o potencial de um banco de dados estruturado de *press releases* (comunicados de imprensa) para a construção de indicadores contextualizados, que chamaremos aqui de "indicadores qualificados de atenção social da ciência". Para isso, foi realizada a coleta de comunicados de imprensa de três agências de notícias de ciência: EurekAlert! (EUA), AlphaGalileo (Reino Unido) e Agência BORI (Brasil). Essas agências disponibilizam nos seus *sites press releases* sobre artigos científicos; no entanto, informações importantes como título do artigo, conteúdo, data da publicação, URL, DOI, não estão disponíveis de forma estruturada. Para essa estruturação, dados foram coletados por meio de técnicas de *web scraping*, organização e análise dessas informações. A utilização de *web scraping* com o armazenamento em banco de dados MySQL mostrou-se eficaz para coletar e gerenciar informações de páginas *web* dessas agências, o que possibilita análises

[a] Laboratório de Inclusão na Comunicação e na Ciência, Laboratório de Estudos Avançados em Jornalismo, Universidade Estadual de Campinas, Campinas, SP, Brasil.

[b] Automação e Sistemas para Agricultura 4.0 do Departamento de Recursos Naturais e Proteção Ambiental da Universidade Federal de São Carlos, UFSCar, Araras, SP, Brasil

[c] Departamento de Política Científica e Tecnológica, Instituto de Geociências, Universidade de Campinas, Campinas, Brasil.

[d] Centre for Science and Technology Studies (CWTS), University of Leiden, Leiden, Países Baixos.

[e] School of Publishing, Simon Fraser University, Burnaby, BC, Canada.

[*] Correspondência para/*Correspondence to* Germana Barata. E-mail: germana@unicamp.br .

Recebido em/*Received*: 05/05/2024; Aprovado em/*Approved*: 22/07/2024.

abrangentes e contextualizadas.Enquanto o comunicado de imprensa não é a única forma de qualificar e entender a atenção social da ciência, a construção do banco de dados e testes realizados demonstram que se trata de um material relevante nessa compreensão.

Palavras-chave: Agência de Notícias de Ciências; Press Releases; Acesso Aberto; Indicador Social da Ciência; Processamento de Linguagem Natural.

ABSTRACT: In the last decade, altmetrics has opened space for indicators of attention and social impact of science and for revamping scientometrics. However, the centrality of traditional indicators in scientific production persists. There is a demand for data sources that may provide additional context and that can form the basis for indicators that can meaningfully represent the interface between science and society. This study will demonstrate the potential of a structured database of press releases for developing more contextualized indicators, which we will refer to as "qualified social attention indicators". To this end, press releases from three science news agencies were collected: EurekAlert! (USA), AlphaGalileo (United Kingdom) and Agência BORI (Brazil). These agencies provide press releases about scientific articles on their websites; however, important information such as article title, content, publication date, URL, DOI, are not available in a structured form. Web scraping techniques were used to collect, organize and analyze this information. The use of web scraping, combined with MySQL database storage proved to be effective for collecting and managing information from these agencies' web pages, enabling comprehensive and contextualized analyses to be carried out. While press releases are not the only way to qualify and understand the social attention of science, the construction of the database and tests carried out have demonstrated that it can be a relevant source for this understanding.

Keywords: Science News Agencies; Press Releases; Open Access; Indicator of Social Attention; Natural Language Processing

INTRODUÇÃO

Um artigo científico pode ter parte de seus usos rastreado em plataformas digitais não acadêmicas, como notícias jornalísticas, postagens em redes sociais, verbetes da Wikipedia, vídeos no Youtube e por outros meios. São os chamados indicadores de impacto social na ciência, que tiveram interesse catapultado com a pandemia da Covid-19, com a produção científica mais acessada por não acadêmicos e instituições de ciência cobradas por mais diálogos com a sociedade.

Técnicas para esse tipo de rastreamento de atenção social da ciência estão concentradas na altmetria, área capaz de medir parte destes usos e da atenção social que o artigo científico recebeu. Com mais de uma década de existência, a altmetria tem rastreado a circulação de artigos científicos em fontes online variadas, e é complementar a métricas tradicionais como as citações (Thelwall 2021).

Entretanto, reconhecendo que a área contribuiu para pensar a circulação do material científico para além da comunidade científica, o método foi acusado de supervalorizar métricas e produzir efeitos similares à cientometria, não contrabalanceando a obsessão com indicadores (Haustein 2016).

As críticas à altmetria recaem sobre a possibilidade da produção artificial de indicadores, sobre a falta de contexto em que as os artigos são compartilhados e sobre a impossibilidade de fornecer retorno qualificado sobre a pesquisa circulando nas redes: sua qualidade ou uma real dimensão do seu impacto social (García-Villar 2021; Thelwall 2021).

Contudo, é um desafio ir além das limitações da altmetria e, de fato, qualificar o modo como o material científico circula socialmente. A começar por uma aparente separação entre métrica e contexto, que a utilização de métodos mistos na pesquisa científica

demonstra estar sendo superada, com o uso de dados, métodos e análises quali-quantitativos (Tashakkori, Creswell 2007).

Uma maneira de superar essa dificuldade é a busca por bancos de dados que unam contextos de divulgação de artigos científicos com possibilidades de rastreamento. Com bancos de dados mistos, é possível a construção de indicadores baseados em informações qualitativas e associá-los com o potencial da altmetria.

Promover maior contextualização à circulação do conhecimento é uma maneira de chamar a atenção da comunidade científica sobre sua inserção social e seu papel no desenvolvimento de cultura científica (Vaccarezza 2009). A construção de indicadores contextualizados e qualificados é um dos caminhos para avaliar se há direcionamentos e esforços de instituições e da comunidade científica nesse sentido.

Assim, este artigo explora a possibilidade de construção de um banco de dados estruturado de *press releases* ser um caminho para a análise de métricas qualificadas de atenção social da ciência. *Press releases*, ou comunicados de imprensa, são textos produzidos por departamentos de relações públicas de empresas e instituições, que têm como objetivo prioritário o relacionamento com a mídia; e, com isso, garantem a publicidade e divulgação gratuita de produtos e iniciativas. A proposta é pela construção de "indicadores de atenção social qualificada de ciência", tendo o *press release* como suporte.

Na comunicação de ciência, *press releases* têm especial relevância pela dependência que jornalistas possuem desse tipo de material, seja por dificuldades técnicas específicas de cobertura da área, seja por mais confiança de profissionais de comunicação nas instituições de ciência (Autzen 2014) ou por falta de jornalistas produzindo conteúdos inéditos.

Esse gênero de texto chegou a ser classificado como uma "tendência" importante na comunicação de ciência, iniciando uma era de "copia-e-cola" na área. Além da reprodução muitas vezes na íntegra desses materiais por jornalistas (Autzen 2014), também há evidências de trechos de *press releases* compartilhados nas redes sociais (Verstappen et al. 2022), o que demonstra a extensão da disseminação desses comunicados, indo além das instituições e das agências em que são compartilhados.

Agências de notícia de ciência

A disseminação de *press releases* na comunicação científica é resultado de um longo processo, que tem seu formato mais elaborado na criação de instituições exclusivamente dedicadas para esse fim. Agências especializadas na divulgação em massa de *press releases* de ciência começaram a surgir a partir dos anos 1990, a exemplo da agência de alcance internacional EurekAlert!. Fundada em 1996, a agência é mantida pela AAAS (sigla para *American Association for the Advancement of Science*), associação também responsável pela revista *Science* (Seijo 2016).

O alcance dos materiais do EurekAlert! é uma amostra da força dessas agências, com textos chegando a um milhão de *views* (EurekAlert!, 2020). Comunicados de imprensa do EurekAlert! também são reproduzidos na íntegra por jornalistas, com pesquisas verificando altas semelhanças entre o material divulgado pela agência e textos jornalísticos. Um estudo verificou que um terço das notícias na área médica tem por base unicamente esse material (Holshue 2015; de Vrieze, J. 2018).

A relevância que essas agências têm para a divulgação científica, seja pelo apoio a comunicadores, seja via acesso direto pelo público, as colocam como um ator prioritário de análise sobre como a ciência chega à sociedade. Elas podem fornecer dados qualificados de atenção social da ciência, um potencial que tem sido explorado por iniciativas de construção de bancos de dados (Orduña Malea, Costas 2022).

Contudo, apesar de muitas agências de notícias disponibilizarem *press releases* em seus sites, informações importantes como título do artigo, conteúdo, data da publicação, URL, DOI (*Digital Object Identifier,* identificador persistente do artigo), impressão da pesquisa, relevância científica, área de conhecimento, autores, gênero, citações, e tipo de acesso não estão organizados em uma base de dados disponível para *download.* A disponibilidade das informações não é feita de forma organizada ou de fácil acesso, o que seria fundamental para a realização de estudos e de análises mais abrangentes.

*Assim, este artig*o pretende responder à seguinte pergunta de pesquisa: É possível rastrear *press releases* de agências de notícias de ciência a partir de seus *websites* de forma automatizada e aberta para que sejam uma fonte de atenção social qualificada da ciência?

METODOLOGIA

O processo para a construção do banco de dados envolveu dois passos principais: 1) encontrar as melhores técnicas de coleta de *press releases* (e dados associados) do *site* das agências de notícia de ciência; e 2) deixá-los estruturados para análise.

As principais etapas — compostas pela coleta, organização, análise e visualização de dados — foram realizadas utilizando bibliotecas dedicadas a cada uma dessas tarefas, disponíveis na linguagem de programação (Python). Em programação, uma biblioteca contém um conjunto de funções e códigos já referendados que facilitam o trabalho a ser realizado.

Inicialmente, para a coleta dos *press releases,* foram utilizadas técnicas de *web scraping*, também conhecidas como raspagem de dados na *web*. Para isso, foram escolhidas as bibliotecas *Selenium* e *BeautifulSoup*, que contêm uma série de funções que permitem a raspagem de dados de várias páginas, com a biblioteca Selenium também possibilitando o controle do navegador *web* de forma automatizada. O *web scraping* possibilitou a extração das informações de *press releases* e dados diversos, transformando-os em um formato adequado para análise posterior. A raspagem foi feita, a partir das seguintes etapas:

- Identificação do alvo: seleção dos *websites* das agências de notícias de ciência escolhidas para extração dos dados;

- Análise da estrutura: estudo minucioso da estrutura das páginas *web*-alvo para identificar a localização das informações desejadas e elementos HTML relevantes;

- Extração dos dados: utilização de bibliotecas de programação para automatizar o processo de extração dos dados, com a identificação das informações desejadas, como textos, links e tabelas;

- Transformação e armazenamento: realização de transformações para limpar, formatar e organizar as informações. Também foi realizado o armazenamento de dados de forma estruturada no formato CSV, em arquivo XLSX e em banco de dados MySQL.

Após a coleta das informações, utilizou-se a biblioteca de análise de dados (Pandas) para transformar os dados em uma estrutura tabular, o que torna possível a aplicação de operações matemáticas e estatísticas sobre diversos dados em poucas linhas de código. Desta forma, foi possível armazenar os dados corrigidos no *BigQuery*, um *data warehouse* em nuvem que permite o armazenamento e consultas a grandes quantidades de dados a preços baixos — se comparado aos custos relativos à compra e manutenção de uma infraestrutura própria (Lakshmanan, 2019). Por fim, foi utilizada uma biblioteca de visualização (Plotly), um complemento essencial à biblioteca de análise, pois transforma dados tabulares em gráficos e imagens, o que torna o processo de compreensão do conjunto de dados mais simples e rápido.

O banco de dados foi complementado com informações não presentes diretamente nas páginas acessadas, a partir da correspondência do DOI (informação que alguns *press releases* disponibilizam) com dados do OpenAlex (banco de dados aberto de artigos científicos). A partir dessa correspondência, por exemplo, é possível verificar se os artigos divulgados pelos comunicados de imprensa estão em acesso aberto ou não, que tipo de acesso aberto possuem e quantas citações os artigos receberam. O uso do OpenAlex foi possível com o código disponibilizado em github.com/insyspo/openalex (InSySPo 2024).

Algumas informações também foram recuperadas com o uso de *Application Programming Interfaces* (APIs), que permite o acesso prático a dados disponíveis em bancos de dados externos sem a necessidade de acessar sua estrutura diretamente. Para a Agência BORI, a identificação do gênero de autores e a localização geográfica das instituições foi possível por esse método.

A construção do banco de dados envolveu a coleta de informações nas seguintes agências de notícias de ciência:

- **EurekAlert!** (EUA) lançada em 1996, publica *press releases* elaborados por instituições universitárias de diferentes países e idiomas;

- **AlphaGalileo** (Reino Unido), criada em 1998, reúne mais de 7.000 jornalistas, e inclui *press releases* de diferentes países em inglês, francês, alemão e espanhol;
- **Agência BORI** (Brasil), lançada em 2020, é uma agência que dá visibilidade à ciência nacional. Reúne mais de 3.000 jornalistas e seu conteúdo é produzido por equipe própria e também de parceiros a partir, principalmente, de revistas científicas de acesso aberto.

O processo considerou a coleta de *press releases* de 2018 a 2023. O período escolhido tem por objetivo gerar análises comparativas entre o período pandêmico e o pré-pandêmico. Uma exceção ao período escolhido foi a coleta realizada na Agência BORI (2020-2023), uma vez que a agência foi criada em 2020 (tabela 1).

Tabela 1. Coleta de dados de agências de notícias de ciência para a construção de banco de dados de *press releases* .

Agência	Período coletado	Status da coleta
Agência BORI	(2020-2023)	Finalizado
AlphaGalileo	(2018-2023)	Finalizado
EurekAlert!	(2018-2023)	Finalizado

RESULTADOS

O processo de *web scraping* utilizado para extrair informações de páginas *web* específicas de agências de notícias de ciências envolveu a seleção de *sites*-alvo relevantes e a aplicação de técnicas de extração automatizada de dados, inicialmente armazenados CSV e XLSX. Posteriormente, foi desenvolvido um banco de dados MySQL para armazenar as informações coletadas, com a estrutura de tabelas, colunas e restrições necessárias para garantir a integridade dos dados.

Foram coletados das agências: o texto completo do *press release*, o título, a URL, o DOI do artigo divulgado, a data de publicação, bem como *tags* associadas à área disciplinar ao qual o material está associado. A estrutura da página de cada uma das agências foi analisada e, na presença de outras informações, houve coleta personalizada. Assim, no caso da EurekAlert!, alguns *press releases* contêm ficha técnica destacando o tipo de metodologia usada no artigo científico; na Agência Bori, um selo identifica se o artigo tem como origem o SciELO; na AlphaGalileo, *tags* informam se o *press release* é derivado de um artigo que passou por revisão por pares.

Após a coleta, foram realizadas análises para avaliar a completude dos dados. Em algumas agências, o *scraping* precisou ser realizado em etapas para garantir a maior coleta possível de informações. Para um banco de dados viável da AlphaGalileo, por

exemplo, foi necessário contato com a equipe para a disponibilidade de um *login* de acesso para a coleta.

A coleta de dados da Agência BORI foi concluída. No total, foram coletados 560 *press releases*. A primeira coluna da tabela 2 mostra as informações coletadas e a completude de cada item na BORI. Vê-se que o DOI está disponível em 40.4% dos *press releases*. O dado reflete tanto a variedade de textos (alguns comunicados são de projetos ou iniciativas não publicadas em periódicos), quanto ao fato do material não conter o link para o artigo científico.

Além do material na tabela 1, outras informações foram recuperadas de forma complementar. A partir do DOI coletado no *press release*, foi possível obter informações de outros bancos de dados, como o OpenAlex e SciELO (Mazoni 2024). Assim, foi possível definir:

- Autores citados: o nome, posição e gênero provável dos autores citados nos press releases foram recuperados. O nome e posição foram encontrados no OpenAlex, enquanto que o gênero foi recuperado a partir de uma API (no caso da BORI) e também com o apoio da biblioteca Python gender-guesser (no caso da EurekAlert! e AlphaGalileo).
- Instituições: o nome, ROR (*Research Organization Registry*) e estado (somente BORI) de instituições de pesquisa foram recuperados. O nome e ROR foram encontrados no OpenAlex, e o estado (que posteriormente foi categorizado em regiões) por meio de uma API. O ROR é uma iniciativa que visa fornecer identificadores únicos para instituições de pesquisa em todo o mundo.
- Publicações indexadas no SciELO: alguns *press releases* já dispunham dessa informação (BORI), outros foram analisados a partir do banco de dados do SciELO disponibilizado no BigQuery.
-

Tabela 2. Completude de cada dado analisado na amostra de press releases da Agência BORI (2020 a 2023)*, AlphaGalileo (2018 a 2023) e EurekAlert! (2018 a 2023).

	Completude (%)		
Nome da coluna	**Agência BORI**	**AlphaGalileo**	**Eurekalert!**
Total de *press releases* coletados	560	60.663	204.430
titulo / title (título do *press release*)	100	100	100
url (*link* do *press release*)	100	100	100
data_publicacao / date (data de publicação)	100	100	100

conteúdo / content (conteúdo do *press release*)	100	99.9	99.9
instituicão / institution (instituição a qual a pesquisa divulgada está vinculada)	97.5	100	100
periódico / journal	68.2		71.5
todos_DOIs / DOI (DOI do artigo divulgado)	40.4	54	59.2
tipo / type (se é *artigo revisado por pares*)		68	100
palavras_chave / keywords (do press release)	82.1	99.8	99.6

***** A Agência BORI foi criada em 2020; por isso, a coleta se iniciou neste ano.

A coleta para a AlphaGalileo (segunda coluna, tabela 2) também foi finalizada, com um total de 60.663 *press releases* encontrados para o período de 2018 a 2023. Os *releases* publicados por esta agência já dispõem de uma grande variedade de dados, reduzindo a necessidade de recuperação destes por métodos externos (como foi feito para a Agência BORI). Os dados sobre instituições e regiões, por exemplo, são frequentemente publicados em conjunto com o conteúdo do material.

Por fim, realizou-se uma coleta de dados para a EurekAlert! (terceira coluna, tabela 2). Foram coletados dados de 240.430 *press releases*. Todos os dados necessários foram encontrados diretamente nos comunicados de imprensa, mas há planos para uma análise futura em conjunto com os dados do OpenAlex, como foi feito para a Agência BORI.

Dada a importância do *press release* na divulgação científica e o alcance desse conteúdo para além do jornalismo, a coleta do texto completo do *press release*, associada ao identificador do artigo científico, oferece oportunidades para informações qualificadas sobre a circulação do material.

DISCUSSÃO

O banco de dados apresenta ao menos três potenciais de análise: 1) processamento de palavras-chave no texto coletado e análise qualitativa; 2) cruzamento de dados coletados de *press releases* com outros bancos de dados, como o OpenAlex; e 3) análises combinadas com a altmetria.

Tais potenciais podem ser constatados por meio de amostras de *press releases* extraídos da base construída. Um dos *press releases* do EurekAlert!, por exemplo, divulgava um estudo em acesso aberto publicado no periódico *Endocrinology* (DOI: 10.1210/endocr/bqz044). A partir de testes em camundongos, o artigo conclui que o óleo de soja se mostrou capaz de alterar a expressão do gene hipotalâmico e do

fenótipo metabólico. Segundo o artigo, isso potencialmente contribuiria para o desenvolvimento de condições associadas a alterações metabólicas, como a diabetes, por exemplo (Deol et al. 2020). O artigo tem 16 coautores da Universidade da Califórnia.

Dados altmétricos mostram que o artigo recebeu 12 citações, mais de 49 mil visualizações no site da revista e acumulou 854 pontos no indicador de atenção social, com 65 menções em sites de notícias (a menção com pontuação mais alta) e 476 tweets no X (ex-Twitter). Ou seja, o artigo recebeu atenção social muito superior à acadêmica, sendo que a atenção social é proveniente de 86% do público em geral.

Uma análise qualitativa do texto permite compreender o seu desempenho altmétrico. Observam-se diferenças entre o tom do artigo e do *press release* divulgado. Enquanto o título do artigo científico descreve: "Desregulação da expressão gênica hipotalâmica e do sistema oxitocinérgico por dietas com óleo de soja em camundongos machos" (Deol et al. 2020), o press release traz a seguinte titulação: "Óleo mais consumido na América causa alterações genéticas no cérebro" (*University of California* 2020).

É possível considerar que o título hiperbólico do *press release*, bem como uma associação causal que extrapola os resultados do artigo científico podem estar associados ao seu desempenho altmétrico. O estudo em camundongos foi veiculado com uma associação causal subentendida em humanos no título do comunicado de imprensa.

Outro exemplo é o artigo sobre a descoberta de um fóssil de dinossauro no Brasil de título "*New reptile shows dinosaurs and pterosaurs evolved among diverse precursors*" (DOI: 10.1038/s41586-023-06359-z) com 10 citações recebidas e pontuação de atenção social ainda maior do que o exemplo anterior, totalizando 872, com 88% da atenção sendo proveniente do público em geral.

O artigo em acesso aberto foi mencionado em 50 sites de notícias (com grande diversidade de países), 663 tweets, e o material aparece em 6 páginas da Wikipedia. O texto envolveu 4 coautores do Brasil, 3 da Argentina e 2 dos Estados Unidos. O *press release* "Descoberta de fóssil no Sul do país muda paradigma da origem dos dinossauros e pterossauros", divulgado pela Agência BORI, traz grande ênfase na contribuição da pesquisa para quebrar um paradigma na paleontologia, e valoriza a contribuição brasileira em nível mundial.

Além do potencial de análise de casos isolados — como demonstrado acima —, o conjunto de dados de press releases, associados ao artigo científico, pode contribuir para a identificação de padrões de divulgação em áreas disciplinares e como esses padrões se relacionam com dados altmétricos. Outras associações podem indicar como o valor-notícia do jornalismo, conceito que define critérios utilizados na área para definir o que será ou não veiculado, exerce influência sobre a performance de artigos científicos.

Investigações adicionais podem comparar autores divulgados no *press release* com os presentes no artigo, bem como sua nacionalidade, filiação institucional e gênero. O uso de palavras-chave e processamento de textos permite análises temáticas. Por exemplo, podem ser avaliados o padrão de divulgação de artigos relacionados às mudanças climáticas, à Covid-longa, a vacinas contra Covid-19, etc, bem como a circulação de temas socialmente relevantes, como os que ganharam mais atenção da mídia ou que mais geraram produção científica.

Pode-se ainda analisar o impacto do acesso aberto de artigos divulgados em *press releases* para a sociedade, se contribuem para a inclusão de temas relevantes para o debate social, ou na maior visibilidade de pesquisadoras mulheres e de países do Sul Global, por exemplo.

Os potenciais acima representados pela construção do banco de dados de *press releases* demonstram que a atenção social da ciência demanda investimento interdisciplinar, cruzamento de bancos variados, e de métodos quali-quantitativos para não ficar restrita a métricas de performance e explicações parciais.

Os dados coletados, contudo, possuem algumas limitações. Analisamos *press releases* que citam artigos científicos que possuem DOI (a tabela 2 mostra o percentual de textos com esse dado: BORI 40,4%; AlphaGalileo: 54%; EurekAlert: 59.2%). Além de entender melhor a representatividade desse percentual, análises que incluam os demais materiais, não necessariamente de divulgação de *papers,* poderão dar uma visão mais complexa sobre as estratégias de divulgação da mídia.

A coleta também foi realizada com o uso de login no caso da agência AlphaGalileo, já que apenas uma pequena parcela do material estava disponível publicamente — e isso pode dificultar o uso de dados para análises futuras. O mesmo ocorre com conteúdos completos da EurekAlert! que não podem ser compartilhados publicamente, por questões de direitos autorais, que variam de acordo com a instituição ou empresa autora do conteúdo.

É preciso ainda relativizar a representatividade das agências ditas internacionais. No caso da EurekAlert!, fica claro que o enfoque segue na ciência produzida nos Estados Unidos, país sede da agência, e o mesmo ocorre com a AlphaGalileo na Europa. Apesar da escolha da Agência BORI (Brasil) servir como um contraponto, a inclusão de outras agências locais poderá conferir maior diversidade ao banco de dados.

Pesquisas como a presente poderão contribuir para que as próprias agências visualizem suas estratégias de divulgação científica e verifiquem adequação à diversidade de áreas do conhecimento, regiões geográficas, instituições, gênero de autores, entre outras. De modo geral, a proposta aqui de indicadores "qualificados de atenção social" pode contribuir para que a comunidade científica compreenda, de forma contextualizada, como está sua relação com a sociedade de maneira a fomentar valores importantes, como a promoção da inclusão e o desenvolvimento de cultura científica.

CONCLUSÕES

A utilização de *web scraping* em conjunto com o armazenamento em banco de dados MySQL mostrou-se eficaz para coletar e gerenciar informações provenientes de páginas web de agências de notícias de ciências. Essas técnicas proporcionam a automatização do processo de coleta de dados, permitindo a atualização contínua das informações disponíveis. Além disso, o armazenamento em um banco de dados MySQL e utilização de linguagem de consulta estruturada (SQL), possibilitam a realização de consultas complexas e análises mais aprofundadas dos dados.

A disponibilização desses dados coletados a partir de três agências de notícias de ciência em uma base de dados estruturada e acessível para *download* facilitaria o acesso e a análise por parte da comunidade científica, permitindo a realização de estudos quali-quantitativos, a identificação de tendências, a criação de métricas contextualizadas para avaliação da produção científica e até mesmo a criação de ferramentas de busca e de recomendação mais eficientes.

O entendimento da atenção social da ciência provavelmente será realizado com a combinação de diversas ferramentas e a construção de banco de dados capazes de fornecer maiores contextos sobre métricas já consolidadas. Enquanto que o comunicado de imprensa não é a única forma de qualificar e entender a atenção social da ciência, a construção do banco de dados e testes realizados demonstram que se trata de uma fonte relevante nessa compreensão.

A partir dessas investigações, demonstramos que um banco estruturado de dados de press releases tem potencial para ser uma fonte de indicadores qualificados de atenção social da ciência, contribuindo para a contextualização e maior compreensão de dados altmétricos e cientométricos.

AGRADECIMENTOS

Os autores agradecem a parceria com o Projeto InSySPo Fapesp pelo apoio no uso do BigQuery e pela hospedagem e gestão dos dados. Agradecemos ainda os debates realizados após apresentação em Temuco (Chile), no Latmétricas (2023).

FINANCIAMENTO

Auxílio pesquisa Fapesp (no. Processo 2021/07577-8) do projeto VOICES: o valor da abertura, inclusão, comunicação e engajamento pela ciência no mundo pós-pandemia. AM agradece a Fapesp pelo financiamento como parte do projeto InSySPo (processo no. 2021/05823-1, e vinculado no. 2019/04300-5).

DISPONIBILIDADE DE DADOS

O conjunto de dados que ampara os resultados deste estudo, com exceção das informações da agência AlphaGalileo (não acessíveis publicamente), foi disponibilizado no GitHub e pode ser acessado em: https://github.com/thaiscsp/labincc_agencias

REFERÊNCIAS

AUTZEN, Charlotte, 2014. Press releases — the new trend in science communication. *Journal of Science Communication*. 2014. vol. 13, no. 3, p. C02. DOI 10.22323/2.13030302.

DE VRIEZE, J., 2018. EurekAlert! Has spoiled science news. Here's how we can fix it. [em linha]. 2018. Disponível em: https://medium.com/@jopdevrieze/eurekalert-has-spoiled-science-news-heres-how-we-can-fix-it-851ce5c00c9a

DEOL, Poonamjot et al., 2020. Dysregulation of Hypothalamic Gene Expression and the Oxytocinergic System by Soybean Oil Diets in Male Mice. *Endocrinology*. 1 fevereiro 2020. vol. 161, no. 2, p. bqz044. DOI 10.1210/endocr/bqz044.

EUREKALERT, 2020. No. 1 news release on EurekAlert!'s 2020 Trending List smashes previous all-time record for visits. *EurekAlert!* [em linha]. 23 dezembro 2020. [Acesso em 11 novembro 2022]. Disponível em: https://www.eurekalert.org/news-releases/616312

GARCÍA-VILLAR, Cristina, 2021. A critical review on altmetrics: can we measure the social impact factor? *Insights into Imaging*. 2 julho 2021. vol. 12, no. 1, p. 92. DOI 10.1186/s13244-021-01033-2.

HAUSTEIN, Stefanie, 2016. Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics*. 1 julho 2016. vol. 108, no. 1, p. 413–423. DOI 10.1007/s11192-016-1910-9.

HOLSHUE, Jennifer, 2015. Using EurekAlert! as a News Reporting Resource. *2015 Mass Media Fellows* [em linha]. Washington. 2015. Disponível em: https://www.aaas.org/programs/mass-media-science-engineering-fellows/2015-fellows

INSYSPO, 2024. OpenAlex upload to BigQuery. [Acesso em 13 de maio de 2024]. Disponível em: github.com/insyspo/openalex.

LAKSHMANAN, Valliappa, e TIGANI, Jordan. Google Bigquery: the definitive guide: data warehousing, analytics, and machine learning at scale. O'Reilly Media, 2019.

MAZONI, A. F.; MARICATO, J. de M.; ARAÚJO, R. F. de; COMESAÑA, R. C. Challenges in cloud infrastructure and scientific data: technical proposals and tools applied to the SciELO database. Anais do Workshop de Informação, Dados e Tecnologia - WIDaT, [S. l.], v. 6, 2023. DOI: 10.22477/vi.widat.44. Disponível em: https://labcotec.ibict.br/widat/index.php/widat2023/article/view/44. Acesso em: 13 maio. 2024.

ORDUÑA MALEA, Enrique e COSTAS, Rodrigo, 2022. *The Eurekalert! project: dataset of mentions to press releases* [em linha]. 30 setembro 2022. Universitat Politècnica de València. [Acesso em 19 dezembro 2022]. Disponível em: https://riunet.upv.es/handle/10251/186769

SEIJO, Bibiana Campos, 2016. Happy 20th birthday, EurekAlert. *C&EN Global Enterprise*. 30 maio 2016. vol. 94, no. 22, p. 1–2.

TASHAKKORI, Abbas e CRESWELL, John W., 2007. Editorial: The New Era of Mixed Methods. *Journal of Mixed Methods Research*. janeiro 2007. vol. 1, no. 1, p. 3–7. DOI 10.1177/2345678906293042.

THELWALL, Mike, 2021. Measuring Societal Impacts of Research with Altmetrics? Common Problems and Mistakes. Journal of Economic Surveys. 2021. vol. 35, no. 5, p. 1302–1314. DOI 10.1111/joes.12381.

UNIVERSITY OF CALIFORNIA, 2020. America's most widely consumed oil causes genetic changes in the brain. *EurekAlert!* [em linha]. 17 janeiro 2020. [Acesso em 8 maio 2024]. Disponível em: https://www.eurekalert.org/news-releases/562654

VACCAREZZA, Leonardo Silvio, 2009. Estudios de cultura científica en América Latina. *Redes* [em linha]. 2009. [Acesso em 21 agosto 2023]. Disponível em: https://www.redalyc.org/pdf/907/90721335004.pdf

VERSTAPPEN, M. et al., 2022. Van persbericht tot Facebookpost. *Tijdschrift voor Communicatiewetenschap*. 2022. vol. 50, no. 3, p. 210–230. DOI 10.5117/tCW2022.3.005.VErs. EVE, Martin Paul, 2020. Some tips on writing a data management plan for the humanities. *Twitter: @martin_eve* [em linha]. 11 fevereiro 2020. [Acesso em 11 fevereiro 2020]. Disponível em: https://twitter.com/martin_eve/status/1227240006344835073