

INFORMAÇÃO CIENTÍFICA NO GITHUB: possibilidades de rastreamento e Altmétrie por meio de identificadores persistentes

Rafael Gutierrez Castanha

<https://orcid.org/0000-0002-3117-1780> | r.castanha@gmail.com
Universidade de Marília (UNIMAR)

Francielle Franco

<https://orcid.org/0000-0002-3686-595X> | francielle.franco@unesp.br
Universidade Estadual Paulista (Unesp)

RESUMO:

Este estudo analisa a circulação da informação científica no GitHub por meio de identificadores persistentes para avaliar seu potencial altmétrico. Para isso, realizaram-se buscas por diferentes identificadores persistentes em referência a circulação de documentos, autores e instituições nas onze categorias nativas de filtragem da plataforma. Os resultados revelam uma predominância de menções em códigos, evidenciando a circulação científica na plataforma, além de expressiva ocorrência em discussões e envios, o que reflete avaliação e curadoria. Conclui-se que o GitHub transcende sua função técnica, consolidando-se como promissora fonte altmétrica dada a presença expressiva dos objetos científicos.

Palavras-chave: Altmétrie. GitHub. Indicadores Persistentes. Informação científica.

EIXO TEMÁTICO:

Altmétrie e Webometria

MODALIDADE:

Resumo expandido

DOI: 10.22477/x.ebbc.973

COMO CITAR ESTE ARTIGO:

CASTANHA, Rafael Gutierrez; FRANCO, Francielle. Informação científica no GitHub: possibilidades de rastreamento e altmétrie por meio de identificadores persistentes. In: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.973. Disponível em: <https://doi.org/10.22477/x.ebbc.973>.



1 INTRODUÇÃO

A altmetria refere-se a indicadores que medem o engajamento com a produção científica em plataformas digitais, oferecendo uma visão instantânea da circulação da informação científica. Esta abordagem permite compreender melhor a visibilidade, relevância e engajamento da produção científica além do meio acadêmico (Bornmann, 2014; Araújo, 2014). Os ambientes digitais favorecem dinâmicas emergentes de disseminação, tornando os indicadores altmétricos fundamentais para captar o impacto em múltiplas esferas. Nesse sentido, destaca-se que os indicadores altmétricos oferecem uma avaliação científica mais ágil e diversificada (Haustein, 2016) e que os agregadores altmétricos tornam-se indispensáveis para análises comparativas e identificação de tendências (Ortega, 2020; Rasuli e Nabavi, 2024).

Dentre os novos canais complementares, o GitHub¹ destaca-se por suas características sociais e colaborativas durante o desenvolvimento de códigos. É importante notar que funcionalidades como estrelas² e *forks*³ refletem o impacto de projetos e que tais métricas fornecem dados relevantes sobre o alcance e a colaboração em torno do desenvolvimento de códigos (Safiullina *et al.*, 2024; Braga *et al.*, 2023). A infraestrutura do GitHub favorece a ciência aberta, funcionando como um espaço que une desenvolvedores através de recursos sociais. No que tange ao contexto científico, a plataforma facilita a conexão entre repositórios e publicações científicas (Escamilla *et al.*, 2022; Färber, 2020), embora se reconheça que a manutenção da conexão entre código e artigo pode ser desafiadora (Wattanakriengkrai *et al.*,

¹ Vale ressaltar que o GitHub se consolida como a principal plataforma de desenvolvimento colaborativo de código de colaboração, reunindo mais de 180 milhões de desenvolvedores e 630 milhões de repositórios (GITHUB, 2025), justificando assim a sua escolha para aplicação deste estudo.

² A funcionalidade de estrela em um repositório do GitHub indica que um usuário favoritou um repositório podendo indicar a utilidade ou popularidade de um repositório.

³ Essa funcionalidade permite criar uma cópia do repositório na conta do usuário permitindo edições no código.

2022), impactando a rastreabilidade a longo prazo.

A representatividade dos dados no GitHub configura um desafio, visto que muitos repositórios são experimentais e a API da plataforma impõe limitações na coleta. Ainda assim, defende-se que o acesso aberto e a integração com dados bibliográficos tornam o GitHub um recurso promissor (Escamilla *et al.*, 2022; Wattanakriengkrai *et al.*, 2022). Interações sociais e a possível circulação de informações científicas em projetos computacionais, podem sugerir a plataforma como potencial fonte altmétrica ao evidenciar como seus 180 milhões (aproximadamente) de usuários relacionam-se com informações científicas diversas. Nesse sentido, é sabido que, para compreensão de informações científicas, aqueles que possuem algum identificador persistente, facilitam sua rastreabilidade, como por exemplo, o rastreamento de artigos com DOI (Digital Object Identifier) utilizado pela Altmetric.com. Assim, a presença de identificadores como DOI (artigos), ORCID (autores) ou ROR (instituições) no GitHub, poderá suscitar a plataforma como uma fonte de informações científicas alternativas as tradicionais.

Diante do exposto, este estudo busca responder ao seguinte problema: em que medida a presença de identificadores persistentes em diferentes níveis no GitHub pode revelar padrões de circulação, apropriação e reconhecimento da informação científica? Tendo por objetivo analisar a circulação da informação científica no GitHub por meio da identificação e caracterização de menções a objetos científicos representados por identificadores persistentes em nível de documentos (DOI), autores (ORCID) e instituições (ROR), visando avaliar o potencial da plataforma como fonte de indicadores altmétricos. Assim, tem-se como premissa de que o GitHub configura-se como um ambiente relevante para a disseminação de conhecimento científico e tem potencial para ser explorado como fonte altmétrica complementar às bases tradicionais.

Destaca-se ainda que este trabalho pretende preencher uma lacuna na literatura ao realizar uma análise exploratória da presença de identificadores persistentes no GitHub, identificando padrões de circulação, apropriação e reconhecimento em ambientes de desenvolvimento de código que não são capturadas pelos indicadores altmétricos convencionais nem abrangidas de forma ampla por outros estudos.

2 METODOLOGIA

Com intuito de analisar a circulação da informação científica no GitHub, por meio da identificação e caracterização de menções a objetos científicos representados por identificadores persistentes, este estudo analisa o uso deste recurso dentro do ecossistema de desenvolvimento colaborativo do GitHub. Para tanto, a coleta de dados é centralizada na ferramenta *GitHub Search* (<https://github.com/search>), sem necessidade de login, utilizando como termos de busca “doi.org” (para captar menção a documentos científicos), “orcid.org” (captar menção a autores) e “ror.org” (captar menção a instituições). Destaca-se que a coleta foi realizada através da busca simples, sem a necessidade de realizar o login na plataforma. Esta escolha metodológica permite observar as menções a diferentes objetos científicos nos seguintes níveis: documentos, autores e instituições.

O procedimento de busca, via *GitHub Search*, segmenta os resultados obtidos nas 11 categorias nativas de filtragem da plataforma, o que possibilita uma análise do contexto de circulação da informação científica, sendo elas: 1) *Code* (Código): Buscar por termos dentro de arquivos de qualquer linguagem de programação, além de arquivos de documentação como README.md; 2) *Repositories* (Repositórios): Busca o termo no nome do projeto ou na descrição curta (“About”) do repositório;

3) *Issues* (Tarefas/Problemas): Localiza termos em títulos e comentários ou discussões abertas por usuários; 4) *Pull Requests* (Solicitações de Alteração): Localiza por termos em propostas de mudanças no código; 5) *Discussions* (Discussões): Localiza o termo em discussões dentro dos repositórios (circulação de linguagem menos formal); 6) *Users* (Usuários): Localiza o termo no nome ou descrição de perfil de usuários ou organizações; 7) *Commits* (Envios): Localiza termos inseridos pelos desenvolvedores ao salvar alterações; 8) *Packages* (Pacotes): Busca em registros de *software* publicados e hospedados no GitHub (varre "About this package"); 9) *Wikis*: Localiza o termo em conteúdo de páginas de documentação extensas (wikis) que acompanham alguns repositórios; 10) *Topics* (Tópicos): Localiza o termo em tags (etiquetas) que os autores atribuem aos projetos; 11) *Marketplace*: Localiza o termo em extensões e aplicativos que podem ser instalados no GitHub para automatizar fluxos de trabalho. Toda a coleta foi realizada em 20 de fevereiro de 2026.

A partir da busca pelos três termos ("doi.org", "orcid.org" e "ror.org"), construiu-se o Quadro 1 contendo os resultados quantitativos de cada busca perante cada uma das categorias do GitHub. Com o intuito de propor uma interpretação altmétrica fundamentada para cada um dos resultados, a caracterização de cada categoria foi embasada nas classificações de engajamento *online* propostas por Haustein, Bowman e Costas (2016) e Bornmann (2014). Enquanto o modelo de Haustein e colaboradores permite agrupar as interações nos eixos de acesso, avaliação e aplicação de documentos e agentes científicos, a proposta de Bornmann (2014) detalha os níveis desse impacto em categorias como salvo, discutido e citado. Para revisão gramatical deste manuscrito, utilizou-se o modelo inteligência artificial generativa Gemini Google em sua versão 3 Flash.

3 RESULTADOS

Os dados obtidos foram segmentados nas 11 categorias nativas de filtragem da plataforma, o que possibilitou uma análise direta do contexto de circulação da informação científica. Conforme apresentado no Quadro 1, os resultados quantitativos evidenciam o volume de ocorrências distribuídas nas categorias de filtragem. Importa mencionar que não há uma precisão exata sobre alguns resultados em que a plataforma fornece os valores como sufixos “K” para milhão e “M” para milhar, após os valores. A partir dessa caracterização, delinea-se a possível interpretação altmétrica correspondente a cada categoria, demonstrando em que medida o GitHub revela padrões de apropriação e reconhecimento, avaliando seu potencial como fonte altmétrica complementar às fontes altmétricas conhecidas. A partir do Quadro 1, tem-se que em 7 das 11 categorias, encontrou-se resultados para os três identificadores.

Quadro 1 – Resultado das buscas por categorias de conteúdo no GitHub.

CATEGORIA	TERMOS BUSCADOS			INTEPRETAÇÃO ALTMÉTRICA
	doi.org	orcid.org	ror.org	
Code	4.5 M	860 mil	230 mil	Menção direta a objetos científicos no processo de desenvolvimento de códigos ou na documentação principal (README).
Repositories	5 mil	61	95	A menção no "About" ou no nome do repositório suscita uso direto do objeto científico, em que o repositório representa um projeto diretamente associado a este objeto
Issues	60 mil	26 mil	27mil	Captura, possivelmente, um debate técnico, análise e/ou resolução colaborativa de parte do projeto. O objeto científico poderá ser algum tipo de argumento técnico, comentário ou parte do próprio código.
Pull Requests	14 mil	3 mil	316	Pode capturar dinâmicas de revisão no desenvolvimento de código seja como argumento técnico, seja como parte do código.

Discussion	1.000	140	20	Captura a discussão, não necessariamente em linguagem formal, como geração de ideias e disseminação social do objeto científico, a partir de menções a estes objetos.
Users	34	266	0	Captura o uso da plataforma para autoapresentação e construção de identidade online de pesquisadores e instituições a partir de objetos científicos.
Commits	378 mil	3 mil	227	Indica que um objeto científico motivou e/ou fez parte de uma alteração ou evolução direta e rastreável do projeto em desenvolvimento.
Packages	3	0	0	Suscita atribuição formal de crédito científico a objetos científicos em registros de softwares, reconhecendo que aquele objeto faz parte do desenvolvimento deste projeto.
Wikis	5 mil	290	67	Indica o uso do objeto científico para fins instrutivos, como um guia de referência, de forma mais alongada, para os usuários.
Topics	0	0	0	Reflete o uso de etiquetas para organização da informação, facilitando a categorização e a descoberta de projetos que venham a fazer menção a objetos científicos.
Marketplace	0	0	0	Potencial uso de objetos científicos para descrever aplicações de fluxo de trabalho.

Fonte: elaboração dos autores (2026).

A categoria *Code* concentra a maioria das menções no GitHub, com 4,5 milhões de DOIs. Segundo Haustein, Bowman e Costas (2016), essa presença em códigos e arquivos README.md pode configura-se como um ato de aplicação. Contudo, essa magnitude quantitativa exige cautela analítica: como alertam Kalliamvakou *et al.* (2014), o mecanismo de *fork* (cópias de projetos) pode replicar o mesmo arquivo contendo um DOI milhares de vezes, inflando os resultados. Ainda assim, evidencia-se a materialização técnica e reutilização ativa do conhecimento científico. Já as categorias *Issues*, *Pull Requests* e *Discussions* sugerem o uso do GitHub como um dinâmico espaço de análise. Sob a tipologia de Haustein, Bowman e Costas (2016),

refletem atos de avaliação da informação. Para Bornmann (2014), equivalem à dimensão de debate, capturando a atenção social da pesquisa. Ferramentas como *Issues* e *Pull Requests* funcionam como mecanismos colaborativos assemelhados à revisão por pares. Nesses locais, o documento científico pode atuar como argumento técnico ou debater ideias antes de sua efetiva implementação no código.

O volume expressivo de DOIs em *Commits*, aliado a *Repositories* e *Wikis*, ilustra as funções de curadoria e rastreabilidade da plataforma. Para Haustein, Bowman e Costas (2016), menções em *Wikis* e descrições representam atos de acesso, enquanto os *Commits* podem traduzir-se em uma aplicação granular. No modelo de Bornmann (2014), espaços descritivos operam na camada de armazenamento. O GitHub ao abarcar milhões de usuários e repositórios, sugere que, ao referenciar um artigo (observado pela presença do termo “doi.org”), concretiza a relação entre pesquisas formais e o desenvolvimento de projetos computacionais, e seu potencial altmétrico, com rastreabilidade histórica e documentada de modificações nestes projetos.

No que diz respeito aos identificadores de autoria (ORCID) e afiliação (ROR), principalmente na categoria *Users*, atende ao que Haustein, Bowman e Costas (2016) denominam como atos de acesso a agentes. Isso reflete o uso do ambiente para a construção e gestão de identidade online. Conforme destaca Araújo (2014), perfis em plataformas colaborativas são vitrines essenciais para a autoapresentação e visibilidade acadêmica. Os resultados sugerem, portanto, que o GitHub fornece interações ricas que complementam a avaliação do impacto da pesquisa científica nos ambientes digitais.

4 CONSIDERAÇÕES

A pesquisa apresentou uma análise da circulação da informação científica no GitHub a partir da menção a objetos científico representados por identificadores persistentes, a fim de avaliar seu potencial enquanto fonte altimétrica. Foi possível observar que GitHub apresenta indícios de que pode ultrapassar sua função técnica de desenvolvimento consolidando-se como ambiente potencialmente altimétrico dada a presença dos objetos científicos buscados (DOI, ORCID e ROR), evidenciando novas dinâmicas de circulação na Ciência 2.0.

A identificação das menções à objetos científicos em diferentes níveis no GitHub revela dimensões de uso que indicadores tradicionais não alcançam diretamente. Embora existam desafios e limitações metodológicas, como os *forks*, que ao gerar cópias de repositórios, podem inflar os resultados, repositórios inativos, limites de extração via API, e por hora, a não verificação da origem exata de cada elemento recuperado, observa-se forte condições de rastreabilidade de informações científicas. Assim, sua integração como fonte complementar pode ampliar a mensuração do impacto científico e o debate altimétrico.

AGRADECIMENTOS

Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), no âmbito do Programa de Doutorado Sanduíche no Exterior (PDSE), Processo nº 88881.125548/2025-01.

REFERÊNCIAS

ARAÚJO, Ronaldo Ferreira de. Ciência 2.0 e a presença online de pesquisadores: visibilidade e impacto. **Ciência da Informação em Revista**, Maceió, v. 1, n. 3, p. 32–40, 2014. DOI: <https://doi.org/10.28998/cirev.%25y132-40>. Acesso em: 24 fev. 2026.

BORNMANN, Lutz. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. **JOI** Amsterdam, v. 8, n. 4, p. 895–903, 2014. DOI: <https://doi.org/10.1016/j.joi.2014.09.005>. Acesso em: 24 fev. 2026.

BRAGA, Pedro Henrique Pereira; HÉBERT, Katherine; HUDGINS, Emma J.; SCOTT, Eric R.; EDWARDS, Brandon P. M.; SÁNCHEZ REYES, Luna L.; GRAINGER, Matthew J.; FOROUGHIRAD, Vivienne; HILLEMANN, Friederike; BINLEY, Allison D.; BROOKSON, Cole B.; GAYNOR, Kaitlyn M.; SHAFIEI SABET, Saeed; GÜNCAN, Ali; WEIERBACH, Helen; GOMES, Dylan G. E.; CRYSTAL-ORNELAS, Robert. Not just for programmers: how GitHub can accelerate collaborative and reproducible research in ecology and evolution. **Methods in Ecology and Evolution**, London, v. 14, n. 6, p. 1364–1380, 2023. DOI: <https://doi.org/10.1111/2041-210X.14108>. Acesso em: 24 fev. 2026.

ESCAMILLA, Emily; KLEIN, Martin; COOPER, Talya; RAMPIN, Vicky; WEIGLE, Michele C.; NELSON, Michael L. The rise of GitHub in scholarly publications. In: SILVELLO, Gianmaria; CORCHO, Óscar; MANGHI, Paolo; DI NUNZIO, Giorgio Maria; GOLUB, Koraljka; FERRO, Nicola (ed.). **Linking theory and practice of digital libraries**. Cham: Springer, 2022. p. 187–200. DOI: https://doi.org/10.1007/978-3-031-16802-4_15. Acesso em: 24 fev. 2026.

FÄRBER, Michael. Analyzing the GitHub repositories of research papers. In: *PROCEEDINGS OF THE ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES (JCDL '20)*, 2020. Anais [...]. New York: ACM, 2020. p. 491–492. DOI: <https://doi.org/10.1145/3383583.3398578>. Acesso em: 10 fev. 2026.

HAUSTEIN, Stefanie. Grand challenges in altmetrics: heterogeneity, data quality and dependencies. **Scientometrics**, Cham, v. 108, p. 413–423, 2016. DOI: <https://doi.org/10.1007/s11192-016-1910-9>. Acesso em: 24 fev. 2026.

HAUSTEIN, Stefanie; BOWMAN, Timothy D.; COSTAS, Rodrigo. Interpreting ‘altmetrics’: viewing acts on social media through the lens of citation and social theories. In: SUGIMOTO, Cassidy R. (ed.). **Theories of Informetrics and Scholarly Communication**. Berlin; Boston: De Gruyter Saur, 2016. p. 372–406. DOI: <https://doi.org/10.1515/9783110308464-022>. Acesso em: 24 fev. 2026.

GITHUB. Octoverse: a new developer joins GitHub every second as AI leads Type-

Script to #1. **GitHub Blog**, 2025. Disponível em: <https://github.blog/news-insights/octoverse/octoverse-a-new-developer-joins-github-every-second-as-ai-leads-typescript-to-1/>. Acesso em: 28 abr. 2026.

KALLIAMVAKOU, Eirini; GOUSIOS, Georgios; BLINCOE, Kelly; SINGER, Leif; GERMAN, Daniel M.; DAMIAN, Daniela. The promises and perils of mining GitHub. In: *PROCEEDINGS OF THE 11TH WORKING CONFERENCE ON MINING SOFTWARE REPOSITORIES (MSR 2014)*, 2014, Hyderabad. Anais [...]. New York: ACM, 2014. p. 92–101. Disponível em: <https://doi.org/10.1145/2597073.2597074>. Acesso em: 24 fev. 2026

ORTEGA, José Luis. Altmetrics data providers: a meta-analysis review of the coverage of metrics and publication. **El Profesional de la Información**, Barcelona, v. 29, n. 1, e290107, 2020. DOI: <https://doi.org/10.3145/epi.2020.ene.07>. Acesso em: 24 fev. 2026.

RASULI, Bahram; NABAVI, Mohammad. Altmetric.com or PlumX: does it matter? **Learned Publishing**, Hoboken, v. 37, n. 4, e1631, 2024. DOI: <https://doi.org/10.1002/leap.1631>. Acesso em: 24 fev. 2026.

SAFIULLINA, Guzel; GUMEROV, Aidar, DLAMINI, Gcinizwe; SUCCI, Giancarlo. Preliminary study: exploring GitHub repository metrics. In: ARAI, Kohei (ed.). **Advances in Information and Communication**. Cham: Springer, 2024. v. 919, p. 579–591. DOI: https://doi.org/10.1007/978-3-031-53960-2_38. Acesso em: 24 fev. 2026.

WATTANAKRIENGKRAI, Supatsara; CHINTHANET, Bodin; HATA, Hideaki; KULA, Raula Gaikovina; TREUDE, Christoph; GUO, Jin; MATSUMOTO, Kenichi. GitHub repositories with links to academic papers: public access, traceability, and evolution. **JSS**, Amsterdam, v. 183, 111117, 2022. DOI: <https://doi.org/10.1016/j.jss.2021.111117>. Acesso em: 24 fev. 2026.